

The Igbo Language in an AI World: Assessing its Current Status

Dereck-M. A. Orji & Ijeoma Jennifer Korie

Department of Linguistics

Nnmadi Azikiwe University, Awka

Abstract

With the advent of Artificial Intelligence (AI) in these past few years, there has been a lot of revolution on various aspects of human living and culture, and an important aspect of human existence; language, has had its own share of this impact. This research paper explores and reviews the current state and prospects of the Igbo language in an AI-driven world. The Igbo language is one of the major languages spoken in Nigeria, with a rich cultural heritage and a significant number of speakers. However, like many indigenous languages around the world, Igbo faces challenges in the digital age, where dominant languages and technologies often overshadow lesser-known languages, as well as opportunities in the context of increasing technological advancements. This paper examines the impact of artificial intelligence on the Igbo language, including language preservation efforts, machine translation, natural language processing, and the role of AI in language revitalization. The paper argues that AI has the potential to revolutionize the way we interact with the Igbo language, preserving the language and making it more accessible and easier to learn. However, it also highlights the need for a balanced approach that blends technology and human preservation of language and cultural heritage. By assessing these factors, we aim to provide insights into the current situation of the Igbo language and its future in the era of AI.

Keywords: The Igbo language, Artificial Intelligence (AI), Language, Nigeria, Technology

Introduction

Language forms a pivotal part of cultural identity and continuity for communities globally. With tens of thousands of languages in existence today, linguistic diversity represents an invaluable collective heritage of humanity (Nettle & Romaine, 2000). However, in the evolving landscape of technological advancement and digital integration, numerous minority languages face the threat of extinction or endangerment (Moseley, 2010). Already, UNESCO (2022) warns that almost 43% of the estimated 6000 languages spoken today across the world are endangered. With over 520 native languages, Africa exhibits vast linguistic diversity. However, in the evolving technological landscape dominated by languages like English, Mandarin, French, Spanish and Hindi, African indigenous

languages face the risk of digital extinction. This "digital language extinction crisis" stems from the economic prioritization and computational favoring of these dominant languages within major technology ecosystems and applications (Joshi et al., 2020). The marginalization extends to major domains like education, media & entertainment, digital interfaces, search engines and voice assistants (Dearden & Rizvi, 2021).

The risks of digital extinction are especially prominent for indigenous and minority languages spoken in previously colonized regions like Africa and Asia. Of the over 2500 languages native to Africa, experts have flagged several to be vulnerable within this century if adequate digital preservation measures are not implemented for their continuity (Anderson, 2022). One such affected language is Igbo, spoken by nearly 45 million people predominantly belonging to the Igbo ethnic group of Nigeria (Ohia, 2022).

The Igbo Language in Nigeria

Nigeria, a multilingual nation prominent for her rich linguistic diversity houses about 774 ethnic groups with approximately 514 indigenous languages (Onyemelukwe, 2019). Out of these, the major languages (by the number of native speakers) are Igbo, spoken in southeastern Nigeria; Yoruba, spoken in the Western part of Nigeria; and Hausa, spoken in Northern Nigeria. The Igbo language is a tonal language belonging to the kwa group of the Niger-Congo language family. It is spoken by the Igbo people, primarily residing in southeastern Nigeria but also present in other parts of the country and in various diaspora communities. The Igbo language has about 250 dialects (Emenanjo, "Central Igbo"; Onyemelukwe, "Polyglottism for Regional" 107-108) and this variance is so much that certain dialects of this same Igbo language are unintelligible to each other, hence the Standard Igbo that serves as a central dialect for Ibo people. However, like many other indigenous languages, the Igbo language faces challenges, including a decline in the number of speakers, a lack of resources, and a lack of awareness of its importance especially with the growing and continued dominance of international languages like English (the lingua franca in Nigeria), French, Mandarin Chinese, Spanish, etc. these indigenous languages are being overshadowed.

