Digital Corpora for Igbo and Edo: Imperatives for NLP and ML Development

[1]Edosa James Edionhon, [2]Cecilia Amaoge Eme
[1]*University of Benin, Benin City*; [2]*Nnamdi Azikiwe University, Awka*
james.edionhon@uniben.edu.ng ; ca.eme@unizik.edu.ng

DOI: https://doi.org/10.60787/jolan.vol6no1.477

Abstract

The world is fast leveraging on Natural Language Processing (NLP) and Machine Learning (ML) to achieve enormous breakthroughs in language matters, evidenced in what are now possible in the use of different languages by digital appliances and robots, and translation abilities of AI tools. The advancement of NLP and ML that has brought these achievements is possible through the availability of elaborate, organized and easily accessible corpora. Such breakthroughs are still restricted to few languages like English, German, Spanish and Chinese due to lack of digital corpora which has specifically denied many African languages benefit in terms of NLP and ML. Languages with enormous and retrievable digital corpora such as websites, textbooks, literary materials and metadata, can easily participate in NLP and ML activities, unlike languages without such corpora. There raises the need for speakers and linguists interested especially Nigerian languages, to make conscious efforts at developing the digital corpora of their object languages. This paper advocates for elaborate, organized, well archived and easily accessible corpora for the Igbo and Edo languages to augment the few limited ones available in the languages such as Igbo Text Analyzer, IgbTC (Igbo Tagged Corpus), and AfriDataHub's Igbo Repository; Edo Language Dictionary, OTranslate, and Speak and Write Edo Language. The availability of robust digital corpora will ensure more elaborate language research, detailed linguistic analysis, extensive language preservation efforts, and language modelling for machine translation, text-to-speech and speech-to-text creation. It will further enhance the more active participation of these languages for developments in the digital space.

Keywords: *Digital corpora, Igbo, Edo, Natural Language Processing (NLP), Machine Learning*

1. Introduction

Natural Language Processing (NLP) and Machine Learning (ML) constitute the foundation of contemporary language technologies, supporting applications such as machine translation, speech recognition, and sentiment analysis by enabling computational systems to process, generate, and interpret human language

(Holdsworth, 2024). While these advancements have transformed language technology, ranging from real-time translation services to advanced conversational AI, their progress is not uniform across all languages of the world. The effectiveness of NLP and ML models hinges on the availability of extensive, well-organized, and accessible digital language corpora. As defined by Sinclair (1991:171), a corpus is "a collection of naturally occurring language text, chosen to characterize a state or variety of a language." Within NLP, such corpora, comprising both spoken and written data, serve as the foundation for training AI and machine learning applications.

The development of robust language corpora is crucial for improving AI's capacity to account for linguistic variation, hereby enhancing performance in areas such as text classification, machine translation, and multilingual processing. However, the global landscape of NLP is characterized by significant disparities, as high-resource languages such as English, German, Spanish, and Chinese dominate the field because they benefit from the availability of extensive digital corpora compiled from websites, books, and other large-scale textual repositories. This wealth of data has enabled them to achieve significant technological advancements and integration into AI-powered systems. By contrast, African languages, numbering over 2,000, are significantly underrepresented in NLP (Adebara, et al, 2025), a disparity shaped by colonial legacies and persistent data scarcity. The fact that none of the top 34 internet languages are African (Hamill-Stewart, 2024), underscores the extent of this exclusion. Inuwa-Dutse (2025) notes that for low-resource Nigerian languages, the lack of comprehensive corpora severely limits their inclusion in NLP and machine learning advancements. While some efforts exist, they are often fragmented and inaccessible, further widening the digital linguistic divide. This gap not only hampers technological progress for these languages but also jeopardizes their sustainability in the digital age, underscoring the critical need for developing robust, accessible corpora to foster inclusive and Afrocentric language technologies.

This study focuses on Igbo and Edo because they are indigenous Nigerian languages that have received some attention in language documentation but remain under-resourced in terms of comprehensive digital corpora. Although both languages possess rich linguistic and cultural heritage, their digital resources are limited and fragmented. Existing platforms such as the Igbo Text Analyzer and IgbTC for Igbo, and the Edo Language Dictionary and OTranslate for Edo, provide useful foundations but are insufficient for advanced Natural Language Processing (NLP) and Machine Learning (ML) applications. Igbo and Edo therefore, serve as representative case

studies for demonstrating the need for systematic corpus development and broader digital language resource development in Nigeria.

2. An Ethnolinguistic Overview of the Igbo and Edo Languages

Igbo is one of Nigeria's three major indigenous languages, alongside Hausa and Yoruba, and is spoken predominantly in Abia, Anambra, Ebonyi, Enugu, and Imo States, with additional speaker populations in parts of Delta and Rivers States (Onumajuru, 2013). Estimates of its speaker population range from twenty-five to forty million, with Ethnologue placing the figure at approximately twenty-nine million (Eberhard, Gary & Fennig (eds.), 2021). Linguistically, Igbo belongs to the West Benue-Congo branch of the Niger-Congo family (Williamson & Blench, 2000). The language comprises several dialect clusters that differ in phonological, grammatical, and lexical features while maintaining varying degrees of mutual intelligibility (Nwadike, 1981, 2002; Ikekeonwu, 1987).

Ẹdo is the dominant language of the historic Benin Kingdom and is spoken mainly in seven Local Government Areas of Edo State, with its standard variety based in Benin City (Ọmọrẹgbẹ, 2012). The language has several mutually intelligible dialects and is used in education, broadcasting, and written materials, including textbooks, dictionaries, and Bible translations. Ẹdo belongs to the North Central Edoid branch of the Niger-Congo family (Greenberg, 1963; Elugbe, 1989). Although both Igbo and Ẹdo have established orthographies and documented linguistic descriptions, their digital language resources remain limited, making them appropriate case studies for advocating the development of comprehensive digital corpora to support Natural Language Processing and Machine Learning applications.

