Datasets Creation and Empirical Evaluations of Cross-Lingual Learning on Extremely Low-Resource Languages: A Focus on Comorian Dialects

Abdou Mohamed Naira[1,2], **Abdessalam Bahafid**[1,2], **Zakarya Erraji**[1], **Imade Benelallam**[1,2]
[1] INSEA, Rabat, Morocco [2] ToumAI Analytics, Rabat, Morocco
{nabdoumohamed,i.benelallam,a.bahafid,zerraji}@insea.ac.ma
{naira,imade,abahafid}@toum.ai

Abstract

In this era of extensive digitalization, there are a profusion of Intelligent Systems that attempt to understand how languages are structured for the aim of providing solutions in various tasks like Text Summarization, Sentiment Analysis, Speech Recognition, etc. But for multiple reasons going from lack of data to the nonexistence of initiatives, these applications are in an embryonic stage in certain languages and dialects, especially those spoken in the African continent, like Comorian dialects. Today, thanks to the improvement of Pre-trained Large Language Models, a spacious way is open to enable these kind of technologies on these languages. In this study, we are pioneering the representation of Comorian dialects in the field of Natural Language Processing (NLP) by constructing datasets (Lexicons, Speech Recognition and Raw Text datasets) that could be used on different tasks. We also measure the impact of using pre-trained models on languages closely related to Comorian dialects to enhance the state-of-the-art in NLP for these latter, compared to using pre-trained models on languages that may not necessarily be close to these dialects. We construct models covering the following use cases: Language Identification, Sentiment Analysis, Part-Of-Speech Tagging, and Speech Recognition. Ultimately, we hope that these solutions can catalyze the improvement of similar initiatives in Comorian dialects and in languages facing similar challenges.

1 Introduction

The Comoros are an archipelago composed of four islands in the Indian Ocean. Approximately 850,000 people are living there (Worldometers), speaking four dialects belonging to the Bantu Language family (Atlasocio). These dialects are consequently impacted by geo-spatial features that progressively increase or eliminate similarities between them as shown in (Maurizio and Michele, 2021; Chamanga, 2022) and according to the ORELC[1] lexicon (See Fig. 1) in which we can observe that in a dictionary of 7,386 entries, 15.38% of the words are shared by all the dialects, 6.47% by three and 16.10% by two dialects. Indeed, these dialects can be divided into two groups: Eastern group (ShiNdzuani and ShiMaore) and Western group (ShiNgazidja and ShiMwali). Moreover, a part of the experiments conducted in (Maurizio and Michele, 2021) has shown through lexical distances calculation that these dialects could be classified into two other different groups, the first one composed of the ShiNgazidja while the second one contains the other three dialects.

The arrival of Transformers (Vaswani et al., 2017) was a real breakthrough in Artificial Intelligence (AI). This architecture allows us to better represent the context within texts which is a major spearhead in Language Understanding. Pre-trained Language Models like Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) have encapsulated this architecture, paving the way to better representation of multiple languages in all around the world (Pikuliak et al., 2021) through Cross-Lingual Learning. In multilingual scenarios, this latter allows languages that suffer from data scarcity to learn from the others owing to a sort of transfer learning. This becomes more interesting when working on close languages as demonstrated in (Tebbifakhr et al., 2020) where a Machine Translation system was adapted to a language close to the source language used on training.

The aim of this work is to contribute on the pioneering of Natural Language Processing (NLP) on Comorian dialects by (a) constructing datasets that could be used on different downstream tasks for future works and (b) experimenting the impact of using a cross-multilingual approach on close lan-

[1] https://orelc.ac/academy/ShikomoriWords/?i=kmWords

Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII), pages 140–149
March 22, 2024 ©2024 Association for Computational Linguistics

