Decolonizing NLP for “Low-resource Languages”

Applying Abebe Birhane’s Relational Ethics

By

Tolúlọpẹ́ Ògúnrẹ̀mí

Stanford University

Wilhelmina Onyothi Nekoto

Masakhane

And

Saron Samuel

Stanford University

Reflecting on the remarkable persistence of African languages, the great Kenyan writer Ngũgĩ wa Thiong'o, who formerly wrote in English and who now writes primarily in Gikuyu, asserts: “these languages have developed despite all the odds set against them by the historical experience of the plantation, the colony and the neo-colony…. Languages meant to die have simply refused to die,” (wa Thiong'o, 2005). Today African languages are spoken by more than a billion people, yet in the world of machine translation and natural language processing (NLP), these are considered “low-resource languages” (LRLs) because they lack the same level of data, linguistic resources, computerization, and researcher expertise as “high resource languages” such as French and English (Cieri, 2016). The reasons African languages remain still “low resource,” however, extend far beyond issues of data availability and instead reflect marginalization in a global society dominated by Western technology (Nekoto et al., 2020). Indeed, of the 7000 languages currently in use worldwide, over 2000 of these are African languages, yet machine

translation focuses on a mere 20 global languages (Joshi, et al. 2020). As Africans build data sets for their languages, they continue to struggle to gain agency over their own data and stories (Abebe, et al., 2021). Given the history of African colonialism and its linguistic domination, Dr. Abebe Birhane's article, "Algorithmic Injustice: A Relational Ethics Approach," (Birhane, 2020) offers an important framework for developing machine translation for "low-resource" African languages. Our response to Birhane considers the impact of NLP on Africa, and applies Birhane's ethics to support the project of decolonization of African data and data subjects.

To center African languages in NLP, especially in an era when generative models are hungry for African data, it is first important to consider the persistence of linguistic colonization long after nations have gained their sovereignty. Once colonialism punished native African language use, now it would like to monetize it to create ever greater markets for global goods. Then as now, the paradigm remains the same, controlling language that ought to belong to its linguistic communities, for whom that language also offers a road to liberation. Over thirty years ago, Ngũgĩ wa Thiong'o described education under European colonization as "foreign," with a "foreign language" so that "[t]he colonial child was made to see the world and where he stands in it as seen and defined by or reflected in the culture of the language of imposition" (wa Thiong'o, 1992). Ngũgĩ's decision to write in his native tongue, Gikuyu, endeavors to affirm his culture while offering an alternative to European language among his people. In their own languages Africans can gain greater autonomy over their own cultural production. Likewise, Martinique-born philosopher and psychologist, Frantz Fanon criticizes using the language of the colonizer, which he asserts exacerbates both the inferiority complex of the colonized and the situation of being silenced in one's own community. Instead, Fanon argues for the importance of

a shared linguistic tradition in which African members of the diaspora can envision liberation together. Many theorists of African languages imagine recovering their languages for decolonial identities. For example, Nigerian feminist writer Ifi Amadiume, argued that language before colonialism once united African communities and has the power to do so again in new ways if Africans have sovereignty over their language (Amadiume, 1987).

Considering that African languages are key to African history and identity, and that Big Tech corporations see these languages as monetizable data, Africans must gain the ability to build new NLP models and data sets from within their own communities. History has shown that when non-native speakers, no matter how well-intentioned, attempt to describe African languages, they often impose incompatible categories that erase linguistic complexities and disfigure the language. Kiswahili offers a clear example of such colonialism, where missionaries attempted to "standardize" multiple dialects into single language, ending up with a product that better reflected their perspectives than those of the speaking communities (Mazrui, 1992). Western-designed NLP replicates these same problems when trying to develop models for African languages, particularly regarding diacritics, orthography, and dialects (Adebara & Abdul-Mageed, 2022). The technology may be new, but it causes old-fashioned colonial problems and harm (Derczynski, 2022). Big Tech's models can make some sense of African languages but they will be much more accurate and equitable with African ownership over the data and say in the design (Abeba et al. 2021). When generative models use an existing model and fine-tune it on some low resource language data, they need to give those language speakers agency in the process (Mahelona et al., 2023).

Harms from misuse of African languages are well documented. NLP critics have already shown how Big Tech already scrapes unconsented data collection from platforms around the world. Emily M. Bender, Timnit Gebru, Margaret Mitchell, Angelina McMillan-Major and many others have shown that this process of data collection and performance of language models deployed in NLP cause a multitude of harms including bias, exclusion, error, and environmental harm (Bender et al. 2020). This technology further exacerbates marginalization of "low resource languages" when it collects unconsented data from Africans and excludes them from algorithmic design. Such an approach makes African language communities the "product" rather than the producers; the "researched" not the researchers. Birhane and others remain skeptical whether Big Tech is willing to grant such agency to African NLP workers and researchers. Thus far, tech corps have shown they remain less likely to benefit Africa, than "favor the needs of research communities and large firms over broader social needs," without "acknowledging critiques or alternatives," (Birhane, 2021).

