



YorùbáÒwe: A Rule-based Bidirectional Web System for Yorùbá–English Idiom and Proverb Translation with Context-Aware Disambiguation

A Lexicon-Matching Computational Linguistics Approach to Non-Compositional Text Translation for a Low-Resource Nigerian Language

O.A. Odeniyi

Department of Computer Science
Osun State College of Technology
Esa-Oke, Nigeria

O.O. Ogunsuyi

Department of Computer Science
Osun State College of Technology
Esa-Oke, Nigeria

A.R. Akinwumi

Department of Computer Science
Osun State College of Technology
Esa-Oke, Nigeria

ABSTRACT

Idioms and proverbs (òwe) are central to Yorùbá discourse, yet computational Yorùbá NLP has concentrated on compositional tasks, leaving non-compositional expressions largely unaddressed. This study presents YorùbáÒwe, a publicly deployed web application for bidirectional Yorùbá↔English idiom and proverb translation. Its engine combines input normalization, sliding-window proverb detection, exact lexicon matching, near-match retrieval, rule-based context disambiguation, and word-level fallback. The engine's production source code was evaluated directly through a multi-layer evaluation. An initial evaluation against a 13-entry seed lexicon traced every Yorùbá-side miss (69.23% exact-match recall) to a hard-coded 12-token segmentation ceiling. The lexicon was then expanded to approximately 578 entries from a 568-pair Yorùbá–English proverb corpus, which was divided deterministically (80/20, random seed 42) into a 454-record Development/Lexicon Set and a 114-record Held-Out Test Set. All 13-unit tests passed, and the production build succeeded. On the Development/Lexicon Set the system retrieved the matching entry for 451/454 records (99.34% exact-match rate) and handled 806/818 controlled variants (98.53%); these figures describe lexicon-supported behavior, not general translation accuracy. On the Held-Out Test Set, evaluated with the held-out entries removed from the lexicon, reference-overlap scores were very low (BLEU 0.26, chrF 6.64, ROUGE-L F1 6.41; 0/114 exact textual matches), whereas a ChatGPT baseline on the same 114 records scored BLEU 38.42, chrF 58.77, and ROUGE-L F1 66.00 (4/114 exact textual matches). These metrics measure similarity to a single reference translation and do not establish translation quality. A human evaluation of 20 selected proverbs by three evaluators (60 judgments) gave mean ratings of 4.72/5 for meaning preservation, 4.72/5 for English naturalness, and 4.83/5 for cultural appropriateness, with 59/60 (98.33%) judged acceptable; an earlier 77-responder evaluation of the 13 seed entries is also reported. YorùbáÒwe is therefore characterised as a lexicon-based, retrieval-oriented system whose coverage depends on its stored proverb knowledge.

General Terms

Algorithms, Design, Experimentation, Languages, Verification.

Keywords

Yorùbá language; idiom translation; proverb translation; òwe; natural language processing; multiword expressions; low-

resource language; rule-based lexicon matching; web-based system; held-out evaluation; LLM baseline; human evaluation.

1. INTRODUCTION

1.1 Background to Study

Computational treatment of Yorùbá, a tonal Niger-Congo language spoken by an estimated 50 million people primarily in Nigeria and its diaspora, has expanded considerably over the past decade, spanning diacritic restoration, part-of-speech tagging, morphological analysis, sentiment analysis, speech corpora, and machine translation [1, 2]. The great majority of this work, however, is compositional: it treats meaning as derivable, word by word or morpheme by morpheme, from a sentence's constituent parts. Yorùbá discourse relies heavily on òwe — proverbs and idiomatic expressions — whose meaning is non-compositional: a literal, word-for-word rendering of an òwe typically fails to convey, and can actively obscure, its figurative sense. Ademowo and Balogun [3] and Owomoyela [4] both document the pervasiveness of proverbs across Yorùbá social, educational, and conflict-mediation contexts, while Ehineni [5] shows that proverbs perform identifiable discourse functions — warning, rebuke, consultation, endorsement — that a native or fluent listener recognises from structural and contextual cues rather than from the individual words alone.

In the wider natural language processing literature, expressions of this kind are studied under the heading of multiword expressions (MWEs): idiosyncratic, non-compositional word sequences that resist literal translation [6]. Carpuat and Diab [7] and, more recently, Liu, Ghosh, and Jiang [2] demonstrate empirically that verbal and idiomatic multiword expressions measurably and consistently degrade machine translation quality relative to compositional text, and that dedicated detection and handling — rather than reliance on general-purpose translation — is necessary to preserve figurative meaning. For Yorùbá specifically, dedicated idiom- or proverb-level translation tooling has been almost entirely absent from the published record; existing rule-based Yorùbá machine translation systems target verb phrases [8], noun phrases, and numerals, while the one recent computational treatment of Yorùbá proverbs identified in the literature, Olusanya, Masakhane NLP, and Abolade's OWE-YOR [9], addresses proverb classification (is this sentence a proverb or not) rather than translation of a proverb's figurative meaning.

This study addresses that gap by presenting and empirically



evaluating YorùbáŌwe, a live web-based system deployed at <https://yorubaidiom.onrender.com>, which translates Yorùbá idioms and proverbs into their English figurative equivalents and vice versa, using a segmentation-and-lexicon-lookup engine with rule-based context disambiguation.

1.2 Statement of the Problem

General-purpose machine translation systems, whether rule-based or neural, are optimized for compositional accuracy and consequently tend to translate idiomatic Yorùbá expressions word-for-word, producing output that is grammatically well-formed but semantically nonsensical or misleading to an English-speaking reader. A learner, researcher, or cultural-heritage application that needs the figurative sense of an òwe — not its literal gloss — has no dedicated, openly available tool to consult. Where computational resources for Yorùbá proverbs do exist, they are oriented towards classification or corpus-building rather than bidirectional, user-facing translation with an explanation of both the literal and figurative reading. This leaves a practical gap between the descriptive linguistic literature on Yorùbá proverbs [3, 4, 5] and the applied NLP tooling available to translate them.

1.3 Aim and Objectives of the Study

The aim of this study is to design, deploy, and empirically evaluate a web-based system capable of bidirectionally translating Yorùbá idioms and proverbs into English (and vice versa) with figurative-sense fidelity. The specific objectives are to:

formalise a segmentation-and-lexicon-matching approach capable of locating idiomatic spans of variable length within free-form input text;

design and implement a bidirectional lexicon data structure and a rule-based context disambiguation module capable of distinguishing figurative from literal usage of a matched span;

design and implement a full-stack web-based system architecture, comprising a React 19/Vite browser client and an Express.js/TypeScript application layer, with a curator workflow for lexicon maintenance;

evaluate the correctness, robustness, and computational performance of the deployed system's own production source code through direct execution of its test suites; assess its behaviour on held-out proverbs that are not represented in its lexicon, alongside a general-purpose language-model baseline on the same items; complement the automatic evaluation with human judgements of meaning preservation, naturalness, and cultural appropriateness; and identify and diagnose any systematic sources of translation failure.

1.4 Significance of the Study

This study contributes an openly reproducible, code-verified case study of idiom-level (as distinct from word- or numeral-level) machine translation for a low-resource Nigerian language, complementing the growing body of Yorùbá computational linguistics scholarship documented by Jimoh, De Wille, and Nikolov [1] and Inuwa-Dutse [10]. By evaluating the actual deployed engine rather than a theoretical design, the study also demonstrates a methodology — direct import-and-execute evaluation of production source code — that other student and early-stage NLP projects working with small, evolving lexicons can adopt to obtain honest, falsifiable performance figures rather than projected ones. The specific root-cause diagnosis of the segmentation span cap offers a directly actionable engineering finding rather than a general observation.

The contributions of this study, each limited to the evidence

reported in Section 4, are: (i) a specialized Yorùbá proverb translation system that combines structured lexical knowledge with automated translation processing; (ii) a documented, deterministic evaluation split (80/20, random seed 42) of a 568-pair corpus into a 454-record Development/Lexicon Set and a 114-record Held-Out Test Set; (iii) evaluation under both lexicon-supported and held-out conditions, rather than on lexicon-supported examples alone; (iv) a ChatGPT baseline evaluated on exactly the same 114 held-out records; (v) human-centred evaluation addressing meaning preservation, English naturalness, cultural appropriateness, and overall acceptability; and (vi) a multi-layer evaluation framework combining software verification, development-set performance, automatic reference-overlap metrics, baseline comparison, and human judgement.

1.5 Scope and Limitations

The system evaluated is restricted to text-based Yorùbá↔English idiom and proverb translation; it does not address speech input or output, dialectal variation, or grammatical tenses/aspect beyond the idiom span itself. The evaluation was conducted in two phases. In Phase 1 (Sections 4.1–4.9), the engine was evaluated against a 13-entry seed lexicon whose entries were all explicitly marked as unverified placeholder data in the project's repository documentation; the recall and robustness figures from that phase characterise the engine's matching logic and should not be read as a claim about the linguistic accuracy of the entries themselves. A structured native-speaker lexicon-rating workbook was also completed for those 13 entries (Study 1): 77 fully completed responses were retained for analysis, returning consistent output-quality agreement across all rated dimensions alongside a notable false-positive and embedded-detection divergence from the code-level tests, reported in Section 4.9. Three of the 77 respondents were non-native speakers (self-rated comprehension "Good" for two and "Limited" for one); restricting analysis to the 74 native-speaker respondents changed the grand mean Part A rating by 0.02 points (Section 4.9). This sample size exceeds the 8–15 range typically recommended for usability-style pilot studies at this lexicon scale. In Phase 2 (Sections 4.10–4.17), the lexicon was expanded to approximately 578 entries from a 568-pair Yorùbá–English proverb corpus, which was divided into a 454-record Development/Lexicon Set and a 114-record Held-Out Test Set (Section 3.10), and the system was evaluated through software verification, development-set evaluation, held-out evaluation, a ChatGPT baseline on the same held-out records, and a second human evaluation of 20 selected proverbs by three evaluators (Study 2). Important limitations remain and are set out in Section 4.17: the entries added from the corpus are unverified and carry no literal-gloss annotation; development-set figures describe lexicon-supported behaviour and not general translation accuracy; the held-out set contains incomplete source records; automatic metrics depend on a single reference translation per proverb; the false-positive control set remains small; and the human evaluation covers 20 proverbs and three evaluators. An optional large-language-model-based second disambiguation pass (Gemini 2.5 Flash) exists in the codebase but is feature-flagged off by default and was not enabled for the Phase 1 evaluation, which accordingly characterises the rule-based path only.

2. LITERATURE REVIEW

2.1 Yorùbá Proverbs: Structure, Culture, and Discourse Function

Owomoyela [4] provides the most comprehensive documented survey of Yorùbá proverbs, cataloguing their forms and social functions across contexts ranging from dispute settlement to



moral instruction. Ademowo and Balogun [3] analyse a sample of Yorùbá proverbs specifically for their role in conflict management, arguing that proverbial expressions encode values — cooperation, warning, restraint — that are recognised and invoked strategically by speakers. Ehineni [5], working from a corpus of oral interviews and native Yorùbá texts, shows that proverbs are structurally and lexically distinguishable and are deployed to achieve specific discourse acts; critically for a computational system, Ehineni's structural observations imply that a proverb is not always a fixed, single canonical string but can appear with connective and clause-level variation depending on interactional context. This has direct implications for any system, such as the one evaluated here, that relies on exact-match lookup against a canonical lexicon form: variation beyond simple orthography (word order, elision, or clause reordering) would not be captured by diacritic- and case-tolerant normalisation alone.