While Igbo is recognized as one of the four largest native languages of Africa, its relative exclusion from high performing natural language processing (NLP) datasets and marginalization across digital interfaces is well-documented (Adelani et al., 2021; Ogueji & Magu, 2021). This paper seeks to conduct an in-depth assessment regarding the penetration and representation of the Igbo language across prominent AI applications, platforms and services. Through an analysis of its current computational support, inclusion within language technologies and NLP research efforts, as well as a review of digital

preservation initiatives, the overarching challenges, priorities and strategies for ensuring Igbo language vitality within an AI dominant world will be highlighted.

This paper examines the role of digital technologies and Artificial Intelligence (AI) systems in the preservation and survival of the Igbo language. We will explore various research, documentation projects, initiatives to see how the Igbo language is faring with AI, as well as the challenges/threats AI poses to the Igbo language. With the rapid growth of artificial intelligence and its increasing integration into various aspects of our lives, it is crucial to understand how the Igbo language is faring in the face of these advancements.

Thematic Review

A review of existing scholarly literature reveals the prominence of European and Asian languages in AI systems while African indigenous languages are highly under-resourced. Ogueji & Magu (2021) highlighted the lack of linguistic datasets which makes African languages almost invisible to major machine translation services. Similar marginalization exists in speech recognition tools and conversational interfaces (Ajama, 2021). However, growing research in African language technology indicates rising awareness.

The Trajectory of the Igbo Language in Nigeria

The Igbo language holds deep cultural and socio-historical significance in Nigeria. With between 24-33 million speakers concentrated mainly across South-eastern Nigeria, it is one of the four largest languages of the region alongside Yoruba, Hausa and Nigerian English (Hanani et al., 2022). The language bears traces of heavy British influence within some vocabulary and syntax due to lasting imprints of colonialism.

Post Nigeria's independence, the ethnic Igbo population faced tension and violent conflict with other linguistic groups vying for greater representation and privileges across education, governance and economy. Such divisions have influenced uneven prioritization and allocation of resources towards Igbo language development initiatives compared to more politically dominant languages like Hausa and Yoruba over the last few decades. For instance, in comparing Wikipedia article depth and page updates across key Nigerian languages, the Igbo Wikipedia currently contains just over 2000 articles while Yoruba and Hausa boast almost 15,000 and 30,000 pages respectively indicating sparse growth.

However, contemporary demands for inclusion and calls for promoting Nigeria's cultural diversity have reignited governmental and private interest in preserving and advancing minority ethnic languages facing digital extinction risks (Ugonna, 2019). This emerging focus offers scope for comprehensive inquiry assessing the trajectory of languages like Igbo

within AI systems determining global digital communication, knowledge storage and accessibility in the 21st century.

The Emergence/Design of AI Systems

In recent times, the emergence, proliferation, and advancement of digital technology and online language assistants strongly favor global languages like English, French, Chinese, Spanish, etc. While it is not very clear the datasets used to train these AI models, one thing is clear: the founders of these technologies are Europeans, not Africans and certainly, not Nigerians (Igbo). Collen Bjork (2022) expressed this concern that the system is dominated by white men and there is “a risk of (AI) making everyone sound like white men”. This invariably means that the prominent languages in these digital technologies would be/are European and/or more well-known languages, such as English, French, Chinese, German, Spanish, Russian, etc. The implication is a threat to the preservation and promotion of indigenous languages like Igbo, even though they also have a great influence on their documentation and accessibility. AI technologies have the potential to both support and threaten the survival of minority languages such as Igbo. On the one hand, AI-powered translation and natural language processing tools have made it easier to create digital content in Igbo, facilitating communication, education, and cultural preservation. These tools enable automatic translation, speech recognition, and text-to-speech synthesis, which can enhance the availability and accessibility of Igbo language resources.

On the other hand, the dominance of major languages in AI development and deployment poses challenges for Igbo. Most AI models and datasets are trained primarily on dominant languages, resulting in a lack of adequate resources and support for Igbo-specific applications. As a result, AI systems may struggle to accurately understand and generate Igbo text or speech, which hinders the development of Igbo language technologies.

