© IIJP 2026 | Volume 2, Issue 3, pp. 471-484 | ISSN: 2636-4484

IGBO DIALECTS AND AI; TOWARDS A DIALECT-AWARE FRAMEWORK

Ejiogu Dinah Adaoha
Benjamin Uwajumogu State College of Education, Ihitte/Uboma, Nigeria
adaohadinah@gmail.com

ARTICLE INFO
Article No.: 0451
Accepted Date: 10/07/2026
Published Date: 29/07/2026
Type: Research

ABSTRACT

Artificial intelligence (AI) systems developed for Igbo are trained mainly on Standard (Central) Igbo, although everyday communication occurs through numerous regional varieties. This paper examines how that standardization bias affects the digital vitality and representation of Igbo dialects. It uses a structured descriptive-theoretical review of 18 sources on Igbo dialectology, language ecology, digital vitality, natural language processing, machine translation, and community data governance. The sources are analyzed thematically against four research questions concerning dialect documentation, available digital resources, system performance, and strategies for dialect-aware development. The review finds that Igbo dialects are extensively described in print scholarship but remain poorly represented in machine-readable corpora, speech datasets, input tools, and automated processing systems. IgboAPI provides an important partial lexical resource, yet no complete publicly accessible text-and-speech corpus exists for any dialect. Evidence reviewed also indicate that systems trainedon Standard Igbo perform less effectively on dialectal input, thereby limiting access for dialect speakers. In response, the paper proposes a four-tier framework that combines dialectological classification, language ecology, digital vitality indicators, community-led data collection, and isomorphism-guided transfer learning. The framework prioritizes practical development while retaining every documented variety within the long-term planning horizon. The study concludes that inclusive Igbo AI requires coordinated resource creation, transparent performance testing, and community control over linguistic data.

Keywords: Igbo dialects; dialectal NLP; low-resource languages; artificial intelligence; language development.

Introduction

Language carries communication, cultural knowledge, identity, and collective memory. In south-eastern Nigeria, Igbo is best understood as a cluster of related regional varieties rather than a single homogeneous code. As AI increasingly mediates reading, writing, speech, translation, and information access, the varieties represented in digital systems acquire greater practical and symbolic value. Dialectological research demonstrates substantial variation across Igbo-speaking communities. Nwaozuzu (2008), for example, documents differences among varieties associated with areas such as Mbieri, Owerri, Onitsha, Nsukka, and Afikpo. Speakers may share broad mutual intelligibility while differing in lexical choices, tone patterns, phonological processes, and grammatical preferences. These documented contrasts provide a formal linguistic basis for examining dialect representation in AI. A dialect is a regional or social variety that differs from related varieties in pronunciation, vocabulary, grammar, and usage while remaining part of a broadly shared language system (Chambers &Trudgill, 1998). Igbo dialects have been classified into several clusters, although scholars differ on the precise number and boundaries of those clusters (Ikekeọnwụ, 2001; Nwaozuzu, 2008). This classificatory variation reflects the complexity of the dialect continuum rather than an absence of systematic structure. Standard Igbo is the codified variety used extensively in education, formal writing, broadcasting, and public communication. Its institutional role supports literacy and interregional communication, but it is not the home variety of many speakers (Emenanjo, 1996). Chinagọrọm (2010) therefore argues that the internal structural relations among dialects, described through linguistic isomorphism, must be considered when explaining the organization of Igbo variation.

AI systems learn linguistic patterns from the data used to train and evaluate them. A system trained predominantly on Standard Igbo is therefore likely to represent that variety more effectively than underrepresented dialects. Emezue et al. (2024) show the importance of dialect-labelled lexical resources and report performance limitations when language technologies encounter dialectal forms that are poorly represented in available datasets. This imbalance has practical consequences. When a translation, speech, or text-processing system does not recognize a speaker's home variety, the speaker may shift towards Standard Igbo or English in digital communication. Such pressure extends the long-standing tension between the scalable benefits of standardization and the exclusionary effects that may arise when one variety is treated as the language as a whole (Emenanjo, 1996; Nwaozuzu, 2008).

