Power Asymmetries, Bias, and AI: A Reflection of Society on Low-Resourced Languages – African Languages as a Case Study

Simbiat Ajao
Masakhane
University of Lagos
simbiatajao18@gmail.com

Abstract

In recent times, artificial intelligence (AI) systems have become the primary intermediary to information access, services, and opportunities. Currently, there are growing concerns as to how existing social inequalities are reproduced and amplified through AI. This is significantly evident in language technologies, where a small number of dominant languages or what we'll refer to as big languages and cultural contexts shape the training, design, and evaluation of models. This paper examines the intersections of power asymmetries, linguistic bias, and cultural representation in AI, with a major focus on African languages and communities. We argue that current Natural Language Processing (NLP) systems reflect a high level of global imbalances in the availability of data, infrastructure, and decision making power, often marginalizing low-resourced languages and cultural peculiarities. It is important we know that how these data are structured is a great determinant in what their outcome will be. With reference to examples from speech recognition, machine translation, and large language models, we highlight the social and cultural consequences of linguistic exclusion, including reduced accessibility, misinterpretation, and digital invisibility. Finally, we identify and discuss pathways toward more equitable language technologies, emphasizing community-led data practices, interdisciplinary collaboration, and context-aware evaluation frameworks. By foregrounding language as both a technical and political concern, this work advocates for African-centered approaches to NLP that promote fairness, accountability, and linguistic justice in AI development (Bender et al., 2021b) (Masakhane Community, 2020–2025).

Keywords: Power asymmetries, Bias, AI, Low-resourced, African natural language processing.

1 Introduction

African Languages continually tops the charts as to the intensity of work being done in Artificial Intelligence (AI) to drive inclusivity, particularly with the various approaches for culturally embedded or nuanced data/models. Natural Language Processing (NLP) systems increasingly shape how we, as humans, communicate, access information, and participate in digital life. From the birth of voice assistants and machine translation to automated content moderation and large language models, language technologies are no longer secondary; they are infrastructural. However, despite their global deployment, these systems are not built on equal linguistic ground.

In the context of NLP, these inequalities are especially visible in the uneven representation of the worlds languages. Although a small number of dominant languages, especially English, receive disproportionate attention in data collection, model training, and evaluation, the vast majority of the over 7,000 languages in the world remain underrepresented or completely excluded from modern language technologies (Joshi et al., 2020).

The majority of contemporary NLP models are trained on data drawn disproportionately from a small number of dominant languages, particularly English. As a result, African languages, despite representing immense linguistic diversity and millions of speakers, remain underrepresented, underperforming, or entirely absent in many AI systems. This imbalance is not merely technical; it reflects deeper asymmetries of power, knowledge production, and cultural authority.

Colonial language policies, limited research infrastructure, and the concentration of AI development in the Global North have all contributed to the marginalization of African languages in digital spaces (UNESCO, 2021)

Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), pages 235–242
March 28, 2026 ©2026 Association for Computational Linguistics

This paper argues that language bias in AI is inseparable from global power structures. We position African NLP not only as a technical challenge of low-resource languages, but as a site where questions of equity, representation, and justice must be confronted. By examining how power operates through data, models, and evaluation practices, we aim to reframe linguistic inclusion as a core requirement of ethical AI.

2 Related Work

Research on bias and inequality in artificial intelligence has expanded significantly over the past decade, with scholars documenting how AI systems reproduce and amplify existing social hierarchies. Foundational work by (Noble, 2018) demonstrates how search and ranking systems encode racial and gendered bias, while (Benjamin, 2019) situates such systems within broader structures of power, arguing that technological harms are inseparable from social context. These perspectives provide an essential grounding for understanding bias in NLP beyond purely technical metrics.

