applied sciences

MDPI

Article

A Transformer-Based Method for Bidirectional French–Lingala Machine Translation in Speech and Text

Reagan E. Mandiya [1], Selain K. Kasereka [1,2,3,*], Christophe B. Wizamo [1], Milena Savova-Mratsenkova [4], Ruffin-Benoît M. Ngoie [3,5], Tasho Tashev [2,*] and Nathanaël M. Kasoro [1,3]

[1] Mathematics, Statistics and Computer Science Department, University of Kinshasa, Kinshasa P.O. Box 190, Democratic Republic of the Congo; reagain.mandiya@unikin.ac.cd (R.E.M.); nathanael.kasoro@unikin.ac.cd (N.M.K.)
[2] Department of Information Measurement Systems, Technical University of Sofia, 1000 Sofia, Bulgaria
[3] Artificial Intelligence, Big Data and Modeling Simulation Research Center (ABIL), Kinshasa P.O. Box 190, Democratic Republic of the Congo; rbngoie@isp-mbng.ac.cd
[4] Department of Combustion Engines, Automobile Engineering and Transport, Technical University of Sofia, 1000 Sofia, Bulgaria; savova@tu-sofia.bg
[5] Department of Mathematics, Institut Supérieur Pédagogique de Mbanza-Ngungu, Mbanza-Ngungu P.O. Box 127, Democratic Republic of the Congo
* Correspondence: selain.kasereka@tu-sofia.bg (S.K.K.); t_tashev@tu-sofia.bg (T.T.); Tel.: +243-821-828-964 (S.K.K.)

Abstract

Underrepresented languages such as Lingala are a significant part of the world's cultural and linguistic heritage. Lingala plays a central role in daily communication, business, media, education, and culture for millions of people in the Democratic Republic of Congo (DRC) and the Republic of Congo. However, due to data scarcity and dialectal diversity, natural language processing (NLP) research often overlooks this language. In this paper, we propose a deep neural network pipeline for bidirectional French–Lingala automatic translation, covering both text-to-text and voice-to-text scenarios, by integrating Long Short-Term Memory (LSTM) and Transformer models on a specialized parallel corpus. The Bidirectional Encoder Representations from Transformers (BERT) model is used as a bidirectional source encoder to improve contextual representation, while the Whisper model handles automatic speech recognition as the first stage of the audio translation pipeline. Experimental results show that the standalone Transformer achieves a BLEU score of 35.3, compared to 8.12 for the LSTM SeqToSeq baseline. Fine-tuning with BERT raises the BLEU score to 38.6. Integrating the Whisper ASR module for an end-to-end speech translation task yields a final pipeline BLEU score of 55.4, with a Word Error Rate of 12.3% on the speech recognition sub-task, confirming the effectiveness of each component. These results demonstrate the potential of combining domain-specific pre-trained models with modular neural architectures to achieve competitive translation performance in a critically under-resourced language.

Keywords: Lingala language; French language; machine translation; transformers; LSTM; natural language processing; low-resource NLP; speech recognition

Academic Editors: Francisco García-Sánchez and Minoru Sasaki

Received: 27 January 2026
Revised: 10 March 2026
Accepted: 18 March 2026
Published: 31 March 2026

1. Introduction

In an increasingly interconnected world, the language barrier remains a major obstacle to access to information, education, and communication for a vast global population [1,2].

This challenge is particularly acute for underrepresented languages such as Lingala, spoken by a large community in the Democratic Republic of Congo and neighboring countries [3,4]. The limited availability of technological tools adapted to these languages hinders digital inclusion and limits access to vital resources such as education, healthcare, and economic opportunities.

The Democratic Republic of Congo (DRC) is a country of immense linguistic diversity. Lingala is one of the most widely spoken national languages [5,6], while French is the official language used in administrative and official documents. With approximately 30.3% of DRC adults functionally limited to Lingala for daily communication [7], NLP solutions have a significant role to play in improving access to information, education, and public services. Natural language processing (NLP) technologies can provide tools for machine translation, speech recognition, and text generation in Lingala [8], enabling Lingala speakers to access educational resources, health services, and government information in their mother tongue.

