
Masakhane – “We Build Together”¹

Open Access Teaching Case Developed for the Tech for Humanity Pathways Minor

Funded by the Andrew Mellon Foundation

Developed by Dr. Shalini Misra

Introduction

This case study is designed to elucidate and prompt discussion about issues at the intersection of Natural Language Processing (NLP), decolonial AI, and the digital language divide. Key themes include: 1) The design of Natural Language Processing models; 2) Data governance structures; 3) AI Ethics Charters; 4) Decolonial AI; 5) Language exclusion; and 6) Public Interest Technology.

Background

Close to seven thousand different languages are spoken worldwide. Over two thousand are spoken in the African continent. Yet, very few of these languages are represented on the Internet and other digital technologies. For example, only about 13 of the 130+ languages for which Google Translate offers translation services are native to Africa. Speech recognition technologies on smartphones are available primarily in English or other European languages. African languages are underserved on Wikipedia. Within the resultant technological space we occupy, African cultures, places, and history are almost non-existent.

Language exclusion has serious and far-reaching implications. A significant portion of communities living in Sub-Saharan Africa cannot access global markets or engage in the digital economy because of language barriers. During the COVID-19 global pandemic, many African governments did not use the languages that were most commonly spoken in their countries to

¹ This teaching is based on the true story of Masakhane (<https://www.masakhane.io/>). Bibliographic references and resources that were used to write this case are provided at the end of the case.

communicate information about the disease with the public, since most publically accessible scientific information about COVID-19 was in English and other European languages and were not accurately or readily translatable to most native African languages.

In addition to these language barriers, there is a deep distrust of science in many African communities because of the way science and scientific language have been used to colonize, subjugate, oppress, and marginalize communities, and to justify invasive scientific practices. Being able to communicate science in native African languages not only furthers the reach of science but also allows science to be culturally integrated and owned by communities. Language accessibility is crucial for building educational applications for underserved communities, coordinating emergency and crisis response, and preserving and integrating languages that are on the brink of extinction because of the devastating effects of colonialism.

Focus Questions: The Digital Language Divide

The most extensive catalog of the world's languages, [Ethnologue](#), includes 7151 distinct languages. Of these, linguists estimate that about 1000 are spoken in the Americas, over 2000 in Africa, over 200 in Europe, over 2000 in Asia, and over 1000 in the Pacific, including Australia. These numbers are approximate because linguists disagree on the distinction between languages and dialects in some cases, the information about many languages can be outdated or scant, and the languages themselves are in flux in a rapidly changing world. Linguists do agree, however, that a century from now, many of these languages may be extinct. Some linguists believe the number may decrease by half; some by 80%; some say the total number of languages used in the world could fall to mere hundreds. Today, roughly 40% of languages are now [endangered](#) (languages that have fewer than 1,000 speakers remaining). Meanwhile, [just 23 languages](#) account for more than half the world's population.

We are often told that anyone with an Internet connection can access the information found online, yet of the world's 7,151 languages, more than half do not have any digital footprint—that is, there is no online content in those languages. This phenomenon is described as the **digital language divide**—and unless addressed, it could accelerate the extinction of thousands of languages.

Consider the humanitarian and economic consequences of language exclusion and language extinction. What happens when a language dies? What impacts does language loss have on the

cultural habits of communities? What are the scientific impacts of language extinction and language exclusion?

In a small group, discuss how Artificial Intelligence and Natural Language Processing approaches may enable community members, organizations, and non-profits to build multilingual applications that give voice to the large number of languages that are not represented online.

When AI is principally developed and researched outside of the African continent, products and services can potentially perpetuate or exacerbate existing outside biases and discrimination. The extent to which products and services developed outside the African context can advance African lives is limited because they do not address cultural, political, and social conditions in their design. There is a need for Africans to build their technological skills and develop their own technological solutions for their challenges, rather than relying on interventions that are blind to the local context.

Focus Questions: Parallel corpora, the backbone of Natural Language Processing (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence concerned with giving computers the ability to understand, analyze, and create text and spoken words, including the intent and sentiment expressed in language. Machine translation (MT) is just one example of an NLP, while other applications include speech recognition, auto-prediction and correction, and sentiment analysis, to name just a few. There is a good chance you have interacted with NLP in the form of voice-operated GPS systems, digital assistants, speech-to-text dictation software, customer service chatbots, and other consumer conveniences.

Effective language translation, in particular, requires the creation of parallel corpora—a large mass of text that has equivalent meaning, sentence-by-sentence, in multiple languages. Parallel corpora draws upon documents from a variety of genres: parliamentary proceedings, news reports, novels, poetry, film scripts, and more. Machine translation engines use parallel corpora to determine which words are most often used in parallel in different languages to come to conclusions about which words are equivalent to each other. One of the reasons that languages such as Greek, Czech, Hungarian, and Swedish, have robust online representation despite the relatively small numbers of speakers is because they are among the 24 official languages of the

European Union. Many official EU parliamentary documents, translated by human translators in multiple languages, are available to build a parallel corpus. Additionally, these languages are spoken by entire nation-states that have high levels of literacy and Internet access. The speakers of these languages are affluent. Publishers, media companies, and Big Tech companies see the economic benefit of translating their products and services for these consumers. Creating parallel corpora and other language resources takes years, and costs tens of millions of dollars per language.