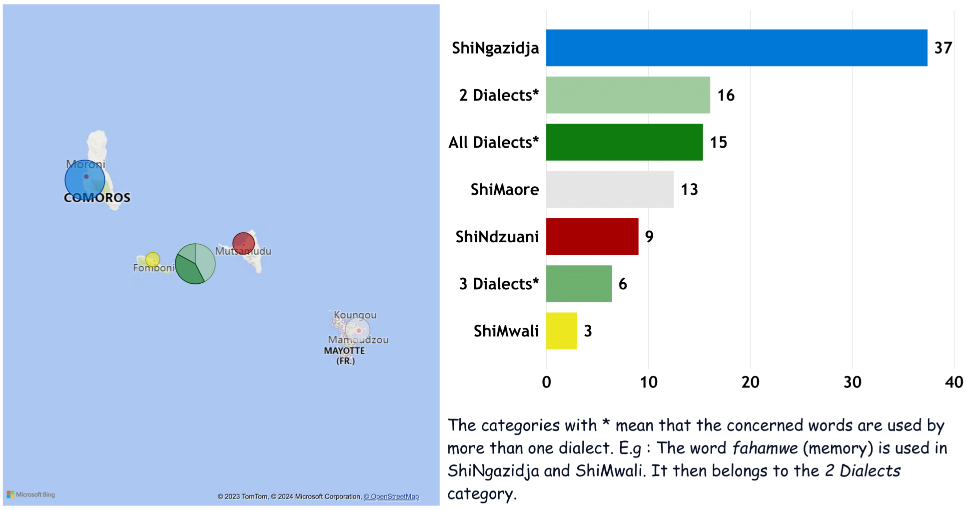

Figure 1: Dialects Varieties.

guage to leverage NLP solutions on low-resource scenarios. The rest of this study is structured as follows: We first present in Section 2 notable previous works in these dialects, then we describe in Section 3 the data collection methodologies that we adopted to collect the different datasets, Section 4 shows the experiments that we conducted to evaluate the constructed datasets while Section 5 presents the future work that could result from this study following a final conclusion.

2 Related Work

One thing to know about these works is that only few of them use NLP approaches and the data used or resulted from them are not publicly available. This make it more interesting the data retrieving and the resorting to recent NLP solutions in order to contribute to the digital representation of these dialects, hoping that this could be helpful in different upcoming use cases by researchers, institutions, companies or particulars.

2.1 Old Resources

In this section, we compile studies that have significantly influenced the advancement of Comorian dialects processing. These earlier resources predominantly employed linguistic and manual methodologies, primarily focusing on proposing structural frameworks for the written forms of these dialects and translations into foreign languages like French:

- **The Kamar-Eddine system**: In the 1960s, as described in (Lafon, 2007), Said Kamar-Eddine proposed a writting system of Comorian dialects using Arabic scripts. This notable work allowed several people to learn how to write their language and is used until now.

- **French-Comorian dictionnaries**: These dictionaries were published in 1979 (Sacleux et al., 1979) and 1997 (Chamanga, 1997). Other initiatives like ORELC followed and allow until now many people to learn these idioms.

- **Introduction to Shikomori**: A structural grammar books (Ahmed-Chamanga, 2010; Chamanga and national de documentation et de recherche scientifique , Comoros) written by the linguist Mohamed Ahmed Chamanga.

2.2 Modern NLP-specific Resources

After the democratization occuring since these recent years of solutions based on recent information technologies, the necessity to resort to these approaches for low-resource languages has become apparent. For Comorian dialects, among the few solutions that consider them, we emphasize:

- **Machine Translation dataset**: To the best of our knowledge, the work described in (Abdourahamane et al., 2016) is one of the first attempts to manage Comorian dialects through NLP. The corpus was created based on a Transfer Learning from Swahili due to the similarities between these languages.

- **Language Identification**: In (Adebara et al., 2022), Comorian dialects were added to a massive corpus for Language Identification in several African Languages.

3 Datasets

Ensuring the quality of data has always been at the center of concerns when designing AI solutions, especially in NLP (Sonntag, 2004; Nesca, 2021). This is more important in low-resource scenarios to the point that before trying to understand which model architecture could be more appropriate for a given task in a given language, ensuring quality and sufficiency of data is crucial.

Following the experiments conducted in (Artetxe et al., 2022), interesting propositions were advanced. In fact, the experiments consisted of measuring the impact of focusing on data processing in the Basque language. They first estimated with native Basque speaker the quality of three datasets (mC4, CC100 and EusCrawl) then trained different models (Topic Classification, Sentiment Analysis, Stance Detection, Named Entity Recognition and Question Answering) with the same parameters for each dataset. One of the main conclusions resulted from this study was that in language understanding on low-resource scenario, the quantity of data could be more helpful than its quality, even if this latter is a crucial feature to take into account when managing natural language.