The issue also has a sociolinguistic dimension. Public attitudes may assign greater prestige to standardized forms and treat rural or local varieties as informal or deficient. Recent institutional discussion has similarly linked indigenous-language development with social attitudes in the AI era (University of Lagos, 2026). AI systems that recognize only Standard Igbo may reproduce this hierarchy within an increasingly influential communicative domain.

Statement of the Problem

The significant unresolved problem addressed in this study is that AI tools for Igbo are developed mainly around Standard Igbo, while the language's dialectal diversity remains weakly represented in digital and computational resources. Although Igbo dialectology and computational linguistics have each produced important scholarship, their intersection has not been synthesized sufficiently to show how dialect documentation can inform inclusive language technology. Three related knowledge gaps follow. First, the digital and AI-ready resources available for documented Igbo dialects have not been systematically inventoried. Second, evidence on the performance of existing systems across dialectal varieties remains fragmented. Third, there is no established development framework that balances the practical value of a

standard variety with the need to represent dialectal diversity. Without an integrated account, decisions about data collection, model design, evaluation, and funding may continue to privilege Standard Igbo by default.

The consequences extend beyond technical performance. Obiwuru (2025) argues that AI translation systems may reproduce existing linguistic inequalities when they exclude varieties associated with rural, older, or less formally educated speakers. This paper addresses the three gaps through a structured review of dialectological scholarship, available language resources, reported system performance, and relevant theoretical approaches.

Research Questions

The following research questions guide the study:

1. What is the documented dialectal landscape of Igbo, and which dialects have been described in the linguistic literature?
2. What digital and AI resources currently exist for Igbo dialects?
3. What performance gaps have been documented when existing AI systems process Igbo dialectal varieties?
4. What theoretical frameworks and practical strategies can support dialect-aware AI development for Igbo in Africa?

Objectives of the Study

The study pursues the following objectives:

1. To examine major dialectological descriptions and classifications of Igbo dialects.
2. To identify available digital resources for Igbo dialects and assess their readiness for AI development.
3. To analyze documented performance gaps in existing AI systems when they process Igbo dialectal variation.
4. To propose a dialect-aware framework that combines practical standardization with inclusive representation.

Theoretical Framework

The study combines three complementary perspectives: dialectology and linguistic isomorphism, language ecology, and digital language vitality.

1 Dialectology and Language Ecology

Dialectology describes and maps regional linguistic variation (Chambers &Trudgill, 1998). For Igbo, Nwaozuzu (2008) provides the most extensive dialect inventory, while Ikekeọnwụ (2001) offers an alternative classificatory account. Chinagọrọm (2010) develops linguistic isomorphism to explain structural relationships among the Nsḳụka and Owerri clusters. Together, these approaches identify the varieties to be represented and indicate which structurally related dialects may share computational resources. Language ecology examines the social environments in which a language or dialect survives, changes, expands, or declines (Haugen, 1972; Mühlhäusler, 1996). A dialect remains stronger when it is used across domains such as the home, education, work, media, and technology and when speakers maintain positive attitudes towards it. AI now forms part of this ecology. Persistent absence from AI-supported communication may narrow a dialect's functional range and increase pressure towards Standard Igbo or English.

2 Digital Language Vitality

UNESCO's (2003) framework treats language vitality as multidimensional, including intergenerational transmission, institutional support, and media use. Bird (2022) extends the

discussion to digital participation. For the purposes of this study, digital vitality at dialect level is assessed through four practical criteria:

a) the presence of digital texts and corpora;
b) the availability of input tools, including keyboards and speech interfaces;
c) the availability of automated processing tools, including machine translation, speech recognition, text-to-speech, and spelling support; and
d) active use in social-media and other digital communication.

A dialect that has meaningful resources in these four areas is treated as digitally vital. The criteria are measurable and action-oriented because they show where data, tools, or community participation remain absent. The findings in this paper indicate that documented Igbo dialects have very limited formal resources in the first three areas, although some dialectal use occurs informally on social media.

The three perspectives form a unified analytical framework. Dialectology supplies the descriptive inventory and structural relationships; language ecology explains the consequences of inclusion or exclusion across communicative domains; and digital vitality provides criteria for assessing present resources and future progress. Their combined use links each research question to both linguistic description and technological planning.