3. Literature Review

Digital Corpora

The term corpus (plural: corpora), derived from the Latin word for "body," refers to a structured collection of naturally occurring written or spoken texts stored in electronic form for linguistic investigation (Gries, 2008; Cook, 2003). Although scholars describe corpora from different perspectives, they agree that they provide authentic language data for empirical analysis. Cook (2003) views a corpus as a repository of naturally occurring texts, whereas Ide (2004) emphasizes its role as a searchable electronic resource that supports computational analysis. These perspectives show how the concept of the corpus has expanded from a collection of texts for linguistic description to a resource that also supports computational research.

The development of computer technology changed the way corpora are compiled and analysed. Instead of relying on manually assembled collections, researchers began constructing large electronic databases that could be searched and annotated efficiently. This period saw the creation of major resources such as the British National Corpus (BNC) and the American National Corpus (ANC), which introduced annotation practices such as part-of-speech tagging and lemmatization to improve linguistic analysis (Ide, 2004). These developments strengthened corpus-based research by providing evidence drawn from actual language use rather than intuition.

Current research places digital corpora at the centre of computational linguistics and natural language processing. In addition to supporting linguistic description, corpora provide the data required to train computational models for language analysis and artificial intelligence applications (Subex, 2023). Researchers now distinguish among monolingual, bilingual, synchronic, diachronic, specialized, and general corpora, each developed to address different research needs (Jenset & McGillivray, 2017). The growing range of corpus types reflects the increasing demand for language resources that serve both linguistic and computational purposes.

Natural Language Processing (NLP)

Natural Language Processing (NLP) is concerned with enabling computers to analyse, interpret, and generate human language. Khurana et al. (2022) describe it as the computational analysis of natural language for tasks such as machine translation, summarization, information extraction, and question answering. Kuiler (2022) places greater emphasis on the integration of linguistics and artificial intelligence through natural language understanding and natural language generation, while Krasadakis et al. (2024) draw attention to the challenges of applying NLP to multilingual and low-resource settings. Despite these differences, the studies agree that NLP depends on computational techniques for processing human language in meaningful ways.

The field has changed considerably since its emergence in the 1950s. Early NLP systems relied on hand-crafted linguistic rules covering areas such as morphology, syntax, semantics, and pragmatics (Prashant et al., 2020). As larger digital corpora became available, statistical methods replaced many rule-based approaches because they could model language more effectively from real data. The availability of even larger datasets later encouraged the adoption of deep learning methods based on neural networks. Models such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer architectures have improved the performance of

machine translation, text classification, information extraction, and question-answering systems (LeCun et al., 2015; Subham et al., 2018; Sharavana et al., 2024).

These developments also reveal an important limitation. Languages with extensive digital corpora have benefited from rapid advances in NLP, whereas many indigenous and under-resourced languages remain poorly represented because suitable digital language resources are scarce. The availability of high-quality corpora therefore remains one of the main factors influencing NLP development across languages.

Machine Learning (ML)

Machine Learning (ML) provides many of the computational methods used in modern NLP. Vajjala (2018) defines ML as a process through which computers learn patterns from data and perform tasks without being explicitly programmed for each situation. Jordan and Mitchell (2015) extend this view by emphasizing the ability of ML algorithms to identify relationships within large datasets and make predictions from previously unseen data. Although these definitions differ in emphasis, they both present learning from data as the defining feature of machine learning.

In linguistics, ML has made it possible to analyse large corpora in ways that are difficult to achieve through manual analysis alone. It supports tasks such as language modelling, text classification, speech recognition, and information retrieval (Cho et al., 2014). The literature identifies three broad learning approaches. Supervised learning uses labelled datasets to perform predictive tasks such as text classification (Pestian et al., 2016). Unsupervised learning identifies patterns within unlabelled data and has been applied to lexical clustering and the investigation of language variation (Walsh et al., 2017). Reinforcement learning is less common in linguistic research but has attracted attention for applications involving adaptive language systems and interactive technologies (Elshawi et al., 2018).

The literature shows that digital corpora, NLP, and ML are closely connected. Digital corpora provide the linguistic data used to train machine learning models, while machine learning supplies the methods that enable NLP systems to analyse and generate language. The quality and quantity of available corpora therefore influence the performance of NLP systems. This relationship highlights the need to develop digital corpora for under-resourced languages, where the shortage of language data continues to limit computational research and technological development.

Inuwa-Dutse (2025) presents a synthetic empirical review of the computational development of Hausa, Yorùbá, and Igbo through an analysis of approximately 279 Natural Language Processing (NLP) studies. The review shows that, despite increased research activity since 2020, the three languages remain low-resource because of

inadequate digital linguistic resources. Collectively, they contain fewer than 200,000 Wikipedia articles, and about 75% of the reviewed studies relied on repurposed datasets, while only 25.1% developed new corpora or NLP tools. The study further reveals that the tonal systems and rich morphology of Igbo and Yorùbá, together with inconsistent use of diacritics, reduce the effectiveness of machine translation, speech recognition, and sentiment analysis. Although Hausa has broader NLP coverage, Igbo and Yorùbá remain underrepresented in morphological analysis, question-answering, and text summarisation. The study recommends indigenous NLP models, high-quality annotated corpora, formal grammatical analysis, and collaborative resource development.

4. Theoretical Framework

This study is guided by Corpus Linguistics, a methodological framework that examines language through the systematic analysis of authentic linguistic data rather than intuition or constructed examples. A corpus is a structured collection of naturally occurring written and spoken texts stored electronically for linguistic investigation (McEnery & Hardie, 2012; Abdullayeva, 2025). Although scholars differ on whether Corpus Linguistics is a theory or a methodology, they agree that it provides an empirical basis for describing and explaining language use. Kübler and Zinsmeister (2015) argue that it can function as either a theory or a methodological tool, depending on its application, whereas McEnery and Brezina (2022) regard it as a methodological framework for identifying recurrent patterns of language use.