Current Status of Igbo Language Technology/In AI Systems

Of over 500 languages spoken in Africa, Igbo is among the four largest languages by number of speakers (Eberhard et al., 2021). However, Ezeani et al. (2021) analyzed the support for Igbo across various AI applications and found it almost non-existent. For instance, Igbo is not supported on voice assistants like Siri, Alexa or Google Assistant. Though it exists on Google Translate, the translation quality is poor. Social media apps like Facebook and Twitter still do not allow interface localization into Igbo. This exemplifies the peripheral status of the language despite its millions of speakers (Ohia, 2022). There are some explanations for why this is the case. First, there is limited digital content as only around 0.01% of content on the internet is in Igbo (Decolonial AI Lab, 2022). This limits

data available for training AI models. This is because there are few Igbo speech and text datasets. These AI models work with data that are already available in the digital space and does not automatically generate data. For example, only 919 Igbo sentences available on Leipzig Corpora Collection as of 2022 (Goldhahn et al., 2012). Another explanation for the current status of the Igbo language in the digital space is the poor Igbo keyboard support. Igbo contains unique tones, vowels and alphabets like *ĩ* and *Ọ*, which is lacking on most keyboards. This hampers creation of digital Igbo content (Ahia, 2022).

1. Language Preservation Efforts (AI, Nlp)

The advent of AI has brought both challenges and opportunities for language preservation. Efforts for preserving and digitizing Igbo language are evident through projects like the development of the Igbo Wikipedia, Online Igbo dictionary, and Igbo language learning mobile apps (Iruoma et al., 2021). These have made the Igbo language become more accessible through digital platforms, including mobile applications, websites, and social media, thereby enabling Igbo speakers to learn, practice, and engage with their language. Datasets and machine translation models targeting the Igbo language have also emerged focusing on improving Igbo linguistic AI (Adelani et al., 2021). Additionally, initiatives such as Igbo language revitalization programs, community organizations, and educational institutions play a vital role in promoting and preserving the language. However, such initiatives currently exist as isolated projects by research groups and private organizations rather than concerted, large-scale efforts.

Machine translation and natural language processing (NLP) technologies have made significant strides in recent years. These advancements can positively impact the Igbo language by facilitating communication across different languages. However, the development of accurate and reliable Igbo language processing tools remains a challenge due to limited linguistic resources and data. Collaborative efforts between linguists, AI researchers, and Igbo speakers are essential to improve machine translation and natural language processing for Igbo. Machine translation services like Google Translate and Microsoft Translator have also expanded their language catalogue beyond just commercially prioritized languages. Google's linguistic datasets now cover over 135 languages including minority ones like Igbo, Zulu, Amharic and Khmer (Google, 2022). However, the depth of translation support varies greatly across high resource and low resources languages.

2. Initiatives for Digital Inclusion of African Languages

In response to the threats of unequal technological development posing risks of digital language extinction, several initiatives have emerged across academia, government and industry seeking to promote inclusion and survival of ethnic languages. UNESCO's official

position through its AlgoRhythm recommendation archives stresses the need to “incorporate pluralistic approaches to the relationship between AI systems and languages” and prevent marginalization. The UN body has funded research grants studying implementation frameworks prioritizing support for indigenous languages within conversational agents, recommender systems and semantic engines (UNESCO, 2021). UNESCO's official report (2021) highlighted those algorithmic systems largely exclude content in African languages. Efforts are being made through research projects like *Masakhane* (Kreutzer et al., 2022) to improve NLP support for such languages by creating annotated datasets.

Grassroots academic efforts like the *Masakhane* open-source project have focused on creating annotated datasets for lower-resourced languages in Africa to mitigate risks of their exclusion from cutting-edge NLP research which is conventionally centered on English or Chinese datasets (Neubig et al., 2022). Such pioneering initiatives underscore the need for dedicated collaborations between language communities, tech companies and universities to prevent the destructive consequences of unequal AI development resulting in digital language extinction of thousands of ethnic languages.