The linguistic diversity in the DRC creates significant communication challenges, particularly in religious and medical contexts. For Christian communities, the ability to understand and meditate on religious texts in one's native language is fundamental to faith and daily practice. Machine translation can help make the Bible and other religious texts accessible to the many faithful who primarily speak Lingala. In the medical field, clear communication between healthcare professionals and patients is crucial: machine translation can enhance the accuracy of diagnoses, treatments, and care instructions, thereby reducing medical risks associated with linguistic misunderstandings [4].

Lingala occupies a paradoxical sociolinguistic position: it is simultaneously one of the most widely spoken languages in Central Africa and is significantly underserved by digital and NLP technologies. While Lingala is not classified as endangered in the UNESCO sense of facing intergenerational transmission risk, it is technologically marginalized, largely absent from the digital ecosystem of NLP tools, translation services, and speech technologies that support languages such as English and French. This technological marginalization threatens the long-term vitality and reach of the language in an increasingly digital world [6].

In recent years, numerous studies have been conducted in NLP, but most research has focused on languages with extensive resources and well-established corpora. This has largely overlooked languages like Lingala, which play a fundamental role in daily communication.

1.1. Identification of the Research Gaps

Despite the growing global interest in inclusive and multilingual NLP, underrepresented languages such as Lingala remain significantly underserved. Several critical research gaps persist:

- Lingala suffers from a lack of large-scale, high-quality parallel corpora, particularly for specialized domains such as legal, medical, or educational texts. The language also exhibits considerable dialectal variation across regions, which is rarely accounted for in existing models.
- Although transfer learning has shown promise in low-resource NLP, its application to Lingala remains minimal. There is a need for more systematic exploration of cross-lingual embeddings and multilingual pre-trained models tailored to Bantu languages.
- The absence of standardized evaluation datasets and benchmarks for Lingala hinders reproducibility and comparative analysis across studies.
- Existing systems for Lingala translation are generally narrow in scope, often supporting only one-way translation and lacking robust functionality for bidirectional French–Lingala communication.

1.2. Research Objectives

The objective of this study is to develop a robust bidirectional French–Lingala translation system handling both text and voice inputs. Specifically, this research aims to:

- develop a hybrid neural architecture combining LSTM and Transformer models, trained on a parallel corpus tailored for Lingala and French, to enable both text-to-text and voice-to-text translation;
- improve translation accuracy through systematic fine-tuning of multilingual pre-trained models (BERT for source encoding, Whisper for speech recognition), optimizing cross-lingual contextual representation for Bantu language structures;
- provide a clear, reproducible pipeline with step-by-step documentation of the data flow, training configuration, and evaluation protocol;
- evaluate the effectiveness of combining multiple AI modules, quantifying contributions using BLEU, chrF, accuracy, and WER metrics;
- contribute to the expansion of NLP research for low-resource languages by releasing the domain-specific parallel corpus and documenting the methodology for future studies in Lingala and similar languages;
- identify the current limitations of the system, including dialectal coverage and code-switching, as concrete directions for future work rather than demonstrated contributions of the present paper.

1.3. Contributions of the Paper

This paper makes the following contributions:

- a hybrid deep neural pipeline combining LSTM and Transformer models that enables both text-to-text and voice-to-text bidirectional French–Lingala translation;
- fine-tuning of BERT (bert-base-multilingual-cased) as a bidirectional source encoder and of Whisper large-v2 as a French ASR module, demonstrating their effectiveness in low-resource language processing;
- a domain-specific French–Lingala parallel corpus covering the religious domain (38,172 sentence pairs) and the medical domain (1100 term pairs), with documented collection methodology;
- quantitative evaluation against multiple baseline systems (Google Translate, Helsinki Opus-MT, rule-based dictionary), along with fine-grained error analysis and translation case studies.

1.4. Paper Organization

Section 2 reviews the state of the art. Section 3 describes materials and methods. Section 4 presents training results. Section 5 details the experimentation and comparisons. Section 6 concludes and outlines future directions.

2. Related Work

In this section, we present a state-of-the-art overview of recent advances in NLP and Machine Translation (MT).