We consider two observations (the data quality and quantity importance) when constructing Comorian datasets for the aim to manage different NLP tasks. For that we resort to different methodologies depending on the task, the nature of data and from each source the data was initially retrieved. We also investigate the effectiveness of using advance processing approaches like transfer learning from close languages and data augmentation in possible cases.

3.1 Lexicons

3.1.1 Lexicon Processing Pipeline

The pipeline described in Figure 2 aims to process a given lexicon in order to make usable in different downstream task like Sentiment Analysis (SA) and Part-Of-Speech (POS) tagging.

To enhance SA tasks, we employ pseudo-labeling using the Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto and Gilbert, 2014). This is an English lexicon-based SA model constructed using existing human-validated sentiment lexicons, to which additional lexicons used in Social Media, such as emoticons, slang, etc., were added. The annotation was done through a wisdom-of-the-crowd approach, involving human raters who rated each lexicon on a scale ranging from -4 (extremely negative) to 4 (extremely positive), with 0 indicating a neutral sentiment. The average sentiment of the words within a given text is considered as the sentiment of that text. VADER has proven to be more efficient than many state-of-the-art models.

If, instead of utilizing a lexicon with English translations, we have a different language, such as French, two approaches can be employed: adopting a method similar to VADER for this language or translating the lexicon into English and then applying VADER. For instance, when dealing with French, we opt for the latter approach due to the absence of a cost-effective solution in lexicon-based SA. Our suggestion is to leverage NLLB (Team et al., 2022) for translating French words into English. NLLB, short for No Language Left Behind, is an extensive multilingual machine translation model that supports pairs of 200 languages. We simply configure French as the input language and English as the output in its parameters.

The last module of the pipeline comes into play to complete the outputs and to enrich the dataset. In fact, at the end of the previous modules, we observe that some words are not mapped to a tag. We retrieve some of these tags using the Swahili POS dataset proposed in (Dione et al., 2023) by simply searching the non-mapped words in this dataset. Moreover, since in the dictionary names, places and punctuation are nonexistent, we add to the dataset all the corresponding entries in the Swahili dataset.

3.1.2 Bahari Foundation

We use here the ShiNgazidja-English dictionary (Thrower) written by Bahari Foundation. After transforming the PDF file into text, we apply several processing procedures. In fact, for some entries, we can found the three particularities: the existence of words variants (madjana, madjanaza, etc.), variant spellings (djando → mdjando) or implosive consonants (ɗ and ɓ). For word plurals and variant spellings, we simply consider them as new entries that taking the same translations as their associated words. As for the implosive consonants, we add new entries by just replacing them with their similar letters (ɗ → d, ɓ → b). In fact, despite the fact that these consonants are the correct spellings, they are infrequently used. For example, they are not used in the JW datasets (Section 3.2.1). We then consider the two orthographies, with and

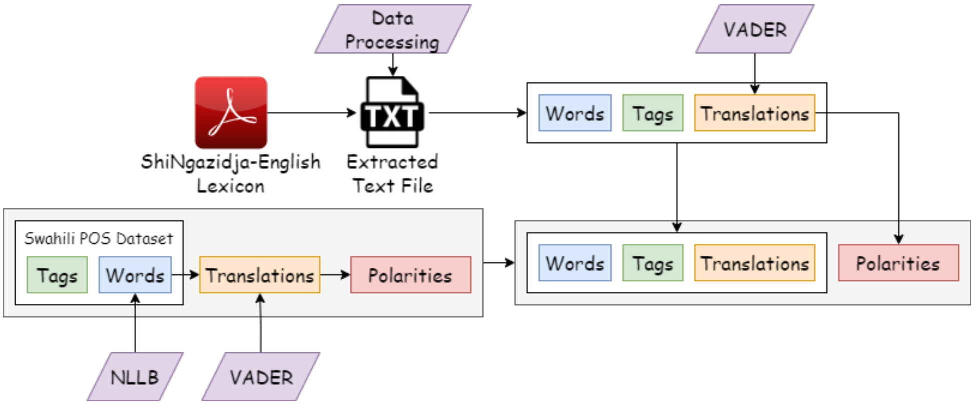

Figure 2: Lexicon Processing Pipeline.

without the implosive consonants.