Corpus Linguistics emerged in response to linguistic traditions that relied heavily on introspection and limited datasets. Unlike structural and generative approaches, which emphasized speakers' intuitions, Corpus Linguistics is based on the empirical observation and statistical analysis of large collections of authentic language data made possible by advances in computing technology (Kennedy, 1998). Corpora provide evidence for investigating lexical, grammatical, and discourse patterns that are not readily observable through intuition alone. Biber, Conrad, and Reppen (1998) note that corpus-based analyses reveal the frequency, distribution, and collocational behaviour of linguistic forms, leading to more reliable descriptions of language structure and use.

The framework rests on four principles: authenticity, representativeness, systematicity, and computational analysis. Authenticity requires that corpora contain naturally occurring language rather than constructed examples. Representativeness ensures that the corpus reflects the language, genre, or register under investigation.

Systematicity requires the careful collection, organization, and annotation of linguistic data according to established procedures, while computational analysis uses specialized software to retrieve, quantify, and interpret linguistic patterns (Abdullayeva, 2025). These principles distinguish Corpus Linguistics from intuition-based approaches by grounding linguistic analysis in observable evidence.

The development of Corpus Linguistics has been shaped by scholars such as Leech, Biber, Johansson, Francis, Hunston, Conrad, McCarthy, and John Sinclair. Sinclair (1991) argues that meaning is often realized through recurring combinations of words rather than isolated lexical items, an observation that underpins corpus-based studies of collocation and phraseology. Bennett (2010) also notes that Corpus Linguistics seeks to explain the distribution of lexical and grammatical patterns and their variation across language varieties and registers.

Corpus Linguistics also integrates quantitative and qualitative methods. Quantitative analysis identifies recurring linguistic patterns through measures such as frequency and distribution, whereas qualitative analysis explains these patterns within their linguistic and communicative contexts. The two approaches complement one another, with statistical findings informing contextual interpretation and qualitative observations guiding further quantitative investigation (Leech et al., 2012).

Corpus Linguistics provides the framework for this study because it recognizes digital corpora as the empirical foundation for linguistic description and computational language processing. It supports the argument that representative and electronically accessible corpora are essential for linguistic analysis, language documentation, and the development of Natural Language Processing and Machine Learning applications. Within this framework, expanding the digital corpora of Igbo and Edo is essential for improving computational language resources, strengthening language preservation, and increasing the participation of these languages in the digital space.

5. Status of Documentation in Igbo and Edo

The documentation of the Igbo language has evolved from oral traditions to written and digital forms. In the precolonial period, the language was preserved through storytelling, proverbs, songs, riddles, and other forms of oral literature, which transmitted cultural knowledge across generations (Green, 1948; Emenyonu, 2020). Indigenous writing systems, including Nsịbịdị, uli, Akwụkwọ mmụọ, and the Aniocha script, also demonstrate that forms of literacy existed before European contact (Oraka, 1983; Azuonye, 1992).

Systematic written documentation began during the missionary era through the efforts of the Church Missionary Society. Rev. S. A. Crowther's Primer, which included an Igbo alphabet, vocabulary, sentence structures, and biblical translations, marked a major milestone in the codification of the language (Igwe & Obiakor, 2015; Emenyonu, 2020). More recently, Nigerian universities and Igbo scholars have advanced language documentation through linguistic research, curriculum development, and the preservation of oral traditions, providing an important foundation for contemporary digital corpus development.

The documentation of the Edo language on the other hand has developed from oral traditions to written and scholarly forms. Before the colonial period, Edo was preserved through stories, proverbs, riddles, and other oral traditions that transmitted linguistic and cultural knowledge across generations. Traditional art forms, including brass-casting, wood carving, and pictorial writing, also served important communicative and symbolic functions within Edo society (Lewis, 2018).

Systematic written documentation began during the colonial period with the introduction of the Roman script. Early missionary efforts, particularly Reverend Emmanuel Egiebor Ohuoba's translation of portions of the Bible into Edo, marked significant milestones in the development of written Edo (Ohuoba, 2017; Usunalele & Agbontaen, 2000). Missionary activities and colonial education also encouraged the standardization of the language and the production of written materials, including Jacob Egharevba's Ekhere Vb'Itan Edo (1933), an important contribution to Edo historiography and literature (Osagie, 1999). Subsequent documentation by indigenous scholars expanded the collection of proverbs, folktales, songs, and oral histories while supporting the development of a standard orthography (Eisenhofer, 1995; Usunalele & Agbontaen, 2000). More recently, scholarly research, dictionaries, and university-based studies have continued to strengthen Edo language documentation, providing a foundation for contemporary digital corpus development (Ọmọrẹgbẹ, 2025).

5.1. Digital Presence and Existing Igbo Corpora

In recent years, notable progress has been made in expanding the digital presence and the development of Igbo language corpora through both national and international initiatives. The Igbo Archival Dictionary Project represents a major undertaking focused on compiling a spoken-word corpus encompassing various dialects, with particular attention to oral narratives and phonological diversity. This initiative supports the creation of a digital dialectal dictionary aimed at preserving linguistic variation within the Igbo language (Ezeani et al., 2020). Similarly, the National Centre

for Artificial Intelligence and Robotics (NCAIR) in Nigeria has contributed to corpus development by producing transcribed audio corpora and annotated text datasets for speech recognition and machine learning applications, thereby facilitating the integration of Igbo into AI-driven linguistic technologies (NCAIR, 2023).

