Journal of Linguistics, Literary and Communication Studies
Vol. 5 No. 1 (2026): ISSN (Online): 2957-8477
DOI: https://doi.org/10.58721/jllcs.v5i1.1736
Voicing the Voiceless: Developing a Dholuo Parallel Corpus for Natural Language Processing

Voicing the Voiceless: Developing a Dholuo Parallel Corpus for Natural Language Processing

Quin Elizabeth Awuor
United States International University-Africa, Kenya

Article History
Received: 14.09.2025
Revised: 18.04.2026
Accepted: 04.05.2026
Published: 07.05.2026

Keywords
Dholuo
Dictionaries
Low-resource languages
Parallel corpus

How to cite:
Awuor, Q. E. (2026). Voicing the Voiceless: Developing a Dholuo Parallel Corpus for Natural Language Processing. *Journal of Linguistics, Literary and Communication Studies, 5*(1), 73-87.

Abstract
Despite the exponential growth of Natural Language Processing (NLP) technologies worldwide, African indigenous languages remain severely underrepresented in digital language resources. Dholuo, a Western Nilotic language spoken by approximately four million people in Kenya and Tanzania, lacks the structured parallel corpora needed to support machine translation, speech recognition, and other AI-driven language tools. This paper reports on the design, methodology, and preliminary findings of a corpus development initiative funded by United States International University-Africa (USIU-Africa) to build a Dholuo–English parallel corpus comprising 20,000 translated sentence pairs and 30 hours of transcribed speech data. Preliminary findings indicate that approximately 18,400 parallel sentence pairs have been collected, of which 31% contain at least one figurative element, confirming the centrality of figurative expression in authentic Dholuo discourse and validating the native-speaker-led, community-engaged collection methodology. Drawing on community-driven data collection, crowdsourced translation platforms, and the Living Dictionaries digital lexicography tool (livingdictionaries.app), this study integrates culturally embedded linguistic features, including metaphors (*weche mitiyo kodo kakaranyisi*), proverbs (*Ngeche*), riddles (*Ponge*), and folktales (*sigendni mochuogi*), into the corpus architecture. The paper situates the initiative within broader debates about digital linguistic equity, FAIR data principles, with text data deposited in CoNLL-U format and speech archived as WAV with TextGrid annotation files to ensure interoperability, the role of indigenous languages in sustainable development, and the decolonisation of computational linguistics. It argues that Dholuo corpus construction must go beyond lexical tokenisation to capture the pragmatic and metaphorical richness that characterises oral-indigenous discourse; figurative items are tagged with semantic domain labels in the Living Dictionaries companion platform. The findings demonstrate that native-speaker validation using an 80% accuracy threshold, combined with the Living Dictionaries platform for orthographic consistency checking and open-access deposition on Zenodo and Mozilla Common Voice, constitutes a replicable and community-empowering

model for enabling NLP in low-resource African languages. The corpus is partially multimodal: the Mozilla Common Voice speech recordings are sentence-aligned with the text pairs, while the interview and narration recordings constitute a separate spontaneous speech sub-corpus intended for ASR development.

Introduction

The twenty-first century has, in large part, been defined by the power of language technologies to mediate access to information, services, and social participation. Voice assistants, machine translators, spell-checkers, and sentiment analysis engines now form an invisible infrastructure through which billions of people navigate the digital world. Yet this infrastructure is built almost entirely on the linguistic labour of a handful of dominant languages, principally English, Mandarin, Spanish, and French, leaving the majority of the world's more than 7,000 languages functionally invisible in the digital domain (Joshi et al., 2020; Blasi et al., 2022).

Dholuo, also known as Luo and predominantly spoken in the Nyanza region of Western Kenya and the Mara region of Tanzania, is a case in point. A Western Nilotic language within the Nilo-Saharan family, Dholuo is the mother tongue of the Luo people and is spoken by an estimated 4 million native speakers (Ethnologue, 2022). The language traces its origins to the migration of Nilotic peoples from southern Sudan, who settled in the Great Lakes region of East Africa around the fifteenth century (Mbogho et al., 2025a). It is rich in oral literary tradition: its proverbs, riddles, folktales, and songs encode sophisticated philosophical and ecological knowledge accumulated over generations. Yet when a Dholuo speaker opens a smartphone, searches the internet, or navigates a health service portal, there is virtually nothing that speaks back in Dholuo. The language, as it were, exists outside the digital commons.

The consequences of this digital exclusion extend far beyond inconvenience. Awuor (2024) argues that indigenous languages in Kenya are essential pathways to sustainable development, enabling communities to access information, participate in civic life, and transmit knowledge across generations in culturally grounded ways. When these languages are absent from digital platforms, the communities that speak them are denied the benefits of an increasingly digital public sphere, including e-government services, digital health information, and online education. The language gap is thus not merely a technological problem but a development-equity issue with direct implications for achieving the Sustainable Development Goals in multilingual societies (Awuor, 2024).

The primary cause of Dholuo's digital exclusion is the absence of structured linguistic datasets. NLP systems, from simple autocomplete functions to large language models, learn from vast corpora of text and speech. Where such corpora do not exist for a given language, the technology does not work. For Dholuo, neither monolingual text corpora nor English-aligned parallel corpora have been developed at the scale required to train contemporary NLP models (Mbogho et al., 2025a; 2025b). This paper addresses this gap directly by presenting the design, methodology, and initial outcomes of a Dholuo parallel corpus project undertaken in Homabay County, funded through a USIU-Africa internal research grant.