Review of Related Literature

The literature relevant to this study falls into three intersecting areas: Igbo dialect description, computational treatment of Igbo, and theories of language vitality, ecology, and AI equity. Rather than presenting the sources as isolated summaries, this section compares their assumptions, contributions, and limitations in relation to the research questions.

1 The Dialectal Landscape of Igbo: What Has Been Documented

Nwaozuzu (2008) offers the broadest inventory of Igbo dialects, describing more than 40 varieties within eight major clusters. Ikekeọnwụ (2001) and Ezikeojiakụ (1998), by contrast, use six-group classifications. The difference does not imply that one account is unusable; it shows that cluster boundaries depend on the linguistic features and level of internal variation selected for analysis. For AI resource planning, Nwaozuzu provides wider coverage, while the alternative classifications help identify areas in which taxonomy remains contested. Emenanjo (1996) concentrates more directly on phonological and lexical variation. Forms such as gịnị in Standard Igbo and nịnị in Mbieri or Onitsha illustrate that variation is systematic rather than random. Unlike a broad cluster inventory, such feature-level description indicates the specific contrasts that lexicons, speech models, and translation systems must learn. Classification identifies where variation occurs; feature analysis explains what a model must represent.

Chinagọrọm (2010) moves beyond both inventory and isolated feature comparison by emphasizing linguistic isomorphism between the Nsụkka and Owerri clusters. Whereas Nwaozuzu and Ikekeọnwụ primarily classify varieties, Chinagọrọm foregrounds structural relationships among them. This relational approach is especially relevant to transfer learning because data from one well-resourced variety may support adaptation to a structurally related variety, provided the relationship is empirically tested rather than assumed.

Taken together, the dialectological studies offer substantial knowledge of Igbo phonology, lexicon, morphology, and classification. Their principal limitation for AI development is not descriptive weakness but format: much of the knowledge remains in print-oriented scholarship rather than machine-readable, dialect-labelled datasets. The literature therefore reveals a conversion gap between linguistic documentation and computational infrastructure.

2 The Standardization Debate in Igbo Linguistics

Emenanjo (1996) presents standardization as necessary for coordinated literacy, education, orthography, and wider communication. From this position, Standard Igbo supplies a shared target around which dictionaries, curricula, keyboards, and language technologies can be developed efficiently. The argument is strongest where limited resources make immediate support for every variety impracticable.

Nwaozuzu (2008) and Ikekeọnwụ (2001), however, demonstrate that the standard does not exhaust the linguistic reality of Igbo speakers. Their dialect classifications make visible the varieties that standardization may marginalize. Unlike Emenanjo's emphasis on coordination, their work draws attention to representational costs, especially when speakers are encouraged to regard home varieties as deficient or non-prestigious.

AI intensifies this debate because digital systems do more than provide a common written form: they determine which input is recognized, translated, corrected, or synthesized. Obiwuru (2025) argues that systems centered on Standard Igbo may reproduce existing inequalities affecting rural and less formally educated speakers. Standardization may therefore improve scalability while simultaneously restricting participation when dialectal performance is not measured or addressed.

The present study adopts a qualified position. It accepts Standard Igbo as a practical computational foundation but rejects its treatment as the only legitimate target. The proposed tiered framework therefore retains common infrastructure while requiring dialect-labelled data, cross-dialect evaluation, transfer learning, and long-term coverage of smaller varieties.

3 Digital Resources for Igbo Dialects: The Current Landscape

The surveyed literature indicates a marked imbalance between resources for Standard Igbo and those for dialectal varieties. Published NLP work commonly uses “Igbo” as an operational label for data that is predominantly standardized, leaving the dialectal composition of training and evaluation sets unclear.

Inuwa-Dutse (2025) reports that only 25.1% of the Nigerian-language NLP studies reviewed introduced new linguistic resources, while 74.9% reused existing resources. This pattern is especially restrictive for dialect research because reuse repeatedly concentrates development around the same standard-oriented datasets. The survey establishes the broader scarcity of resource creation, while the present study identifies dialect labelling as an additional, under-examined deficit.