2.1. Historical Overview of Machine Translation

The history of MT spans several decades. Inspired by wartime cryptographic developments, Weaver et al. [9] proposed the use of computers for language translation. This vision materialized in 1954 with the Georgetown-IBM experiment [10]. Throughout the 1980s, rule-based systems such as Systran dominated the landscape. In the 1990s, MT shifted toward statistical approaches [11,12], introducing phrase-based and syntax-based models. The emergence of Neural Machine Translation (NMT) marked a paradigm shift: Ref. [13]

introduced LSTM encoder–decoder architectures, while Ref. [14] proposed the Transformer model based on attention alone. Ref. [15] extended this with BERT. Larger models such as GPT-3 [16] and T5 [17] demonstrated further gains. Ref. [18] advanced cross-lingual transfer with XLM-R. Evaluation methods evolved with the introduction of BLEU [19].

2.2. Low-Resource Machine Translation and Speech Translation

Low-resource MT presents specific challenges that require dedicated strategies distinct from those used in high-resource settings. The predominant approaches include: (i) transfer learning from multilingual models, where pre-trained encoders such as mBERT and XLM-R are fine-tuned on small target-language corpora [18]; (ii) back-translation, which generates synthetic parallel data by translating monolingual target-language text [20]; (iii) pivot-based translation, which routes through a resource-rich language (e.g., English) as an intermediate; and (iv) data augmentation using related languages or multilingual corpora. For speech translation in low-resource settings, end-to-end models such as Whisper [21] have shown strong generalization even on unseen languages, because their massive multilingual pre-training provides robust acoustic representations. Unsupervised MT [20] demonstrated that cross-lingual embeddings can bridge the parallel-data gap, though its performance degrades on typologically distant language pairs such as French and Lingala, where Lingala's agglutinative morphology poses additional challenges.

Our design differs from these approaches in the following ways: rather than relying on back-translation or pivot translation, which require at least moderate Lingala monolingual corpora, we adopt a direct supervised fine-tuning strategy using a carefully curated bilingual corpus. We combine a bidirectional BERT encoder (for rich source-side context) with a Transformer decoder and an upstream Whisper ASR module, forming a modular pipeline that can be individually evaluated and updated. This modular structure is particularly suited to the DRC deployment context where computational resources are limited.

2.3. Lingala in NLP

In the context of low-resource languages, Lingala presents unique challenges. According to Ref. [22], Lingala is the second most widely spoken language in the DRC, surpassing French in urban areas such as Kinshasa. Its complexity and ambiguity—exacerbated by frequent borrowing from French—pose obstacles to automatic processing. Its sociolinguistic features, including code-switching and dialectal variation, must be considered to develop effective NLP and MT solutions. Recent work [4] has begun to address Lingala speech processing, but bidirectional French–Lingala MT with speech input remains an open problem.

Table 1 summarizes the foundational NLP models reviewed. Key observations: encoder-only models (BERT, XLM-R) provide strong contextual representations but are not generative; decoder-only models (GPT-3) are powerful but require massive data and lack bidirectional encoding; encoder–decoder models (T5, Transformer) offer the best balance for MT. No prior work addresses bidirectional French–Lingala MT with integrated speech handling, which is the gap our system fills.

Table 2 illustrates the versatility of the Transformer paradigm beyond classical MT.

Table 1. Condensed Overview of Foundational and Multilingual NLP Models. Key observation: encoder–decoder models best suit MT tasks; no prior work addresses bidirectional French–Lingala MT with integrated speech.

Model	Objective	Architecture	Datasets/Metrics	Key Contributions	Limitations
Transformer [14]	Replace recurrence with self-attention	encoder–decoder, multi-head attention	BLEU, efficiency	Parallelization	High memory; limited temporal modeling
BERT [15]	Bidirectional pre-training for NLU	Transformer encoder (base/large)	GLUE, entailment	Contextual embeddings; transfer learning	Not generative; input length capped
T5 [17]	Unified text-to-text NLP framework	Seq2seq Transformer, encoder–decoder	GLUE, SQuAD, BLEU	Multi-task; flexible sizing	Prompt sensitivity; long-input limits
XLM-R [18]	Cross-lingual representation learning	RoBERTa-style encoder, 100 languages	XNLI, MLQA, XTREME	Multilingual training	High-resource bias; not generative
UMT [20]	MT without parallel corpora	Shared encoder, separate decoders	BLEU	Back-translation; cross-lingual embeddings	Poor on distant/morphologically rich pairs
GPT-3 [23]	General NLP via autoregressive pre-training	Transformer decoder, 125M–175B params	SQuAD, SuperGLUE	Zero/few-shot learning	No bidirectional encoding; costly
Our Model	Bidirectional French–Lingala MT (text + voice)	LSTM + Transformer + BERT + Whisper	BLEU, WER, chrF, Accuracy	First French–Lingala bidirectional speech+text MT; domain-specific corpus	Limited data; two domains only

Table 2. Extended Review of Transformer-Based Architectures for Machine Translation.