The project's dual objectives are: (1) to construct a Dholuo–English parallel text corpus of 20,000 sentence pairs, and (2) to compile 30 hours of transcribed speech data. A distinctive feature of this initiative is its integration of oral literary forms, including metaphors (*weche mitiyo kodo kakaranyisi*), proverbs (*ngeche*), riddles (*ponge*), and folktales (s*igendni mochuogi*), as primary data sources, alongside

the use of the Living Dictionaries platform (*https://livingdictionaries.app/*) as both a lexicographic reference and a community documentation tool. The paper proceeds as follows. Section 2 reviews the relevant literature on NLP for African languages and the specific challenges facing Dholuo. Section 3 outlines the theoretical framework. Section 4 describes the methodology. Section 5 presents the results and discussion. Section 6 concludes with implications and recommendations.

NLP and the Low-Resource Language Problem

The phrase "low-resource language" has become technical shorthand in computational linguistics for languages that lack the extensive digital datasets required to train machine learning models (Joshi et al., 2020). The designation, however, is politically freighted: many so-called low-resource languages are spoken by millions of people who are resource-rich in oral tradition, cultural knowledge, and communicative vitality. The resource deficit is not intrinsic to the language but an artefact of historical, economic, and geopolitical inequalities that have concentrated language technology investment in already-dominant languages (Bird, 2020). Colonialism imposed foreign languages, including English, French, and Portuguese, as the official media of communication across Africa, and these languages continue to dominate education, governance, and technology at the expense of indigenous languages, marginalising the majority of the African population from access to essential information and digital tools (Mbogho et al., 2025a).

Joshi et al. (2020) developed an influential taxonomy that classifies the world's languages into five tiers based on the availability of NLP resources. Languages in the lowest two tiers, including Dholuo, are characterised by a near-total absence from digital platforms, scholarly publications, and AI applications. Blasi et al. (2022) documented systematic performance gaps across NLP tools for different language groups, showing that speakers of under-resourced languages receive qualitatively inferior technological services, with downstream consequences for healthcare access, civic participation, and economic opportunity. In Africa, Adebara and Abdul-Mageed (2022) showed that even widely spoken languages such as Yoruba, Igbo, and Twi remain severely underrepresented relative to their speaker populations.

Parallel corpora, aligned datasets in which the same content appears in two or more languages, are the bedrock of machine translation systems. Their absence for a given language pair makes it technically impossible to build reliable translation tools. Hu et al. (2020) and Tiedemann (2018) have shown that even modest parallel corpora of 20,000–50,000 sentence pairs can serve as viable training data for neural machine translation when combined with transfer learning from related higher-resource languages. The Kencorpus project (Wanjawa et al., 2022) offered an early proof of concept for Dholuo, collecting limited text-and-speech data for Dholuo, Luhya, and Swahili, but did not produce a parallel Dholuo–English corpus at the scale required for downstream NLP model development. The present study builds on this and subsequent work to address that gap at a larger scale and with greater attention to figurative linguistic registers.

Indigenous Languages, Education, and Sustainable Development in Kenya

Kenya is a linguistically diverse nation with approximately 68 living languages, of which 61 are indigenous (Eberhard et al., 2022). Despite this diversity, English and Swahili dominate official communication, higher education, and formal governance, while indigenous languages remain largely confined to informal and domestic domains (Ogechi, 2019; Wamalwa & Oluoch, 2018). This hierarchy has profound consequences for sustainable development. Awuor (2024) argues compellingly that indigenous languages are not peripheral to Kenya's development agenda but central to it: they are the primary means through which the majority of citizens, particularly in rural

areas, understand their rights, make health decisions, access education, and transmit knowledge across generations. Treating them as obstacles to modernisation rather than as assets for development is a fundamental policy error.

Kenya's Competency-Based Curriculum (CBC), implemented in 2017, mandates the use of mother-tongue languages as the primary medium of instruction in pre-primary education and Grades 1–3, reflecting a growing policy recognition of the pedagogical and cognitive benefits of mother-tongue education (Mukuthuria, 2020). However, Awuor (2024) notes that this policy commitment has not been matched by the development of adequate instructional materials, teacher training, or digital resources in indigenous languages, a gap that corpus development projects of the kind reported in this paper directly address. Without Dholuo textbooks, digital content, and NLP-enabled tools, the CBC's mother-tongue provisions risk remaining aspirational rather than operational.

The COVID-19 pandemic brought these inequities into sharp relief. Critical public health announcements were disseminated almost exclusively in English and Swahili, leaving millions of speakers of indigenous languages in rural areas without timely, comprehensible health information (Ngugi, 2021; Mbogho et al., 2025a). This experience galvanised a broader conversation about the structural inadequacies of Kenya's language technology infrastructure and underscored the urgency of developing digital resources to translate critical communications into indigenous languages at speed and scale.

African Language NLP: Community-Driven Approaches

The Masakhane project, launched in 2019, marked a turning point in African language NLP by demonstrating that community-driven, distributed collaboration could rapidly produce usable machine translation models for more than 30 African languages (Nekoto et al., 2020). Masakhane's model, which engaged native speakers and diaspora communities through GitHub, shared data repositories, and workshops, showed that language data extraction need not follow the extractive logic of earlier colonial linguistic documentation and could be genuinely participatory. Orife et al. (2020) extended this analysis to highlight structural challenges, including data scarcity, orthographic inconsistency, dialectal diversity, and the prevalence of code-switching, all of which complicate corpus construction for African languages.