Emezue et al. (2024) partly address this deficit through IgboAPI, a multi-dialectal Igbo-English lexical dataset developed with dialect experts. Its dialect labels and lexical coverage make it more useful for dialect-aware work than standard-only wordlists. However, a dictionary records lexical correspondences rather than extended language use. Sentence corpora, conversational data, and speech recordings are still required for models to learn syntax, discourse, pronunciation, and contextual meaning.

Speech resources are more limited. The reviewed record contains no complete, publicly accessible dialect-specific corpus for automatic speech recognition or text-to-speech, and no established dialect-sensitive keyboard, spelling checker, or voice-input system. Under the four digital-vitality criteria adapted from Bird (2022), formal dialect support is therefore minimal.

Social media presents a contrasting pattern. Content creators and community members use dialectal forms in informal digital communication even where formal datasets and tools are absent. This activity demonstrates continuing social vitality and offers potential source material,

but its reuse for AI requires informed consent, careful dialect labelling, privacy protection, and community agreements on ownership and benefit.

4 AI Performance Across Igbo Dialectal Variation

Evidence on cross-dialect performance remains limited, but the available studies point in a consistent direction. Emezue et al. (2024) report that systems developed primarily with Standard Igbo data perform less effectively on dialectal inputs. The limitation is therefore a data-distribution problem: model performance declines when the input differs systematically from the variety represented during training.

Comparative African-language research reports similar difficulties with non-standard varieties (Adelani et al., 2022). These comparisons support the inference that the Igbo case is part of a wider problem in low-resource language technology. However, Igbo requires its own dialect-labelled benchmarks because performance patterns observed in Yorùbá or Hausa cannot be assumed to transfer directly to Igbo.

Tone and diacritics add another layer of complexity. Research on neural tone prediction for Igbo shows that tonal marking is a computational task in its own right (Onyenwe et al., 2020). Removing tone marks may simplify surface processing, but it does not resolve dialectal data mismatch and may erase contrasts required for accurate interpretation. Dialect-aware systems therefore need appropriately annotated data rather than reduced linguistic representation.

Obiwuru (2025) situates these technical limitations within a social analysis. When poor performance is concentrated among rural, older, or less formally educated dialect speakers, the system reproduces pre-existing inequalities. The literature therefore converges on two related gaps: insufficient dialect data and insufficient attention to the distribution of technological benefits and errors.

5 Dialect-Aware AI: Theoretical Proposals and Emerging Approaches

The technical literature proposes several routes towards dialect-aware AI. Ezulu (2025) outlines an architecture that identifies a dialect from text or speech and then routes the input to a variety-specific model or an adapted shared model. The proposal addresses model selection and adaptation more directly than the dialectological literature, which was not designed for computation.

Its central limitation is data dependence. Dialect identification and adaptive routing require labeled examples for training and evaluation, yet the resource review shows that such data remain scarce. Ezulu's model-level proposal therefore cannot be implemented at scale without prior corpus development. In this respect, technical design has advanced further than foundational resource creation.

Chinagọrọm's (2010) linguistic isomorphism offers a possible response to this constraint. If structural relationships among dialects are mapped carefully, a model trained on a better-resourced variety can serve as the starting point for a related under-resourced variety. This approach does not remove the need for local data; it reduces the amount required for adaptation and directs collection towards the features that differ most.

Technical efficiency must also be governed ethically. Dialect classification may flatten local distinctions if speakers are assigned automatically to the “closest” available model. Obiwuru (2025) further warns against extractive data practices in which communities provide linguistic material without control or benefit. A viable framework must therefore combine architecture, corpus development, community validation, data ownership, and transparent limits on model use.