2.2 Multiword Expressions and Idiom Translation in NLP

Baldwin and Kim [6] define multiword expressions as lexical items spanning multiple orthographic words whose properties — syntactic, semantic, or both — are not fully predictable from those of their components and survey the principal computational strategies for handling them: listing known forms as 'words with spaces', hierarchically organized lexicons, and statistical affinity measures. Carpuat and Diab [7] provide an early task-based demonstration that explicit multiword expression handling measurably improves downstream statistical machine translation quality relative to leaving such expressions to compositional translation. More recently, Liu, Ghosh, and Jiang [2] conduct a systematic, multi-language evaluation confirming that verbal multiword expressions — including verbal idioms — consistently and measurably degrade neural machine translation quality, and that the degradation is attributable to the non-compositional expression itself rather than to general sentence-level difficulty; their proposed mitigation, LLM-based paraphrasing into a literal equivalent prior to translation, is architecturally analogous to YorùbáOwe's own literal-gloss-plus-figurative-sense output design, though YorùbáOwe generates the equivalence from a curated lexicon rather than a paraphrasing model.

2.3 Computational Treatments of Yorùbá Idioms, Proverbs, and Related Rule-Based Translation Systems

Direct computational precedent for Yorùbá idiom or proverb translation is scarce. Olusanya, Masakhane NLP, and Abolade [9] introduce OWE-YOR, a balanced dataset of 15,925 sentences (7,963 proverbs, 7,962 non-proverbs) and train a Multinomial Naive Bayes classifier alongside fine-tuned AfroLM and multilingual BERT models to detect whether a given sentence is a Yorùbá proverb; their Naive Bayes classifier achieved 85% accuracy on the classification task. This is a materially different task from the one addressed here — OWE-YOR determines proverb-hood, whereas YorùbáOwe determines and translates a specific proverb's figurative meaning — but the existence of a 15,925-sentence labelled corpus is directly relevant as a candidate future data source for expanding YorùbáOwe's own lexicon (13 seed entries at the time of the initial evaluation; approximately 578 entries in the expanded release) well beyond its current scale.

Adjacent rule-based Yorùbá machine translation work provides useful precedent on architecture and evaluation methodology even though it does not target idioms specifically. Eludiora and

Ayoade [8] describe an English-to-Yorùbá verb phrase translation system using dictionary lookup and a Context-Free Grammar validated with finite-state automata, evaluated by expert fluency scoring; related published work in the same research programme reports rule-based English↔Yorùbá noun-phrase translation accuracy in the region of 91% against a manually curated test set. These figures were achieved on grammatically regular noun- and verb-phrase constructions with dictionary-scale coverage; the lower Yorùbá-side recall (69.23%) observed in the initial seed-lexicon evaluation of the present study is, as demonstrated in Section 4, attributable not to weaker matching logic but to a fixed span-length parameter interacting with the substantially greater average length of proverbial compared to noun-phrase input.

2.4 Rule-Based Versus Neural Approaches for Low-Resource Idiom Translation

Jimoh, De Wille, and Nikolov [1], synthesizing a decade of Yorùbá NLP research, and Inuwa-Dutse [10], surveying Nigerian low-resource languages more broadly, both identify a persistent shortage of annotated datasets and dedicated tooling as the central bottleneck for Yorùbá NLP, a diagnosis directly consistent with YorùbáOwe's own small, unverified lexicon. Given this data scarcity, a rule-based, exact-match lexicon lookup strategy — as used by YorùbáOwe — has the practical advantage of requiring no training data beyond the lexicon entries themselves and of producing fully traceable, explainable output (a literal gloss and a figurative sense alongside the translation), at the cost of recall being strictly bounded by lexicon coverage and by any structural limits, such as a segmentation window, placed on the matching algorithm. YorùbáOwe's own architecture anticipates this trade-off by including a feature-flagged optional large-language-model disambiguation pass (Gemini 2.5 Flash), left disabled in the present evaluation, as a documented pathway toward hybrid rule-based/neural handling of the kind Liu, Ghosh, and Jiang [2] find effective for verbal multiword expressions in mainstream machine translation.

2.5 Web Application Development Approaches

From a software engineering standpoint, Al-Saqqa, Sawalha, and Abdel-Nabi [12] and Molina Ríos and Pedreira-Souto [13] document the prevailing use of agile, iterative methodologies for web-based software development, characterised by short development cycles, continuous testing, and incremental feature delivery. YorùbáOwe's own repository — a TypeScript codebase with an existing automated unit test suite (vitest), a feature-flag mechanism for the optional LLM pass, and an explicit curator workflow for incremental lexicon growth — is consistent with this iterative, test-driven development pattern, and its continuous-deployment target (Render) is typical of contemporary lightweight web-application hosting for prototype-stage NLP tools.

2.6 Summary and Identified Gap

The reviewed literature establishes, first, that Yorùbá proverbs are structurally and functionally well documented from a linguistic standpoint [3, 4, 5]; second, that multiword and idiomatic expressions are a recognised, empirically demonstrated source of machine translation error in the general NLP literature [2, 6, 7]; and third, that computational Yorùbá NLP resources remain data-scarce and, where proverb-specific resources exist at all, are oriented towards classification [9] rather than figurative-sense translation. No prior published system was identified that performs bidirectional Yorùbá↔English idiom/proverb translation with an explicit



literal-gloss/figurative-sense distinction and context-aware disambiguation. This study addresses that specific gap and, distinctively among the reviewed rule-based Yorùbá translation systems, evaluates the system by direct execution of its live production source code rather than by illustrative example.

3. MATERIALS AND METHODS

3.1 Research Design

This study adopted a Design Science Research (DSR) methodology, appropriate for studies whose primary output is a working software artefact evaluated against its own stated design objectives. The DSR cycle followed here comprised: (i) problem identification from the literature gap established in Section 2.6; (ii) artefact design and implementation, documented as the deployed YorùbáOwe system; and (iii) artefact evaluation, conducted not through simulated or hypothetical data but through direct importation and execution of the artefact's own production TypeScript source code, its existing automated test suite, and seven additional evaluation batteries constructed specifically for this study and executed against the live lexicon and engine logic. The evaluation was subsequently extended to a held-out test set, a general-purpose language-model baseline, and a human evaluation of selected translations (Sections 3.10–3.13).

3.2 System Requirements

Functional requirements, derived from the objectives in Section 1.3 and as implemented in the evaluated codebase, were: accept free-form Yorùbá or English text and identify idiomatic spans within it; return, for each matched span, a literal gloss, a figurative sense, and a best-fit target-language equivalent; fall back to word-level dictionary translation for spans that do not

match a known idiom; log fallback and low-confidence translations for curator review; and provide a curator interface for adding and editing lexicon entries. Non-functional requirements included diacritic- and case-insensitive matching (given that Yorùbá orthography is frequently typed without full tone marking), sub-100-millisecond translation latency suitable for interactive use, and input sanitization against malformed API requests.

3.3 System Architecture

The system implements a four-layer architecture (Figure 1). The presentation layer is a React 19/Vite/Tailwind CSS single-page application comprising a Translator view with a Yorùbá diacritic quick-palette and matched-span highlighting, a Lexicon Explorer/Editor, and a Curator dashboard for reviewing logged fallback translations. The application layer is an Express.js/TypeScript server exposing a translation endpoint and a set of lexicon-management endpoints, guarded by input-sanitization and validation middleware. The idiom translation engine — the core logic layer evaluated in Section 4 — comprises five internal modules: a segmented that tokenizes input and generates sliding-window candidate spans; a normaliser that performs diacritic- and case-folding for matching purposes only (the original text is preserved for display); a lexicon repository that performs exact-match bidirectional lookup; a disambiguator that applies rule-based (and, optionally, LLM-based) context checks to matched spans; and a fallback module that performs word-level dictionary translation and orthographic post-processing for unmatched tokens. The data layer persists the lexicon and a rolling log of the last 100 translations as JSON files.

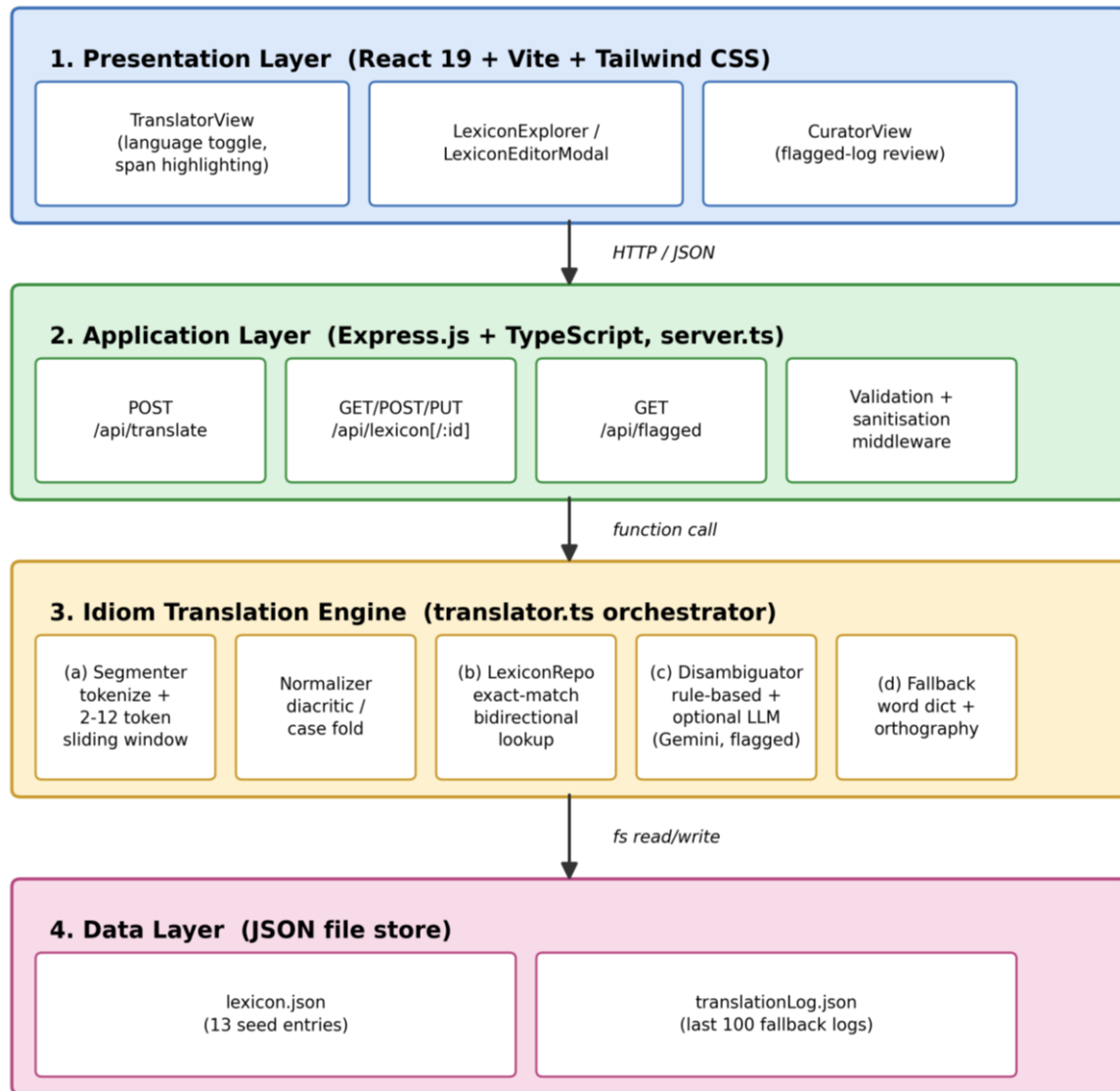


Figure 1: Four-Layer Reference Architecture of the Deployed YorùbáÒwè System

3.4 Live System Interface

Figure 2 shows the deployed system's live translator interface, captured directly from the production deployment at <https://yorubaidiom.onrender.com>. The interface presents a source-language toggle (Yorùbá òwè ↔ English), a Yorùbá

diacritic quick-palette for entering tone-marked characters without a Yorùbá keyboard layout, and a row of sample-proverb shortcuts. The Translator, Lexicon, and Curator views are accessible from a persistent top navigation bar, corresponding to the three presentation-layer components shown in Figure 1.