Finally, in Table 1, a sample of the dictionary can be found. The lexicon contains three columns: the ShiNgazidja word, the Noun Class, and the English translation. In the latter, we can find the POS of the word such as adjective, adverb, noun, etc. We separate the POS tags from the English words, and then we apply the pipeline to these to generate the sentiments.

3.1.3 Ylangue e-langue

Ylange e-langue is an online ShiMaore-French lexicon[2]. We manually concatenate all the entries into the same text file then we apply the previous pipeline. We use NLLB to translate the French lexicon into English so that we can apply VADER and proceed to the rest of the pipeline.

3.2 Parallel Texts

3.2.1 Jehovah Witnesses

For Machine Translation, we retrieve data from the Jehovah Witnesses website[3]. ShiNgazidja and ShiMaore are one of the languages present in this platform. We can find there different PDF files containing texts in these two dialects. We can also find the French corresponding PDF (a sentence-by-sentence translation) when filtering on French data. After converting the documents into text files, we chunk the dialectal and French texts into sentences by considering dots as separators. Finally, we map each sentence to its French translation and we end with approximately 4,000 sentences for ShiNgazidja and 2,000 sentences for ShiMaore.

3.2.2 Bloom Library

The Bloom Library (Leong et al., 2022) is a multilingual dataset covering 363 languages and 32 language families. An educational web platform resulted from this initiative, in which 15 ShiNdzuani books[4] translated into English can be found. We concatenate the content of these books and we finally end with a corpus of 1,000 sentences with their translations.

3.2.3 Bible.com

The Bible.com website[5] contains all of the bible books translated into several languages including ShiMaore. In the website there are the possibility to visualize at the same time the bible translated verse by verse in two languages as we can see in Figure 3. We use Selenium[6] to perform bitext mining then we end with a total of 7,643 verses.

3.3 Speech Recognition

The Pangloss Collection[7] is a project initiated to archive speech documents in different languages with a special focus on the low-resourced ones. The initiative covers 43 countries and contains 1,120 hours of audios spread over 240 languages and dialects. The corpus contains 1h30min of ShiMaore audios and 12min of ShiNgazidja with their transcriptions. We apply a speech processing pipeline (See Fig. 4) to make the dataset easily manageable using two task: (a) audio segmentation and down-sampling. In fact, the audio transcriptions are stored in XML files with the timestamps of

[2] http://ylangue.free.fr/lexique/index-french/main.htm
[3] https://www.jw.org
[4] https://bloomlibrary.org/language:wni
[5] https://www.bible.com/
[6] https://selenium-python.readthedocs.io/
[7] https://pangloss.cnrs.fr/?mode=normal&lang=en

Table 1: ShiNgazidja-English Dictionary.

ShiNgazidja	Noun Class	English
-a ɗaima	-	Adj. eternal
-a ɗini	-	Adj. religious
-a ɗiwara	-	Adj. round
-adabisha	15	V. to correct a child, to punish
-adabishiwa	15	V. to be punished
-adhini	15	(ar.) V. to call to prayer
-adiana	15	V. to promise
-airisha	15	V. to postpone, to delay, to bargain
-alfuɓesha	15	(fr. alphabétiser) V. to teach literacy
(...)	(...)	(...)
djana (madjana)	5-6	N. one-hundred (number)
djanaza (madjanaza)	5-6	N. board for carrying dead body
djando (madjando) / mdjando (midjando)	5-6, 3-4	N. deceit