Methodology

The study uses a structured descriptive-theoretical review rather than fieldwork or experimental model evaluation. It synthesizes secondary sources to connect Igbo dialectology with computational language research and to develop an analytical framework for dialect-aware AI. The review covers publications from 1972 to February 2026. The starting date includes foundational language-ecology theory, while the end date captures the most recent sources available when the manuscript was revised. Searches were conducted in Google Scholar, ACL Anthology, arXiv, LREC proceedings, and relevant institutional repositories. The exact search strings were "Igbo dialects", "Igbo dialect classification", "Igbo NLP", "Igbo machine translation", "Igbo speech recognition", "Igbo tone prediction", "dialect-aware Igbo AI", "low-resource African language AI", "digital language vitality", and "language ecology and AI". Reference lists of relevant sources were also checked for earlier dialectological and theoretical works. Sources were included when they: (a) described or classified Igbo dialects; (b) documented Igbo text, speech, lexical, or input resources; (c) reported AI or NLP performance relevant to Igbo or comparable African-language variation; or (d) provided a theoretical or ethical framework directly applicable to dialect vitality and community data governance. Duplicate records, items without substantive relevance to dialect or language technology, and unsupported opinion pieces were excluded. The verifiable final analytical corpus comprised the 18 sources listed in the reference section. The original search notes did not retain reliable initial retrieval and screening totals; the study therefore reports the confirmed final corpus rather than reconstructing unverified counts.

Analysis proceeded in two stages. First, the sources were coded thematically under four categories corresponding to the research questions: dialect documentation, digital resources, system performance, and dialect-aware strategies. Second, each documented cluster was assessed against the four digital-vitality criteria adapted from Bird (2022). The dialect classifications of Nwaozuzu (2008), Ikekeọnwụ (2001), and Chinagọrọm (2010) were then compared to identify coverage gaps and possible structural relationships for transfer learning. The tiered framework in Section 4.4 was developed from this synthesis. The method has two limitations. It depends on published and publicly described sources, so unpublished community archives and ongoing projects may be absent. It also does not test a new model or independently verify every reported performance result. The findings should therefore be interpreted as an account of the documented evidence and as a framework for subsequent empirical research.

Findings and Discussion

Finding 1: The Documented Dialectal Landscape of Igbo

The published record contains a substantial but uneven map of Igbo dialectal variation. Table 1 reorganizes the major documented clusters, representative varieties, and distinguishing features reported principally by Nwaozuzu (2008), with supporting classifications from Ikekeọnwụ (2001) and Chinagọrọm (2010).

Table 1: Documented Igbo Dialect Clusters

Cluster	Representative dialects	Key distinguishing features	Primary source
North-Eastern	Izzi, Ezza, Igbo-Eze, Oshiri	Distinct phoneme inventories and lexical items	Nwaozuzu (2008)
East-Central	Mbieri, Owerri, Oguta, Izombe, Mbaise, Mbano, Okigwe, Isu, Orlu, Mbaitoli	Consonantal features distinct from Standard Igbo; partial basis of the codified standard	Nwaozuzu (2008)
Western Niger	Agbor, Ogwashi-Ukwu, Ibusa	Contact influence from Edo languages; distinctive negation and consonantal variants	Nwaozuzu (2008)
East Niger	Ihiala, Ọgbaru, Agụata, Nnewi, Nsugbe, Agụleri, Achịna	Eight- to ten-vowel systems and distinctive morphological patterning	Nwaozuzu (2008)
Cross-River	Amamgbelu, Ndịokpo, Obitoo, Atani, Ubụrụ, Ọhafịa, Abịrịba, Ebem	Nine-vowel systems in some varieties; contact features near the Igbo-Efik boundary	Nwaozuzu (2008)
South-Western	Etche, Ahụda, Tai, Andọni, Opobo, Ikwerre	Distinct negative markers (lam, lem) and contact influence from Niger Delta languages	Nwaozuzu (2008)
South-Eastern	Ngwa, Ụmụosu, Nsulu, Uhie, Ụmụmpkei	Distinct tonal patterns and morphophonological rules	Nwaozuzu (2008)
Northern	Isakpụ, Iheagụ, Ozalla, Ogboli, Ngwo, Nenwe, Ụzo-Uwani	Predominantly nine-vowel systems, including schwa in some varieties	Nwaozuzu (2008)

Three observations follow from the inventory. First, documentation is uneven: some clusters have detailed phonological and grammatical descriptions, while others are represented by brief lists or isolated features. Second, descriptive coverage does not amount to computational readiness because the materials are rarely digitized, dialect-labelled, or structured for model training. Third, Chinagọrọm's (2010) isomorphism framework offers a basis for testing which related clusters may share representations or adaptation strategies, although such transfer must be validated with local data.