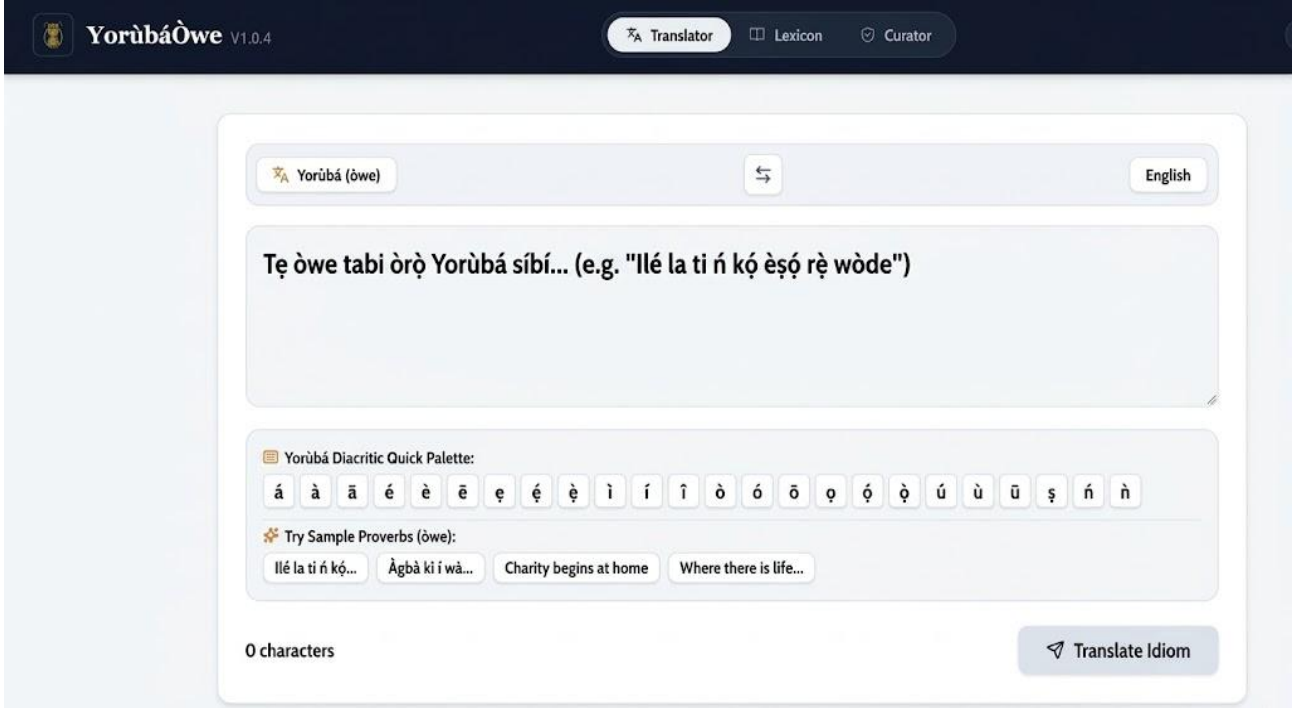


Figure 2: Live Translator Interface of the Deployed YorùbáÒwè System

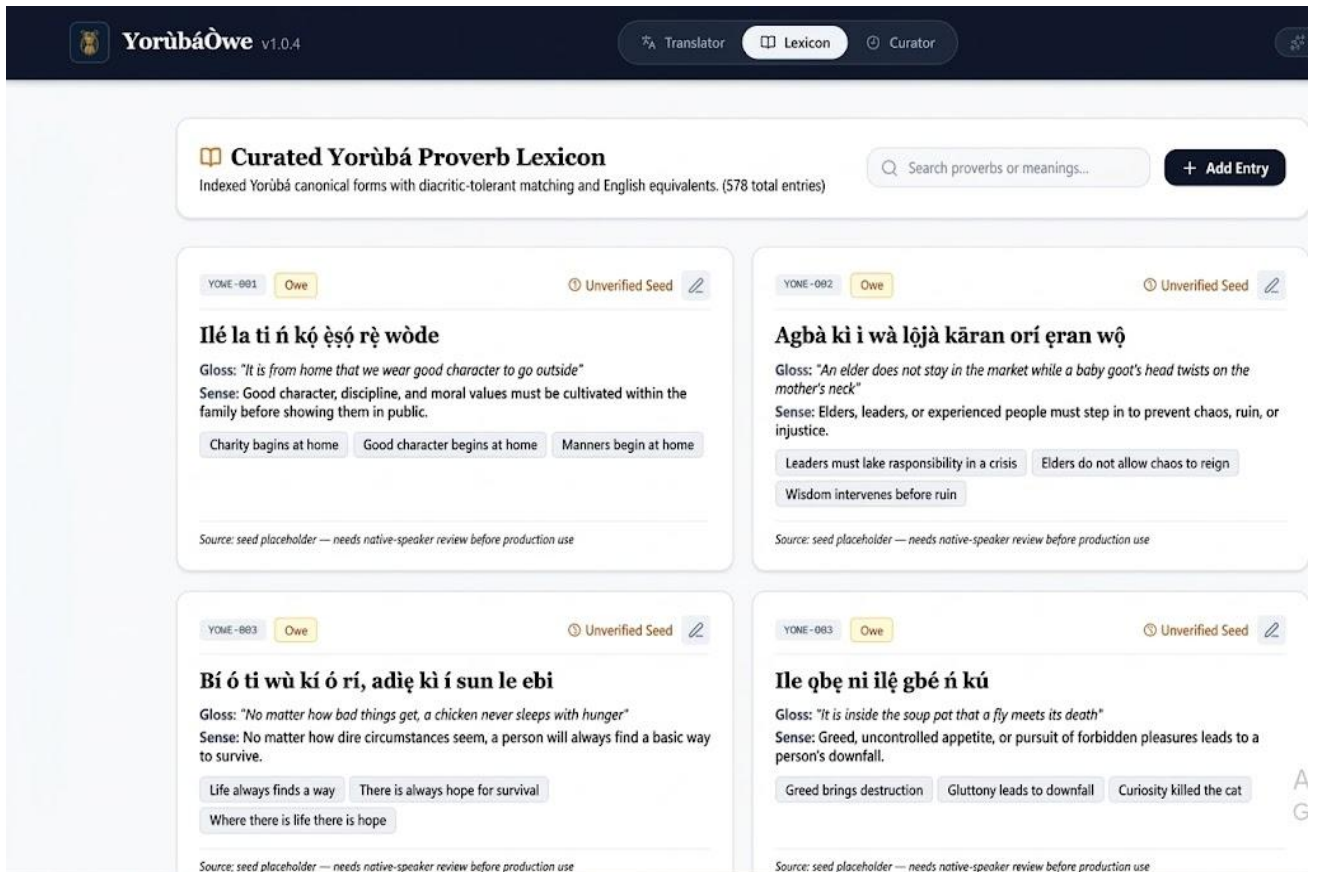


Figure 3: Live Translator Interface of the Deployed YorùbáÒwè System (lexicon)

3.5 Segmentation and Matching Algorithm

Given an input string, the engine first tokenizes it into a sequence of tokens $T = (t_1, \dots, t_n)$, preserving original character offsets for span highlighting. It then generates the set of candidate spans S

$= \{ (i, j) : 0 \leq i < n, 2 \leq j-i \leq 12, j \leq n \}$, i.e. every contiguous run of between 2 and 12 tokens starting at every position. Starting from the leftmost unconsumed token index i , the engine evaluates candidate spans starting at i in descending order of length (greedy longest-match-first), normalizing each candidate span's text —

lower-casing; substituting ẹ, ọ, ẹ with their base Latin letters; decomposing to NFD and stripping combining tone/diacritic marks; and collapsing whitespace and punctuation — and querying the lexicon repository for an exact match of the normalized string against either the normalized canonical Yorùbá form (when translating Yorùbá→English) or any normalized English equivalent (when translating English→Yorùbá). On the first match found, the matched entry and surrounding token

context are passed to the disambiguator; if classified as figurative, the span is recorded as an idiom match, the cursor advances past the matched span, and the loop continues. If no candidate span starting at i matches, or the top match is classified as literal, the single token at i is passed to the word-level fallback translator and the cursor advances by one token. Figure 3 presents this control flow as a flowchart, and it is formalized in Algorithm 1.

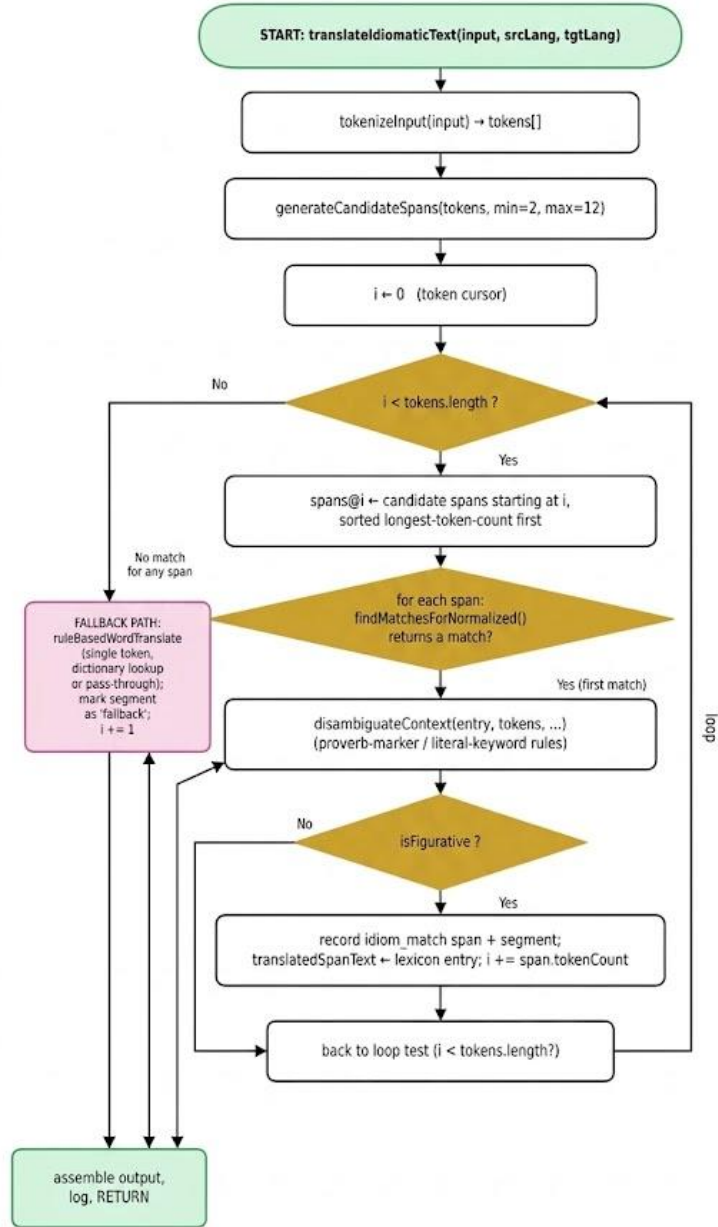


Figure 4: Flowchart of the translateIdiomaticText() Segmentation–Match–Disambiguate–Fallback Loop

