Corpus Development of Kiswahili Speech Recognition Test and Evaluation sets: Preemptively Mitigating Demographic Bias Through Collaboration with Linguists

Kathleen Siminyu* **Kibibi Mohamed Amran**† **Abdulrahman Ndegwa Karatu**‡
Mnata Resani$^\#$ **Mwimbi Makobo Junior**$^\alpha$ **Rebecca Ryakitimbo*** **Britone Mwasaru***
*Mozilla Foundation †County Government of Mombasa
‡Hekaya Arts Initiative $^\#$Dodoma University
$^\alpha$Independent
kathleensiminyu@gmail.com

Abstract

Language technologies, particularly speech technologies, are becoming more pervasive for access to digital platforms and resources. This brings to the forefront concerns of their inclusivity, first in terms of language diversity. Additionally, research shows speech recognition to be more accurate for men than for women(Tatman, 2017) and more accurate for individuals younger than 30 years of age than those older(Sawalha and Abu Shariah, 2013). In the Global South where languages are low resource, these same issues should be taken into consideration in data collection efforts to not replicate these mistakes. It is also important to note that in varying contexts within the Global South, this work presents additional nuance and potential for bias based on accents, related dialects and variants of a language. This paper documents: i) the designing and execution of a Linguists Engagement for purposes of building an inclusive Kiswahili Speech Recognition dataset, representative of the diversity among speakers of the language, ii) the unexpected yet key learning in terms of socio-linguistcs which demonstrate the importance of multidisciplinarity in teams developing datasets and NLP technologies, iii) the creation of a test dataset intended to be used for evaluating the performance of Speech Recognition models on demographic groups that are likely to be underrepresented.

1 Introduction

Language technologies, particularly speech technologies, are becoming more pervasive for access to digital platforms and resources. This brings to the forefront concerns of their inclusivity, first in terms of language diversity. Additionally, research shows speech recognition to be more accurate for men than for women and more accurate for individuals younger than 30 years of age than those older. In the Global South where languages are low resource, these same issues should be taken into consideration in data collection efforts to not replicate these mistakes. It is also important to note that in varying contexts within the Global South, this work presents additional nuance and potential for bias based on accents, related dialects and variants of a language.

Kiswahili is a language widely spoken in East Africa and is one of the official languages of the East African Community in addition to being a national language in Tanzania, Kenya, the Democratic Republic of Congo and Uganda. Kiswahili has over 200 million speakers[1]. It is the most widely spoken African language. In 2021, Mozilla Foundation kicked off efforts to build a Kiswahili dataset on Common Voice. Common Voice(CV) (Ardila et al., 2019) is a massively multilingual speech corpus developed for Automatic Speech Recognition purposes but can be useful in other domains such as language identification. Common Voice 8[2], the latest release of CV as of February 2022, is the most diverse multilingual open speech corpus in the world. It is now 18,000 hours, and 13 million voice clips - generated entirely by 200,000+ volunteer contributors around the world.

The inclusion of Kiswahili on CV is intended to democratise and diversify voice technology. Beyond the effort to include a language community previously left out of voice technology development, we are sensitive to the fact that even among marginalised communities, there is the possibility of having subsets of the entire population excluded based on characteristics such as age, gender, accent and dialect and we are working to mitigate these possible effects from the outset. This is the main reason we sought to include linguists in the planning and development stages of our work.

This paper documents:

1. the designing and execution of a Linguists

[1] Swahili gaining popularity globally

[2] Mozilla Common Voice dataset grows by 30% and reaches 87 languages

May 26-27, 2022 ©2022 Association for Computational Linguistics

Engagement for purposes of building an inclusive Kiswahili Speech Recognition dataset, representative of the diversity among speakers of the language

2. the unexpected yet key learning in terms of socio-linguistcs which demonstrate the importance of multi-disciplinarity in teams developing datasets and NLP technologies

3. the creation of a test dataset intended to be used for evaluating the performance of Speech Recognition models on demographic groups that are likely to be underrepresented

2 Linguists Engagement

2.1 Preliminary Preparation

In order to understand how best to invite linguists' participation, we took stock of some of the things we knew, in addition to drawing up what outputs we wanted to get from the process.