Research Question 1 is therefore answered in two parts. Igbo has a well-established broad dialectal landscape in the linguistic literature, but the depth of description varies considerably and no cluster is comprehensively represented in machine-readable, AI-ready form.

Finding 2: Digital and AI Resources for Igbo Dialects

The resource survey shows that no documented Igbo dialect has complete support across the four digital-vitality categories adapted from Bird (2022). Table 2 distinguishes the comparatively stronger position of Standard Igbo from the limited formal support available to dialectal varieties.

Table 2: Digital Vitality Status of Igbo Dialects

Vitality criterion	Standard Igbo status	Dialect status
Presence in digital texts	Substantial: multiple corpora and parallel texts	Minimal: mainly informal social-media use
Input tools: voice and keyboard	Partial: some voice tools and keyboard layouts	No established dialect-specific tools identified
Automated processing: MT, ASR, TTS	Partial and growing	No complete publicly accessible dialect-specific systems identified
Active use on social media	Moderate and growing	Active informally among creators and communities
Lexical datasets	Multiple dictionaries and wordlists	IgboAPI provides partial multi-dialectal coverage
Speech and audio corpora	Some standard-oriented resources	No complete publicly accessible dialect-specific corpus identified

IgboAPI is the principal partial exception in the lexical category because it records dialect-labelled words from several varieties (Emezue et al., 2024). Its contribution is important, but lexical entries alone cannot represent connected speech, syntax, discourse, or contextual meaning. Full sentence corpora and speech datasets remain necessary.

Informal social-media use presents the strongest evidence of current dialectal digital activity. Creators and community members use local forms in skits, commentary, and everyday interaction. This material may support corpus development only when collection follows consent, privacy, attribution, dialect validation, and community-governance requirements.

Research Question 2 is therefore answered by a clear resource imbalance: Standard Igbo has partial and growing infrastructure, while dialectal varieties remain largely absent from formal corpora, input tools, speech resources, and automated processing systems.

Finding 3: Documented AI Performance Gaps Across Igbo Dialects

As reviewed in Section 2.4, available evidence indicates that systems trained mainly on Standard Igbo perform less effectively on dialectal input (Emezue et al., 2024). The contribution of the present analysis is to identify three implications of that performance gap:

1. Dialectal failure is an access problem, not a marginal reduction in accuracy. A system that cannot process a speaker's variety may be functionally unusable for that speaker.
2. Improvement on Standard Igbo does not automatically improve dialect performance. Separate dialect-labelled training data and cross-dialect evaluation are required.
3. Simplifying orthography or removing tone marks cannot substitute for representative data.

Accurate systems need dialect-sensitive text and speech resources with appropriate annotation.

These implications connect the technical evidence to language ecology. Persistent exclusion narrows the domains in which dialect speakers can use their varieties and may increase reliance on Standard Igbo or English. It may also distribute AI errors disproportionately towards communities already affected by social and educational inequalities (Obiwuru, 2025).

Finding 4: Frameworks and Strategies for Dialect-Aware Igbo AI

Research Question 4 asks which theories and strategies can support dialect-aware Igbo AI. The response combines the three analytical perspectives with a four-tier development framework and five implementation strategies.

1 Applying the Theoretical Frameworks

Language ecology (Haugen, 1972; Mühlhäusler, 1996) explains why the absence of dialects from AI matters: exclusion from an expanding communicative domain may accelerate language shift and reduce functional vitality. The framework diagnoses the social consequence but does not prescribe a technical architecture.

Digital language vitality (Bird, 2022; UNESCO, 2003) provides measurable indicators for planning and monitoring. Digital texts, input tools, automated processing tools, and social-media use form a practical checklist. The reviewed evidence suggests that most Igbo dialects have minimal formal support in the first three categories, while informal social-media use is more visible.

Ezulu (2025) supplies a technical direction through dialect identification and adaptive routing, while Chinagorọm (2010) offers a linguistic basis for deciding where transfer learning may be appropriate. Their combination links structural similarity with model adaptation, but implementation still depends on dialect-labelled data and community validation.