Algorithm 1: translateIdiomaticText(input, sourceLang, targetLang)

1. BEGIN
2. tokens ← Tokenize(input)
3. spans ← GenerateCandidateSpans(tokens, minLen=2, maxLen=12)
4. i ← 0
5. WHILE i < |tokens| DO

6. candidates ← spans starting at i, sorted by length DESC
7. matched ← FALSE
8. FOR each span IN candidates DO
9. entry ← LexiconRepo.FindExactMatch(Normalize(span.text), sourceLang)
10. IF entry != NULL THEN
11. result ← Disambiguate(entry, tokens, span, input)
12. IF result.isFigurative THEN



```

13. RECORD idiom_match segment; i <- i + span.tokenCount;
   matched <- TRUE; BREAK
14. END IF
15. END IF
16. END FOR
17. IF NOT matched THEN
18. RECORD fallback segment for tokens[i] via word-dictionary
   lookup; i <- i + 1
19. END IF
20. END WHILE
21. output <- ApplyOrthographyRules(segments); Log(output);
   RETURN output
22. END

```

Algorithm 1 and Figure 3 describe the pipeline as evaluated in Phase 1, with a fixed maximum span of 12 tokens. Two changes were made to the expanded release evaluated in Phase 2. First, the maximum span length is now set from the longest lexicon entry (never below 12 tokens) instead of being fixed at 12. Second, when no exact idiom match is found for an input of at least three tokens, a near-match step compares the normalised input with each lexicon entry using character-trigram Sørensen–Dice similarity and accepts the best match at or above a threshold of 0.75; an accepted near-match is flagged in the output together with its similarity score. The translation pipeline therefore comprises input normalisation, segment detection, exact lexicon matching, near-match retrieval, context disambiguation, and sentence-level fallback translation.

3.6 Context Disambiguation

A matched span is not automatically treated as idiomatic. The rule-based disambiguator first checks the full input sentence for proverbial discourse markers (e.g. explicit framing such as 'owe', 'as the saying goes', 'àwọn àgbà'); if present, the span is classified figurative. It then checks, per matched lexicon entry, a hand-authored list of literal-context keywords (e.g. school/uniform vocabulary for the 'charity begins at home' entry); if present without a proverbial marker, the span is classified literal and routed to word-level fallback instead. In the absence of either signal, the default classification for a canonical idiom-form match is figurative. An optional second pass, gated by the ENABLE_LLM_DISAMBIGUATION environment flag, would additionally query Gemini 2.5 Flash to judge figurative versus literal usage from full sentence context; this flag was left at its default (disabled, no API key configured) throughout the present evaluation, so all results in Section 4 characterise the rule-based path exclusively.

3.7 Lexicon Schema and Data

Each lexicon entry is a JSON object with the following fields: a unique identifier; the canonical Yorùbá surface form with full tone marking; a literal (word-for-word) English gloss; a plain-English figurative sense; a ranked array of English idiomatic equivalents; a register tag (e.g. proverbial, formal); a usage note; an isProverb boolean; a sourceNote; and a verified boolean recording whether a native speaker has confirmed the entry. At the time of the initial (Phase 1) evaluation the production lexicon.json contained 13 entries (11 tagged isProverb = true), all with verified = false and sourceNote = 'seed placeholder — needs native-speaker review before production use', exactly as declared in the project's own repository documentation. The expanded release used in Phase 2 contains approximately 578 entries: the 13 seed entries plus 565 entries derived from the 568-pair corpus described in Section 3.10, after three malformed source rows (a footnote line, a record whose English side had been replaced by

a footnote marker, and an incomplete fragment with a placeholder translation) were excluded. Entries derived from the corpus store the corpus translation as both the figurative sense and the sole English equivalent, have an empty literalGloss field, are tagged isProverb = true, and carry verified = false; none has been reviewed by a native speaker.

3.8 Technology Stack and Implementation

The deployed system uses React 19 with Vite and Tailwind CSS for the client single-page application; Express.js and TypeScript for the server and API layer; a file-backed JSON repository pattern for lexicon and translation-log persistence; and vitest for automated unit testing. An optional dependency on Google's google/genai SDK provides the feature-flagged LLM disambiguation pass. The production build compiles the client with Vite and bundles the server with esbuild; the live deployment is hosted on Render at <https://yorubaidiom.onrender.com>.

3.9 Evaluation Method

Consistent with the direct-execution evaluation methodology used in prior work on a related numeral-translation system [11], the deployed engine was evaluated by cloning its public source repository, installing its declared dependencies, and importing and executing its production translateIdiomaticText(), JsonFileLexiconRepository, and normalize Text functions directly — the identical code paths used by the live HTTP API — rather than by re-implementing or simulating the matching logic. The existing vitest unit suite (13 tests spanning the normaliser, segmenter, lexicon repository, disambiguator, fallback translator, and full end-to-end engine) was first executed unmodified. Seven additional evaluation batteries were then constructed and executed against the live 13-entry lexicon: (i) Yorùbá→English exact-match recall on each entry's canonical form; (ii) English→Yorùbá exact-match recall on each entry's first-ranked equivalent; (iii) robustness under three orthographic perturbations (uppercasing, full diacritic stripping, and irregular whitespace) applied to every entry; (iv) false-positive control on five genuinely non-idiomatic Yorùbá sentences; (v) idiom-span detection when a canonical form is embedded mid-sentence rather than presented alone; (vi) latency, measured as wall-clock time for 200 in-process calls to the translation function; and (vii) a static lexicon-collision check for duplicate normalised Yorùbá forms and duplicate English equivalents across distinct entries. Because tests (i)–(v) and (vii) exercise the engine's core matching and disambiguation logic in-process, they are exact and repeatable; the latency figures in (vi) measure engine compute time only and exclude the network round-trip of the deployed HTTP API, a distinction made explicit in Section 4.8. A human evaluation was conducted as an eighth battery: a structured rater workbook (Google Form) covering four parts — Part A: all 13 lexicon entries, each rated on meaning accuracy, figurative meaning, English fluency, and cultural appropriateness (1-5 scales, "1 - Very poor" to "5 - Excellent"); Part B: five genuinely non-idiomatic control sentences, each judged as an ordinary sentence, an idiom/proverb, or "not sure"; Part C: six idioms embedded mid-sentence, each judged as correctly identified and translated (Yes/No/Not sure); and Part E: five overall system-usability statements (ease of understanding, meaning preservation, fluency, usefulness, and willingness to reuse), each rated 1-5 — was distributed to Yorùbá speakers as a web form together with three screening questions (native-speaker status, self-rated comprehension, and frequency of Yorùbá use). All 77 responses received were fully completed across all four parts and were retained for analysis (74/77 self-identified as native speakers; the remaining three self-rated their comprehension as



"Good" (two) or "Limited" (one)), as reported in Section 4.9. This sample exceeds Nielsen's usability-testing heuristic that five participants surface most usability issues in a study of this kind [14] and is large enough to support rating-scale means with reasonable precision rather than being read as an indicative pilot alone.

The initial evaluation described above (Phase 1) was extended by a second phase (Phase 2) against the expanded lexicon. Phase 2 comprised software verification (the unit suite, a production build, and an end-to-end evaluation script); evaluation on the Development/Lexicon Set (exact retrieval and controlled variants); evaluation on the Held-Out Test Set with automatic reference-overlap metrics; a ChatGPT baseline on the same held-out records; and a human evaluation of 20 selected proverbs. Sections 3.10–3.13 describe the dataset partition, the metrics, the baseline protocol, and the human-evaluation protocol.

3.10 Dataset, Development/Lexicon Set, and Held-Out Test Set

The primary corpus comprises 568 Yorùbá–English proverb pairs. Phase 1 of this study evaluated only a 13-entry seed lexicon and did not prescribe any partition of this corpus. The corpus was deterministically divided using an 80/20 split with random seed 42 into a Development/Lexicon Set of 454 proverb pairs and a Held-Out Test Set of 114 proverb pairs (Table 1). The 454 records are called the Development/Lexicon Set, and not a training set, because YorùbáÒwè is not presented as a model trained on these examples: they serve as the source of lexicon entries and for development-time evaluation.

Table 1: Corpus Partition and Lexicon Versions Used in the Phase 2 Evaluation (80/20 Split, Random Seed 42)

Component	Records	Role
Full corpus	568	Complete Yorùbá–English proverb corpus
Development/Lexicon Set	454 (80%)	Source of lexicon entries; development-set evaluation
Held-Out Test Set	114 (20%)	Held-out evaluation (lexicon-unseen condition)
Released lexicon	≈578 entries	13 seed entries plus 565 corpus-derived entries; contains entries for 113 of the 114 held-out records
Held-out evaluation lexicon	465 entries	Released lexicon with the 113 held-out entries removed (heldout-lexicon.json)

The held-out evaluation did not use the released lexicon. It used heldout-lexicon.json, from which the entries corresponding to the held-out records had been removed, so that the 114 proverbs were not represented in the lexicon being queried. The released lexicon, in contrast, contains entries for 113 of the 114 held-out records; the remaining held-out record (Ìwòsì ní, with reference translation "X") is an incomplete source row that has no lexicon entry. Consequently, "Held-Out Test Set" is the name of the dataset, and "lexicon-unseen" describes the relationship of that same set of 114 records to the held-out evaluation lexicon; there is no second 114-item dataset.

3.11 Automatic Evaluation Metrics

Held-out outputs were scored against the reference translations with four measures, applied with the same methodology to every system evaluated. BLEU was computed with SacreBLEU using

13a tokenization, case-sensitive scoring, and exponential smoothing; it measures n-gram overlap between a candidate and its reference. chrF was computed with a character n-gram order of 6, a word n-gram order of 0, and $\beta = 2$; it measures character-level similarity. ROUGE-L F1 measures sequence overlap based on the longest common subsequence; it was computed at sentence level on lowercased, Unicode-normalized text and averaged across 114 items. Exact textual match counts the items whose output is identical to the reference after NFC Unicode normalization, lowercasing, and whitespace normalization, and measures literal textual agreement only. All four are reference-overlap measures. A translation can be semantically correct while using different wording from the single reference, so none of them is a direct judgement of semantic or cultural translation quality, and exact textual match must not be read as translation accuracy. A human evaluation (Section 3.13) therefore complements the automatic metrics.

3.12 ChatGPT Baseline Protocol

A ChatGPT baseline translated the same 114 held-out Yorùbá proverbs, to provide a contemporary general-purpose language-model reference point. Alignment between the reference file and the baseline output was verified before scoring: 114 reference records, 114 baseline records, and 0 Yorùbá-side alignment mismatches. Baseline outputs were scored with the metric methodology of Section 3.11. The baseline is included as a descriptive reference point on the same held-out items and not to rank the systems.