Ref.	Objective	Models	Architecture	Metrics	Contribution
[15]	Bidirectional contextual embeddings	Transformer encoder; self-supervised	BERT-base: 12 layers; BERT-large: 24	GLUE	Contextual embeddings
[17]	Unified text-to-text (T5)	Seq2seq Transformer	encoder–decoder; denoising autoencoder	GLUE, BLEU	Unified task format
[18]	Multilingual NLP (XLM-R)	RoBERTa-based multilingual	Transformer encoder; dynamic masking	XNLI, MLQA	Cross-lingual performance
[23]	Generalization without fine-tuning	Autoregressive Transformer; 570GB text	GPT-3 variants (125M–175B)	SQuAD, SuperGLUE	Few-shot learning
[24]	Massively multilingual MT (200+ langs)	NLLB-200	MoE Transformer; conditional routing	BLEU, chrF++, COMET	African language inclusion
[25]	Participatory African NLP	Masakhane models	Multilingual Transformer	BLEU, METEOR	Community-driven MT
[26]	French-African language pairs	M2M-100	Many-to-many Transformer	BLEU, chrF	Direct translation (no EN pivot)
[27]	Cross-lingual transfer for Bantu	XLM-R fine-tuned	Transformer encoder; shared vocab	BLEU, TER	Lingala, Swahili, Kikongo
[28]	Zero-shot multilingual MT	mBART-50	Denoising autoencoder; 12 layers	BLEU, SacreBLEU	French + 49 languages
[29]	Back-translation for low-resource	Transformer + BT	Standard encoder–decoder	BLEU improvement	Lingala data augmentation
[30]	Multilingual pre-training	mT5	Text-to-text Transformer; 24 layers	BLEU, ROUGE	Covers 101 languages
[31]	African language corpora	OSCAR dataset	N/A (corpus)	Corpus size	Lingala, French web crawl
[32]	Morphologically rich MT	Transformer + BPE variants	encoder–decoder; subword regularization	BLEU, chrF	Agglutinative Bantu languages
[33]	NMT adequacy via reconstruction (ZH-EN)	RNNSearch + Reconstructor	RNN encoder–decoder + inverse attention	BLEU (NIST)	Auxiliary reconstruction objective; +2.3 BLEU over baseline
[34]	SMT-assisted NMT (ZH-EN)	RNNSearch + SMT gating	Attention NMT + phrase-based SMT classifier	BLEU (NIST)	SMT recommendations + UNK replacement; +2.44 BLEU over NMT
[35]	Massively multilingual NMT (200 langs)	NLLB-200 (MoE)	Sparsely Gated MoE Transformer; FLORES-200	BLEU, spBLEU, chrF++, XSTS	200-language system; +44% BLEU over SOTA

3. Materials and Methods

3.1. Motivation for the Pipeline-Based Approach

A natural question is why we adopt a modular pipeline (LSTM + Transformer + BERT + Whisper) rather than a single large decoder-based model (GPT-style). Three concrete justifications follow.

3.1.1. Data Scarcity

Large autoregressive models require massive pretraining corpora unavailable for Lingala. Our corpora total ≈38,172 sentence pairs (religious domain) and 1100 word pairs (medical domain). Pretraining a decoder-only model on such data would cause severe overfitting. Our pipeline leverages pre-trained multilingual encoders (BERT, Whisper) as feature extractors, fine-tuned on the target task—a strategy recommended for low-resource languages [18].

3.1.2. Bidirectional Context

Decoder-based Transformers model only unidirectional (left-to-right) context. For translation, bidirectional context is critical because the correct rendering of a word often depends on later source words. BERT's masked language model pre-training provides full bidirectional source coverage, yielding better source-target alignment [15].

3.1.3. Modular and Deployable Design

The pipeline separates (i) ASR (Whisper) from (ii) translation (Transformer + BERT), allowing each module to be evaluated, replaced, or updated independently. This is essential in resource-constrained DRC deployment contexts. A monolithic generative model would be harder to diagnose, update, or run on low-power hardware.