2 A Tiered Framework for Dialect-Aware Igbo AI Development

The proposed tiered framework allocates resources according to present infrastructure and documentation while retaining all documented dialects within the development horizon. It is intended as a sequencing mechanism, not a ranking of linguistic worth.

Table 3: Proposed Tiered Framework for Dialect-Aware Igbo AI Development

Tier	Target varieties	Priority actions	Indicative resource level
Tier 1: Standard	Standard (Central) Igbo	Deepen existing development and establish cross-dialect performance benchmarks	High: existing infrastructure
Tier 2: Major dialects	Mbieri, Owerri, Onitsha, Nsukka	Community-led text and speech corpora; dialect labels; lexicon expansion	High: new investment required
Tier 3: Other documented dialects	Remaining documented clusters	Isomorphism-guided transfer learning; community input tools; lightweight adaptation	Medium: transfer reduces data requirements
Tier 4: Endangered or very small varieties	Highly divergent or small speaker communities	Machine-readable archiving, transcription, annotation, and preservation for future adaptation	Lower immediate cost; archival priority

The principal ethical concern is that Tier 3 and Tier 4 communities may interpret delayed investment as abandonment. The framework addresses this through concrete adaptation and preservation obligations. For example, where structural mapping shows that a Tier 3 variety is closely related to a Tier 2 variety, developers can initialize the smaller-variety model from the Tier 2 model and fine-tune it with a smaller, locally collected and community-validated sample of speech and text. Speakers then test recurrent errors and approve the conditions of deployment. For Tier 4 varieties, archival work should produce transcribed, annotated, machine-readable materials so that preservation creates a usable foundation for future tools rather than a closed repository.

Community governance applies to every tier. Data collection requires informed consent, meaningful participation, and written agreements on ownership, access, reuse, attribution, and

benefit sharing. Without these safeguards, dialect-aware AI may reproduce extractive research practices in which communities supply linguistic data without control or practical return (Obiwuru, 2025). Community participation is therefore both an ethical requirement and a condition for reliable labelling, evaluation, and adoption.

3 Five Practical Implementation Strategies

The tiered structure can be implemented through five coordinated strategies:

1. Multi-dialectal lexicon expansion. Extend the IgboAPI model to underrepresented varieties and recruit trained lexicographers from smaller dialect communities (Emezue et al., 2024).
2. Grassroots corpus collection. Develop community-led text and speech projects with native speakers, local validators, consent procedures, and community-owned data agreements.
3. Ethical use of informal digital material. Collect and label dialectal social-media content only with appropriate permission, privacy protection, and community governance.
4. Isomorphism-guided transfer learning. Test structural relationships identified by Chinagọrọm (2010) and adapt models from better-resourced to related under-resourced varieties using locally validated data.
5. Dialectal performance benchmarking. Require Igbo NLP studies to report results separately for Standard Igbo and represented dialects so that performance gaps remain visible and accountable.

The strategies are mutually reinforcing. Lexicons and community corpora provide the data foundation; ethically governed digital material expands coverage; transfer learning reduces adaptation costs; and dialect-specific benchmarks show whether the resulting systems improve access in practice.

Conclusions

The future of Igbo dialects in AI will be shaped by present decisions about data, model targets, evaluation, and governance. Systems built around a single standard variety may provide useful common infrastructure, but they cannot represent the language adequately when dialectal inputs remain absent from training and testing. The study establishes three main conclusions. First, Igbo dialectology contains substantial descriptive knowledge, yet this knowledge is poorly represented in AI-ready formats. Second, performance gains for Standard Igbo do not guarantee access for dialect speakers. Third, the linguistic and technical foundations for dialect-aware development already exist, but they require coordinated investment in data creation, transfer-learning research, transparent evaluation, and community governance. Dialect inclusion is therefore both a technical and a sociolinguistic responsibility. Linguists, computer scientists, developers, funding institutions, and Igbo-speaking communities must determine collaboratively which varieties are represented, how data are collected, who controls them, and how benefits are returned to participating communities. The proposed tiered framework provides a practical sequence for this work without treating smaller varieties as linguistically less valuable. The central constraint is not the absence of dialect knowledge but the absence of coordinated and adequately resourced action. Converting established scholarship and living community practice into inclusive language technology can prevent current resource gaps from becoming permanent patterns of digital exclusion.