Governments are also launching digital inclusion initiatives like South Africa's 10-year project to uplift marginalized languages through language technologies (Pearce, 2021). Similarly, several Big Tech firms have launched localized language interfaces of their mobile applications and voice assistants targeting ethnic language speakers in emerging growth markets across Asia, Africa and Latin America. This indicates economic drivers can potentially accelerate computational support for minority languages where adequate financial incentives exist (Jia et al., 2022). Social media platforms also play a key role within public discourse and the user numbers corresponding to ethnic languages guide corporate policy around computational priorities for them (Perez, 2021).

3. Key Challenges

Funding research and development: Financial support is critical for advancing Igbo language technology (Struyf, 2022). Overcoming data scarcity: Data hungry AI applications require large amounts of quality Igbo training data (Ogueji et al., 2021). Better dataset collection and organization: Creating larger, well-organized Igbo datasets will help train AI models for translation, speech recognition etc. Also through increased digital content creation, having more online content, books, news, podcasts etc., available in Igbo will provide data for these models to work with. That way standardizing the language: Multiple dialects make standardizing Igbo spelling and speech difficult (Anyanwu, 2021). *(Get evidence that a major work has been done to have a standard dialect of Igbo).* After

this, development of bespoke keyboards that fully support the Standard Igbo alphabet will make typing and sharing Igbo content easier (Ahia, 2022).

Findings and Discussion

Through an analysis of the status of Igbo language across AI systems, applications and platforms, and considering the major challenges and prospects, some notable findings can be highlighted:

- a. Marginalization of minority languages like Igbo persists amidst unprecedented growth of AI across public and private sectors. The underlying challenge seems to stem from the precedence given to widely spoken global languages like English in building, training and deploying AI models (Joshi et al., 2020). Economic drivers also contribute to the favoring of languages with the largest set of users and consumers. Such forms of "algorithmic exclusion" which amplify existing inequalities, call for urgent policy interventions before such languages fade into silent digital extinction (Mohamed et al., 2022).
- b. Igbo is largely excluded from prominent AI applications like virtual assistants, language translation services, etc.
- c. Scarce availability of annotated Igbo datasets results in poor performance of AI algorithms for the language.
- d. With digital preservation efforts like NLLb-200, the Igbo language has a hope for the future if indigenous scholars rise up to the challenge of filling the internet massively with resources that future development in AI and digital technology can rely on. Steps need to be taken by platforms like Facebook and organizations like AFRIKALIA to improve Igbo linguistic datasets and establish language technologies like optical character recognition, predictive text and speech interfaces for supporting ongoing preservation efforts by startups.
- e. Smaller digital preservation projects for Igbo exist through individual efforts but lack organizational backing and funding which they need if it is going to be effective. Increased R&D investments combined with public awareness campaigns can also accelerate technological support for the continuity of this rich language spoken by millions globally.

Conclusion and Recommendations

In summary, this paper reveals that despite growing risks of digital language extinction and erosion due to global tech hegemony, minority languages like Igbo currently occupy sparse presence within AI applications and linguistic datasets. Preservation attempts exist but

appear fragmented, lacking coordination between public and private stakeholders. Strategic policy changes emphasizing digital inclusion of indigenous languages within AI systems seem necessary to prevent algorithmic imperialism and extinction of linguistic diversity. Some recommendations going forward include:

- (i) Mandating standardized Igbo dataset contributions from Big Tech firms operating in core languages.
- (ii) Partnerships between universities and Indigenous Language Boards for language documentation, advanced digital archiving and annotation of the Igbo language.
- (iii) Multi-sector R&D projects focusing on machine translation, predictive text and speech synthesis for marginalized languages sponsored by government STEM programs.
- (iv) Platform localization regulations requiring social media companies and voice assistant providers to support languages like Igbo in regions within their growing user bases.