Birhane, like other African decolonial NLP practitioners, see a path forward from alienation through a relational ethics which involves community ownership over models and data. The first step requires viewing data subjects as members of connected communities, reflecting on African history and interests before gathering content, data, hiring native speakers for translation, tagging, and assessing. Relational ethics, which Birhane defines as a series of "habits" that allow researchers "to rethink the nature of data science through a relational understanding of being and knowing" present an alternative. Rejecting Western concepts of individualist rationality as a definition for personhood, Birhane demonstrates how relational ethics, which link one's personhood to the personhood of others, foregrounds the "historical

injustices and the currently tangible impact of AI systems on vulnerable communities." Alongside the many efforts "to shift social reality at scale, advancing social justice by promoting linguistic justice," (Nee et al., 2021) relational ethics draws on the African tradition of decolonization and promotion of African languages because it asks those language speakers how their living language works and what ends they want it to serve. Many in the Global North following Birhane's lead, like Virginia Dignum, affirm that "rethinking AI from a relational, feminist, non-Western perspective is not a fad or a thought experiment for philosophers. It is ultimately the only way forward, for AI as a scientific field, and more importantly for all of us, and for the world," (Dignum, 2022).

The inclusive global imperative that Dignum declares can also be implemented in everyday practice. Approaching the model design and building cycle from a relational ethics perspective can help non-Africans better understand the communal bonds and linguistic cultures in Africa and support Africans in developing NLP for their own datasets. Centering African NLP workers will more effectively tackle language-specific issues, like getting the diacritics and dialects right. In the case of Yorùbá, the recognition of the importance of diacritics in the written form of the language led to work on Automatic Diacritic Restoration (ADR) which can bootstrap "standard" NLP tasks to improve results with the same amount of data due to increased quality of the data. There are several active African communities elaborating this research including Masakhane, Deep Learning Indaba, Knowledge 4 All Foundation Ltd (K4A), Zindi and ALTI. Their projects are not merely technological in practice, fixing exclusions for dialect and the like, they also tackle the more challenging task of translation between African languages. A particularly successful intervention can be seen in AI4D - African Language Dataset Challenge

which invites community production of African language datasets and submission of annotated datasets for training task-specific supervised machine learning models language. Intentional in its outreach to African language communities, this challenge provides an exemplary approach to African data as well as its critical use of available resources such as the JW300 corpus which derives from Jehovah Witness publications and contains a large collection of parallel texts which can enable cross-lingual processing between African languages. Though freely available for non-commercial use, this corpus remains limited by its largely religious content and what its authors call "known biases." (Željko Agic, & Ivan Vulic 2019). AI4D asks reviewers to consider how well JW300 compares against other sources like literature and poetry in the target languages. Such questions strongly align with relational ethics that closely examine the language as it is practiced among community members.

Likewise, Masakhane, an open-source, continent-wide, distributed, online research effort for machine translation for African languages focuses on building the community and research as well as addressing protection of data subjects and offers a highly promising alternative to corporate interventions (Orife, 2020). MasakhaNER demonstrates the progress that can be made with collaborators who are native language speakers, dataset curators, NLP practitioners, and evaluation experts from the communities (Adelani, 2021).

In all these cases, endeavors to increase access to "low-resource languages" commit to deploying machine learning technologies already notorious for as much peril as promise. Following Birhane, some African and diasporic communities may want to opt out of such projects to maintain control over their data and unique linguistic communities. Others may attempt to promote "habits" as Birhane suggests, to understand "low-resource languages" in

terms of relational ethics that prioritize community needs as they work to gain access to the technology to influence the design and better serve their communities. Those trying to improve performance of LRLs in their own linguistic and communal contexts find there are still numerous ethical issues in their best local efforts to create not just English to African translation, but multilingual translators between languages(Fan et al. 2021; Martinus & Abbott 2019) Considering "many-to-many" models that can translate directly between any pair of 100 languages and open-source a training data set that covers thousands of language directions with parallel data, created through large-scale mining, requires caution (Kreutzer, 2022). Often small datasets have produced better results for African languages than training with large high-resource datasets (Ogueji, 2021). XLM-R, which is pretrained on over 100 languages performs about as well as AfriBERTa, a multilingual language model trained on only low-resource languages with less than 1 GB of text (Oladipo, et al., 2022). MasakhaNER and AfriBERTa both aim at bringing more African languages to NLP with different approaches. MasakhaNER, which is by Africans, for Africans, is a high-quality dataset for NER in ten African languages for public use. This dataset fills the gap that is needed to bring some of the most popular African languages to NLP. AfriBERTa, trained on a small dataset that utilizes MasakhaNER, is successful at text classification and relatively competitive at NER, often outperforming mBERT and XLM-R. For low-resource languages, AfriBERTa presents a step towards more inclusive multilingual language models as well as opportunities for smaller curated datasets and reduced environmental impact. This project differs in ethos and institutional motivation from Western technology projects that attempt to carefully examine communal conversations as they build their corpora, for example, Mozilla's Common Voice project, (Ardila, 2019) which collects recordings of