Within the NLP community, several studies have examined linguistic bias and uneven language representation. (Joshi et al., 2020) provide a comprehensive analysis of linguistic diversity in NLP, showing how benchmark datasets and research attention overwhelmingly favor a small number of high-resource languages. Their work highlights how structural research practices, not linguistic properties, drive disparities in model performance across languages. Similarly, (Bender et al., 2021b) critiques the prevailing modeling assumptions in NLP, arguing that many architectures implicitly treat English as a linguistic norm, thus marginalizing typologically diverse languages. Research focusing specifically on African contexts further underscores the role of power asymmetries. (Birhane, 2020) frames the marginalization of African languages and communities as a form of algorithmic colonization, highlighting how data extraction and model deployment often occur without a meaningful local agency. This perspective aligns with broader critiques of data colonialism articulated by (Couldry and Mejias, 2020), who argue that data-driven systems replicate historical patterns of resource extraction and control. Empirical studies have also documented the real-world consequences of linguistic bias. (Koenecke et al., 2020) show that commercial speech recognition systems perform significantly worse for non-dominant accents and dialects, illustrating how linguistic marginalization translates into measurable performance gaps. More recently, African-led NLP initiatives have begun to challenge dominant research paradigms. The Masakhane research community promotes collaborative and community-driven approaches to NLP for African languages, emphasizing capacity building and equitable knowledge production. Complementary efforts such as Lanfrica focus on surfacing and organizing resources for African languages, addressing visibility gaps in the research ecosystem. While existing work has made significant contributions to understanding bias, language diversity, and African NLP, much of the literature treats these issues in isolation. This paper builds on and connects these strands by explicitly foregrounding power as a unifying analytical lens. By situating linguistic bias within historical and institutional asymmetries, we extend prior work and argue for a more holistic, justice-oriented approach to NLP for African languages.

3 Power Asymmetries in the AI andăNLP Ecosystem

Power asymmetry in AI refers to the unequal distribution of control, influence, and benefits in the development of language technologies (Benjamin, 2019); (Birhane, 2020). AI development is characterized by a pronounced concentration of resources, expertise, and decision-making authority.

English operates as the default language of AI research, reinforcing Anglophone norms in datasets, benchmarks, and evaluation practices. Such dynamics reflect broader patterns of unequal knowledge production, where communities most affected by AI systems often have limited influence over their design and deployment (Benjamin, 2019). In the context of African languages, this results in technologies that poorly reflect local linguistic practices, including dialectal variation, tone, and code-switching.

These asymmetries are not new but are rooted in longer histories of how African knowledge, languages, and perspectives have been systematically marginalized. As (Badawi, 2024) argues, dominant global narratives have often spoken *about* Africa rather than *from* Africa, privileging external interpretations over indigenous voices and epistemologies. This historical pattern is reproduced in contemporary AI systems, where African languages and communicative practices are frequently underrepresented or framed as peripheral. Situating NLP within this broader historical context highlights that linguistic exclusion in AI is not merely a technical oversight, but part of a continuing struggle over whose knowledge is valued, preserved, and amplified in global technological systems.

3.1 Language as a Site of Power in AI Systems

Language is both a tool of communication and a repository of culture, worldview, and identity. In AI, it becomes a site where power is exercised. When AI systems privilege certain languages, they also privilege the worldviews embedded in those languages. In NLP, dominant modeling choices such as tokenization schemes, pretrained embeddings, and benchmark datasets are optimized primarily for Indo-European languages. African languages, many of which are tonal, agglutinative, or exhibit complex morphology, are poorly served by these assumptions (Bender et al., 2021b).

/ In the Yorùbá corpora sampled, this fragmentation typically results in 23Œ more tokens per sentence compared to English sentences expressing equivalent semantic content. Standard subword tokenizers, such as **BPE** and **WordPiece** disproportionately fragment Yorùbá text because they ignore morphemic structure and tone-bearing units, producing subword sequences that do not align with linguistically meaningful representations. This fragmentation collapses tone-dependent distinctions as represented in table 1, into overlapping token patterns, weakening lexical representations, and increasing semantic ambiguity in downstream tasks including machine translation and named entity recognition. At the same time, over-fragmentation inflates token counts for semantically equivalent content relative to English, leading to higher training and inference costs in transformer-based models due to the quadratic complexity of self-attention and more rapid exhaustion of fixed context windows. Consequently, token-level evaluation metrics and fixed-context benchmarks systematically penalize Yorùbá, making models appear less efficient or low-performing when the observed disparities primarily arise from tokenizerlanguage mismatch rather than inherent linguistic complexity.

As a result, African languages are often treated as deviations from a presumed linguistic norm rather than as central objects of study. This reinforces a hierarchy in which some languages are considered "standard" technologically, while others are framed as difficult, noisy, or peripheral.