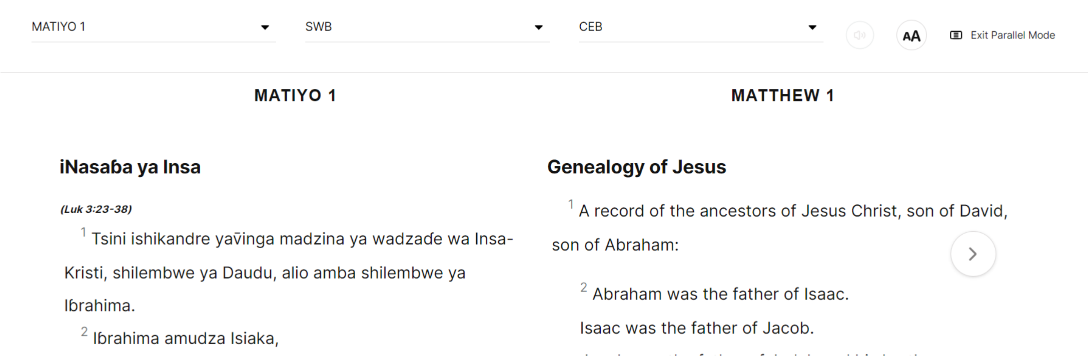

Figure 3: ShiMaore and English translations of the Bible.

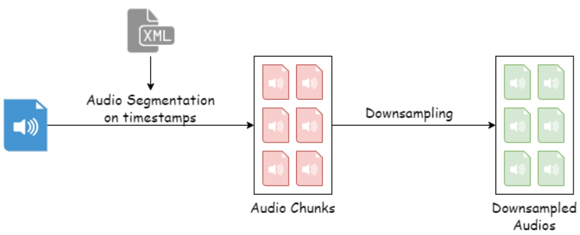

Figure 4: Speech Processing.

each sentence. We then use The AudioSegment[8] module of the Python package Pydub to segment the audios into different chunks. The final dataset contains 1.9 hours and 800 sentences.

3.4 Data-Centric Experiments

3.4.1 Sentiment Pseudo-Labeling

Here, we are proposing to create from scratch supervised datasets using transfer learning from various existing works. For that we consider the Jehovah Witnesses and Bloom Library datasets. Additionally, we employ an approach close to the one used in the lexicons to obtain the sentiments associated to each sentence using VADER. But since VADER works only on English, we translate before pseudo-labeling the French translations in the Jehovah Witnesses dataset into English.

One thing to notice here is that the choice of VADER was because of its ability to detect sentiment on single words. But when dealing with long texts, attention-based model like BERT perform generally well (Devlin et al., 2019) precisely because of its ability to understand the text. For that we use an SA fine-tuned BERT model[9] to detect the polarities of the English translated sentences. We finally consider the average sentiments between VADER and BERT for sentence labeling.

3.4.2 Audio Data Augmentation

In Automatic Speech Recognition (ASR), Data Augmentation on audio allows to better improve the models performances (Rebai et al., 2017) especially in low-resourced languages (Bartelds et al., 2023). We use the SpeechBrain toolkit (Ravanelli

[8] https://audiosegment.readthedocs.io/en/latest/audiosegment.html

[9] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

et al., 2021) to augment the speech dataset that we constructed by corrupting the audio following four steps:

- **Speed Perturbation**: We change a bit the sampling rate to make the audio a bit slower or faster than the original audio.
- **Time Dropout**: This consists of replacing random chunks within the raw waveform of audio by zeros. The idea is to allow the Neural Network to better process the data even if such information are not found.
- **Frequency Dropout**: Here, the zeros are added into the frequency domain.
- **Clipping**: It is a saturation effect that is added to the signal.

3.5 Data Availability

We leave all the datasets that we have constructed throughout this work available to the public. They can be found on Table 2.

4 Evaluations

4.1 Evaluation Metrics

We assess the Text and Token Classification model performances using these classical four metrics classification problems: Accuracy, F1-score, Recall and Precision. For ASR, we resort to the Word Error Rate (WER) and Character Error Rate (CER). WER measures the percentage of word errors in the generated transcription compared to a reference transcription. It is calculated by comparing the total number of substitutions, deletions, and insertions needed to align the generated to the reference transcriptions. CER operates similarly but measures the percentage of character errors rather than word errors. It is often used to assess the quality of transcriptions on a character-by-character basis.

4.2 Models

We conduct all the training model experiments in Google Colaboratory[10] on a machine with 12GB of RAM, powered by a Tesla T4 GPU. Since the default data storage is not persistent, we connect the environment to a Google Drive storage. During the model training process, we first shuffle the datasets (sentences, words, or audio) before splitting into training and testing sets. We then set 80% for training and we test on the remaining 20%.