Such initiatives could catalyze the continuity and survival of Igbo as a human language and within AI through enhanced representation, accessibility and engagement among current and future speakers globally.

References

- Adelani, D. I., Abbas, S., Nehemiah, Y., Atilola, O., Uchechukwu, A., Gethers, M., & Olabiyi, O. S. (2021). MasakhaNER: Named Entity Recognition for African languages. arXiv preprint arXiv:2103.11811.
- Ahia, O. (2022). Towards Igbo Language Digitalization: Igbo Keyboard Technology Concepts. *Journal of Information Technology and Digital World*, 3(2), 32-39.
- Ajama, J. (2021). Digital language vitality: Assessing the viability of indigenous languages on digital platforms. *Language Policy*, 1-19.
- Anyanwu, P. (2021). Standard Igbo Language in Formal Communication. *International Journal of Communication*, 25(1), 159-170.
- Dearden, A., & Rizvi, H. (2021). AI and African languages. *Transactions of the Royal Society A*, 379(2198), 20200341.
- Decolonial AI Lab. (2022). Assessing Igbo language data scarcity in machine learning. <https://decolonialai.lab/igbo-data-scarcity>
- Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2021). *Ethnologue: Languages of the World*. Twenty-fourth edition. Dallas, Texas: SIL International. Online version: <http://www.ethnologue.com>.
- Ezeani, I., Hepple, M., & Onyenwe, I. (2021, June). Igbo Language in the AI Age-An Audit. In 2021 International Conference on Artificial Intelligence for Industries (AI4I) (pp. 22-26). IEEE.
- Iruoma, B., Ezeani, I., & Hepple, M. (2021). Digital preservation of African languages: Igbo language case study.

- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (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* (pp. 6282-6293).
- Kreutzer, J., Daniel, F., Sibanda, N., Ungurean, C., Renduchintala, A., Ray, S., ...&Ogueji, K. (2022). Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. arXiv preprint arXiv:2208.11604.
- Mathias S. (2023). How worried are universities about ChatGPT? Retrieved on 22nd November, 2023 from How worried are universities about ChatGPT? |
- The Spinoff Mir S.A (2023). Digital Technology and AI in the Survival of Regional Languages in South Asian Countries. In book: Language and Digital Media (pp.180). Publisher: Radhika Books. https://www.researchgate.net/publication/373900112_Digital_Technology_and_AI_in_the_Survival_of_Regional_Languages_in_South_Asian_Countries
- Mohamed, S., Otterbacher, J., & Talias, M. (2022). Deconstructing "Bias" in Natural Language Processing. *ACM Transactions on Information Systems (TOIS)*, 40(3), 1-36.
- Ogueji, K., & Magu, R. (2021). Comparing European, Asian and African Language Representation in Machine Translation. In *Proceedings of the Sixth Widening Natural Language Processing Workshop* (pp. 26-36). Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Ohia, I. (2022). *Searching for Mother Tongues: Locating African Languages like Umbundu and Kikongo on the Internet*. Cambridge, UK: Open Book Publishers.
- Onyemelukwe I.F. (2019). LANGUAGE ENDANGERMENT: THE CASE OF THE IGBO LANGUAGE. *IMPACT: International Journal of Research in Humanities, Arts and Literature (IMPACT: IJRHAL)* ISSN (P): 2347-4564; ISSN (E): 2321-8878 Vol. 7, Issue 1, Jan 2019, 213-224
- Pearce, M. (2021). Developing marginalised languages in the age of globalisation and technology: the example of South Africa. *Globalisations*, 18(4), 540-558.
- Ross E.A.S. (2023). A New Frontier: AI and Ancient Language Pedagogy. Research article in *The Journal of Classics Teaching*, 24, 143–161. <https://doi.org/10.1017/S2058631023000430>. Published online by Cambridge University Press.
- UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Paris, France: UNESCO. <https://unesdoc.unesco.org/ark:/48223/pf0000381137>.
- <https://punchng.com/meta-begins-translation-of-igbo-other-african-languages/>