3.13 Human Evaluation Protocol (Study 2)

Twenty selected Yorùbá proverbs and their YorùbáÒwè translations were rated by three evaluators, giving 60 proverb-level judgments. Each translation was rated on meaning preservation, English naturalness, and cultural appropriateness (each on a 1–5 scale) and on overall acceptability (yes/no). The evaluators also gave overall ratings of translation quality and of meaning preservation and judged practical suitability (yes/no). Rater 1 is a native Yorùbá speaker, fluent in English, with extensive experience with Yorùbá proverbs. Rater 2 is a native Yorùbá speaker, highly proficient in English, with extensive experience with Yorùbá proverbs. Rater 3 is a basic Yorùbá learner, fluent in English, with extensive experience with Yorùbá proverbs. No inter-rater agreement statistics or significance tests were computed. This evaluation (Study 2) is distinct from the 77-responder evaluation of the 13 seed entries reported in Section 4.9 (Study 1).

4. RESULTS AND DISCUSSION

Results are reported in two parts. Sections 4.1–4.9 report Phase 1: the evaluation of the engine against the 13-entry seed lexicon, together with the 77-responder human evaluation of those entries (Study 1). Sections 4.10–4.17 report Phase 2, a multi-layer evaluation of the expanded system. Sections 4.18 and 4.19 summarise the Phase 1 results and place the study among related systems.

4.1 Unit Test Verification

The project's existing vitest suite, covering the normaliser, segmenter, bidirectional lexicon repository, rule-based disambiguator, fallback translator, orthography post-processor, and full end-to-end engine, passed in full: 13 of 13 tests (100%), executed in 26 ms. One non-fatal implementation issue surfaced as console output during this run — the translation-logging function throws and catches a JSON parse error on a freshly created empty log file — which does not affect translation

correctness but is noted in Section 6 as a minor robustness fix.

4.2 Lexicon Composition

The evaluated production lexicon contained 13 entries: 11 tagged as proverbs (isProverb = true) and 2 as non-proverbial idioms. All 13 entries were marked verified = false, consistent with the project's own documented disclaimer that the seed data is a placeholder pending native-speaker linguistic review. Canonical Yorùbá surface-form length ranged from 5 to 19 tokens (mean ≈ 10.3 tokens), a range that proves directly consequential in Section 4.3.

4.3 Exact-Match Recall: Yorùbá→English

Translating each lexicon entry's own canonical Yorùbá form and

Green = matched by the engine · Red = missed (entry exceeds the 12-token candidate-span cap)

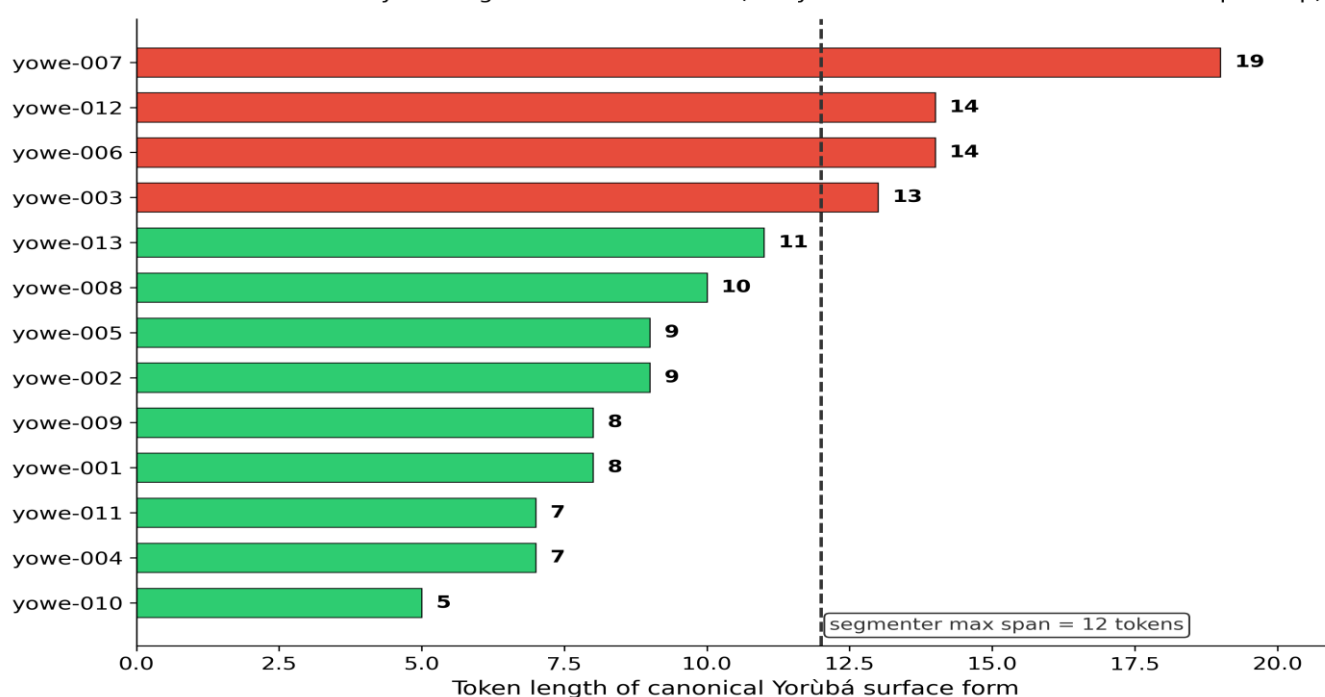


Figure 5: Lexicon Entry Length vs. Exact-Match Outcome (Yorùbá→English), Showing the 12-Token Segmentation Ceiling as the Sole Determinant of Failure

4.4 Exact-Match Recall: English→Yorùbá

Translating each entry's first-ranked English equivalent back to Yorùbá, the engine achieved 12/13 correct matches (92.31%) — markedly higher than the Yorùbá-side figure, because English equivalents are short (2–8 tokens) and none approach the 12-token ceiling. The single miss was diagnostic of a distinct, second defect rather than a repeat of the span-cap issue: entries yowe-003 and yowe-011 independently list 'Where there is life there is hope' as an English equivalent — yowe-003 as its third-ranked equivalent, yowe-011 as its first-ranked equivalent. Because findMatchesForNormalized() returns match in lexicon array order and the translator accepts the first match returned (matches [0]), a query of exactly that phrase is always resolved to yowe-003, and yowe-011 is unreachable from that input regardless of its own ranking. This is a lexicon data-integrity issue — a duplicate value across two distinct entries — rather than an engine logic defect and was confirmed independently via a static scan of all English Equivalents arrays in the lexicon.

4.5 Robustness Under Orthographic Perturbation

Each of the 13 entries' canonical Yorùbá form was tested under three perturbations — full uppercasing, full diacritic/tone-mark

checking whether the engine returned exactly the corresponding entry, the engine achieved 9/13 correct matches (69.23%). All four misses (yowe-003, yowe-006, yowe-007, yowe-012) were spans of 13, 14, 19, and 14 tokens respectively — everyone exceeding the segmenter's hard-coded maximum candidate-span length of 12 tokens (Section 3.5). Because generateCandidateSpans() never constructs a span longer than 12 tokens, these four entries are structurally unreachable by exact-match lookup regardless of lexicon correctness or normalization quality; the failure is a matching-window defect, not a data or disambiguation defect. Figure 4 plots this relationship directly: the four misses (red) and nine hits (green) separate perfectly at the 12-token line.

stripping (simulating a user typing without a Yorùbá keyboard), and irregular internal whitespace — for 39 trials in total. The engine correctly matched 27/39 (69.23%). Cross-referencing the 12 failing trials against Section 4.3 shows they are the same four span-cap-limited entries repeated across all three perturbation types ($4 \times 3 = 12$), with zero additional failures introduced by perturbation itself. In other words, once the span-cap confound is controlled for, the normaliser's diacritic-, case-, and whitespace-tolerance was 100% robust (27 of 27 within-window trials succeeded) — a materially different and more favourable conclusion than the raw 69.23% figure suggests in isolation, and an illustration of why root-cause attribution, not just aggregate pass rate, is reported throughout this section.

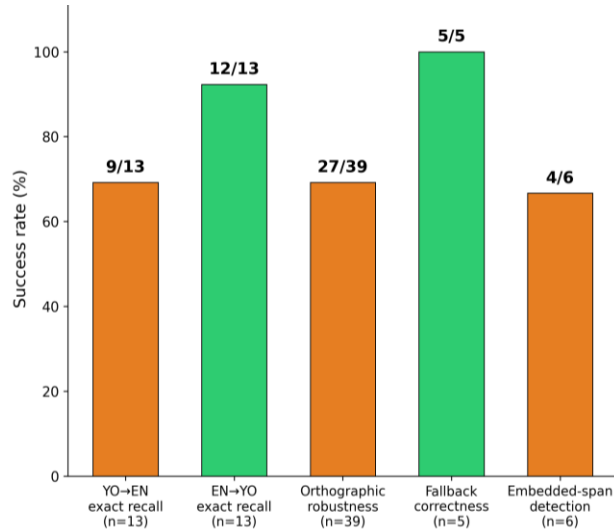
4.6 False-Positive Control on Non-Idiomatic Input

To verify that the engine does not over-trigger idiom matches on ordinary text, five genuinely non-idiomatic Yorùbá sentences (e.g. 'Mo lẹ sí ojà' — 'I am going to the market') were submitted. All five (5/5, 100%) were correctly classified with overallConfidence = 'fallback' and zero idiom spans, indicating no false positives in the tested sample.



4.7 Idiom Detection When Embedded Mid-Sentence

Six of the 13 canonical forms were embedded inside a longer carrier sentence ('Bàbá mi máa n sọ pé, ..., ó sì jẹ òtító.' — 'My father usually says, ..., and it is true.') to test whether the sliding-window segmenter can locate an idiom span that does not span the entire input. 4/6 (66.67%) were correctly detected; the two misses (yowe-003 and yowe-006) were, again, the same span-cap-limited entries identified in Section 4.3, now additionally lengthened by the surrounding carrier text, which only widens the gap between their true token length and the 12-token ceiling.



4.8 Latency

Wall-clock latency for 200 in-process calls to the production `translateIdiomaticText()` function averaged 7.90 ms, with p50 = 5.26 ms, p95 = 16.79 ms, p99 = 22.88 ms, and a maximum of 24.92 ms (Figure 5b). These figures measure engine compute time only — tokenisation, span generation, lexicon lookup, disambiguation, and file-based translation logging — executed directly against the production module; they exclude the HTTP request/response round trip and network latency of the live deployed API, which were not separately measured in this study. All timings are well within the sub-100 ms interactivity requirement stated in Section 3.2.