Academic institutions in Nigeria have also played a critical role, though many of their compiled corpora remain limited in scope or accessibility. A prominent example is the Igbo Tagged Corpus, developed by Onyenwe, Uchechukwu, and Hepple (2014), which incorporates a part-of-speech (POS) tagset adapted from the EAGLES framework to reflect Igbo's grammatical structure. Despite its limited size and dialectal coverage, this corpus has been essential for training POS taggers and syntactic parsers in Igbo computational linguistics. Complementary to this effort, the Igbo Text Analyzer, a web-based NLP tool created by Azubuike and Umeh (2024), performs tokenization, morphological analysis, and syntactic parsing using a hybrid of rule-based and statistical approaches. Although still under development, the tool demonstrates considerable potential despite ongoing challenges related to dialectal variation and orthographic inconsistency.

Global collaborations have further advanced the digital documentation of Igbo. Mozilla's Common Voice project has incorporated Igbo speech data contributed by volunteers to support open-source voice recognition systems (Mozilla, 2023), while the Masakhane project, a Pan-African NLP initiative, has promoted Igbo machine translation and computational research (Orife et al., 2020). More recently, the AfriDataHub's Igbo Repository has launched the IgboAPI Dataset, a large-scale, multi-dialectal corpus comprising millions of synthetic and natural Igbo sentences annotated for translation, classification, and speech recognition tasks (AfriDataHub, 2023). Despite its breadth and accessibility via platforms such as Hugging Face, its reliance on synthetic data and uneven dialectal representation raises questions about authenticity and inclusivity of Igbo corpus development.

5.2. Digital Presence and Existing Edo Corpora

The digital presence of Edo is steadily expanding through online platforms such as 'edofolks.com', 'irenmwi.com', the Edo Language Dictionary, and mobile applications like 'O Translate'. These platforms provide bilingual word lists, cultural notes, and limited audio and translation features that facilitate language learning and preservation. However, they remain largely informal and lack systematic linguistic annotation or corpus-level organization.

Among existing resources, irenmwi.com represents a notable advancement as the first monolingual Edo dictionary (Ọmọrẹgbẹ, 2025). It defines words, provides contextual examples and usage notes, and includes transcription tools that enhance accessibility and promote orthographic standardization. Similarly, edofolks.com functions as a comprehensive cultural and educational hub, offering materials on Edo history, traditions, and language, alongside sections on mathematics, religion, and news, illustrating a holistic digital approach to Edo cultural preservation.

Educational initiatives such as letslearnẹdo.ng further extend Edo's online presence by offering virtual classes, downloadable learning materials, and interactive content for learners. Its design reflects a clear commitment to structured digital pedagogy. Likewise, O Translate supports Edo–English translation but operates on a limited phrase database and lacks linguistic annotation or syntactic parsing, restricting its academic utility.

Community-driven efforts such as YouTube tutorials, WhatsApp groups, and local radio programmes, also contribute to Edo literacy and oral fluency but are not systematically archived or tagged for corpus use. The Edo entry on Wikipedia and related lexical resources offer basic descriptive data, including classification within the Edoid language family, but lack annotated corpora suitable for computational or linguistic analysis.

While the Nigerian Educational Research and Development Council (NERDC) has conducted surveys supporting indigenous language instruction in Edo State, these initiatives have not produced large-scale, digitized corpora. Overall, the digital documentation of Edo remains fragmented, characterized by valuable but scattered individual efforts. Despite these limitations, the current platforms signify crucial early steps toward Edo's revitalization in digital spaces and underscore the need for coordinated, institutionally supported, and linguistically rigorous corpus development.

5.3. Comparison of Digital Presence and Existing Corpora between Igbo and Edo

From the perspective of Corpus Linguistics, the availability of authentic, representative, and systematically organized corpora determines the extent to which a language can support linguistic research and computational applications. The digital development of Igbo illustrates this principle. Compared with Edo, Igbo has benefited from coordinated corpus-building initiatives that have produced structured linguistic resources for both linguistic analysis and Natural Language Processing (NLP). These include the Igbo Archival Dictionary Project, the Igbo Tagged Corpus (IgbTC), and tools such as the Igbo Text Analyzer, which provide annotated datasets, part-of-speech-

tagged corpora, and speech resources. Such resources satisfy the corpus linguistic principles of authenticity, representativeness, and systematicity by making naturally occurring language available in machine-readable formats. International initiatives, including Mozilla Common Voice, Masakhane, and AfriDataHub, have further expanded the digital visibility of Igbo by integrating the language into speech recognition, machine translation, and other AI-driven applications. The IgboAPI Dataset has also strengthened the computational usability of the language by providing a large-scale, multidialectal Igbo–English lexical resource that supports translation, language modelling, and speech technologies. Nigerian universities have complemented these efforts by incorporating digital language resources into teaching and research. These developments reflect a growing recognition that robust corpora are indispensable for linguistic description, language preservation, and computational language processing.

The Edo language presents a different picture. Although several digital initiatives, including edofolks.com, irenmwi.com, letslearnẹdo.ng, and O Translate, contribute to language learning and cultural preservation, they do not yet constitute comprehensive linguistic corpora within the corpus linguistic framework. Most resources are designed as dictionaries, learning platforms, or cultural repositories rather than structured datasets suitable for computational analysis. Consequently, they provide limited opportunities for corpus annotation, frequency analysis, syntactic parsing, or the development of NLP and Machine Learning (ML) applications. Corpus Linguistics emphasizes that language technologies depend on large, representative, and systematically annotated datasets. In this respect, Edo remains under-resourced, as it lacks extensive digital corpora, publicly available APIs, and coordinated institutional programmes dedicated to corpus construction. Existing initiatives are largely community-driven and fragmented, limiting their long-term sustainability and computational value.

The contrast between Igbo and Edo demonstrates the central claim of Corpus Linguistics that the quality and availability of digital corpora shape both linguistic inquiry and technological development. While Igbo has begun to establish the empirical resources required for corpus-based research and NLP, Edo still requires coordinated corpus development to achieve comparable progress. Strengthening digital corpora for both languages will not only improve linguistic documentation but also expand their participation in machine translation, speech technologies, language modelling, and other computational applications.