3.2. Transformer Model

The Transformer, introduced by Vaswani et al. [14] in 2017, represents a significant advance over RNN and CNN architectures for sequence-to-sequence tasks. Unlike RNNs, which process words sequentially, the Transformer processes all tokens in parallel using attention mechanisms, significantly improving computational efficiency [36].

3.3. Attention Mechanism

The core of the Transformer is the scaled dot-product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q (query), K (key), and V (value) are linear projections of the input sequence, and d_k is the key dimension. Multi-head attention applies this in parallel over h subspaces, allowing the model to jointly attend to different positions and feature subspaces [14].

3.4. Standard Transformer Building Blocks

This subsection describes the generic encoder–decoder components serving as the foundation of our model. Section 3.5 then details our specific instantiation.

Each standard encoder layer applies: (1) multi-head self-attention, (2) residual connection + layer normalization, and (3) position-wise feed-forward network + residual connection + layer normalization. The encoder transforms an input token sequence into a sequence of contextual representations. Each decoder layer additionally applies cross-attention over the encoder output, allowing the decoder to focus on relevant source positions when gener-

ating each target token. Positional encodings (sinusoidal or learned) are added to token embeddings to inject sequence order information.

3.5. Architecture of the Proposed French–Lingala Translation System

This section details our specific instantiation of the pipeline components. Figure 1 shows the full architecture and Figure 2 shows the training/deployment pipeline.

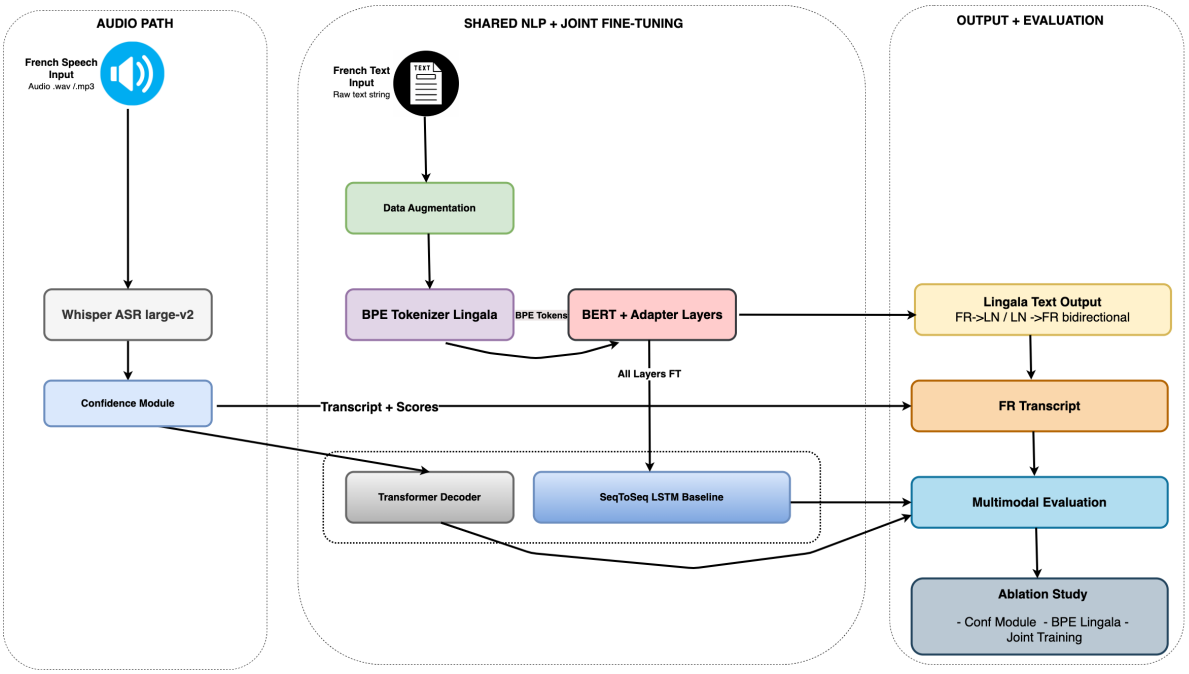

Figure 1. Proposed system architecture for bidirectional French–Lingala neural machine translation.