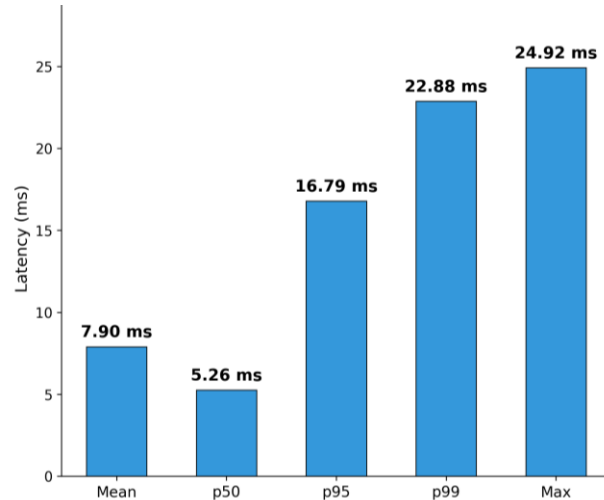


Figure 6: (a) Exact-Match Recall, Robustness, and Fallback Results; (b) Measured In-Process Engine Latency, N = 200

4.9 Human Evaluation Results (Study 1: 13 Seed Entries, 77 Respondents)

Following the completion of the code-execution evaluation (Sections 4.1–4.8), a human evaluation was carried out using a structured rater workbook distributed as a web form to Yorùbá speakers, as described in Section 3.9. All 77 responses received were fully completed across all four parts and were retained for analysis. Respondents were predominantly native speakers with frequent daily exposure to the language: 74/77 (96.1%) self-identified as native Yorùbá speakers, and 49/77 (63.6%)

rated their comprehension as "Fluent" or "Very good", a further 21 (27.3%) as "Good", and 7 (9.1%) as "Fair" or "Limited"; 60/77 (77.9%) reported using Yorùbá every day. Restricting the analysis to the 74 native-speaker respondents changed the grand mean Part A rating from 3.914 to 3.894 — a difference of 0.02 points — indicating the results are not materially sensitive to the three non-native respondents, so the full sample of 77 is reported throughout. Table 2 summarises the results; Figure 6 presents them visually.

Table 2: Human Evaluation Results (n=77 Fully Completed Responses)

Metric	n	Result
Meaning accuracy (Part A, 13 entries x 77 raters)	4004	3.86 / 5 (entry range 3.58–4.32)
Figurative-meaning accuracy (Part A)	4004	3.88 / 5 (entry range 3.61–4.19)
English fluency (Part A)	4004	3.99 / 5 (entry range 3.71–4.26)
Cultural appropriateness (Part A)	4004	3.92 / 5 (entry range 3.74–4.26)
False-positive correct-rejection (Part B, 5 items)	77	71.4–83.1% per item (mean 80.3%)
Embedded mid-sentence correct-detection (Part C, 6 items)	77	76.6–92.2% per item (mean 82.0%)
Overall system usability (Part E, 5 statements)	77	4.00–4.08 per statement (mean 4.04 / 5)

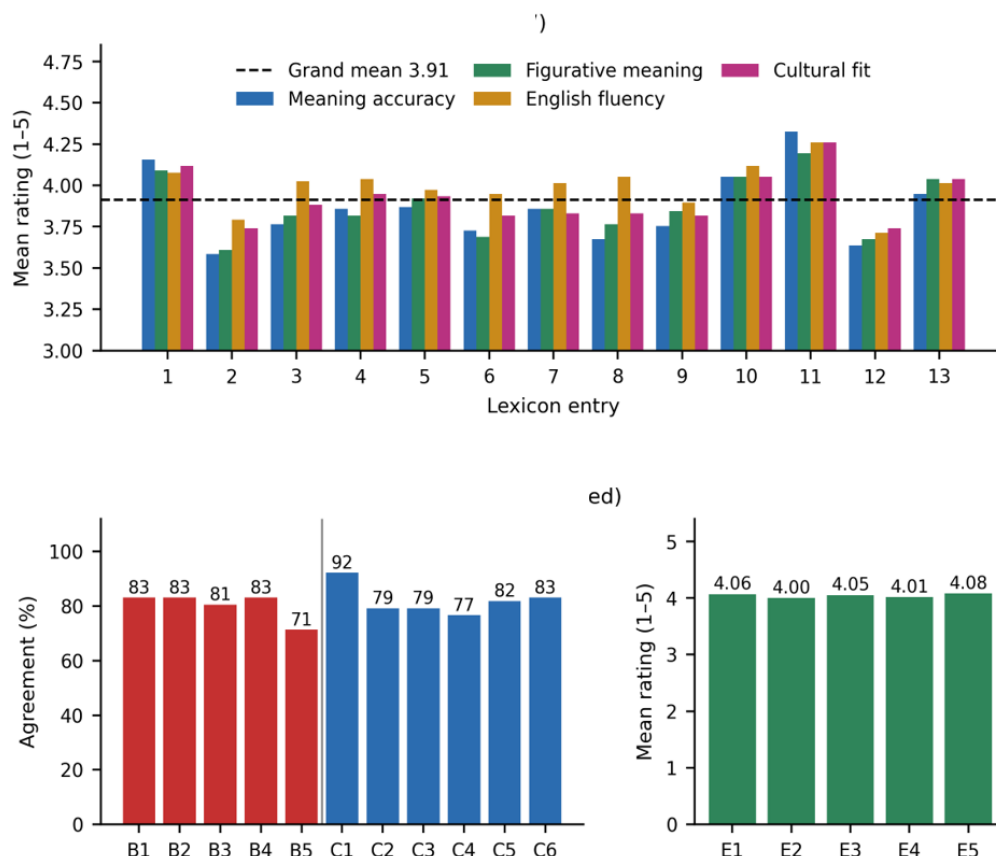


Figure 7: Expanded Human Evaluation Results (n=77): (a) Part A Lexicon-Rating Means; (b) Part B/C Per-Item Agreement (B1–B5 control sentences, C1–C6 embedded idioms); (c) Part E Overall Usability (E1 easy to understand, E2 preserves meaning, E3 sounds natural, E4 useful, E5 would use again)

Part A ratings were consistently strong across all four dimensions (3.86–3.99 out of 5) and across all 13 individual entries (no entry mean fell below 3.68), indicating that raters judged the engine's literal gloss and figurative-sense output to be linguistically sound and appropriately registered across the full lexicon, not just a favourable subset. Part E overall-usability ratings were similarly strong and consistent (4.00–4.08 across all five statements), with "would use a system like this again" averaging 4.08/5. The most notable findings, however, are two divergences between the code-level and human-level results. First, on false-positive control, the automated test (Section 4.6) reported 5/5 (100%) correct rejection of non-idiomatic input, but human item-level correct-rejection ranged from 71.4% to 83.1% (mean 80.3%) across the same five control sentences, with the weakest item (71.4%) falling well short of the automated result. Second, on embedded mid-sentence detection, the automated test (Section 4.7) reported 4/6 (66.67%) correct detection, while human item-level agreement ranged from 76.6% to 92.2% (mean 82.0%) — human raters judged the system to be detecting embedded idioms more often than the fixed automated carrier-sentence test suggested, though still with meaningful per-item variation (Example 4 lowest at 76.6%, Examples 2 and 3 at 79.2%). Taken together, these divergences confirm that the automated tests' fixed control sets do not fully characterise real-world classification behaviour as perceived by Yorùbá speakers in either direction — understating reliability on embedded detection and overstating it on false-positive control — and both are named as priority follow-ups in Section 6.

Table 3 gives the Part A ratings for each of the 13 lexicon entries, numbered in the order in which they appeared in the rating

workbook. Entry means ranged from 3.68 (Entry 2) and 3.69 (Entry 12) to 4.26 (Entry 11). The two weakest entries sit about 0.2 points below the grand mean of 3.91, which makes them the natural first candidates for native-speaker review of their literal-gloss and figurative-sense fields.

Table 3: Part A Mean Ratings by Lexicon Entry (n=77 raters per entry, 1–5 scale)

Entry	Meaning accuracy	Figurative meaning	English fluency	Cultural fit	Mean
1	4.16	4.09	4.08	4.12	4.11
2	3.58	3.61	3.79	3.74	3.68
3	3.77	3.82	4.03	3.88	3.87
4	3.86	3.82	4.04	3.95	3.92
5	3.87	3.92	3.97	3.94	3.93
6	3.73	3.69	3.95	3.82	3.80
7	3.86	3.86	4.01	3.83	3.89
8	3.68	3.77	4.05	3.83	3.83
9	3.75	3.84	3.90	3.82	3.83
10	4.05	4.05	4.12	4.05	4.07
11	4.32	4.19	4.26	4.26	4.26
12	3.64	3.68	3.71	3.74	3.69



Entry	Meaning accuracy	Figurative meaning	English fluency	Cultural fit	Mean
13	3.95	4.04	4.01	4.04	4.01
All	3.86	3.88	3.99	3.92	3.91

Native-speaker-only means for the four Part A dimensions were 3.84 (meaning accuracy), 3.86 (figurative meaning), 3.98 (English fluency), and 3.90 (cultural appropriateness). The three non-native respondents gave higher ratings (mean 4.41) than the native respondents (3.89). With only three such respondents this is descriptive, and it is a further reason to treat the native-speaker figures as the primary result.

Three checks on data quality did not change the conclusions. First, the five Part E usability statements showed high internal consistency (Cronbach's $\alpha = 0.95$). Second, the respondent-level mean Part A rating was 3.91 (95% CI 3.71–4.12, $n = 77$). Third, eight respondents (10.4%) gave the same rating to all 52 Part A items; excluding them lowers the grand mean from 3.91 to 3.85 and excluding them together with the three non-native respondents gives 3.82 ($n = 67$).

Table 4 gives the response distribution behind the Part B and Part C percentages, including the "Not sure" responses. Restricting the analysis to native speakers moves item-level agreement by less than 2 points (Part B: 73.0–83.8% against 71.4–83.1% for all respondents; Part C: 77.0–93.2% against 76.6–92.2%), so the divergences from the automated tests reported above do not depend on the non-native respondents.

Table 4: Item-Level Responses for Part B (False-Positive Control, B1–B5) and Part C (Embedded Detection, C1–C6), $n=77$

Item	Correct response, n (%)	Incorrect, n	Not sure, n
B1	64 (83.1%)	11	2
B2	64 (83.1%)	11	2
B3	62 (80.5%)	10	5
B4	64 (83.1%)	10	3
B5	55 (71.4%)	16	6
C1	71 (92.2%)	4	2
C2	61 (79.2%)	12	4
C3	61 (79.2%)	12	4
C4	59 (76.6%)	13	5
C5	63 (81.8%)	9	5
C6	64 (83.1%)	8	5

4.10 Multi-Layer Evaluation Design

The Phase 2 evaluation is organized in four complementary layers, each addressing a different question (Table 5). Software reliability asks whether the implementation behaves correctly. Development/Lexicon Set performance asks how the system behaves when its structured proverb knowledge covers the input. Held-Out Test Set evaluation asks what happens when the evaluated proverb entries are not represented in the lexicon and sets this beside a general-purpose language-model baseline on the same records. Human-centered evaluation asks how evaluators judge the meaning, naturalness, cultural appropriateness, and acceptability of selected translations, which

automatic reference-overlap metrics cannot fully capture. No single layer is read as an overall accuracy figure for the system.

Table 5: Multi-Layer Evaluation Design and Verified Results (Phase 2)

Layer	Records	Verified results
1. Software reliability	n/a	13/13 tests passed; production build successful; end-to-end evaluation completed
2. Development/Lexicon Set	454	451/454 exact (99.34%); 806/818 variants (98.53%); mean latency 31.3 ms, p95 50.6 ms
3. Held-Out Test Set (lexicon-unseen)	114	YorùbáŌwe: BLEU 0.26, chrF 6.64, ROUGE-L F1 6.41, 0/114 exact. ChatGPT baseline: BLEU 38.42, chrF 58.77, ROUGE-L F1 66.00, 4/114 exact
4. Human evaluation	20 proverbs; 3 raters; 60 judgments	Meaning 4.72/5; naturalness 4.72/5; cultural appropriateness 4.83/5; acceptability 59/60 (98.33%)

Layers 3 and 4 use different samples (the 114-record Held-Out Test Set and 20 selected proverbs, respectively), and the results of different layers are reported side by side and not combined.