For East African languages, Hussein et al. (2023) documented the specific challenges of building NLP resources in multilingual contexts where code-switching between Swahili, English, and indigenous languages is normative. The most directly relevant precedent for the present study is Mbogho et al. (2025a; 2025b), who report on a year-long crowdsourcing initiative at USIU-Africa and partner institutions to develop parallel text-and-speech corpora for three Kenyan languages: Dholuo, Kidaw'ida, and Kalenjin. Their project employed selective crowdsourcing, native-speaker translation workshops, and Mozilla Common Voice for speech collection, and deposited the resulting datasets on Zenodo under open-access licences. For Dholuo specifically, the project demonstrated both the feasibility of community-engaged data collection and the persistent challenges of orthographic standardisation and sustained community participation. The present study builds directly on this foundation while advancing it in two key directions: a dedicated focus on Dholuo figurative language as corpus content, and the integration of Living Dictionaries as a lexicographic companion tool.

Wanjawa et al. (2022) developed the Kencorpus. This earlier Dholuo text-and-speech datasetrved as a proof of concept for Dholuo NLP resource development, though its scope was limited and it did not yield a fully aligned parallel corpus with English. Together, these prior efforts mark a shift in the field

from the question of whether African language corpora can be built to how to build them most effectively, sustainably, and equitably.

Dholuo: Linguistic Features and Digital Status

Dholuo (ISO 639-3 code: luo) is a tonal, agglutinative language characterised by subject-verb-object word order, nominal gender distinctions, and a complex system of verb extensions that encode direction, intensity, and causativity (Tucker, 1994). Its phonological system includes prenasalised consonants, retroflex sounds, and a distinctive voiced dental plosive /ð/, represented orthographically as <dh>, hence the name Dholuo, meaning "the mouth/language of the Luo" (OED, 2024). Tone is phonemically distinctive in Dholuo, a feature that poses significant challenges for automatic speech recognition systems trained primarily on non-tonal European-language data. Mbogho et al. (2025a) note that Dholuo's use is largely restricted to informal settings, as English and Kiswahili dominate formal domains such as education, governance, and commerce in Kenya. This situation compounds the risk of what Calvet (as cited in Mbogho et al., 2025a) termed "glottophagy", the erosion of a language through absorption by a more dominant one.

Beyond its structural features, Dholuo is exceptionally rich in figurative and oral literary registers. Proverbs (Ngeche, Ngero) encode communal wisdom in elliptical, context-dependent formulas. Riddles (ponge) are used in social and pedagogical settings to sharpen inferential thinking. Metaphors (*weche mitiyo kodo kakaranyisi*) pervade everyday speech, grounding abstract concepts in agricultural, pastoral, and lacustrine imagery, with the lake (Nam Lolwe, Lake Victoria) and the homestead (dala) serving as master metaphors for community and belonging. Folktales (*sigendni mochuogi*) transmit historical, moral, and spiritual knowledge across generations. These oral forms are not decorative additions to Dholuo; they are its epistemological infrastructure. A corpus that omits them is, in an important sense, not a corpus of Dholuo at all. Awuor (2024) emphasises that indigenous knowledge systems embedded in such oral forms constitute irreplaceable intellectual heritage, whose digital preservation is inseparable from the broader project of sustainable development and cultural sovereignty.

Despite its vitality as a spoken language, Dholuo has a minimal digital presence. The Living Dictionaries platform (https://livingdictionaries.app/) hosts a growing Dholuo dictionary that includes audio recordings, semantic domain classifications, and example sentences, making it one of the few structured digital lexicographic resources currently available for the language (Living Dictionaries, 2025). The Kencorpus project (Wanjawa et al., 2022) and the subsequent USIU-Africa corpus initiative (Mbogho et al., 2025a) are the only substantive prior attempts to build NLP-ready datasets for Dholuo. As Mbogho et al. (2025b) document, Dholuo's status as a low-resource language means it lacks the comprehensive linguistic data, text corpora, speech recordings, and annotated resources required to develop applications such as machine translation, speech recognition, or language generation at scale. Without sustained and coordinated effort, Mbogho et al. (2025b) warn, Dholuo faces the threat of "digital extinction", a condition in which its speakers are systematically excluded from AI-powered communication technologies.

Theoretical Framework

Two complementary theoretical orientations guide this study. The first is the framework of linguistic decolonisation, articulated by Bird (2020) and further developed in African contexts by Adebara and Abdul-Mageed (2022). This framework insists that language technology development for indigenous languages must be community-led, culturally grounded, and oriented towards the communicative needs and creative norms of speaker communities rather than the convenience of external researchers

or technology corporations. Concretely, this means that the design of a Dholuo corpus cannot simply replicate templates developed for English or French; it must take seriously the oral, figurative, and communal dimensions of Dholuo as a language system. For instance, standard tokenisation pipelines assume whitespace as the primary word boundary. Yet Dholuo's agglutinative morphology and oral literary conventions for riddles and proverbs require boundary-detection strategies that no English-trained model provides out of the box.

The second theoretical anchor is corpus linguistics, a methodology for linguistic description and NLP resource development (McEnery & Wilson, 2001; Tiedemann, 2018). Corpus linguistics provides the technical framework for designing data collection, developing annotation schemas, establishing quality assurance procedures, and evaluating corpus representativeness. The present study adopts a balanced corpus design that seeks proportional representation of different registers, namely conversational, formal, oral literary, and transactional, to maximise the corpus's utility for diverse downstream NLP applications. Mbogho et al. (2025a) demonstrate, through their crowdsourcing methodology, that this balance can be achieved even in low-resource settings when community engagement is systematic and sustained, providing a direct methodological precedent for the approach taken here.

The third theoretical pillar draws on scholarship on language and development in African contexts. Awuor (2024) argues that indigenous languages are not merely cultural artefacts but functional development tools: they are the medium through which grassroots communities access health information, understand legal rights, participate in education, and engage with governance. From this perspective, corpus development is not a narrowly technical exercise but an investment in development, extending digital infrastructure to language communities historically excluded from it. This development-oriented framing of language technology work shapes the domain distribution choices in the corpus (prioritising health, agriculture, and education content) and the participatory approach to data collection and quality assurance.