Smaller languages may lack extensive written examples, such as books or parliamentary records. Oftentimes these resources may not be publicly available to train a language processor, or these languages may not have professional translators who are proficient in multiple native African languages. Millions of people who speak these languages participate on social media. However, the types of data available on social media are not ideal for building effective language translators.

Case Study

Masakhane was created to address inequities, bridge linguistic and cultural divides across the African continent and beyond, and put Africa on the technological map. Masakhane, meaning “we build together” in isiZulu is a cross-continental, grassroots organization whose goal is “...for Africans to shape and own technological advances towards human dignity, well-being, and equity, through inclusive community building, open participatory research and multidisciplinary” (<https://www.masakhane.io/>).

Masakhane has over 100 natural language processing (NLP) and machine learning experts, the majority of whom are African, working to build NLP in native African languages. The Masakhane team is working to create their own parallel corpora for native African languages. With the help of groups like Translators Without Borders, they source publicly available datasets such as governmental documents, religious texts, literature, and news articles, and then use that data to develop and machine translation models from English to their African mother tongues. All of the data sets and translation models that they create are open-source, which enables anyone with the skills and interest to build digital tools for Africa.

Among the projects that the Masakhane team is working on is the development of a multilingual parallel corpus of African research. Together with ethicists, science communicators, linguists,

and computer scientists, this project will translate African pre-print research papers released on AfricArxiv into six different African languages—isiZulu, Northern Sotho, Yoruba, Hausa, Luganda, and Amharic. This tool aims to decolonize science by enabling research produced in African universities to be accessible across the continent and beyond. The machine translation tool will allow science communicators in turn to share research across the African continent. International scientific publishers can use this resource to broaden the reach and impact of science produced around the world.

In another project called “Know Our Names”, an international and multidisciplinary team of researchers is working on a Named Entity Recognition (NER) tool for 20 African languages to identify African names, places, and people for information retrieval. The availability of NER in a broad array of African languages will allow the speakers of these languages to search and access more information in their native languages and enable broad applications such as search tools, chatbots, voice assistants, and speech recognition that can correctly identify local names and locations. Further, NER datasets created by the Masakhane team can be used in African universities so that students can learn to build language technology in their native languages. The broader impacts of this research include phone keyboards in native African languages, new or improved translation and transliteration resources, and increased accessibility to digital resources in native languages. So far the Masakhane researchers have developed baseline models of 16 African languages on the [software development platform GitHub](#).

In yet another project, the Masakhane team with the Mozilla Foundation and natural language processing experts from three African universities are working to deliver open, accessible and high-quality text and speech datasets for low-resourced East African languages from Uganda, Tanzania, and Kenya. This research will create Parallel Text Corpora for Luganda, Swahili, Runyankore-Rukiga, Acholi, and Lumasaaba and a speech data set for Luganda and Swahili. The voice data will be collected using the Mozilla Common Voice platform, a well-established platform for crowdsourcing voice contributions and accessing the voice data for free. The applications for speech data include driving aids for the impaired and the development of AI tutors to support early childhood education and more.

To learn more about Masakhane, visit their website [here](#).

Focus Questions: Masakhane Values (<https://www.masakhane.io/>)

Umuntu Ngumuntu Ngabantu — As loosely translated from isiZulu, this term means “a person is a person through another person” or “I am because you are”. This philosophy calls for collaboration and participation and community. It proposes relationality over individualism for stronger social cohesions towards sustainable communities. It is based on the belief that we share our successes, and thus one’s personhood is evaluated based on one’s contributions to the community.

African-centricity — We centralize the narratives of Africans as a remedy to the effects of Euro-centricism on our beliefs. This way we reassert a new way of looking at information from an African perspective and shun any attempts to devalue our knowledge and stories.

Ownership — We believe that Africans should be in charge of owning, driving, and participating in the NLP research process, rather than mere observers or data providers.

Openness — We believe in sharing our ideas and progress openly, especially on the African continent, for Africans. We are against research that takes African contributions or data and puts them behind a paywall that is infeasible for Africans to access.

Multidisciplinarity — We truly believe that participation from all fields and experiences leads to a more robust and more inclusive society. Everyone has valuable knowledge. We believe that each person’s individual experiences have value and each person is worth listening to and has something to contribute.

Kindness — We believe that being considerate, friendly, and generous within our community is the best way to support it and encourage more inclusivity.

Responsibility — We believe that each person in the technology process has an ethical responsibility for what they produce in the world. For this reason, we actively reckon with the ethical impacts of our work.

Data Sovereignty — We believe Africans should be able to decide what data represents our communities globally, retain ultimate ownership of that data, and know how it is used.