There are nuanced differences that occur in speech which to a native of East Africa, hint to a speaker's ethnic background or where they have spent a considerable amount of time so as to significantly impact how they speak. While these nuances are perceptible to locals, we were interested in determining whether linguists have codified these linguistic differences and, if yes, whether these would potentially be useful labels in a speech recognition dataset.

We considered already known to us that;

- 'Standard' Kiswahili is one of several Swahili dialects which have varying levels of mutual intelligibility, therefore dialectal differences should be considered

- Speakers for whom Kiswahili is a second-language may have their pronunciations affected by their mother-tongue

- Due to the multilingual nature of different geographical contexts within East Africa, code-switching and the influence of other languages spoken has given rise to variations of the language

As output that would be useful in the context of model training and development, we wanted;

- to identify dominant Kiswahili dialects and variants, based on number of existing speakers

- to select several from among these that we would then collaboratively build word lists and sentences for, as resources demonstrative of the dialectal differences

- to identify dominant Kiswahili accents and the features demonstrative of their distinctions

We invited expressions of interest from linguists and language experts within the EA region, looking to create a team that would balance a spread of various factors;

- Demonstration of a familiarity of the content of interest to us with regards to the language

- Geographical spread of Kenya, Tanzania, the Democratic Republic of Congo and possibly the Comoros would be good to ensure we have people connected to the dialects/communities

- Gender diversity

- Their personal contributions to the Kiswahili language community

We identified and worked with a team of 4 from Kenya, Tanzania and the DRC.

2.2 Methodology

2.2.1 Discussions

We had a series of discussions which were an interactive platform where we invited thoughts and opinions from the language experts based on their expertise and experience on a variety of topics. We recorded the discussions to enable us transcribe and extract the information we needed from them. Once data collection had taken place, these discussions were also a platform via which linguists could review and validate each others' work. Each discussion had a topic shared in advance to enable participants do preliminary research and prepare their thoughts. These topics included:

- Introduction to the Common Voice project - so as to introduce linguists and language experts to our work, why their contributions are important and how we will use the outputs

- Dominant Kiswahili Dialects - What are they? Why do they differ? Geographical regions where they are spoken, estimated number of speakers and what is the level of mutual intelligibility between them

Dialect	Region(Originated)	Classification
Kimiini (Mwiini, Barawa)	Sounthern Somalia	Northern dialect
Kitikuu (Bajuni, Gunya)	Border of Kenya and Somalia	Northern dialect
Kisiu	Pate Island	Northern dialect
Kipate	South West region of Pate island	Northern dialect
Kiamu	Northern region of Lamu Island	Northern dialect
Kishela	Lamu Island	Northern dialect
Kimatondoni	Southern region of Lamu Island	
Kimvita	Mombasa and Kilifi	Central dialect
Kijomvu	Mombasa	Central dialect
Kingare	Mombasa	Central dialect
Chifundi (Kisharazi)	Mombasa and Funzi Island	Central dialect
Kivumba (Kivanga)	Border of Kenya and Tanzania	Central dialect
Kichwaka	Shimoni	
Kimtang'ata (Kimrima)	Tanga	Southern dialect
Kipemba	Pemba Island	Southern dialect
Kiunguja (Kimji)	Mjini, Zanzibar	Southern dialect
Kitumbatu	Tumbatu Island	Southern dialect
Kijambiani	Zanzibar	
Kimakunduchi (Kikae)	Southern Zanzibar	Southern dialect
Kimgao	Southern coast of Tanzania	Southern dialect
Kimwani	Northern Msumbiji	
Kingwana	Shaba province of the DRC	

Table 1: Kiswahili dialects and their regions of origin. The 13 highlighted have been most used in writing.