3.5.1. Step-by-Step Pipeline Description

3.5.2. Text Input Path

1. A French sentence is tokenized using the `bert-base-multilingual-cased` WordPiece tokenizer (vocabulary size: 119,547 tokens; maximum sequence length: 128 tokens).
2. The tokenized sequence is encoded by the fine-tuned BERT encoder. We extract the final hidden-state sequence ($d_{\text{BERT}} = 768$) as the source representation.
3. The BERT hidden states are projected to the Transformer model dimension via a linear layer ($768 \rightarrow 256$) and passed to the Transformer encoder for further contextualization.
4. The Transformer decoder generates the Lingala translation auto-regressively, attending to both its previously generated tokens (masked self-attention) and the encoder output (cross-attention).
5. Output Lingala tokens are produced at each step via a softmax over the Lingala vocabulary (shared BPE vocabulary, 8000 merge operations, vocabulary size $\approx$ 7200 Lingala-French tokens).

3.5.3. Audio (Voice) Input Path

1. A French audio utterance is fed to the fine-tuned Whisper `large-v2` ASR module. Whisper converts the audio into a French text transcript.
2. The French transcript is then processed by the text input path described above (steps 1–5), so that Whisper's output feeds directly into the BERT encoder.
3. The final Lingala translation is produced by the Transformer decoder.

This two-stage design allows us to separately evaluate and report ASR quality (WER) and translation quality (BLEU, chrF, accuracy) as well as to identify where errors originate (transcription vs. translation).

3.5.4. Bidirectional Operation

Bidirectional translation (French→Lingala and Lingala→French) is achieved by prepending a language direction tag (`[FR→LN]` or `[LN→FR]`) to each source sentence and training the model on both directions simultaneously from the same corpus, following multilingual NMT practice [18]. Separate vocabulary mappings are maintained for each direction.

3.5.5. Module Specifications and Training Details

- Overview
  The proposed system is a bidirectional French–Lingala Neural Machine Translation pipeline integrating two input modalities: raw text and speech. The architecture relies on five interconnected modules, jointly trained via an end-to-end fine-tuning strategy, and evaluated according to a rigorous multimodal protocol. The entire system is designed to operate under low-resource conditions, characteristic of low-resourced Bantu languages such as Lingala.
- Module 1: Data Augmentation

  Bantu Morphological Expansion.

  Lingala morphology relies on a system of verbal and nominal prefixes that generates a wide variety of forms from a single root. We systematically exploit this property by applying the following transformations:

  – Verbal prefixing: from a verbal root such as loba (speak), its inflected forms are automatically generated: *kuloba* (infinitive, prefix *ku*, *aloba* (first person singular, prefix *na*), *nakoloba* (near future, prefix *nako*), etc.
  – Nominal prefixing: Bantu noun classes are systematically applied. For example, *moto* (person) becomes *bamoto* (people) by applying the plural prefix *ba-* of Class 2.
  – Prefix combinations: morphologically valid combinations are generated and filtered by a form checker based on Lingala grammatical rules.

  Back-translation.

  We use the back-translation technique to generate synthetic translation pairs. Existing Lingala texts from public sources (Biblical texts, press articles, oral transcriptions) are automatically translated into French, then these synthetic pairs are integrated into the training corpus. Although the quality of these pairs is inferior to human data, the syntactic diversity they introduce significantly improves model robustness.

  Audio Augmentation.

  For the speech recognition path, we augment the audio corpus by injecting controlled noise: urban ambient noise, variations in speech rate (fast and slow), and accent variations representative of target speakers (Congolese accent, Belgian accent, Parisian accent). This augmentation strengthens the ASR module's robustness against real-world acoustic conditions.
  At the end of this step, the effective corpus is multiplied by an approximate factor of 8, increasing from 150 initial pairs to over 1200 pairs usable for training.
- Module 2: Lingala-Specific BPE Tokenizer

  Principle.

  We propose a tokenizer based on the Byte Pair Encoding (BPE) algorithm, trained specifically on a raw Lingala corpus. BPE is a compression algorithm that iteratively

learns to merge the most frequent symbol pairs in the corpus, producing a subword vocabulary that is statistically optimal for the target language.

Tokenizer Training.