4 Bias Across the NLP Pipeline

AI research and innovation are structurally biased toward actors with the greatest computational, financial, and institutional resources, leading to unequal influence over whose objectives are prioritized and who benefits the most from AI technologies (Ahmed and Wahed, 2020)

4.1 Data Bias

Training data for NLP systems are heavily skewed toward web-based sources dominated by English and other high-resource languages. Even when data on African languages exist, it is often fragmented, inconsistently annotated, or stripped of sociolinguistic context. These limitations directly affect model performance and robustness.ă

4.2 Model Bias

Models trained on skewed datasets inherit their biases. Previous work has shown that NLP systems exhibit reduced accuracy for non-dominant dialects and accents, leading to systematic performance gaps (Koenecke et al., 2020). For African users, this often results in unreliable speech recognition and culturally inappropriate text generation.

4.3 Evaluation Bias

Standard evaluation metrics prioritize surface-level accuracy while neglecting cultural appropriateness, pragmatic meaning, and code-switching features central to everyday language use across Africa. This misalignment reinforces the invisibility of African communicative practices in NLP benchmarks.

4.4 Algorithmic bias

Models amplify the patterns present in the data, often reinforcing stereotypes. Example: Speech recognition systems perform poorly for African-American Vernacular English or Nigerian English because the training data are skewed. AI bias is not accidental; it results from systemic imbalances in data collection, design, and validation.

Sentence	Tokenizer	Tokens Produced	Impact
I bought a black cap (English)	WordPiece	I / bought / a / black / cap	These are mainly whole words or sub-words with meaning
Mo ra fìlà dúdú kan	WordPiece	Mo / ra / fì / là / dú / dú / kan	Diacritics force arbitrary splits, breaking lexical units

Words	Gloss	BPE / WordPiece Output	Linguistic Implication
mlúàbí	person of good character	/ m / lú / à / bí	Morphemes fragmented; cultural concept split into ambiguated pieces
w	hand	/ w /	Loses distinction from owo (money) when diacritics are removed or mismatched

Table 1: Comparative Analysis: Subword Tokenization Effects in Yorùbá

5 Linguistic and Sociocultural Context of African Languages

Africa is the continent with the longest human history in the world (Badawi, 2024) . In this book, Badawi went on to say that history is not only about the past, it also informs our present and shapes our future, which is why the drive for a digitally and infrastructurally inclusive AI system constantly remains valid. African languages represent one of the richest and most diverse linguistic ecologies worldwide, comprising more than 2,000 languages that span multiple language families, including Niger-Congo, Afroasiatic, Nilo-Saharan, and Khoisan (Guthrie, 1967); (Heine and Nurse, 2000). This diversity is characterized not only by the number of languages, but also by the extensive variation in morphology, syntax, phonology, and pragmatics. Almost, if not all African languages are tonal, morphologically rich, and rely heavily on contextual meaning, posing distinct challenges for standard NLP pipelines developed primarily for a different specific language class or family.

5.1 Sociocultural Embeddedness of Meaning

Meaning in African languages is often shaped by cultural norms, social roles, politeness strategies, and shared communal knowledge rather than by literal semantic content alone (Bamgbose, 1991). Pragmatic features such as honorifics, kinship terms, indirectness, and metaphor are central to interpretation, but are often flattened or lost in surface-level NLP representations. Literal translations may therefore misrepresent intent, especially in culturally dense expressions such as idioms or proverbs (Bender and Koller, 2020).

5.2 Power, Representation, and Digital Marginalization

Language technologies developed without local participation risk misrepresenting speakers or excluding entire communities. Minority languages, rural dialects, and non-elite speakers are often absent from datasets, leading to AI systems that encode urban, male, and elite speech as normative. This exclusion has tangible consequences, particularly when AI systems are deployed in sensitive domains such as education, healthcare, or civic participation.

6 Empirical Case Studies and Illustrative Evidence from African NLP

6.1 Community-Driven Machine Translation: The Masakhane Paradigm

Languages and Tasks: Yorùbá-English, Hausa-English, Kiswahili-English machine translation (MT) Recent work within the Masakhane community provides a concrete illustration of how community-centered research practices improve NLP outcomes for African languages. Rather than relying exclusively on web-scraped or legacy parallel corpora often dominated by religious or colonial texts, Masakhane emphasizes participatory data creation involving native speakers, linguists, and translators. Empirically, MT systems trained on community-curated datasets demonstrate measurable improvements over baselines trained on noisier corpora. For example, Yorùbá-English systems that preserve diacritics and enforce orthographic consistency outperform non-diacritized baselines, as tone-sensitive distinctions reduce semantic ambiguity. Similarly, Hausa MT models that explicitly account for Boko and Ajami orthographies generalize better across domains than models trained on homogenized text.