4.11 Software Testing and System Verification

The unit suite of 13 tests passed in full (13/13) against the expanded release, the Vite production build completed successfully, and the end-to-end evaluation script ran to completion. These results show software-level reliability; they are not translation-quality measures. A short non-idiomatic Yorùbá sentence ("Mo ló sí ojà") was correctly classified as requiring fallback translation, but the fallback output was poor ("Know ló and market"). This is a limitation of the word-level fallback translation and not of the matching step, and it is recorded in Section 4.17.

4.12 Development/Lexicon Set Evaluation

The 454-record Development/Lexicon Set was evaluated against the released lexicon in two tests (Table 6). In the exact-match test, each record's Yorùbá text was submitted, and the system was checked for returning the corresponding lexicon entry: 451 of 454 records (99.34%) were retrieved. Three records of the Development/Lexicon Set have no byte-identical entry in the lexicon: two are incomplete source rows that were excluded when the lexicon was built (a record whose English side is a footnote marker, and a footnote line), and one is a record whose lexicon entry differs from the dataset string at the character-encoding level.

In the variant test, two controlled variants were generated for each of the 409 eligible records, one with a single word deleted and one truncated to its first 80% of words, giving 818 variant cases. The system returned the correct entry for 806 of them (98.53%), and all 806 were resolved through the near-match step, as expected for inputs that are not exact lexicon strings. Mean latency was 31.3 ms and p95 latency 50.6 ms over 150 samples; these figures are not directly comparable with the Phase 1 latency (mean 7.90 ms, 13-entry lexicon) because the lexicon size and

the measurement conditions differ.

Table 6: Development/Lexicon Set Evaluation (454 Records)

Metric	Result
Exact retrieval	451/454
Exact-match rate	99.34%
Variant-eligible records	409
Correct variants	806/818
Variant correctness	98.53%
Variants resolved by near-match	806
Mean latency (150 samples)	31.3 ms
p95 latency	50.6 ms

On the Development/Lexicon Set, YorùbáÒwè achieved 451 exact retrievals out of 454 records (99.34% exact-match rate), and 806 of 818 variants were correctly handled (98.53% variant correctness). Because the lexicon entries were derived from the same corpus and each returned translation is the stored corpus translation, these tests check retrieval and normalisation and not the translation of unseen text. They therefore describe performance within the development/lexicon-supported workflow; they are not the overall translation accuracy of YorùbáÒwè and must not be generalised to the full 568-pair corpus. The variants are controlled and synthetic and are not variants collected from speakers.

4.13 Held-Out Test Set Evaluation

The 114-record Held-Out Test Set was evaluated with heldout-lexicon.json, from which the entries corresponding to the held-out records had been removed (Section 3.10). This is the held-out, lexicon-unseen condition. All 114 records were processed successfully (114/114). Automatic reference-overlap scores were very low: BLEU 0.26, chrF 6.64, and ROUGE-L F1 6.41, with 0/114 exact textual matches (Table 7). The matching step accepted one of the 114 held-out records as a match to an existing lexicon entry (0.88%).

Table 7: YorùbáÒwè on the Held-Out Test Set (114 Records, Lexicon-Unseen Condition)

Metric	Result
Records	114
Successfully processed	114/114
BLEU	0.26
chrF	6.64
ROUGE-L F1	6.41
Exact textual match	0/114
Records accepted by the matching step	1/114 (0.88%)

These figures are not an overall accuracy of YorùbáÒwè and are not described as 0% accuracy. They show how the system behaves when its structured proverb knowledge does not contain the evaluated entries: translation of a proverb that matches the lexicon is retrieval of a stored entry, so when the entry is absent only near-match retrieval and word-level fallback remain, and their output overlaps little with the reference translations. The result demonstrates a strong dependence on lexicon coverage, which is the distinction between structured proverb retrieval and generalization to lexicon-unseen proverbs. One held-out record

(Ìwòsì ní, with reference translation "X") is an incomplete source row; it is retained in all 114-item scores.

4.14 ChatGPT Baseline on the Same Held-Out Records

A ChatGPT baseline translated the same 114 held-out records (Section 3.12). Its outputs obtained BLEU 38.42, chrF 58.77, and ROUGE-L F1 66.00, and 4 of the 114 outputs (3.51%) were textually identical to the reference after normalisation (Table 8). The 3.51% figure is the rate of exact textual agreement with the reference wording and is not the translation accuracy of ChatGPT: a translation can preserve the meaning of a proverb while using different words. For illustration (a constructed example, not an item from the evaluation set), the reference "A cow may not boast in the presence of a horse" and the translation "A cow cannot boast before a horse" differ textually and would not count as an exact match, although they may express essentially the same meaning.

Table 8: Automatic Reference-Overlap Scores on the Same 114 Held-Out Records (Descriptive Comparison)

Metric	YorùbáÒwè (held-out)	ChatGPT baseline
BLEU	0.26	38.42
chrF	6.64	58.77
ROUGE-L F1	6.41	66.00
Exact textual match	0/114 (0.00%)	4/114 (3.51%)

On the same 114 records, the ChatGPT baseline obtained substantially higher automatic reference-overlap scores than the YorùbáÒwè held-out condition. These metrics measure lexical and sequence similarity to a single reference translation and do not by themselves establish semantic or cultural translation quality, so the comparison is descriptive and the systems are not ranked. It provides a contemporary reference point for the held-out items, and the contrast is consistent with the lexicon dependence described in Section 4.13. Human evaluation (Section 4.15) was included to assess meaning preservation, English naturalness, cultural appropriateness, and overall acceptability.

4.15 Human Evaluation of Selected Translations (Study 2)

Three evaluators rated YorùbáÒwè translations of 20 selected proverbs (Section 3.13), giving 60 proverb-level judgments. Table 9 reports the aggregate results and Table 10 the item-level results. Mean ratings were 4.72/5 for meaning preservation, 4.72/5 for English naturalness, and 4.83/5 for cultural appropriateness, and 59 of the 60 judgments (98.33%) were acceptable. The evaluators' overall ratings were 5.00/5 for translation quality and 5.00/5 for meaning preservation, and all three (3/3, 100%) judged the system practically suitable. Cultural appropriateness received the highest aggregate rating, which illustrates why human-centred assessment matters for culturally embedded proverb translation.

Table 9: Aggregate Human Evaluation Results (20 Selected Proverbs, 3 Evaluators, 60 Judgments)

Dimension	Result
Meaning preservation (1–5)	4.72/5
English naturalness (1–5)	4.72/5



Dimension	Result
Cultural appropriateness (1–5)	4.83/5
Overall acceptability (yes/no)	59/60 = 98.33%
Overall translation quality (1–5)	5.00/5
Overall meaning preservation (1–5)	5.00/5
Practical suitability (yes/no)	3/3 = 100% yes

At item level (Table 10), the lowest meaning rating was for Item 18 (3.67), followed by Item 19 (4.00), and Item 19 was the only item not judged acceptable by all three evaluators (66.7%, two of three). These results apply to the 20 selected items and their 60 judgments; they do not represent the 568 corpus items. Inter-rater agreement and statistical significance were not computed.

Table 10: Item-Level Human Evaluation Results (Mean of Three Evaluators per Item)

Item	Meaning	Naturalness	Cultural appropriateness	Acceptability
1	5.00	4.67	5.00	100%
2	4.33	4.33	5.00	100%
3	5.00	5.00	5.00	100%
4	5.00	5.00	5.00	100%
5	5.00	5.00	5.00	100%
6	5.00	4.33	4.67	100%
7	5.00	5.00	5.00	100%
8	5.00	5.00	5.00	100%
9	5.00	4.67	5.00	100%
10	4.67	5.00	5.00	100%
11	5.00	4.67	5.00	100%
12	5.00	5.00	5.00	100%
13	4.67	5.00	5.00	100%
14	4.67	5.00	4.67	100%
15	4.67	4.67	5.00	100%
16	4.67	4.00	4.00	100%
17	4.67	5.00	5.00	100%
18	3.67	4.33	4.33	100%
19	4.00	4.33	4.00	66.7%
20	4.67	4.67	5.00	100%

4.16 Integrated Discussion

The four layers provide complementary evidence about different aspects of YorùbáŌwe and are not combined into a single score. The software layer shows that the implementation is reliable. The Development/Lexicon Set results show strong performance when the structured proverb knowledge in the lexicon covers the input, including controlled variants of known entries. The Held-Out Test Set results show the challenge of translating proverb entries that are not represented in the lexicon. The contrast between these two layers is the contrast between structured proverb retrieval and generalization to lexicon-unseen proverbs.

The ChatGPT baseline provides a contemporary general-purpose

reference point on the same 114 records. Its higher automatic overlap scores should be read descriptively: BLEU, chrF, ROUGE-L, and exact textual match are reference-overlap measures, which can penalize valid translations whose wording differs from the single reference, and no human ratings of the baseline's output were collected here. The comparison therefore does not establish which system produces better translations, and no ranking is claimed.

The human evaluation complements the automatic metrics by assessing dimensions that they cannot capture (meaning preservation, English naturalness, cultural appropriateness, and acceptability) on the selected proverbs, for which ratings were high. Because it covers only 20 items and three evaluators, and because ratings were collected for YorùbáŌwe translations only, it does not settle the held-out question and cannot be generalized to the full corpus. Taken together, the evidence characterizes YorùbáŌwe primarily as a lexicon-based, retrieval-oriented idiom translation system: its output for proverbs in its lexicon is grounded in curated entries, and its coverage depends on that lexicon. It is not presented as a general-purpose Yorùbá proverb translation model. The held-out and baseline evaluations therefore quantify, and do not conceal, the system's dependence on its lexicon, which a self-test on lexicon entries alone could not reveal.

4.17 Limitations of the Present Evaluation

The following limitations qualify the Phase 2 results. (i) The corpus entries are unverified and have no literal-gloss annotation. (iii) Some records are incomplete: one held-out record (Ìwòsì ní, reference "X") and some other held-out references are problematic or truncated. (iv) Development-set figures are lexicon-supported and use controlled synthetic variants, not naturally occurring variants. (v) Automatic metrics depend on the single available reference translation for each proverb. (vi) The ChatGPT baseline is a single system; the composition of its training data is not known to the authors, so prior exposure to the source proverbs or their translations cannot be excluded; and its outputs were not rated by humans. (vii) The human evaluation covers 20 proverbs and three evaluators, one of whom is a basic Yorùbá learner, and no agreement statistics were computed. (viii) The false-positive control remains small (five sentences in Phase 1 and one short sentence in Phase 2), and the fallback output for that short sentence was poor. (ix) Latency was measured on a single configuration. These limitations do not invalidate the results, but they bind what the results can be taken to show.

4.18 Summary of Phase 1 (Seed-Lexicon) Findings

Table 11 summarises the Phase 1 results against the 13-entry seed lexicon. The Phase 2 results are summarised in Table 5.