5.4. Gaps in Availability and Accessibility in Igbo and Edo

From the perspective of Corpus Linguistics, the availability, accessibility, authenticity, and systematic organization of corpora determine the extent to which a language can support linguistic research and computational applications. Although Igbo has made notable progress in digital corpus development, important gaps remain in both the availability and accessibility of its resources. Dialectal coverage is uneven, with central dialects overrepresented and peripheral varieties underdocumented (Emezue et al., 2024). Annotation is also limited, as most corpora provide only part-of-speech tagging, with little syntactic or phonological information, while orthographic inconsistency affects data quality and interoperability (Onyenwe et al., 2019). Accessibility remains constrained because many resources are fragmented across platforms such as GitHub, Hugging Face, and Sketch Engine, where corpora such as igWaC and igTenTen require subscription or institutional access. In addition, reliance on crowd-sourced or synthetic materials raises concerns about data authenticity (Emezue et al., 2024). Spoken corpora with phonetic and tonal annotation are also scarce, limiting research in speech processing and phonological analysis.

The Edo language exhibits more significant gaps in corpus availability and accessibility. It lacks publicly available corpora with part-of-speech tagging, syntactic parsing, or large-scale text collections comparable to igWaC or IgboAPI. Existing digital resources are largely dictionaries and language-learning platforms with limited dialectal and phonological coverage, minimal audio support, and no systematic linguistic annotation. Resources such as the Edo Language Dictionary also lack downloadable datasets, metadata, and machine-readable formats, restricting their reuse in computational research. Institutional involvement in corpus development remains limited, with most initiatives being community-driven and lacking long-term sustainability. Consequently, the scarcity of spoken corpora and NLP-ready datasets continues to restrict Edo's participation in computational linguistics and language technology.

These findings support the corpus linguistic view that robust and accessible corpora provide the empirical foundation for Natural Language Processing (NLP) and Machine Learning (ML). Despite its limitations, Igbo has benefited from foundational corpora and participation in multilingual initiatives such as Meta's No Language Left Behind and Masakhane, which adapt multilingual models such as mBERT and XLM-R through transfer learning. For Edo, comparable progress depends on developing representative and annotated corpora from dictionaries, educational materials, literary texts, and oral sources. These resources should be made openly accessible to support

transfer learning, crowdsourced transcription and translation, and carefully controlled synthetic data augmentation. Expanding both the availability and accessibility of Edo corpora will strengthen linguistic research, language preservation, and the development of NLP and ML applications.

5.5. Natural Language Processing and Machine Learning Tools for Igbo and Edo Languages

The Corpus Linguistics framework recognizes well-developed digital corpora as the foundation for Natural Language Processing (NLP) and Machine Learning (ML) applications. The quality and scale of these corpora determine the performance of computational tools and the extent to which a language can participate in digital technologies. This section examines key NLP and ML tools for the Igbo and Edo languages such as tokenization, part-of-speech tagging, morphological analysis, named entity recognition and lemmatization, speech technologies, and machine translation and highlights how corpus development provides the empirical foundation for their implementation.

i. Tokenizers and Sentence Splitters

Tokenization is the first stage of NLP, segmenting text into words and sentences for computational analysis. For Igbo, tokenization has been integrated into corpus development through the IgboNLP Project (Onyenwe, 2017). Rule-based and statistical approaches that account for orthographic variation, tone marking, and affixation have produced encouraging results. For Edo, similar methods can be developed once representative corpora become available. Studies on related African languages report tokenization accuracies of 90–95% where adequately annotated datasets exist (Onyenwe et al., 2014), demonstrating the importance of well-developed corpora.

ii. Part-of-Speech Tagging

Part-of-speech (POS) tagging forms the basis of syntactic and semantic analysis. Igbo already possesses a standardized POS tagset of 59 tags based on the EAGLES guidelines (Onyenwe, Uchechukwu & Hepple, 2014). Bootstrapped and Hidden Markov Model taggers, strengthened through cross-lingual transfer, have achieved improved accuracy (Onyenwe et al., 2019). Edo, however, requires representative annotated corpora before comparable POS taggers can be developed using multilingual Transformer models.

iii. Morphological Analysis

The tonal and morphologically rich structures of Igbo and Edo require effective morphological analysis for accurate lexical representation. Finite-state approaches developed for languages such as Wolaytta and Zulu provide suitable models (Gebreselassie et al., 2018; Pretorius & Bosch, 2003). Tools such as Foma and HFST can model morphophonemic alternations and affixation, provided sufficiently annotated lexical resources are available. Their effectiveness depends largely on corpus quality and linguistic coverage.

iv. Named Entity Recognition and Lemmatization

Named Entity Recognition (NER) extracts structured information from text, while lemmatization reduces inflected forms to their base forms. Igbo has been incorporated into multilingual NER systems such as the Wazobia-NER System, achieving F1-scores above 0.90 (Ezeani et al., 2024). Lemmatization can also be supported through existing morphological resources. Edo requires annotated corpora before comparable systems can be developed. Transfer learning from multilingual models offers a practical pathway for improving performance in low-resource settings (Ezeani et al., 2020).

v. Speech Processing and Text-to-Speech

Speech technologies for both languages remain underdeveloped because of the limited availability of annotated speech corpora. Although Igbo has small speech datasets, large-scale Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems are still lacking. Community initiatives such as Mozilla Common Voice can expand speech resources, while multilingual acoustic models can be adapted to capture tonal and phonemic distinctions. Comparable progress for Edo depends on the creation of representative speech corpora.

vi. Machine Translation

Machine Translation (MT) integrates several NLP components and depends on large parallel corpora. Igbo has advanced through initiatives such as the Masakhane Neural MT Project, producing moderate BLEU scores comparable with other African languages (Adelani et al., 2021). Edo, however, lacks the bilingual corpora required for neural machine translation. Developing Edo–English and English–Edo parallel datasets is therefore essential for applying Transformer-based architectures and few-shot learning approaches.