This case study empirically supports the argument that data quality, linguistic fidelity, and community participation are critical determinants of model performance in low-resource settings ((Orife et al., 2020; Adelani and Alabi, 2021)

6.2 Orthography and Diacritics in Yorùbá NLP

Language and Tasks: Yorùbá - POS tagging, named entity recognition (NER), machine translation, and ASR preprocessing

Yorùbá is a tonal language in which diacritics encode essential lexical and grammatical distinctions. However, many publicly available datasets omit tone marks and underdots, collapsing distinct lexical elements into identical surface forms. This practice introduces systematic ambiguity into downstream tasks. Empirical experiments across multiple Yorùbá NLP tasks show that models trained on fully diacritized text consistently outperform those trained on stripped text. In POS tagging and NER, diacritics-aware models achieve higher accuracy and more interpretable error patterns. In MT, diacritics preservation reduces mistranslations arising from homographs that differ only in tone or vowel quality. This case highlights how seemingly minor preprocessing decisions can materially affect model performance and reinforces the need for language-specific data handling strategies rather than language-agnostic pipelines ((Alabi et al., 2020; Adelani and Alabi, 2021).

6.3 Named Entity Recognition in Morphologically Rich African Languages

Amharic (Morphological Complexity) In Amharic NER, person and location names frequently appear with attached prefixes, suffixes, and case markers. For example, the surface form (*b-Addis Ababa*, in Addis Ababa) combines a preposition (, in) with the name of the location . Standard whitespace tokenization treats this as a single token, causing baseline NER systems to miss or mislabel the entity boundary. (Gashaw et al., 2020) show that segmenting functional morphemes from named entities either through rule-based morphological preprocessing or morphology-aware tokenization leads to substantial gains in NER performance. When combined with gazetteers of Ethiopian place and personal names, F1 scores improve markedly compared to models trained on raw text, demonstrating that the error source lies in tokenization and annotation mismatch rather than model capacity.

Hausa (Honorifics and Titles) In Hausa, personal names are commonly preceded by titles and honorifics such as *Alhaji*, *Mallam*, *Dr.*, or *Sarki*. In sentences like *Alhaji Musa ya isa Abuja* (Alhaji Musa arrived in Abuja), baseline NER models often label *Alhaji Musa* as a single entity or misclassify *Alhaji* as part of the name. Empirical annotation studies reported by (Adelani and Alabi, 2021) demonstrate that explicitly separating honorifics from name spans in annotation guidelines reduces entity boundary errors and improves consistency across annotators. Models trained on such linguistically informed annotations achieve higher precision in PERSON entities, particularly in news and administrative text.

Igbo (Compounding, Prefixation, and Semantic Heads) Igbo exhibits productive nominal compounding and derivational morphology, which frequently affects named entities. Institutional and location names often include common nouns such as *l* (house), *obodo* (town/city) or *mahadum* (university) followed by a proper name, e.g., *l Akwkw Nnamdi Azikiwe* (Nnamdi Azikiwe Library) or *Obodo Owerri* (Owerri city). In standard NER pipelines, these constructions pose two challenges. First, whitespace-based tokenization fragments semantically unified entities, causing systems to label only the proper-name head (*Nnamdi Azikiwe*, *Owerri*) while excluding the institutional or locative marker. Second, English-centric annotation schemes often misclassify the common-noun component (*l*, *obodo*) as non-entity context rather than part of the named entity span. Empirical annotation studies on Igbo NER reported by (Adelani and Alabi, 2021) show that the consistency of the entity span improves when annotation guidelines explicitly treat such nounproper-name compounds as single named entities. Models trained on these revised annotations achieve higher recall for ORGANIZATION and LOCATION entities, particularly in news and educational-domain corpora, where such constructions are frequent.

Across Amharic, Hausa, and Igbo, these examples demonstrate that NER errors are not primarily due to low data volume but to misaligned linguistic assumptions embedded in standard NER pipelines. Incorporating morphology-aware tokenization, culturally grounded annotation schemes, and locally relevant gazetteers yields consistent performance gains, validating the need for language-specific NER design in African NLP.