Bridging these frameworks is the concept of culturally-situated NLP, which holds that effective language technology for oral-indigenous languages requires not only technical adequacy and sufficient training data, but also cultural adequacy: sensitivity to the pragmatic, figurative, and social dimensions of language use that determine whether a technology is genuinely useful to a speaker community (Nekoto et al., 2020; Awuor, 2024). This principle underpins every methodological decision in the present study, from deliberately including proverbs and riddles in the corpus to using the Living Dictionaries platform as a community-facing documentation interface.

Methodology

Research Design

This study adopts a mixed-methods corpus linguistics design, combining quantitative data collection (sentence pairs and hours of speech) with qualitative analysis of figurative language patterns and community feedback. The research is framed as a participatory, community-engaged project in which native Dholuo speakers serve not only as data contributors but also as translation validators and quality reviewers, consistent with the community-driven methodology shown to be effective by Mbogho et al. (2025a) in their prior Dholuo, Kidaw'ida, and Kalenjin corpus work. Ethics approval was obtained from the USIU-Africa Institutional Review Board (IRB), and NACOSTI registration was completed before data collection. All participants provided written informed consent before contributing to the project.

The study builds directly on the methodological architecture developed in Mbogho et al. (2025a; 2025b) while extending it in scope and depth. Whereas the prior project collected data across three languages simultaneously, the present study concentrates exclusively on Dholuo, allowing greater attention to linguistic nuance, figurative register, and oral literary documentation. The selective crowdsourcing methodology, which recruits committed native speakers rather than attempting broad, undifferentiated crowd engagement, is adopted from Mbogho et al. (2025a), whose experience showed that quality and consistency are better achieved through a smaller cohort of engaged contributors than through large-scale, lightly supervised crowdsourcing.

Text Data Collection

English source texts were drawn from Creative Commons–licensed repositories, including African Storybook (https://www.africanstorybook.org/), public domain works, open-access county government documents, and materials from USIU-Africa's language department. Priority was given to texts with communicative relevance to Dholuo-speaking communities: public health advisories, agricultural extension materials, primary school readers, civic communications, and transcripts of locally produced oral literature. A dedicated online crowdsourcing platform was developed to enable community members to translate English sentences into Dholuo. Non-English-proficient speakers were invited to contribute original Dholuo texts, which bilingual participants then translated into English, following the bidirectional collection approach used in Mbogho et al. (2025a).

A target of 20,000 parallel sentence pairs was set, drawing on precedent from comparable low-resource corpus projects (Mbogho et al., 2025a; Hu et al., 2020). Sentences were selected to represent diverse registers and domains: health, agriculture, education, community governance, and oral literature. Special effort was made to include sentence pairs featuring Dholuo figurative language, enabling future NLP researchers to develop metaphor-processing and pragmatics-aware translation models. Community workshops were organised in Dholuo-speaking counties, Siaya, Kisumu, Homa Bay, and Migori, where participants from diverse age groups were invited to contribute original content. The inclusion of older community members and oral tradition custodians was a deliberate methodological priority, as figurative registers and archaic vocabulary are most concentrated in this demographic.

Speech Data Collection

Speech data were collected through two complementary approaches. First, structured interviews and guided conversations were conducted with native Dholuo speakers across both major dialect regions: Alego/Ugenya in the Central and South Nyanza. Participants were invited to narrate in response to image prompts, answer scenario-based questions, and recount proverbs and riddles in their natural discursive contexts. Recordings were made using standard audio equipment, transcribed by trained Dholuo transcribers, and translated into English. Second, the collected text data were uploaded to Mozilla Common Voice, the same open-source platform used by Mbogho et al. (2025a). Dholuo-community contributors were coordinated through Mozilla Pontoon (Mozilla Pontoon, 2025), where the Common Voice recordings are sentence-aligned with the text pairs and constitute the multimodal component of the corpus; the structured interviews and narrations form a separate, non-aligned spontaneous speech sub-corpus. Volunteer readers recorded sentences in Dholuo. A target of 30 hours of transcribed speech across multiple speakers, genders, dialect regions, and age groups was set, ensuring a dataset with sufficient phonological diversity for ASR model development.

The approach to speech data collection drew on lessons from Mbogho et al. (2025a), who found that community members require sustained engagement and reasonable remuneration to maintain

participation over time. Accordingly, all participating speakers received a stipend commensurate with their time contribution, and regular feedback sessions were held to sustain community investment in the project. Speakers were recorded multiple times reading the same sentences to capture dialectal variation and individual phonological differences, a practice Mbogho et al. (2025a) identified as essential for building robust, production-quality ASR datasets.

Living Dictionaries Integration

The Living Dictionaries platform (https://livingdictionaries.app/) was integrated into the project as a lexicographic anchor and a community documentation tool. Living Dictionaries is a free, open-access multimedia platform created by the Living Tongues Institute for Endangered Languages, serving over 200 language communities worldwide with more than 254,000 published entries (Living Dictionaries, 2025). For Dholuo, the platform enables community members and linguists to collaboratively build a digital dictionary that includes audio pronunciations, semantic domain tags, phonetic transcriptions, sample sentences, and images, making it one of the most accessible and feature-rich lexicographic tools for the language.