[10]https://colab.research.google.com/

4.2.1 Language Identification

We use here the dataset described in 3.2.1, not for a Machine Translation task, but rather for a two-classes classification for Language Identification purpose. We compare three models: (a) mBERT, the multilingual version of BERT, trained on 104 languages and introduced in the original BERT paper (Devlin et al., 2019), (b) AfriBERTa (Ogueji et al., 2021), a model designed to understand several African languages and (c) BantuLM (Abdou Mohamed et al., 2023b), a multilingual model oriented towards Bantu languages.

On Table 3, we find the results obtained at the end of the three experiments. Indeed, we see that the BantuLM model trained specifically on Bantu languages returns better performance than the other two, especially mBERTs. This could partially confirms the hypothesis according to which the transfer of knowledge between different languages is quite important when dealing with closely related languages.

4.2.2 Sentiment Analysis

Our approach is inspired by previous work that has used language models based on BERT to improve the state of the art in SA on African languages (Martin et al., 2021; Muhammad et al., 2023). We actually apply the pseudo labeling methodology introduced in 3.4.1 on parallel corpora. We end up with 15,000 sentences and words accompanied by their polarities.

Finally, we train a multilingual SA model that enhance at the same time all the dialects. Table 4 summarizes the final results of the three approaches.

4.2.3 Part-Of-Speech Tagging

To establish the Part-Of-Speech (POS) Tagging experiment in ShiNgazidja, we use two datasets described in the previous sections: the Jehovah Witnesses sentences and the Bahari Foundation lexicon. In POS, the dataset must have several sentences with their tags. For that we use the python-Levenshtein[11] library to match the lexicon entries to each word in the sentences. In fact, the idea is to find the most similar words in the sentences to the ones in the lexicon. For the couples in which we have a mapping ratio more than 80% we attribute the lexicon tag to the word in the sentence and we attribute a default tag (*"n"*, as in *"noun"*) for the rest. The final dataset contains 23,454 tokens structured as presented in Table 5.

[11]https://pypi.org/project/python-Levenshtein/

Table 2: Data Repositories.

URL	Type	Dialects	Size
ShiNgazidja Lexicon	Lexicon	ShiNgazidja	5,714 words
ShiMaore Lexicon	Lexicon	ShiMaore	2,161 words
ShiKomori Sentiment	Raw text	ShiMaore, ShiNgazidja and ShiNdzuani	17,419 sentences+words
ShiKomori ASR	Audios+Transcriptions	ShiMaore, ShiNgazidja	1.9 hours
ShiKomori ASR Augmented	Audios+Transcriptions	ShiMaore, ShiNgazidja	9 hours

Table 3: Language Identification Results.

Model	Accuracy	F1-score	Recall	Precision
mBERT	0.940574	0.932762	0.928240	0.937825
AfriBERTa	0.945403	0.940489	0.926131	0.948652
BantuLM	**0.963798**	**0.959927**	**0.957621**	**0.962378**

Table 4: SA Results.

Model	Accuracy	F1-score	Recall	Precision
mBERT	0.7623	0.7251	0.7227	0.7388
AfriBERTa	**0.7793**	**0.7577**	**0.7580**	**0.7592**
BantuLM	0.7704	0.7366	0.7342	0.7511

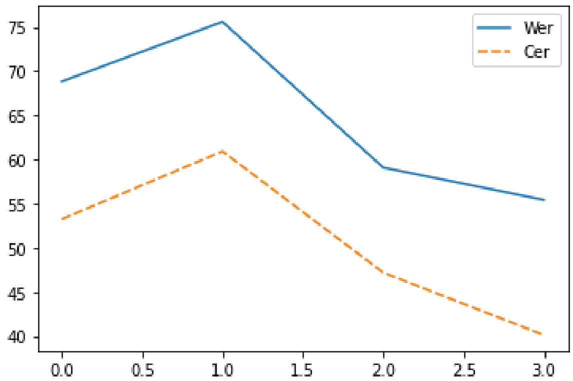

Figure 5: WER and CER Evolutions during Training.