- Dominant Kiswahili Accents, as well as the impact of other languages spoken in Eastern Africa and their impact on the use of the Kiswahili language eg. code-switching and the borrowing of words.
- Use case resource creation
- Validation of data resources created

2.2.2 Linguists' Field Work

The team of linguists and language experts was encouraged to develop the data resources in collaboration with native speakers. They were able to reach out to individuals and hold focus group discussions with groups of people from the relevant dialects, and through these, created the word and sentence lists expected as outputs. Using common words in English as a starting point, the task at hand was for us to identify their equivalents, synonyms or perhaps translations in the various dialects and variants that we selected to work on. These words were then used as a basis for the creation of sentences, with native speakers asked to compose sentences using the words. Discussions on various topics, were also facilitated and later transcribed to create text content. The linguists used various methods of engaging with the local populations;

- Relying on their own subjective experiences having moved from various diverse linguistic spaces. This was employed particularly where the language variants were concerned, as these more commonly vary with geographic location and age
- Watching video content(eg. on YouTube) and listening to audio content that has been created in the respective dialects and variants
- Engaging everyday people belonging to the dialects in conversation, asking them questions and transcribing relevant aspects of the conversation
- Focus groups where groups of people were invited and discussion prompts used to facilitate conversations on certain topics
- The use of communication platforms to reach native speakers in instances where they were not within physical reach eg. WhatsApp

Dialect/Variant	Words and Phrases	Sentences
Kiunguja	904	206
Kitumbatu	295	143
Kiswahili Sanifu	1413	-
Kiswahili cha Bara ya Tanzania	932	205
Kiswahili cha Bara ya Kenya	1311	-
Kipemba	475	183
Kingwana	776	-
Kimvita	2589	665
Kimakunduchi	464	204
Kibajuni	1510	566
Total	10669	2172

Table 2: Resources created for the 10 dialects/variants we focused on.

3 Qualitative Results

3.1 Origin Theories of the Kiswahili Language

Kiswahili is a widely spoken language, in East Africa and beyond. The matter of its origin is still an open question with several existing theories and continues to be a topic of research. There are two main origin stories of the Kiswahili language. The first is that Kiswahili is a pidgin, or creole, of Arabic and Bantu languages and that it came about when the Arabs came to Eastern Africa(EA) for trade purposes and began interacting with locals, who were Bantu speakers in the 19th century. *Linguistic studies show that situations of contact, where two linguistic communities interact, leads to the emergence of pidgins (simplified registers) that allow the two or more distinct linguistic groups to communicate.* (Nesbitt, 2018) Further to this are theories that it is a pidgin or a mixture that includes several other languages, Portuguese, Indian and Persian, as these are some of the other nationalities that were present along the EA coast for trading purposes. The second theory states that the term 'Kiswahili' is what is of Arabic origin, while the language itself is Bantu. That when the Arabs came to EA and found those living there, along the coast, they referred to them as 'Saheel', which is Arabic for 'the coast', and that over time this term evolved to become Kiswahili for the language and Swahili(or Waswahili in plural), referencing the people. (LaViolette, 2008) Further to the claim that Kiswahili is a Bantu language, the researchers support this theory, through demonstrating that linguistic features present in Kiswahili are similar to and also present in many other Bantu languages. Evidence of Kiswahili as a Bantu language dates back to as early as the 2nd century AD in a document called 'Periplus of Erythrean Sea' written by an anonymous Greek author detailing the early expansion of Swahili civilisations towards Somalia, Kenya and Zanzibar. (Maganda and Moshi, 2014)