Recommendations

Based on the findings, the study recommends the following actions:

1. African research councils, universities, foundations, and technology funders should establish dedicated funding streams for dialect-labelled text and speech data, recognizing foundational resource creation as a distinct research output.

2. Researchers developing Igbo NLP systems should test and publish performance separately for Standard Igbo and all dialects represented in their datasets.
3. The Igbo NLP community should coordinate a multi-dialectal corpus program based on established dialect classifications and transparent documentation standards.
4. Community-governance agreements should be completed before large-scale collection and should specify consent, ownership, access, reuse, attribution, and benefit sharing.
5. Transfer-learning studies should test the computational value of Chinagọrọm's (2010) linguistic isomorphism framework using dialect-specific validation data.
6. Existing print grammars, dictionaries, and dialect descriptions should be converted into structured, machine-readable formats that preserve source information and linguistic annotation.

References

Adelani, D. I., Alabi, J., &Nekoto, W. (2022). Low-resource African language AI: Challenges and opportunities. *Journal of African Language Technology, 3*(1), 45–67.

Bird, S. (2022). Decolonising computational linguistics. *Computational Linguistics, 48*(3), 741-768.

Chambers, J. K., &Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge University Press.

Chinagọrọm, L. C. (2010). *Linguistic isomorphism and Igbo dialects: Evidence from the Nsụka and Owere dialect clusters* [Doctoral dissertation, University of Nigeria, Nsukka].

Emenanjo, E. N. (1996). *Olumefula: Asụsụ Igbo na ndụ ndị Igbo*. Assumpta Press.

Emezue, C. C., Okoh, I., Mbonu, C., Chukwuneke, C., Lal, D., Ezeani, I., Rayson, P., Onwuzulike, I., Okeke, C., Nweya, G., Ogbonna, B., Oraegbunam, C., Awo-Ndubuisi, E., Osuagwu, A., &Nmezi, O. (2024). The IgboAPI dataset: Empowering Igbo language technologies through multi-dialectal enrichment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 1–12). European Language Resources Association.

Ezeani, I., Onyenwe, I., &Hepple, M. (2020). Building the Igbo-English machine translation corpus. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)* (pp. 3456–3463).

Ezikeojiakụ, P. A. (1998). *A study of Igbo dialects*. Lingua Publishers.

Ezulu, P. (2025). Dialect-aware neural models for low-resource African languages: A case study on Igbo. *NeurIPS 2025 Workshop on Generative AI for Education*. https://neurips.cc/virtual/2025/loc/san-diego/133953

Haugen, E. (1972). *The ecology of language*. Stanford University Press.

Ikekeọnwụ, C. I. (2001, November 29). *Igbo language and its dialects: A critical appraisal* [Paper presentation]. Ahịajọkụ Lecture Colloquium, Owerri, Nigeria.

Inuwa-Dutse, I. (2025). NaijaNLP: A survey of Nigerian low-resource languages. *arXiv* (arXiv:2502.19784).

Mühlhäusler, P. (1996). *Linguistic ecology: Language change and linguistic imperialism in the Pacific region*. Routledge.

Nwaozuzu, G. I. (2008). *Dialects of the Igbo language*. University of Nigeria Press.

Obiwuru, L. C. (2025). Assessing the impact of AI translation systems on Igbo language: A critical discourse analysis. *EKETE International Journal of Advanced Research, 3*(2).

Onyenwe, I., Uche, C., &Ezeani, I. (2020). Neural tone prediction for Igbo text. In *Proceedings of AfricaNLP 2020* (pp. 45–52).

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003). *Language vitality and endangerment*. UNESCO.

University of Lagos. (2026, February 18). *UNILAG champions indigenous languages in the age of artificial intelligence*. https://unilag.edu.ng/unilag-champions-indigenous-languages-in-the-age-of-artificial-intelligence