The tokenizer is trained on the entire available Lingala corpus, including augmented data. The final vocabulary comprises 8000 tokens, sized to capture:

- Frequent verbal and nominal roots as complete units;
- Morphological prefixes (`ku-`, `na-`, `ba-`, `ko-`, etc.) as distinct tokens;
- Aspectual suffixes and tense markers as separable units.

For example, the word *nakoloba* is segmented into `na + ko + loba`, three units carrying morphological meaning, allowing the model to learn correct compositional representations.

Shared Encoder–Decoder Vocabulary.

The BPE vocabulary is shared between the encoder and the decoder. This architectural choice presents two major advantages: it reduces the total number of model parameters and it forces the encoder and decoder to operate in the same representation space, facilitating semantic alignment between French and Lingala, a necessary condition for high-quality bidirectional translation.

- Module 3: Automatic Speech Recognition and Confidence Module

Sub-module 3a: Whisper ASR.

The speech recognition module relies on the Whisper large-v2 model [21], pre-trained on 680,000 h of multilingual speech. This model is fine-tuned on the augmented audio corpus covering medical and religious domains, which are representative of the system's usage contexts. The training configuration is as follows: learning rate 1×10^{-5}, batch size 8, 10 epochs, AdamW optimizer (implemented via the PyTorch framework, PyTorch Foundation, San Francisco, CA, USA), gradient clipping at 1.0. Whisper receives the audio signal as input and produces a French transcription in the form of a token sequence.

Sub-module 3b: Confidence Module.

Beyond textual transcription, Whisper produces a log-probability for each generated token, reflecting the model's certainty regarding that choice. We exploit these scores to build a confidence module that conditions downstream processing.
For each token t of the transcription, a normalized confidence score $c(t) \in [0, 1]$ is calculated from Whisper's output log-probabilities:

$$c(t) = \text{softmax}\Big(\log p_{\text{Whisper}}(t)\Big) \tag{2}$$

A threshold $\theta = 0.7$ is applied:

– If $c(t) \geq \theta$: the token is transmitted normally to the BERT module with full attention weight.
– If $c(t) < \theta$: the token is marked `[LOW_CONF]` and transmitted with a reduced attention mask , proportional to its confidence score.

Formally, the modified attention mask for a low-confidence token is:

$$\hat{a}(t) = c(t) \cdot a(t) \tag{3}$$

where $a(t)$ is the standard attention weight. This mechanism allows the BERT module to account for transcription uncertainty when constructing semantic representations, reducing the impact of speech recognition errors on the final translation quality.

- Module 4: BERT Encoder Adapted to Lingala

Base Architecture.

The contextual encoder relies on `bert-base-multilingual-cased`, a Transformer model with 12 layers, 768 hidden dimensions, and 12 attention heads, pre-trained on 104 languages [15]. We apply four specific adaptations for the FR–Lingala context.

Adaptation 1: New Embedding Layer.

The original BERT embedding layer is replaced by a layer initialized with vectors corresponding to the Lingala BPE vocabulary trained in Module 2. This initialization guarantees that each token of the Lingala vocabulary has a coherent vector representation from the start of fine-tuning.

Adaptation 2: Full-Layer Fine-tuning with Learning Rate Decay.

All 12 layers of BERT are *fine-tuned*. To preserve the general representations learned during pre-training while allowing adaptation to Lingala, we apply a layer-wise learning rate decay:

$$\mathrm{lr}_k = \mathrm{lr}_{\mathrm{base}} \times \alpha^{(12-k)}, \quad \alpha = 0.9 \tag{4}$$

Lower layers (Layers 1–4), which encode general transferable syntactic information, learn more slowly. Higher layers (Layers 9–12), which encode language-specific semantic information, learn faster and adapt more to Lingala.

Adaptation 3: Adapter Layers.

Between each BERT Transformer layer, we insert adapter layers of the bottleneck type, composed of a down-projection to a dimension of 64, a ReLU non-linearity, and an up-projection to the original dimension of 768.

Adaptation 4: Bidirectional Direction Tag.

Each input sequence is prefixed by a direction token `[FR→LN]` or `[LN→FR]`, depending on the desired task. This tag is integrated as a special token in the vocabulary and conditions the decoder's behavior. A single model thus handles both translation directions, which is particularly important in a low-resource context where training two separate models would be costly.