In this study, Living Dictionaries served three functions: (1) as a reference lexicon to validate the orthographic consistency of the collected corpus data; (2) as a platform for entering and contextualising figurative vocabulary items, proverbs, metaphorical expressions, and riddle lexemes that emerged during corpus collection; and (3) as a sustainability mechanism, ensuring that lexical knowledge documented during the corpus project would remain accessible to the community beyond the project's funded period. This integration represents a methodological innovation over the approach taken in Mbogho et al. (2025a), which focused on depositing datasets on Zenodo and Mozilla Common Voice but did not use a dedicated lexicographic platform for ongoing community vocabulary documentation. The Living Dictionaries integration creates a complementary and mutually reinforcing documentation ecosystem.

Quality Assurance

A three-stage quality assurance process was implemented, adapted, and extended from the expert review procedures described in Mbogho et al. (2025a). In Stage 1, all English source texts were reviewed for clarity and cultural appropriateness before translation. In Stage 2, each Dholuo translation was reviewed by an independent native-speaker validator who had not participated in the initial translation. Translations scoring below 80% on a standardised rubric assessing semantic accuracy, grammatical correctness, idiomatic naturalness, and tonal orthographic consistency were returned to translators for revision. In Stage 3, a panel of three senior Dholuo linguists reviewed a 10% random sample of the corpus for semantic accuracy, tonal orthography, and figurative appropriateness. The involvement of native-speaker linguists at every stage of quality assurance reflects the principle, emphasised across the corpus linguistics and language documentation literature, that insider linguistic competence is irreplaceable in low-resource language data work (Nekoto et al., 2020; Awuor, 2024).

Results and Discussion

Corpus Overview and Quantitative Findings

The project has amassed a corpus of approximately 18,400 parallel sentence pairs at the time of this writing, with collection ongoing towards the 20,000-sentence target. Speech data comprised 24.5 hours of transcribed recordings from 47 unique speakers representing both the Alego/Ugenya and South Nyanza dialect regions, including 23 male and 24 female speakers. The domain distribution in the text corpus reflects the project's theoretical and practical priorities, with oral literature and figurative

language deliberately over-represented relative to typical NLP corpora. Table I below summarises the corpus composition at the current stage of the project.

Table I: Dholuo–English Parallel Corpus: Current Composition

Domain	Sentence Pairs	% of Corpus
Oral Literature & Figurative Language	4,600	25.0%
Public Health	4,048	22.0%
Education	3,312	18.0%
Agriculture & Environment	3,128	17.0%
Civic Communication	1,840	10.0%
General Conversation	1,472	8.0%
TOTAL	18,400	100.0%

These figures are broadly comparable to, and in the Dholuo-specific context, represent a substantial advance over, the datasets reported in the prior USIU-Africa corpus project, which collected data concurrently across three languages. Mbogho et al. (2025b) note that the Dholuo datasets from their project were among the most challenging to produce, owing to the scarcity of existing digital Dholuo texts and the need for specialised expertise in figurative language. By concentrating resources on a single language and extending the methodology to include enhanced oral literature collection, the present project has produced a Dholuo corpus that is both larger and more linguistically representative than prior efforts.

Figurative Language in the Dholuo Corpus

One of the most substantively significant findings of this project concerns the centrality and structural diversity of figurative language in the collected Dholuo data. Across all domains, 31% of the collected Dholuo sentences contained at least one figurative element, a proportion far higher than in comparable English corpus samples drawn from the same source texts. This finding confirms what native-speaker linguists, ethnographers, and scholars of African oral literature have long maintained: that Dholuo is not merely an instrumental medium of communication but a poetic language in which figurative expression is the default rather than an exceptional register. It also reinforces Awuor's (2024) broader argument that indigenous African languages encode ways of knowing that cannot be reduced to simple lexical equivalents in English or Swahili.

Proverbs (sunga) were especially prominent in health communication contexts, where speakers drew on established wisdom formulas. Translation validators cross-referenced figurative expressions with the Glosbe English–Luo dictionary (Glosbe, 2025), a community-built parallel corpus resource that provided contextualised Dholuo translations and audio examples for proverbs and metaphors. Speakers drew on to frame medical advice in culturally resonant terms. For example:

> "*Ng'at ma ok ong'iyo tedo ok nyal chiemo maber.*"
> (Dholuo: "One who does not learn to cook cannot eat well.")

This proverb was used in health education contexts. Translators cross-checked such figurative mappings against the Glosbe English–Luo dictionary (Glosbe, 2025), which provided parallel-corpus evidence of contextualised Dholuo proverb usage. This proverb was used in health education contexts to frame patient compliance with treatment regimes as an act of self-investment, activating a familiar domestic cultural schema to convey a nuanced behavioural message. The metaphorical mapping of health management onto the domestic skill of cooking exemplifies the kind of pragmatic complexity that standard machine translation pipelines, trained on formal written corpora, are poorly equipped to handle. For NLP systems attempting to translate such utterances, the challenge is not merely lexical substitution but the preservation of pragmatic force, cultural resonance, and communicative intent.

Riddles (wicho) presented a distinctive computational challenge. Dholuo riddles are structurally two-part: a question-like prompt (wicho) is followed by an answer (*duok wicho*) that resolves the riddle's productive ambiguity through an unexpected conceptual mapping:

> *Wicho:* "*Ng'at ma wuotho ka wuotho to ok chopi.*"
> (Translation: "One who walks and walks but never arrives.")
> *Duok wicho:* "*ndara/apaya.*"
> (Answer: "The road itself.")

The cognitive operation required to resolve this riddle, mapping "walking without arriving" onto the abstract entity of the road, exemplifies what Dholuo speakers describe as understanding that goes beneath the surface meaning of words. Including such data in the corpus makes it available for future research on pragmatics-aware NLP and figurative language processing, areas that are increasingly recognised as critical frontiers in computational linguistics research (Yilmaz et al., 2021).