We then fine-tune BantuLM for a Token Classification task. Table 6 resumes the results obtained throw these experiments. Here, we observe once again that AfriBERTa and BantuLM perform slightly better than mBERT. One thing to notice is that, unlike sentence classification tasks such as Language Identification, the POS tagging process depends particularly on the tokens used in the pre-training of the model. This is because of the fact that out-of-vocabulary words impacts severely the tags recognition (Horsmann and Zesch, 2016). For that, the absence of Comorian dialects in the pre-training data of the three models definitely plays a major role in the token classification. This is why in the examples presented in the Table 6 we can observe wrong words truncation.

4.2.4 Speech Recognition

For Speech Recognition, we also resort to multilingual models to leverage the state-of-the-art NLP on low-resource languages. Notable previous works have used similar approaches on several African languages (Abdou Mohamed et al., 2023a) or on specific language family like Bantu (Elamin et al., 2023). The first approach is based on Wav2vec (Babu et al., 2021), a cross-lingual pre-trained ASR model, while the second resort to Conformer (Gulati et al., 2020). These both models were initially designed like textual pre-trained language model to enhance several languages including ones that are not present on the pre-training corpus.

In our case, after proceeding to Data Augmentation, we use Whisper (Radford et al., 2022), one of the most current innovative ASR solutions. More precisely, we use *whisper-small*[12], a distilled checkpoint that has 244 millions parameters and which is trained on 680,000 hours of labeled data for the aim of enhancing multiple tasks in Speech Processing like ASR, Speech Translation or Speech Generation. Despite the fact that the checkpoint is multilingual, it was not initially trained to enhance Comorian dialects. But Swahili was in the pre-training corpus. When choosing the language on fine-tuning, we then select Swahili.

The WER and CER evolutions shown in Figure 5 are quite interesting knowing the size of the data. In fact, the high final WER of 55.42% is an expected score because of the fact that the transcribed texts are not sufficient to facilitate the model generalization. Indeed, the training dataset has a very limited vocabulary composed of only 2,216 unique words, which leads to this high score. Unlike that, we observe a low final CER of 40.11% suggesting that the model has the ability to detect the granular sounds within the audio.

[12]https://huggingface.co/openai/whisper-small

Table 5: POS Tokens Frequencies.

adv	adj	v	n	loc	prep	int	plac	conj	pron
Adverb 866	Adjective 1,788	Verb 6,669	Noun 11,161	Locution 96	Preposition 677	Interjection 9	Place 960	Conjunction 1,195	Pronoun 33

5 Conclusion and Future Work

The work presented in this article had two main objectives: contributing to the community by proposing datasets that can be used to advance the state-of-the-art on under-represented languages, particularly Comorian dialects and assessing the impact in terms of transfer learning on different pre-trained models. We evaluated the constructed datasets on four tasks: Language Identification, SA, Part-of-Speech Tagging and Speech Recognition. But before we conducted Data-Centric experiments consisting of Pseudo-Labeling and Data Augmentation respectively for SA and Speech Recognition.

Promising results have been obtained, opening the door to the representation of Comorian dialects in the field of Artificial Intelligence. However, a long way remains to be covered in order to make this representation more effective. It would be interesting in future work to experiment with other areas that we have not been able to cover due to lack of data such as Automatic Translation, Speech Generation, etc. For the downstream tasks already supported, we propose in future work to see how we could best refine them by facilitating their generalization. This could be made possible by enriching and diversifying the data or by experimenting with other models.

References

Naira Abdou Mohamed, Imade Benelallam, Anass Allak, and Kamel Gaanoun. 2023a. Multilingual speech recognition initiative for african languages.

Naira Abdou Mohamed, Imade Benelallam, Abdessalam Bahafid, and Zakarya Erraji. 2023b. Bantulm: Enhancing cross-lingual learning in the bantu language family.

Moneim Abdourahamane, Christian Boitet, Valérie Bellynck, Lingxiao Wang, and Hervé Blanchon. 2016. Construction d'un corpus parallèle français-comorien en utilisant de la TA français-swahili. In *TALAf (Traitement Automatique des Langues africaines)*, Paris, France.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages.