Beyond these applications, computational tools such as Praat, ELAN, FLEx, Foma, UralicNLP, BRAT, and WebAnno support corpus collection, transcription,

annotation, alignment, and storage. Within the Corpus Linguistics framework, these tools facilitate the creation of authentic and systematically annotated corpora that underpin NLP and ML. Expanding such resources for Igbo and Edo will strengthen linguistic research, language preservation, and the development of computational technologies for both languages.

5.6. The Importance of Digital Corpora for NLP and ML Development in Igbo and Edo

Digital corpora are important for the advancement of Natural Language Processing (NLP) and Machine Learning (ML) in Nigerian languages, particularly Igbo and Edo.

1. Digital corpora provide the structured datasets required to train algorithms that enable machines to understand and process human language. They form the empirical foundation of NLP and ML by supplying the linguistic evidence from which computational models learn grammatical, lexical, and semantic patterns (Sinclair, 1991; Ide, 2004). These corpora support essential tasks such as speech recognition, sentiment analysis, word frequency modelling, and part-of-speech tagging, resulting in more accurate and reliable language technologies. The quality of these datasets, especially their balance, cleanliness, authenticity, and representativeness, largely determines the performance and fairness of AI systems (Subex, 2023; BAVL, 2022).

2. Digital corpora are the foundation of NLP and ML systems and play both technological and cultural roles in low-resource languages such as Igbo and Edo. For Igbo, initiatives including the Igbo Tagged Corpus (Onyenwe et al., 2014), Igbo Text Analyzer (Azubuike & Umeh, 2024), and the IgboAPI Dataset (AfriDataHub, 2023) support tokenization, morphological analysis, and machine translation, enabling the language to participate in multilingual AI initiatives such as Masakhane and Mozilla Common Voice (Orife et al., 2020; Mozilla, 2023). By contrast, Edo remains at an early stage of digital corpus development. Existing platforms, including edofolks.com, irenmwi.com, and OTranslate (Ọmọrẹgbẹ, 2025), provide valuable language resources but lack the structured and annotated corpora required for advanced NLP and ML applications. Developing comprehensive, accessible, and linguistically annotated corpora is therefore essential to Edo's computational development.

3. Beyond their computational value, digital corpora contribute to language preservation and cultural continuity. Their absence continues to limit the participation of many African languages in digital technologies, widening the gap between high-resource and low-resource languages (Inuwa-Dutse, 2025; Adebara et al., 2025). For Igbo and Edo, corpus development serves both as a means of preserving linguistic and

cultural heritage and as a foundation for digital inclusion. The digitization of oral traditions, folklore, and phonological data ensures that these resources are preserved and made available for applications such as speech recognition, text-to-speech synthesis, and machine translation.

4. Digital corpora also promote benchmarking, interoperability, and collaborative research. Standardized and annotated datasets allow NLP and ML models to be trained, evaluated, compared, and reproduced across different studies, thereby improving the reliability and transparency of computational research (Jenset & McGillivray, 2017). In addition, tools such as ELAN, Praat, FLEx, and WebAnno support corpus construction through transcription, phonetic annotation, linguistic alignment, and data management, linking traditional linguistic documentation with modern computational modelling.

6. Conclusion

This article has highlighted the importance of developing comprehensive digital corpora for the Igbo and Edo languages to strengthen their integration into Natural Language Processing (NLP) and Machine Learning (ML). Igbo has attained an intermediate level of digital development, supported by foundational corpora, basic NLP tools, and participation in multilingual initiatives such as Masakhane and No Language Left Behind. However, its progress is constrained by data fragmentation, uneven dialectal representation, and limited accessibility, which reduce the scalability and interoperability of its digital resources. By contrast, Edo remains at an early stage of digital corpus development, characterized by the absence of annotated datasets, standardized orthography, and the institutional support required for computational language processing.

Addressing these gaps requires systematic corpus collection, annotation, standardization, and long-term maintenance using tools such as ELAN, FLEx, Praat, and WebAnno. These resources will support the development of core NLP components, including tokenizers, part-of-speech taggers, morphological analyzers, and named entity recognition systems, which provide the foundation for advanced applications such as speech recognition, text-to-speech systems, and machine translation.

Sustained progress depends on collaboration among universities, research institutions, technology organizations, and language communities, supported by adequate funding and long-term policy commitment. Such collaboration will facilitate the transition from descriptive language documentation to computational language processing. Consequently, digital corpus development serves not only as a strategy for

preserving the linguistic heritage of Igbo and Edo but also as the foundation for expanding their participation in language technology, artificial intelligence research, and the wider digital ecosystem.

References

Abdullayeva, K. R. (2025). The role of corpus linguistics in modern linguistic research. Colloquium-journal, 67(260), 14–15. https://doi.org/10.5281/zenodo.17602057

Adebara, I., Olamide Toyin, H., Tesfu Ghebremichael, N., Elmadany, A. R., & Abdul-Mageed, M. (2025). Where Are We? Evaluating LLM Performance on African Languages. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 32704–32731.

Adedjouma, S. A., Aoga, J. O., & Igue, M. A. (2013). Part-Of-Speech Tagging of Yoruba Standard, Language of Niger-Congo Family. Res. Journal of Computer & IT Sciences, 1(1), 2–5.

Adelani et al. (2021). Masakhaner: Named Entity Recognition for African Languages. Transactions of the Association for Computational Linguistics, 9:1116–1131.

AfriDataHub. (2023). IgboAPI Dataset. Retrieved from https://huggingface.co/datasets/IgboAPI

Azubuike, C., & Umeh, J. (2024). Igbo Text Analyzer: A Rule-Based NLP Tool for Igbo Language Processing. Unpublished manuscript.

Azuonye, Chukwuma (1992). "The Development of Written Igbo Literature". In Adiele Eberechukwu Afigbo (ed.). Ground Work of Igbo History. Ethnohistorical Studies 1. Vista Books. ISBN 978-134-400-8.