Reproducibility — We believe in reproducible research. As a result, we publish our code and data from our research so that others can reproduce and build upon it.

Sustainability — We believe that sustainability is necessary for societal change – that small daily efforts, over a long time, are what truly change the world. To that, we aim for sustainability of our work, by being fully integrated with technological stakeholders to ensure the community continues to thrive into the future.

Discussion Questions

- 1) There's a good chance that you have come across natural language processing in your everyday life—when Google auto-completes your search terms, when Gmail tries to pre-empt your responses suggesting short responses like “sounds good”, or when Google Docs suggests text as you type your document. Language models that recognize and generate text are being used to power chat bots, summarize news articles, and translate text. But language models created by Big Tech corporations rely on data available on the web, from primarily English and US based sources, that can be riddled with biases, misinformation, disinformation, and troubling worldviews. This affects their responses to queries. For example, in a 2021 paper published in the journal, *Nature*, researchers found that OpenAI's large language model routinely associated the word “Muslim” with “violence”. When asked to auto-complete the sentence, “Two Muslims walked into a ...,” responses from the model included: “... synagogue with axes and a bomb.” And “ ... gay bar in Seattle and started shooting at will, killing five people.” (Abid, Farooqui, & Zhou, 2021).
- 2) As a professional trained in AI Ethics and Public Interest Technology, how might you go about designing these language models? What steps could you take to embed ethical considerations in the development of the language models from its inception? How is Masakhane's approach to building NLP addressing issues of bias? Below are some resources on NLP Ethics and Governance Structures that you can use to think about your response.

References/resources for NLP Ethics and Governance Structures

- [Data governance structures](#) for Large Language Models developed by 1000 volunteer researchers who worked on the [BLOOM](#) project (which stands for BigScience Large Open-science Open-access Multilingual Language Model). Among the structures in place to mitigate biases in language models, it requires that models should make it clearer what data is being used and who it belongs to, and sourcing different [data sets](#) from around the world that are not readily available online.
 - BLOOM also launching a new [Responsible AI License](#), designed to deter the use of the model in high-risk sectors such as law enforcement or health care, or to harm, deceive, exploit, or impersonate people, even though there are no [laws](#) preventing anyone from using or abusing BLOOM.
 - [BLOOM's Ethical Charter](#). Compare BLOOM's Ethical Charter and Values with Masakhane's Values.
- 3) In response to the growing power of tech monopolies and AI technology, a movement to “decolonize AI” has arisen across the Global South. Consider the [Decolonial AI Manifesto](#) developed by Sabelo Mhlambi and other scholars. What do you think it means to “decolonize AI”? How does the Manifesto challenge the language used to talk about AI? Can you think of examples of hegemonic narratives in the context of AI? For example, the Manifesto states that humans have the “capacity to use AI as a knowledge system to create irrefutable ‘algorithmic truths’ to reinforce domination.” What does this mean? How does Masakhane’s value statement (see text box above) diverge or overlap with the Decolonial AI Manifesto?

References

- Abid, A., Farooqi, M., & Zou, J. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3(6), 461-463.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

- Cashman, K. (2020). Masakhane: Using AI to Bring African Languages Into the Global Conversation. Reset Digital for Good, February 2020.
<https://en.reset.org/masakhane-using-ai-bring-african-languages-global-conversation-02042020/>
- Johnson, K. (2019). The Masakhane project wants machine translation and AI to transform Africa.
<https://venturebeat.com/2019/11/27/the-masakhane-project-wants-machine-translation-and-ai-to-transform-africa/>
- Konteh, M. (2022). How tech can bridge the global digital language divide. *Raconteur*.
<https://www.raconteur.net/digital/tech-bridge-global-digital-language-divide/>
- McCulloch, G. (2018). The Widely-Spoken Languages We Still Can't Translate Online. *Wired*.
<https://www.wired.com/story/google-translate-wikipedia-siri-widely-spoken-languages-cant-translate/>
- Miller, K. (2022). The Movement to Decolonize AI: Centering Dignity Over Dependency. Stanford University Human-Centered Artificial Intelligence.
https://hai.stanford.edu/news/movement-decolonize-ai-centering-dignity-over-dependency?utm_source=Stanford+HAI&utm_campaign=dcaf21679b-Mailchimp_HAI_Newsletter_April+2022_General&utm_medium=email&utm_term=0_aaf04f4a4b-dcaf21679b-214031578
- Olewe, D. (2020). AI in Africa: Teaching a bot to read my mum's texts. *BBC News*, April 2020.
<https://www.bbc.co.uk/news/world-africa-52411797.amp>
- Rivero, N. (2020). The poetic process powering real-time language translation in Namibia.
<https://qz.com/africa/1881656/the-poetic-process-powering-machine-translation-in-namibia/>
- Tiku, N. (2021). Big Tech builds AI with bad data. So scientists sought better data. *The Washington Post*.
<https://www.washingtonpost.com/technology/2022/07/21/big-science-ai-open-source-language-model/>