The oral literature data also yielded a rich set of environmental and agricultural metaphors that reflect the Luo community's historically close relationship with Lake Victoria and the surrounding landscape. Expressions such as:

> "*Nam en baba mar jopiny.*"
> (Translation: "The lake is the father/provider of the people of this land.")

This extends the corpus's usefulness beyond NLP into environmental linguistics, climate communication, and cultural knowledge documentation. This aligns with Awuor's (2024) emphasis on the role of indigenous languages in sustainable development: the vocabulary and metaphors through which Dholuo speakers understand their ecological environment constitute a form of indigenous knowledge that corpus documentation can help preserve and transmit to future generations.

Living Dictionaries as a Corpus Companion Tool

The integration of Living Dictionaries (https://livingdictionaries.app/) into the corpus workflow proved more productive than initially anticipated. Throughout the project period, over 350 Dholuo lexical items identified during corpus collection were entered into the Living Dictionaries platform and enriched with audio recordings by native speakers, semantic domain tags, and example sentences drawn from the corpus. This bidirectional exchange, where the corpus informed the dictionary and the dictionary validated the corpus, created a virtuous cycle of documentation that extended the project's reach beyond the immediate dataset deliverables.

In practice, Living Dictionaries served as a real-time quality-control mechanism. When corpus validators encountered orthographic inconsistencies in collected translations, a persistent challenge

given the absence of a fully standardised Dholuo orthography, as documented by Mbogho et al. (2025a), they consulted the dictionary's phonetic transcription fields to adjudicate between variant spellings. The platform's audio functionality enabled validators to verify that written forms corresponded to phonological realities, particularly for tonal distinctions that are typically unmarked in everyday Dholuo writing.

The Living Dictionaries platform also proved valuable as a community engagement tool. Because the dictionary interface is accessible via mobile phones and does not require technical literacy, community members who could not participate in the online translation platform could still contribute to the broader documentation effort by recording words, adding example sentences, or flagging errors in existing entries. This extended the project's reach beyond the formally recruited participant pool and helped build community ownership of the documentation enterprise, a critical dimension of sustainable language revitalisation and one that Awuor (2024) identifies as essential to ensuring that language development initiatives are genuinely empowering rather than extractive.

Methodological Challenges and Adaptations

The project encountered several challenges that required methodological adaptations, echoing and extending those documented in Mbogho et al. (2025a). First, the absence of a standardised Dholuo orthography created inconsistencies in the crowdsourced translation data. Different writing traditions persist across religious and educational communities in Western Kenya. The project responded by developing a minimal orthographic style guide, drawing on Tucker (1994) and the Living Dictionaries Dholuo entries, and distributing it to all translators and validators before data collection began. Post-distribution quality checks confirmed a measurable improvement: orthographic inconsistency rates in collected translations declined from an estimated 18% during the first collection phase to approximately 7% after all translators and validators adopted the guide, indicating that this minimal standardisation intervention was effective, even if it did not eliminate the problem entirely. The style guide prioritised tone marking on minimal pairs, word segmentation conventions for agglutinative verb forms, and consistent treatment of prenasalised consonants, the three areas identified by the quality assurance panel as most prone to inconsistency.

Second, the tonal character of Dholuo posed challenges for transcription quality in the speech corpus. Tone marks are rarely used in everyday Dholuo writing, so transcribers had to make explicit decisions about tonal representation that would not arise when transcribing an atonal language. A supplementary transcription protocol was developed in consultation with the quality assurance panel, establishing conventions for marking High (H), Low (L), and Falling (HL) tones on lexically significant items. This protocol is being shared openly with the Dholuo NLP community to promote consistency across future corpus projects. Given that manual tonal transcription is a known bottleneck in corpus production, the protocol prioritised tone marking on ambiguous minimal pairs and high-frequency lexical items rather than exhaustive item-level annotation, thereby balancing linguistic rigour with the practical demands of large-scale data collection. ('Lexically significant items' are defined here as lexical content words, principally nouns, verbs, and adjectives, in which tonal contrasts are semantically or grammatically distinctive, distinguishing minimal pairs such as kom [chair] vs kom [to plant], as opposed to grammatical function words, where tonal patterns are largely predictable from syntactic context.)

Third, the corpus's figurative language component required specialist cultural knowledge that exceeded the competencies of some recruited translators. Proverbs and riddles, in particular, are highly context-dependent, and their translation into English required not only linguistic but also

ethnographic expertise. A dedicated oral literature consultant, a retired secondary school teacher and community storyteller from Siaya County with deep knowledge of Dholuo oral genres, was recruited to provide specialist validation for the corpus's figurative language subsection. This experience reinforces Awuor's (2024) argument that indigenous language documentation requires not only linguistic training but also cultural insidership, and that the most effective corpus teams will combine both. The specialist validation led to significant revisions: approximately 12% of the figurative language corpus entries required correction or contextual re-annotation, confirming that cultural insidership is not merely desirable but methodologically essential for corpus work involving oral literary genres.

Fourth, sustained community engagement proved challenging throughout the multi-month project, consistent with the findings of Mbogho et al. (2025a), who observed that initial enthusiasm among community volunteers does not automatically translate into long-term participation. The project addressed this by restructuring data collection into shorter, well-compensated engagement sessions rather than extended open-ended contributions, and by maintaining regular communication with participants on the project's progress and impact.

Implications for NLP Applications

The corpus constructed through this project has immediate applications across several NLP use cases. Most directly, it provides training data for Dholuo–English neural machine translation, enabling the future development of translation tools that could transform access to healthcare, government, and educational resources for Dholuo-speaking communities. At 20,000 sentence pairs, the corpus falls within the range identified by Hu et al. (2020) as viable for low-resource machine translation when combined with cross-lingual transfer learning from related languages such as Acholi or Lango. The corpus's domain distribution, with particular strength in health, education, and oral literature, makes it directly applicable to the development-equity use cases that Awuor (2024) identifies as most critical for indigenous language communities.