African voices to better learn about dialects, genders, styles of linguistic expression. At the March 2022 Africa Women in Data Science Conference, Mozilla Foundation researcher Dr. Kathleen Siminyu described the collection methods and their attention to consent, gender, and nuance of African participants (Siminyu, 2022) The Mozilla Foundation, a non-profit, enjoys far more funding than most African originated projects, through the for-profit Mozilla Corporation and its Google search contract. That money raises questions about how to better fund independent projects like Masakhane that help decolonize NLP and ensure that "low resource language" have alternatives to corporations. As projects like MasakhaNER and AfriBERTa continue to work together, we will continue to grow representation of African languages in NLP. Most recently, in February 2023, Lelapa.AI, a socially-grounded Africa-centric AI research & product lab launched. Built with a relational mission, Lelapa.AI centers African wisdom, family, and home: "African AI talent needs a home – an opportunity to contribute to solving problems that represent their best interests, in a place that facilitates them doing their best research, without compromising the kind of remuneration warranted by the application of such critical skills. Lelapa is that home" (Lelapa 2023). Lelapa's call to action, like Masakhane's centers African communities and their well-being. When employers hope to "better resource" African languages, understanding they enter an ethically fraught terrain, they need to pay for researchers to hand curate that data and take time to thoughtfully write out and discuss datasheets and model cards. Funding Africans to shape their own futures with NLP could begin a new chapter in history for Africa– if funders adopt a relational approach and view African language workers as professionals, whose labor and data belong to their communities. Abeba Birhane provides a framework calling for Big Tech to revise its approach to African data and for Africans to retain agency over their own data in a global economy.

References:

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., & Sadagopan, S. (2021, March). Narratives and counternarratives on data sharing in Africa. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 329-341).

Adebara, I., & Abdul-Mageed, M. (2022). Towards afrocentric NLP for African languages: Where we are and where we can go. arXiv preprint arXiv:2203.08351.

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., ... & Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics, 9*, 1116-1131.

Agic, Ž., & Vulic, I. (2019). JW300: A wide-coverage parallel corpus for low-resource languages. *Association for Computational Linguistics*.

Amadiume, I. (1987). *Afrikan matriarchal foundations: The Igbo case*. Karnak House.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... & Weber, G. (2019). Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670.

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).

Birhane, A. (2021). Algorithmic injustice: a relational ethics approach. *Patterns, 2*(2), 100205.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022, June). The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 173-184).

Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016, May). Selection criteria for low resource language programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC'16) (pp. 4543-4549).

Derczynski, L., Kirk, H. R., Birhane, A., & Vidgen, B. (2022). Handling and presenting harmful text. arXiv preprint arXiv:2204.14256.

Dignum, V. (2022). Relational artificial intelligence. *arXiv preprint arXiv:2202.07446*.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., ... & Joulin, A. (2021). Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, *22*(1), 4839-4886.

Fanon, F. (2008). Black skin, white masks. Grove press, 7

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint arXiv:2004.09095.

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. arXiv preprint arXiv:1706.03872.

Kreutzer, J., Caswell, I., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., ... & Adeyemi, M. (2022). Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, *10*, 50-72.

Lelapa. AI homepage. https://lelapa.ai/about/ retrieved February 18, 2023.

Martinus, L., & Abbott, J. Z. (2019). A focus on neural machine translation for african languages. arXiv preprint arXiv:1906.05685.

Mazrui, A. (1992). Roots of Kiswahili: Colonialism, nationalism, and the dual heritage. *Ufahamu: A Journal of African Studies*, *20*(3).

Mahelona, K, Leoni, G., Duncan, S.,Thompson, M. OpenAI's Whisper is another case study in colonization, Papa Reo, January 24, 2023. Retrieved Jan 25, 2023 https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/

Nee, J., Macfarlane Smith, G., Sheares, A., & Rustagi, I. (2021). Advancing social justice through linguistic justice: Strategies for building equity fluent NLP technology. In *Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-9).

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Kolawole, T., Fagbohungbe, T., ... & Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in african languages. arXiv preprint arXiv:2010.02353.

Ogueji, K., Zhu, Y., & Lin, J. (2021, November). 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* (pp. 116-126).

Oladipo, A., Ogundepo, O., Ogueji, K., & Lin, J. (2022). An exploration of vocabulary size and transfer effects in multilingual language models for african languages. In *3rd Workshop on African Natural Language Processing.*

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K., Martinus, L., ... & Bashir, A. (2020). *Masakhane--Machine Translation For Africa*. arXiv preprint arXiv:2003.11529.

Siminyu, K. (2022). *WDS 2022 – Africa Women in Data Science* March 8-10, 2022, Retrieved December 20 2022, idia.ac.za/wds-2022/

wa Thiong'o, N. (1992). *Decolonising the mind: The politics of language in African literature*. East African Publishers.

wa Thiong'o, N. (2003). Makoni, S., Ball, A., Smitherman, G., & Spears, A. K. (Eds.). *Black linguistics: Language, society, and politics in Africa and the Americas.* Psychology Press.