3.1.1 The Politics of Language

The Standard Kiswahili, or Kiswahili Sanifu, we know today was created through the standardisation of a dialect known as *Kiunguja*, which originated from the Zanzibar and Pemba Islands. In the book *'Machozi Yameniishia'*, the poet Mohammed Ghassani, is critical of the choice of Kiunguja as the basis for Kiswahili Sanifu, and many Kiswahili writers and academics share this sentiment. The process was entirely owned by colonial authorities without the involvement of native speakers. The topic of standardising Kiswahili was driven by missionary groups. On one hand were German missionaries who were keen on using the dialects from Mombasa, Pate and Tanga, which are areas where they were stationed. On the other hand were English missionaries keen on using Kiunguja only because it was the language where they were stationed, on Zanzibar and neighbouring islands. In 1930 the Inter-territorial Language Committee chose the Zanzibari Kiswahili dialect, Kiunguja, as the source of Standard Kiswahili (Thomas, 2013), a decision influenced by British colonial rule over East African territories. In his book *Decolonising the Mind: The Language of African Literature*, Ngugi wa Thiong'o talks about the fact that language is an important tool, both for the coloniser and for the colonised. The making of Kiswahili Sanifu was primarily as a tool for the coloniser, so that they could understand the thoughts of the colonised and be understood amongst them. The Inter-territorial Language Committee, to ensure the propagation of Kiswahili Sanifu, would approve textbooks used to teach the language in schools, and this committee was entirely European. Textbooks were written and reviewed by Europeans and through this vocabulary changed, with some words being shortened and completely changing their meaning(Mbaabu, 2007). Therefore the more this language was standardised, the further it drifted away from what native Kiswahili speakers knew as Kiunguja. (Mbaabu, 2007) argues that Europeans completely changed and destroyed the language. Some see the standardisation as a tool to massacre other dialects. Its use and calculated propagation is schools led to reduced use of other related dialects.

Post-independence, Kiswahili Sanifu has been used as a national language in Tanzania, Kenya, the DRC and Uganda, and even as the medium of instruction in schools in Tanzania. The language has enjoyed great government support in the region, particularly in Tanzania. One of the greatest contributions of Julius Nyerere, the first president of Tanzania, was to push for the growth of Kiswahili in East and Central Africa as he believed that it could promote African unity, as it had done in Tanzania. Kiswahili scholars in EA continue to actively grow the language with literature departments at universities and research bodies continuing to publish new editions of Kiswahili dictionaries. Language bodies such as Baraza la Kiswahili la Taifa(BAKITA) in Tanzania and Chama cha Kiswahili cha Taifa(CHAKITA) in Kenya are responsible for the promotion of the Kiswahili language and publishing houses, notably in Tanzania, contribute to a growing body of literary works in circulation in the language.

It is important for us to acknowledge this history and process of standardisation since in our work, we view Kiunguja and Kiswahili Sanifu as two different languages, despite the former being the basis of the latter. Both languages are included in the selected group of languages that we further build upon. Knowledge of this history also justifies the decision to work on dialects related to Kiswahili and to ensure they are able to benefit from the wider work done for Kiswahili Sanifu.

3.2 Kiswahili Dialects

The term dialect refers to a variety of a language that is characteristic of a particular group of the language's speakers. These differences in language use may be caused by differences in age, gender, the clan or lineage of the speakers and geographical separation or distance between the relevant groups.

There are 23 major dialects of Kiswahili. These are listed in Table 1. Of these, 13 dialects have been used widely in writing and therefore more widespread in use. These 13 are highlighted in grey on the table.

Kiswahili dialects are classified into 3 major linguistic categories, clusters which cover the EA coast from north to south. There are Northern dialects, Central dialects and Southern dialects.(Whiteley, 1993) In addition to geographic proximity, there is greater mutual intelligibility within these clusters. This classification of dialects is also indicated in Table 1.