The speech corpus supports the development of automatic speech recognition (ASR) systems. Given the phonological complexity of Dholuo, particularly its tonal distinctions, prenasalised consonants, and the voiced dental plosive /ð/, existing multilingual ASR models perform poorly on Dholuo audio. A dedicated 30-hour speech corpus provides a starting point for fine-tuning existing models and for benchmarking ASR performance. The corpus's dialectal diversity (Alego/Ugenya and South Nyanza speakers) is a particular strength, offering exposure to the phonological variation that any production-quality ASR system would need to handle. Mbogho et al. (2025b) note that multi-speaker, multi-dialect speech datasets are essential for building ASR systems that work across the geographic distribution of a language's speaker community, rather than only for a prestige dialect.

The datasets produced by this project are available on Zenodo under a CC BY-SA licence, following the open-access, FAIR-compliant dissemination model adopted by Mbogho et al. (2025a). Speech data will be mirrored on Mozilla Common Voice to support ongoing community contributions. This dual-platform strategy ensures both academic accessibility (via Zenodo's DOI-linked repository) and community accessibility (via Mozilla's mobile-friendly interface), maximising the corpus's downstream reach and utility. The project team has also engaged with the FORCE11 working group on FAIR data to ensure that metadata standards fully comply with international best practice. Text corpus sentence pairs are deposited in tab-separated UTF-8 plain text and CoNLL-U format, enabling direct compatibility with standard NLP pipelines. Speech data is archived in 16-bit WAV format with corresponding EUDICO Linguistic Annotator (ELAN)/Praat-compatible TextGrid annotation files.

All Zenodo deposits carry JSON-LD metadata conforming to Dublin Core and the CLARIN Component Metadata Infrastructure (CMDI) standard, satisfying the Interoperability requirement of the FAIR principles.

Conclusion

This paper has reported on the design, methodology, and interim findings of an ongoing initiative to build a Dholuo–English parallel corpus as infrastructure for Natural Language Processing applications in one of Kenya's major indigenous languages. Three features distinguish the project. First, it is theoretically grounded in the principle that corpus construction for an oral-indigenous language must centre figurative and communal dimensions of language use, not merely produce tokenizable text. Second, it makes a methodological contribution by integrating the Living Dictionaries platform as both a lexicographic reference and a community documentation tool, one that, with over 350 Dholuo entries enriched with audio and corpus-linked example sentences, proved productive well beyond initial expectations, creating a self-reinforcing documentation ecosystem. Third, all datasets are deposited as open-access, FAIR-compliant resources on Zenodo (in CoNLL-U and WAV/Text Grid formats) and Mozilla Common Voice, maximising downstream reach and long-term sustainability. Taken together, these features advance and extend the foundational corpus work established by Mbogho et al. (2025a; 2025b) in important new directions.

Three findings from this project merit particular attention. First, 31% of the collected Dholuo sentences contain at least one figurative element, a proportion that is far higher than in comparable English source texts drawn from the same domains, confirming that figurative expression is the default communicative register in Dholuo, not an exceptional one. Any corpus that omits oral literary data, proverbs, riddles, metaphors, and folktales systematically misrepresents the language and will produce NLP tools that fail in precisely the contexts where Dholuo speakers communicate most richly. Translators validated figurative vocabulary against both Living Dictionaries and the Glosbe English–Luo dictionary (Glosbe, 2025), which provided community-contributed parallel examples for proverbs and metaphorical expressions. Second, the Living Dictionaries integration demonstrates that lexicographic and corpus-building workflows can be mutually reinforcing: the corpus fed new entries into the dictionary, and the dictionary served as a real-time orthographic validator for the corpus. Third, the methodological challenges encountered, orthographic inconsistency, tonal transcription, and the specialist expertise required for figurative language, offer instructive, replicable lessons for low-resource African language corpus projects more broadly. As Awuor (2024) has argued, indigenous languages are pathways to sustainable development: when they are digitally enabled, the communities that speak them gain access to the information, services, and civic participation that modern life increasingly requires. Corpus development of the kind reported here is therefore not merely a contribution to computational linguistics but an act of development-equity.

Several limitations of the present study should be acknowledged. The 18,400 sentence pairs collected at the time of writing fall approximately 8% short of the 20,000-pair target, a threshold drawn from Hu et al. (2020), who identify corpora in the 20,000–50,000 sentence-pair range as the minimum viable scale for low-resource neural machine translation when supplemented with cross-lingual transfer learning from typologically related languages such as Acholi or Lango. This is not an arbitrary figure but a principled, literature-grounded benchmark; the shortfall means the corpus is functional for exploratory NMT research but should be expanded before deployment in production translation systems. The 24.5-hour speech corpus similarly falls short of the 30-hour target and is sufficient for exploratory ASR research but not for training production-quality models. Despite the orthographic style guide adopted during collection, two specific residual inconsistencies remain: (1) inconsistent

Falling (HL) tone marking on verb extensions, particularly causative and directional suffixes; and (2) vowel harmony representation in loanwords of Swahili and English origin. These are documented in the corpus metadata and earmarked for resolution in future annotation cycles. Future work should prioritise corpus expansion through partnerships with Kenyan community radio stations broadcasting in Dholuo, the Ministry of Education, and diaspora Luo communities. Integrating the corpus with Kenya's Competency-Based Curriculum mother-tongue initiative represents a particularly high-impact application pathway.