M. Ahmed-Chamanga. 2010. *Introduction à la grammaire structurale du comorien: Le shiNgazidja*. Number vol. 1 in Ya Mkobe publications. Komedi.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022. Does corpus quality really matter for low-resource languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Atlasocio. Classement des langues bantoues (et autres langues bantoïdes) par nombre de locuteurs. https://atlasocio.com/classements/langues/familles/classement-langues-bantoues-par-nombre-locuteurs-total-monde.php. (Accessed on 11/06/2023).

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making more of little data: Improving low-resource automatic speech recognition using data augmentation.

M.A. Chamanga. 1997. *Dictionnaire français-comorien: dialecte shindzuani*. Archipel des Comores. CEROI-INALCO.

M.A. Chamanga and Centre national de documentation et de recherche scientifique (Comoros). 2010. *Introduction à la grammaire structurale du comorien: Le shiNdzuani*. Introduction à la grammaire structurale du comorien. Komedit.

Mohamed Ahmed Chamanga. 2022. ShiKomori, the bantu language of the comoros: Status and perspectives. In *Handbook of Language Policy and Education in Countries of the Southern African Development Community (SADC)*, pages 79–98. BRILL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Cheikh M. Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane,

Table 6: Part-Of-Speech Tagging Results (n , prep , v , plac , conj).

Model	Accuracy	F1-score	Recall	Precision	Example: mister said forced the child to go to school to learn
mBERT	0.955857	0.937806	0.930366	0.945369	mze said ha trendeza ye moi na yende shi yoni ili yende ya some
AfriBERTa	**0.971986**	**0.967411**	**0.965182**	**0.969650**	mze said ha trende za ye mo ina yende shiyo ni ili yende ya some
BantuLM	0.957343	0.937451	0.924731	0.950526	mze said ha trendeza ye moin a yende shiyoni ili yende ya some

Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. Masakhapos: Part-of-speech tagging for typologically diverse african languages.

Moayad Elamin, Yonas Chanie, Paul Ewuzie, and Samuel Rutunda. 2023. Multilingual automatic speech recognition for kinyarwanda, swahili, and luganda: Advancing ASR in select east african languages. In *4th Workshop on African Natural Language Processing*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition.

Tobias Horsmann and Torsten Zesch. 2016. LTL-UDE $@$ EmpiriST 2015: Tokenization and PoS tagging of social media text. In *Proceedings of the 10th Web as Corpus Workshop*. Association for Computational Linguistics.

C. Hutto and Eric Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Michel Lafon. 2007. Le système Kamar-Eddine : une tentative originale d'écriture du comorien en graphie arabe. *Ya Mkobe*, 14-15:29–48.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gati L. Martin, Medard E. Mswahili, and Young-Seob Jeong. 2021. Sentiment classification in swahili language using multilingual bert.

Serva Maurizio and Pasquini Michele. 2021. The sabaki languages of comoros | serva | indian ocean review of science and technology (iorst). `http://www.iorst.net/index.php/paper/view/10`. (Accessed on 11/06/2023).

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Marcello Nesca. 2021. *Measuring the quality of unstructured text in routinely collected electronic health data: a review and application*. Ph.D. thesis.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matúš Pikuliak, Marián Šimko, and Mária Bieliková. 2021. Cross-lingual learning for text processing: A survey. *Expert Systems with Applications*, 165:113765.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. Speechbrain: A general-purpose speech toolkit.

Ilyes Rebai, Yessine BenAyed, Walid Mahdi, and Jean-Pierre Lorré. 2017. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 112:316–322.

C. Sacleux, M.A. Chamanga, and N.J. Gueunier. 1979. *Le dictionnaire comorien-français et français-comorien*. Number vol. 1 in Asie et Monde Insulindien Series. SELAF.

Daniel Sonntag. 2004. Assessing the quality of natural language text data. In *GI Jahrestagung*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Amirhossein Tebbifakhr, Matteo Negri, and Marco Turchi. 2020. Machine-oriented NMT adaptation for zero-shot NLP tasks: Comparing the usefulness of close and distant languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 36–46, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Joshua Thrower. Shingazidja english dictionary. `https://fr.scribd.com/document/619345660/ShiNgazidja-English-Dictionary`. (Accessed on 11/06/2023).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Worldometers. Population by country (2023) - worldometer. `https://www.worldometers.info/world-population/population-by-country/`. (Accessed on 11/06/2023).