7 Implications for African NLP Research

Understanding African languages requires moving beyond a purely technical framing of low-resource status toward a sociotechnical perspective that recognizes historical marginalization, data extraction, and uneven power relations. Ethical and effective African NLP should therefore integrate linguistic expertise, community inclusivity, and culturally grounded evaluation frameworks, ensuring that AI systems support linguistic diversity rather than erode it.

8 Re-framing African Languages Beyond the Low-Resource

Describing African languages as low-resource risks naturalizing their marginalization. These languages are not inherently under-resourced; rather, they have been systematically under-supported due to historical and political factors. Re-framing the problem shifts attention from perceived linguistic deficiency to structural inequality.

Crucially, African languages are not inherently under-resourced. Many are spoken by millions of people and have rich oral traditions, established writing systems, and long-standing histories of literacy and scholarship. Their exclusion from mainstream NLP pipelines is better understood as the result of structural neglect rather than linguistic deficiency. Colonial language policies privileged European languages in education, governance, and publishing, shaping which languages were standardized, archived, and later digitized ((UNESCO, 2021). These historical choices continue to influence the availability and research priorities of contemporary data. The low-resource label also narrows the scope of technical innovation by framing African languages primarily as problems to be solved through data augmentation or transfer learning. While such techniques are valuable, they risk reinforcing a deficit-oriented narrative in which African languages are treated as beneficiaries of models trained on dominant languages, rather than as central objects of inquiry in their own right. As (Bender et al., 2021b) argues, modeling choices in NLP often embed assumptions about linguistic structure that align poorly with typologically diverse languages, resulting in systems that perform inadequately despite technical sophistication. An alternative framing emphasizes *resource redistribution* rather than resource scarcity. From this perspective, the central challenge is not the absence of linguistic data, but the lack of sustained investment in community-driven data creation, annotation, and governance. African-led initiatives such as the Masakhane demonstrate how collaborative research models can shift this balance by centering local expertise, shared ownership, and contextual knowledge. These efforts challenge extractive research practices and highlight the importance of building long-term capacity alongside technical outputs. Re-framing African languages beyond the low-resource paradigm also has implications for evaluation. Standard benchmarks often fail to capture key features of African language use, including code-switching, dialectal variation, and culturally grounded meanings. Treating these features as noise rather than signal further marginalizes African communicative practices. More inclusive evaluation frameworks, informed by sociolinguistics and community input, are necessary to ensure that NLP systems meaningfully serve their intended users. Ultimately, moving beyond the low-resource label requires a shift in both language and practice. It calls for recognizing African languages as sites of knowledge, identity, and agency, and for designing NLP systems that reflect this reality. Such a shift aligns with broader calls for linguistic justice in AI, where fairness is understood not only as parity in performance metrics, but as equitable participation in the creation and governance of language technologies (Benjamin, 2019); (Birhane, 2020).

9 Roles of Key Stakeholders in Advancing African NLP

The development of NLP systems for African languages is shaped by the interaction of multiple stakeholders, each contributing distinct forms of expertise, resources, and constraints. Prior work suggests that progress in African NLP is most consistent when responsibilities are distributed across policy institutions, research communities, industry actors, and international funders, rather than concentrated within a single sector (Joshi et al., 2020)(Bird, 2020).

9.1 African Governments: Policy Frameworks and Public Resources

African governments influence NLP research primarily through language policy, public infrastructure, and access to state-held textual resources. Government recognition of indigenous languages in education, administration, and media contributes to the availability and standardization of written materials, which in turn affects data availability for NLP (UNESCO, 2021).

9.2 Local Research Communities: Linguistic Expertise and Task Design

Universities, research institutes, and community-led initiatives contribute linguistic knowledge, annotation expertise, and contextual understanding that are difficult to obtain through large scale automated approaches alone. Empirical evidence from community-driven projects such as Masakhane shows that datasets developed with direct involvement from native speakers and local researchers exhibit higher linguistic fidelity and clearer task definitions (Orife et al., 2020)(Adelani and Alabi, 2021). For languages such as Yorùbá and Hausa, local researchers are particularly well positioned to identify issues related to tone marking, honorific usage, orthographic variation, and code-switching phenomena that directly affect modeling and evaluation. Their role is therefore central to ensuring that research assumptions align with actual language use, rather than inferred abstractions.