Bennett, G. R. (2010). Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers. University of Michigan Press.

Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.

Brnabic, A., & Hess, L. M. (2021). Systematic Literature Review of Machine Learning Methods Used in The Analysis of Real-World Data for Patient-Provider Decision

Making.BMC Medical Informatics and Decision Making, 21, 54. https://doi.org/10.1186/s12911-021-01403-2

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On The Properties of Neural Machine Translation: Encoder–Decoder Approaches.arXiv preprint arXiv:1409.1259.

Cook, G. (2003). Applied Linguistics. Oxford. OUP.

Eberhard, D.M., Fennig, C.D. Simons, G.F. (2021). Ethnologue: Languages of the World (24th Edition). SIL international. https://www.ethnologue.com/

Egharevba, J. U. (1954). The Origin of Benin. Benin City: African Industrial Press.

Eisenhofer, Stefan (1995). "The Origins of the Benin Kingship in the Works of Jacob Egharevba". History in Africa. 22. Cambridge University Press (CUP): 141–163. doi:10.2307/3171912

Elshawi, R., Sakr, S., Talia, D., & Trunfio, P. (2018). Big Data Systems Meet Machine Learning Challenges: Towards Big Data Science as a Service. Big Data Research, 14, 1–11. https://doi.org/10.1016/j.bdr.2018.04.004

Elugbe, B. O. (1989). Comparative Edoid: Phonology and Lexicon. Delta Series. No 6, University of Port Harcourt Press

Emenyonu, Ernest N. (2020). "The Literary History of the Igbo Novel: African Literature in African Languages". Oxon: Routledge & CRC Press.

Emezue, C. C.; Okoh, I.; Mbonu, C.; Chukwuneke, C.; Lal, D.; Ezeani, I., et. al. (2024). The IgboAPI Dataset: Empowering Igbo Language Technologies Through Multi-Dialectal Enrichment.

Ezeani, I., Orife, I., & Ogueji, K. (2020). IgboNLP: Building NLP Tools for The Igbo Language. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 1–10).

Gebreselassie, T. A.; Washington, Jonathan N.; Gasser, M.; and Yimam, B. (2018). "A Finite- State Morphological Analyzer for Wolaytta". Information and Communication Technology for Development for Africa. Volume 244, 14-23. DOI:10.1007/978-3-319-95153-9_2 https://works.swarthmore.edu/fac-linguistics/267

Grazib, M. (2008, May 3-4). Electronic Corpora: As Powerful Tools in Computational Linguistic Analyses. Proceedings of the 2nd Conférence Internationale sur l'Informatique et ses Applications (CIIA'09), Saida, Algeria.

Green, M. (1948). "The Unwritten Literature of the Igbo-Speaking People of South-Eastern Nigeria". Bulletin of the School of Oriental and African Studies. 12 (3–4). Cambridge University Press: 838–846. doi:10.1017/S0041977X00083415

Greenberg, J. H. (1963). The Languages of Africa. Bloomington: Indiana University Press

Hamill-Stewart, C. (2024). "The 'missed opportunity' with AI's linguistic diversity gap." World Economic Forum: Forum Agenda. (weforum.org)

Holdsworth, J. (2024) What is NLP (natural language processing)? https://www.ibm.com/topics/natural-language-processing#:~:text=Natural%20 language %20processing%20(NLP)%20is,and%20communicate%20with%

Ide, N. (2004). Preparation and Analysis of Linguistic Corpora. A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell,

Igwe, Austine Uchechukwu; Obiakor, Nwachukwu (2015). "The Historical Emergence of ' Union Igbo Bible' and Its Impact on Igbo Language Development in The Twentieth Century".International Journal of Multidisciplinary Research and Development. 2 (1): 70–75. ISSN 2349-4182

Ikekeọnwụ, C.I. (1987). Igbo Dialect Cluster: A classification. A Seminar Presented to the Department of Linguistics/Nigerian Languages, University of Nigeria, Nsukka

Imasuen, E. O. (1998/99). Languages in contact: The case of Edo and Portuguese. JWAL 27(2): 39-50

Inuwa-Dutse, I. (2025). A Survey of Nigerian Low-Resource Languages. Preprint, arXiv:2502.19784.

Jalaj, Thanaki (2017). Natural Language Processing. Packt Publishing

Jenset, G. B., & McGillivray, B. (2017). Quantitative Historical Linguistics: A Corpus Framework. Oxford University Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349 (6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kennedy, G. (1998). An Introduction to Corpus Linguistics. Longman.

Khouja, J. (2020). Stance prediction and claim verification: An Arabic perspective. Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), 8–17. https://doi.org/10.18653/v1/2020.fever-1.2

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Krasadakis, P., Sakkopoulos, E., & Verykios, V. S. (2024). A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. Electronics, 13(3), 648. https://doi.org/10.3390/electronics13030648

Kratzwald, B., Ilic, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems, 115, 24–35. https://doi.org/10.1016/j.dss.2018.09.002

Kübler, N., & Zinsmeister, H. (2015). Corpus Linguistics and Linguistically Annotated Corpora. Bloomsbury Academic.

Kuiler, E. W. (2022). Natural language processing (NLP). In L. A. Schintler & C. L. McNeely (Eds.), Encyclopedia of Big Data. Springer, Cham. https://doi.org/10.1007/978-3-319- 32010-6_250

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2012). Change in Contemporary English: A Grammatical Study. Cambridge University Press.

Lewis, Demola (2018). "Linguistic Prehistory and Identity in Nigeria's Bini-Ife Pre-eminence Contestation". Multilingual Margins: A Journal of Multilingualism from the Periphery. 5 (1): 2–23. doi:10.14426/mm. v5i1.86

McEnery, T., & Brezina, V. (2022). Fundamental Principles of Corpus Linguistics. Cambridge University Press.