Table 11: Summary of Phase 1 (13-Entry Seed Lexicon) Evaluation Results Against the Deployed Production Engine

Test	n	Result	Root cause where applicable
Unit test suite (vittest)	13	13/13 (100%)	—
YO→EN exact-match recall	13	9/13 (69.23%)	12-token segmentation span cap

Test	n	Result	Root cause where applicable
EN→YO exact-match recall	13	12/13 (92.31%)	Duplicate English equivalent across 2 entries
Orthographic robustness (case/diacritic/whitespace)	39	27/39 (69.23%); 27/27 (100%) once span-cap-limited entries excluded	12-token segmentation span cap (not normalization)
False-positive control (non-idiomatic input)	5	5/5 (100%)	—
Embedded mid-sentence idiom detection	6	4/6 (66.67%)	12-token segmentation span cap
Lexicon internal Yorùbá-side collisions	13	0 found	—
Latency (in-process, mean / p95)	200	7.90 ms / 16.79 ms	—

4.19 Comparative Discussion

Table 12 situates YorùbáÒwè's evaluation figures against the two most directly comparable prior systems identified in the literature review: Eludiora and Ayoade's [8] rule-based English↔Yorùbá verb/noun-phrase translator, and Olusanya, Masakhane NLP, and Abolade's [9] OWE-YOR proverb classifier.

Table 12: Comparative Summary of Yorùbá Proverb/Idiom-Adjacent Computational Systems

System	Task	Data scale	Reported accuracy / recall
Eludiora & Ayoade [8]	Rule-based EN↔YO verb/noun-phrase translation	160-item curated test set	~91% (noun-phrase translation, related work)
Olusanya et al., OWE-YOR [9]	Proverb vs. non-proverb classification (not translation)	15,925-sentence labelled corpus	85% (Naive Bayes classifier)
YorùbáÒwè (this study, Phase 1)	Bidirectional idiom/proverb translation with figurative-sense output	13-entry unverified seed lexicon	69.23% YO→EN / 92.31% EN→YO (100% once span-cap isolated)

System	Task	Data scale	Reported accuracy / recall
YorùbáÒwè (this study, Phase 2)	Bidirectional idiom/proverb translation with figurative-sense output	~578-entry unverified lexicon; 114-record held-out set	Development/Lexicon Set: 99.34% exact-match. Held-out (lexicon-unseen): BLEU 0.26, chrF 6.64. Not comparable with the figures above

Table 12 makes explicit that YorùbáÒwè's task — bidirectional figurative-sense translation, not classification or phrase-level compositional translation — is not directly commensurable with either precedent, and that the reported figures measure different things: the precedents report accuracy on their own test sets, whereas the Phase 1 figures concern the matching of a 13-entry seed lexicon, the Phase 2 development figures concern lexicon-supported retrieval, and the held-out scores concern unseen proverbs. The expanded lexicon (approximately 578 entries) also remains far smaller than OWE-YOR's labelled corpus of 15,925 sentences. The comparison nonetheless supports two cautious observations. First, once the span-cap confound is isolated (Sections 4.3, 4.5, 4.7), the exact-match-and-disambiguate strategy matched every within-window entry of the seed lexicon; because the inputs were the lexicon's own entries, this is a self-test showing internal consistency, not general accuracy. Second, the raw Phase 1 recall figures (69.23% / 92.31%) would, if reported without root-cause analysis, have understated the engine's matching reliability; conversely, the held-out results show that high lexicon-supported figures must not be mistaken for coverage of unseen proverbs, so that both root-cause analysis and held-out testing are needed to characterise a lexicon-based system.

5. CONCLUSION

This study set out to design, deploy, and empirically evaluate a bidirectional Yorùbá↔English idiom and proverb translation system, addressing a gap left by the compositional focus of prior Yorùbá NLP work and the classification-only focus of prior Yorùbá proverb-specific computational work. The objectives set out in Section 1.3 were addressed: a sliding-window segmentation and exact-match bidirectional lexicon-matching strategy, extended with near-match retrieval, was formalised and implemented; a rule-based context disambiguator distinguishing figurative from literal usage was implemented and unit-tested; a four-layer full-stack web architecture with a curator lexicon-maintenance workflow was designed, implemented, and deployed live; and the system was evaluated through direct execution of its own production source code, a held-out test, a baseline comparison, and human judgement.

The evaluation yields a layered picture and not a single accuracy figure. Within its structured development/lexicon workflow, YorùbáÒwè performed strongly: all 13 unit tests passed, the production build succeeded, and on the 454-record Development/Lexicon Set it retrieved the matching entry for 451 records (99.34% exact-match rate) and correctly handled 806 of 818 controlled variants (98.53%), at a mean latency of 31.3 ms. These figures describe lexicon-supported behaviour and do not



indicate general translation accuracy. The initial seed-lexicon evaluation had also traced every Yorùbá-side miss to a fixed 12-token segmentation ceiling and identified a duplicate English equivalent shared by two lexicon entries as a data-integrity issue. The 114-record Held-Out Test Set, evaluated with the held-out entries removed from the lexicon, revealed the challenge of translating proverb entries that are not represented in the lexicon (BLEU 0.26, chrF 6.64, ROUGE-L F1 6.41, 0/114 exact textual matches). A ChatGPT baseline on the same 114 records obtained BLEU 38.42, chrF 58.77, and ROUGE-L F1 66.00, with 4/114 exact textual matches, which is 3.51% exact textual agreement with the reference wording and not translation accuracy. This comparison is descriptive: the metrics are reference-overlap measures that do not establish semantic or cultural translation quality, and no ranking of the systems is claimed.

Human evaluation addressed what these metrics cannot. In Study 1, 77 respondents rated the 13 seed entries (meaning accuracy 3.86/5; overall usability 4.04/5), and their judgements also diverged from the automated false-positive and embedded detection tests. In Study 2, three evaluators rated 20 selected proverbs at 4.72/5 for meaning preservation, 4.72/5 for English naturalness, and 4.83/5 for cultural appropriateness, with 59 of 60 judgments (98.33%) acceptable; this covers only the selected items and cannot be generalised to the 568-pair corpus. Automatic metrics and human evaluation measure different dimensions of translation quality, and the study supports combining computational evaluation with culturally informed human assessment for Yorùbá proverb translation. Overall, YorùbáOwe is best characterised as a lexicon-based, retrieval-oriented idiom translation system whose performance on the Held-Out Test Set depends on its stored proverb knowledge; it is not claimed to be superior to general-purpose language models. The expanded lexicon remains unverified and without literal-gloss annotation, and its source and licence are to be documented; native-speaker verification, a larger false-positive set, and further human evaluation are the immediate next steps (Section 6).

6. RECOMMENDATIONS

The segmenter's candidate-span ceiling, which was fixed at 12 tokens in Phase 1, has been replaced in the expanded release by a dynamic maximum derived from the longest lexicon entry (Section 3.5); this should be retained and re-validated whenever the lexicon is updated.

A uniqueness constraint (or, at minimum, a curator-facing warning) should be added to the lexicon-entry validation logic to detect when a new or edited entry's English equivalent duplicates an equivalent already present on a different entry, directly addressing the reverse-lookup collision identified in Section 4.4.

The translation-logging function's empty-file JSON parse exception (Section 4.1) should be fixed with an explicit empty-file/empty-array guard before the log is written to for the first time in any environment.

The expanded lexicon should undergo native-speaker verification, with literal-gloss and figurative-sense annotation for the entries imported from the corpus, and the source, provenance, and licence of the corpus should be documented; the 15,925-sentence OWE-YOR corpus [9] remains a candidate source of additional proverb text, though it would require annotation before it could be used for translation (rather than classification) data or lookup.

The false-positive and embedded-detection divergences surfaced by the 77-respondent human evaluation (Section 4.9) should be investigated as a priority: the false-positive control set should be

expanded beyond five sentences and examined for whether the weakest item (71.4% correct-rejection) shares a linguistic pattern the automated test's fixed set did not capture, and the embedded-detection carrier sentences should be reviewed against the lowest-agreement item (76.6% agreement), where human and automated judgements diverged most. A native-speaker linguistic audit of the specific control sentences that produced the widest human/automated gaps would help determine whether the automated test's ground-truth labels, rather than the engine's classification, are the source of the divergence in some cases. The false-positive control set remains small and should be expanded with native-speaker-verified non-proverb sentences, and the fallback output for short non-idiomatic sentences (Section 4.11) should be improved.

The feature-flagged LLM disambiguation pass should be formally evaluated against the rule-based-only path once enabled, given that Liu, Ghosh, and Jiang [2] find LLM-based handling of verbal multiword expressions measurably effective in the general machine translation literature.

The held-out evaluation and ChatGPT baseline (Sections 4.13–4.14) should be extended with reference figurative meanings verified by at least two native speakers, blinded human ratings of the outputs of both systems on the same held-out items, additional baseline systems, and inter-rater agreement, so that automatic reference-overlap scores can be complemented by comparable human judgements of both systems.

The near-match step introduced in the expanded release should be evaluated on naturally occurring variants of proverbs, including the clause-level and connective variation documented by Ehineni [5], and not only on the controlled variants of Section 4.12. Hybrid designs that combine curated lexicon retrieval with generative translation for lexicon-unseen proverbs, evaluated with human judgement, are a further direction suggested by the held-out results.

7. ACKNOWLEDGEMENTS

The authors wish to acknowledge the Institutional Based Research (IBR) Funding of this work via the Tertiary Education Trust Fund (TETFUND).

8. REFERENCES

- [1] Jimoh, T. A., De Wille, T., & Nikolov, N. S. (2025). Bridging gaps in natural language processing for Yorùbá: A systematic review of a decade of progress and prospects. arXiv:2502.17364.
- [2] Liu, L., Ghosh, S., & Jiang, T. (2025). Evaluating the impact of verbal multiword expressions on machine translation. arXiv:2508.17458.
- [3] Ademowo, A. J., & Balogun, N. (2014). Proverbs and conflict management in Africa: A study of selected Yoruba proverbs and proverbial expressions. *International Journal of Literature, Language and Linguistics*, 1(1), 39–44.
- [4] Owomoyela, O. (2005). *Yorùbá proverbs*. University of Nebraska Press.
- [5] Ehineni, T. O. (2016). A discourse-structural analysis of Yorùbá proverbs in interaction. *Colombian Applied Linguistics Journal*, 18(1), 71–83.
- [6] Baldwin, T., & Kim, S. N. (2010). Multiword expressions. In N. Indurkha & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 267–292). CRC Press, Taylor and Francis Group.
- [7] Carpuat, M., & Diab, M. (2010). Task-based evaluation of



multiword expressions: A pilot study in statistical machine translation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 242–245). Association for Computational Linguistics.

- [8] Eludiora, S., & Ayoade, B. A. (2021). Design and implementation of English to Yorùbá verb phrase machine translation system. arXiv:2104.04125.
- [9] Olusanya, J., Masakhane NLP, & Abolade, D. (2025). OWE-YOR: Leveraging transformer-based models for Yorùbá proverb classification. Conference paper, OpenReview.
- [10] Inuwa-Dutse, I. (2025). NaijaNLP: A survey of Nigerian low-resource languages. arXiv:2502.19784.
- [11] Odeniyi, O. A., Akinwumi, A. R., & Ogunsuyi, O. O. (2026). Development of a web-based number-to-Yorùbá name translation system: A rule-based computational linguistics approach to numeral text normalisation for a low-resource Nigerian language. *International Journal of Applied Information Systems (IJ AIS)*, 13(4). Foundation of Computer Science (FCS), New York, USA. www.ijais.org
- [12] Al-Saqqa, S., Sawalha, S., & Abdel-Nabi, H. (2020). Agile software development: Methodologies and trends. *International Journal of Interactive Mobile Technologies*, 14(11), 246–270.
- [13] Molina Ríos, J., & Pedreira-Souto, N. (2019). Approach of agile methodologies in the development of web-based software. *Information*, 10(10), Article 305.
- [14] Nielsen, J. (2000). Why you only need to test with 5 users. Nielsen Norman Group. www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/