9.3 Industry: Engineering Capacity and Deployment Experience

Industry actors contribute engineering expertise, computational infrastructure, and experience with large-scale deployment. These capacities are especially relevant for training and serving models that operate under real-world latency, memory, and cost constraints. In African language contexts, industry-supported systems such as speech recognition or conversational agents for Kiswahili or Hausa have demonstrated the feasibility of deploying NLP beyond research settings.

At the same time, prior studies note that industry-driven development tends to prioritize scalability and reuse, which can benefit from collaboration with linguistically informed research teams to ensure that language-specific characteristics are adequately represented (Bender et al., 2021a). Viewed in this way, industry participation complements academic and community research by translating prototypes into usable systems while benefiting from external linguistic validation.

9.4 International Funding: Enabling Long-Term Research Capacity

International funding organizations play an enabling role by supporting training, access to infrastructure, and sustained research programs, particularly in contexts where local funding for computational research is limited. Beyond individual projects, funding mechanisms influence research practices through requirements related to openness, collaboration, and capacity development (Birhane, 2023).

Evidence from African NLP initiatives suggests that funding models emphasizing local leadership, open resources, and long-term partnerships are associated with broader participation and more reusable research outputs (Adelani and Alabi, 2021). In this sense, international funders act less as directors of research agendas and more as facilitators of stable research ecosystems.

Taken together, the literature indicates that African NLP research benefits from complementary stakeholder roles: governments provide policy and public resources, local research communities contribute linguistic grounding, industry supports scaling and deployment, and international funding enable continuity and capacity building. Rather than attributing challenges to any single group, existing evidence points to coordination and alignment among these actors as a key factor in reducing structural bottlenecks related to data availability, modeling assumptions, and evaluation relevance (Joshi et al., 2020)(UNESCO, 2021).

10 Conclusion

This paper has argued that linguistic bias in AI is fundamentally shaped by power asymmetries in global AI development. African languages are marginalized not because of inherent technical limitations, but because of historical, institutional, and political inequalities embedded in the NLP ecosystem. Building fair and inclusive language technologies requires redistributing power over data, models, and evaluation to include the communities whose languages are being modeled. An ethical future for AI depends on recognizing language as a site of justice, identity, and agency. A truly intelligent AI must be multilingual, culturally grounded, and accountable to all its users.

References

David Ifeoluwa Adelani and Jesujoba Oluwadara Alabi. 2021. Deep learning approaches for low-resource african languages. In *Proceedings of the 3rd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.

Nur Ahmed and Muntasir Wahed. 2020. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *Preprint*, arXiv:2010.15581.

Jesujoba Oluwadara Alabi, David Ifeoluwa Adelani, and 1 others. 2020. Massively multilingual neural machine translation in low-resource settings. In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.

Zeinab Badawi. 2024. *An African History of Africa: From the Dawn of Humanity to Independence*. Penguin Books, London.

Ayo Bamgbose. 1991. Language and the nation : the language question in sub-saharan africa.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610623, New York, NY, USA. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abeba Birhane. 2020. Algorithmic colonization of africa. *SCRIPTed*, 17(2):389–409.

Abeba Birhane. 2023. Algorithmic colonization of africa. In *Imagining AI: How the World Sees Intelligent Machines*. Oxford University Press.

Nick Couldry and Ulises A Mejias. 2020. The costs of connection: How data are colonizing human life and appropriating it for capitalism. *Social Forces*, 99(1):e6–e6.

Ibrahim Gashaw, Solomon Teferra Abate, and Mengistu Tachbelie. 2020. Named entity recognition for amharic using deep learning. In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*, Seattle, USA. Association for Computational Linguistics.

Malcolm Guthrie. 1967. Comparative bantu: An introduction to the comparative linguistics and prehistory of the bantu languages.

B. Heine and D. Nurse. 2000. *African Languages: An Introduction*. African Languages: An Introduction. Cambridge University Press.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Masakhane Community. 2020–2025. Masakhane: Machine translation for africa. Online; accessed 2025-12-22.

S.U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Iroro Orife, David Ifeoluwa Adelani, Jade Abbott, and 1 others. 2020. Masakhane: Machine translation for africa. In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.

UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. UNESCO.