3.3 Kiswahili Variants

Our discussions surfaced the fact that languages are in a constant state of evolution and that for this work to be relevant to current use of Kiswahili in different geographical areas, beyond seeking to be inclusive of dominant and widely spoken (historical) dialects, it was necessary to also identify variants of the language used in different locales. In this work, we use the term linguistic variants to refer to regional, social or contextual differences in the ways that a particular language is used. We identified 5 main variant clusters that are largely based on geographical regions. These are:

- Coastal Kiswahili or Kiswahili cha Pwani, referring to the EA coast where the Swahili people are from.
- Inland Kiswahili in Kenya or Kiswahili cha Bara ya Kenya
- Inland Kiswahili in Tanzania or Kiswahili cha Bara ya Tanzania
- Northern DRC Kiswahili or Kiswahili cha DRC Kaskazini
- Southern DRC Kiswahili or Kiswahili cha DRC Kusini

There are many others, and infact each of these broad categorisations could potentially be further subdivided. However due to limited time and resources, we have chosen to work with these clusters and selected Kiswahili cha Bara ya Kenya and Kiswahili cha Bara ya Tanzania to include in resource development efforts in our work at this stage.

4 Quantitative Results

Our time with the linguists and language experts involved working to develop textual data that is representative of 10 dialects and variants of Kiswahili. In comparison to the work being done for the wider Kiswahili dataset, these subsets will be significantly smaller and our intention is to have the texts and the audios collected from the respective communities, be subsets of the whole.

The dialects and variants we focused on are as listed in Table 2, in addition to the number of resources created for each. We selected these 10 in a bid to balance out several characteristics.

- it is important that the dialects and variants selected have a significant number of speakers as our work is intended to build tools of use in present day settings

- we worked to ensure national representation of dialects and variants, considering Kenya, Tanzania and the Democratic Republic of Congo

- we worked to ensure linguistic diversity by including dialects from each of the 3 major linguistic categories of the language; Northern dialects, Central dialects and Southern dialects

These subsets will have 2 main purposes.

1. to help us quantitatively evaluate how our models and downstream applications perform on related dialects and variants. We would like to work towards models with equal performance across various variant and dialect speakers, not forgetting the gender and age aspects as well, and a first step will be figuring out if there is indeed degraded performance for the different groups

2. In the event that the performance is degraded for different demographics, we would like to make resources available to developers, so that depending on the particular local contexts they are building applications for, they will be able to fine-tune so as to improve performance if necessary.

The texts have been uploaded to GitHub[3] and text as well as audio resources will be made available in future releases of the Common Voice dataset.

5 Future Work

Future work will include further expansion of these text resources, particularly at a sentence level with a target of getting 5,000 unique sentences for each of the dialects and variants of focus. We will then proceed with data collection efforts for the voice component. Beyond the dialects and variants of focus in this work, we would encourage others to replicate these efforts for those that we have not been able to focus on. Additionally, the scope of our work does not include variants that make use of code-switching, such as *sheng*, a slang common among the youth of Nairobi, Kenya that mixes Kiswahili and English, and the variant clusters in the DRC, *Kiswahili cha DRC Kaskazini* and *Kiswahili cha DRC Kusini* which mix Kiswahili and French. The inclusion of these will be key for linguistic equity moving forward.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Adria LaViolette. 2008. Swahili cosmopolitanism in africa and the indian ocean world, ad 600–1500. *Archaeologies*, 4(1):24–49.

DM Maganda and LM Moshi. 2014. The swahili people and their language: A handbook for teaching. *London: Addonis & Abbey*.

Ireri Mbaabu. 2007. *Historia ya usanifishaji wa Kiswahili*. Taasisi ya Uchunguzi wa Kiswahili, Chuo Kikuu cha Dar es Salaam.

Francis Nesbitt. 2018. Swahili creolization and post-colonial identity in east africa. In *Creolization and Pidginization in Contexts of Postcolonial Diversity*, pages 116–131. Brill.

M Sawalha and M Abu Shariah. 2013. The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds.

[3] Kiswahili Dialects Data

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing*, pages 53–59.

Jamie Arielle Thomas. 2013. *Becoming Swahili in Mexico City and Dar es Salaam: Identity in the learning of a globalized language through an African studies program*. Michigan State University.

Wilfred Howell Whiteley. 1993. *Swahili: the rise of a national language*. Gregg Revivals.