The aspiration underpinning this project is expressed most eloquently in the Dholuo proverb:

> "*Wach moro amora nigi chenro ma kele.*"
> ("Every matter has a way that leads to it.")

The path to digital equity for Dholuo speakers begins with data. This corpus is one step on that path, made possible by the knowledge, generosity, and patience of the Dholuo-speaking community members who contributed their voices, words, and stories to this project.

Acknowledgements

The author gratefully acknowledges the support of USIU-Africa through its internal research grant programme. Special thanks are due to Prof. Audrey Mbogho (USIU-Africa), who led the Lacuna Fund-funded project on which this study builds and in which the author served as Co-Principal Investigator and co-author (Mbogho et al., 2025a; 2025b). The Dholuo-speaking community members, translators, and oral literature consultants who generously offered their time, knowledge, and voices are, in the truest sense, the authors of this work. The Living Tongues Institute for Endangered Languages is thanked for the opportunity to contribute to the Living Dictionaries platform. This research was conducted with approval from the USIU-Africa Institutional Review Board and is registered with NACOSTI.

References

Adebara, I., & Abdul-Mageed, M. (2022). The feasibility of NLP for African languages: An empirical study. *Transactions of the Association for Computational Linguistics, 10*, 123–142.

Awuor, O. Q. (2024). Indigenous languages: A pathway to sustainable development in Kenya. *Journal of the African Language Teachers Association, 11*, 59–78.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020* (pp. 3504–3519). Association for Computational Linguistics.

Blasi, D. E., Anastasopoulos, A., & Neubig, G. (2022). Systematic inequalities in language technology performance across the world's languages. *Transactions of the Association for Computational Linguistics, 10*, 82–98.

Capen, C. (1998). *Dholuo–English Dictionary*. Uzima Press.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2022). *Ethnologue: Languages of the World*. SIL International. https://www.ethnologue.com/country/KE

Glosbe. (2025). English–Luo dictionary: Proverb. Glosbe Multilingual Online Dictionary. https://glosbe.com/en/luo/PROVERB

Hinton, L., & Hale, K. (2001). *The Green Book of Language Revitalisation in Practice*. Academic Press.

Hu, J., Tiedemann, J., & Hämäläinen, M. (2020). Advancing parallel corpora for under-resourced languages. *Journal of Computational Linguistics, 46*(2), 345–367.

Hussein, M., Ramesh, M., & Wafula, J. (2023). Building low-resource African language datasets for NLP: Challenges and opportunities. *African Journal of AI Research, 5*(2), 102–118.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in NLP research. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1278–1293). ACL.

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.

Living Dictionaries. (2025). *Living Dictionaries: Language documentation platform*. Living Tongues Institute for Endangered Languages. https://livingdictionaries.app/

Mbogho, A., Awuor, Q., Kipkebut, A., Wanzare, L., & Oloo, V. (2025a). Building corpora for low-resource Kenyan languages. *Journal of the Digital Humanities Association of Southern Africa (DHASA), 6*(2). https://doi.org/10.55492/v6i02.6747

Mbogho, A., Awuor, Q., Kipkebut, A., Wanzare, L., & Oloo, V. (2025b). Building low-resource African language corpora: A Kidawida, Kalenjin, and Dholuo case study. *arXiv preprint*. https://doi.org/10.48550/arXiv.2501.11003v1

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics* (2nd ed.). Edinburgh University Press.

Mukuthuria, M. (2020). Language policy and linguistic rights in Kenya: A historical perspective. *Journal of African Linguistics, 12*(1), 45–59.

Myers-Scotton, C. (2021). *Code-Switching as a Worldview: The Intersection of Language and Identity in Multilingual Societies*. Oxford University Press.

Mozilla Pontoon. (2025). Dholuo (luo): Common Voice contributors. Mozilla's Localisation Platform. https://pontoon.mozilla.org/luo/common-voice/contributors/

Nekoto, W., Orife, I., Kreutzer, J., & Masakhane Community. (2020). Participatory approaches to African language NLP. In *Proceedings of the ACL Workshop on Findings* (pp. 4156–4170). ACL.

Ngugi, J. (2021). Language barriers and public health communication during the COVID-19 pandemic in Kenya. *African Journal of Public Health, 15*(3), 234–249.

Ogechi, N. (2019). *Swahili and Indigenous Kenyan Languages in Public Discourse: A Linguistic Landscape Analysis*. University of Nairobi Press.

Orife, I., Kreutzer, J., & van der Wilk, M. (2020). Low-resource African NLP: Challenges and potential solutions. In *Proceedings of COLING 2020* (pp. 245–259). ACL.

Oxford English Dictionary. (2024). Dholuo, n. & adj. Oxford University Press. https://www.oed.com/dictionary/dholuo_n

Tiedemann, J. (2018). Parallel data, tools, and interfaces in OPUS. In *Proceedings of the 11th International Conference on Language Resources and Evaluation* (pp. 99–105). ELRA.

Tucker, A. N. (1994). *A Grammar of Kenya Luo (Dholuo)*. Rüdiger Köppe Verlag.

Wamalwa, E. W., & Oluoch, R. (2018). Language policy and education in Kenya: Challenges and prospects. *International Journal of Education and Research, 6*(4), 123–137.

Wanjawa, B., Wanzare, L., Indede, F., McOnyango, O., Ombui, E., & Muchemi, L. (2022). Kencorpus: A Kenyan language corpus of Swahili, Dholuo and Luhya for natural language processing tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.2208.12081

Yilmaz, E., Hupkes, D., & Dupoux, E. (2021). Cultural aspects of NLP: Challenges in the preservation of linguistic heritage. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 4156–4170). ACL.