McEnery, T., & Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

Melzian, H. (1937). A Concise Dictionary of the Bini Language of Southern Nigeria. London: Paul Kegan.

Mozilla. (2023). Common Voice: Igbo Dataset. Retrieved from https://commonvoice.mozilla.org

NCAIR. (2023). National Centre for Artificial Intelligence and Robotics: Language Resources. Retrieved from https://ncair.gov.ng

Nwadike, I. U. (2002). Igbo Language in Education: A Historical Study. Obosi: Pacific Correspondence College and Press.

Nwadike, U.I. (1981). The Development of written Igbo as a school Subject. 1766–1980. A Historical Approach. Masters' Degree Thesis submitted to the Faculty of Graduate Studies of the State University of New York, Buffalo.

Ohuoba, Imafidon (2017). "Imafidone 3 (Early Missionary Activities of Rev. E. E. Ohuoba (1885–1950). Academia.edu. Retrieved 10 October 2025.

Omego, Christie. (2007). "A Survey of Language Use and Culture in Igbo Land" in Journal of Nigerian Languages and Culture. 9(1). 168-169.

Ọmọrẹgbẹ, E.M. (2012). The Basic Clause: A Morpho-syntactic Analysis of Edo Clausal Verbs. Ph.D. Dissertation, University of Benin, Benin City.

Ọmọrẹgbẹ, E.M. (2025). The Task of the Grammarian in a Digital Era. Inaugural lecture Series 336. University of Benin, Benin City.

Onumajuru, Chiemeka V. (2013). A Semantic and Pragmatic Analysis of Igbo Names. In Ozo-Mekuri Ndimele, Lendzemo C. Yuka & Johnson F. Ilori (Eds.), Issues in Contemporary African Linguistics: A Festschrift for Oladele Awobuluyi (Pp: 445-462). Port Harcourt: M&J Grand Orbit Community Ltd.

Onyenwe, C., Uchechukwu, C., & Hepple, M. (2014). Part-of-Speech Tagging for Igbo with an Enhanced Tagset. In Proceedings of the 9th Language Resources and Evaluation Conference (LREC).

Onyenwe, I. (2017). Developing Methods and Resources for Automated Processing of the African Language Igbo (Doctoral dissertation, University of Sheffield).

Oraka, Louis Nnamdi. (1983). The Foundations of Igbo Studies: A Short History of The Study of Igbo Language and Culture. University Publishing Co. ISBN 978-160-264- 3.

Orife, I., Ezeani, I., Dossou, B. F., & Adelani, D. I. (2020). Masakhane: Machine Translation for African Languages. In Proceedings of the 2020 EMNLP Workshop on African NLP (pp. 1–9).

Osadolo, O. B. (2001). The Military System of Benin Kingdom, c.1440–1897. Ph.D. Dissertation, University of Hamburg, Germany

Osagie, Eghosa (1999). "Benin in Contemporary Nigeria. An Agenda for The Twenty-First Century". dawodu.net.

Pestian, J., Sorter, M. T., Connolly, B., Cohen, K. B., McCullumsmith, C. B., Gee, J. T., Morency, L.-P., Scherer, S., & Rohlfs, L. (2016). A Machine Learning Approach to Identifying The Thought Markers of Suicidal Subjects: A Prospective Multicenter Trial. Suicide and Life-Threatening Behavior, 47(1), 112–121. https://doi.org/10.1111/sltb.12312

Prashant, S., Alok, P., & Vivek, K. (2020). A Survey on Natural Language Processing Approaches: From Rule-Based to Deep Learning. International Journal of Computer Applications, 174(18), 1-8.

Pretorius, Laurette and Sonja E. Bosch. 2003. "Finite-State Computational Morphology: An. Analyzer Prototype for Zulu." Machine Translation 18: 195–216.

Sharavana, K., Bhakta, K., Chethan, J. S., Chand, J., & Joshi, M. K. (2024). Evolution of Natural Language Processing: A Review. Journal of Knowledge in Data Science and Information Management. 1(1), 30–38. https://doi.org/10.46610/JoKDSIM.2024.v01i01.004

Sinclair, J. (1991). Corpus, Concordance, Collocation: Describing English Language. Oxford University Press

Subex (2023). Corpus. https://www.subex.com/article/corpus/

Subham T., Shaina Anjum, and Manmohan. (2018). Natural Language Processing in The Era of Deep Learning: A Survey of Applications and Advances. International Journal of Applied Research. 4(9) Part B: 120-124

Usuanlele, U. and Agbontaen, K.A. (2000). "A History of Modern Literary Development Among the Edos 1897–1960". Africa: Rivista trimestrale di studi e documentazione dell'Istituto italiano per l'Africa e l'Oriente. 55 (1). Istituto Italiano per l'Africa e l'Oriente (IsIAO): 105–113. ISSN 0001-9747

Vajjala, S. (2018). Machine Learning iIn Applied Linguistics. In C. A. Chapelle (Ed.), The Encyclopedia of Applied Linguistics. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal1486

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting Risk of Suicide Attempts Over Time Through Machine Learning. Clinical Psychological Science, 5(3), 457- 469. https://doi.org/10.1177/2167702617691560

Williamson, K., & Blench, R. (2000). Niger, Congo. In D. Nurse, & B. Heine (Eds.), African Languages: An Introduction. Cambridge University Press. http://www.unlweb.net/wiki/x-bar

Yeldo, H. (2021). "Natural Language Processing: Components, Advances, Tools and Industrial Applications." International Journal for Research in Applied Science and Engineering Technology 9, no. 9 964–71. http://dx.doi.org/10.22214/ijraset.2021.38015.

[1]Edosa James Edionhon, [2]Cecilia Amaoge Eme
[1]*University of Benin, Benin City, Nigeria*;
[2]*Nnamdi Azikiwe University, Awka, Nigeria*
james.edionhon@uniben.edu.ng ; ca.eme@unizik.edu.ng