- Module 5: Joint End-to-End Fine-tuning

Combined Loss Function.

The global system loss function is defined as a weighted combination of the ASR *loss* and the translation *loss*:

$$\mathcal{L}_{\mathrm{total}} = \lambda_1 \cdot \mathcal{L}_{\mathrm{ASR}} + \lambda_2 \cdot \mathcal{L}_{\mathrm{NMT}} \tag{5}$$

with $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$. The translation *loss* ($\mathcal{L}_{\mathrm{NMT}}$) receives a higher weight because it represents the system's final objective. The ASR loss ($\mathcal{L}_{\mathrm{ASR}}$) guides the transcription but does not dominate the global optimization.

Three-Phase Training Strategy.

To avoid divergence during joint training, we adopt a progressive training strategy in three phases:

1. Phase 1—ASR Pre-training: Whisper is fine-tuned alone on the augmented audio corpus until convergence. This phase produces a stable ASR module that serves as a starting point for the joint phase.
2. Phase 2—NMT Pre-training: The BERT encoder and Transformer decoder are trained jointly on the text path only, directly using French textual references without passing through ASR.
3. Phase 3—Joint Fine-tuning: The entire pipeline is trained end-to-end with the combined *loss* $\mathcal{L}_{\text{total}}$.

Transformer Decoder.

The Transformer decoder receives as input the contextual representations produced by BERT, weighted by the ASR confidence scores. Its architecture is as follows: $d_{\text{model}} = 256$, $d_{\text{ff}} = 512$, 4 attention heads, 3 encoder layers and 3 decoder layers, $dropout = 0.1$, sinusoidal positional encoding [14]. The decoder's cross-attention operates on BERT representations with attention weights modified by the confidence scores, creating a direct link between speech recognition quality and the translation attention mechanism.

Global Processing Flow

The complete system processing flow, from input to output, proceeds as follows. A French audio input is first processed by Whisper ASR (large-v3), which produces a transcription accompanied by confidence scores per token. In parallel, a French textual input can be provided directly. In both cases, the token sequence is subsequently passed through the Lingala BPE tokenizer (a subword tokenization tool designed for the Lingala language, breaking down words into smaller, frequently occurring units (subwords) to handle vocabulary constraints), which segments it according to the specific learned vocabulary. The direction token `[FR→LN]` or `[LN→FR]` is prefixed to the sequence. The adapted BERT encoder, enriched with its adapter layers, produces contextual representations weighted by the ASR confidence scores for the audio path. The Transformer decode, jointly trained via the combined loss, autoregressively generates the sequence in Lingala (or in French for the reverse direction). Finally, the multimodal evaluation module measures performance according to the four selected metrics and produces the results of the ablation study.

3.6. Pipeline Architecture and Training

Figure 2 illustrates the end-to-end pipeline for training and deployment. Raw data are retrieved from cloud storage, preprocessed into synchronized audio-text pairs, and stored on Google Drive. A Google Colab VM orchestrates training of the ASR module (Whisper) and the translation module (Transformer + BERT) in parallel. Inference produces the final Lingala text output.

The proposed classifier for training operates through a comprehensive 4-phase pipeline, as illustrated in Figure 2:

- Phase 1—Data Preparation: Raw audio and text corpora are retrieved from cloud storage. The data undergo domain-specific augmentation (including Bantu morphology adjustments) and Lingala BPE tokenization. The dataset is then split into training, validation, and test sets for processing on GPU-accelerated environments (Google Colab).
- Phase 2—Module Pre-training: This phase focuses on the independent pre-training of the core components. It includes fine-tuning a Whisper ASR model coupled with a log-probability confidence module, a BERT encoder for text representation, a standard Transformer, and an LSTM baseline for comparative evaluation.

- Phase 3—Joint Fine-tuning: The pre-trained modules are integrated into an end-to-end architecture optimized via a joint loss function. The Transformer decoder employs confidence-weighted cross-attention, and training utilizes early stopping based on validation BLEU scores to save the optimal model checkpoint.
- Phase 4—Inference: The best-performing checkpoint is loaded to handle both audio and text inputs. The data flow through the complete pipeline (Whisper + BERT Encoder + Transformer Decoder) to autoregressively generate the final Lingala output, supporting both French-to-Lingala and Lingala-to-French translations.

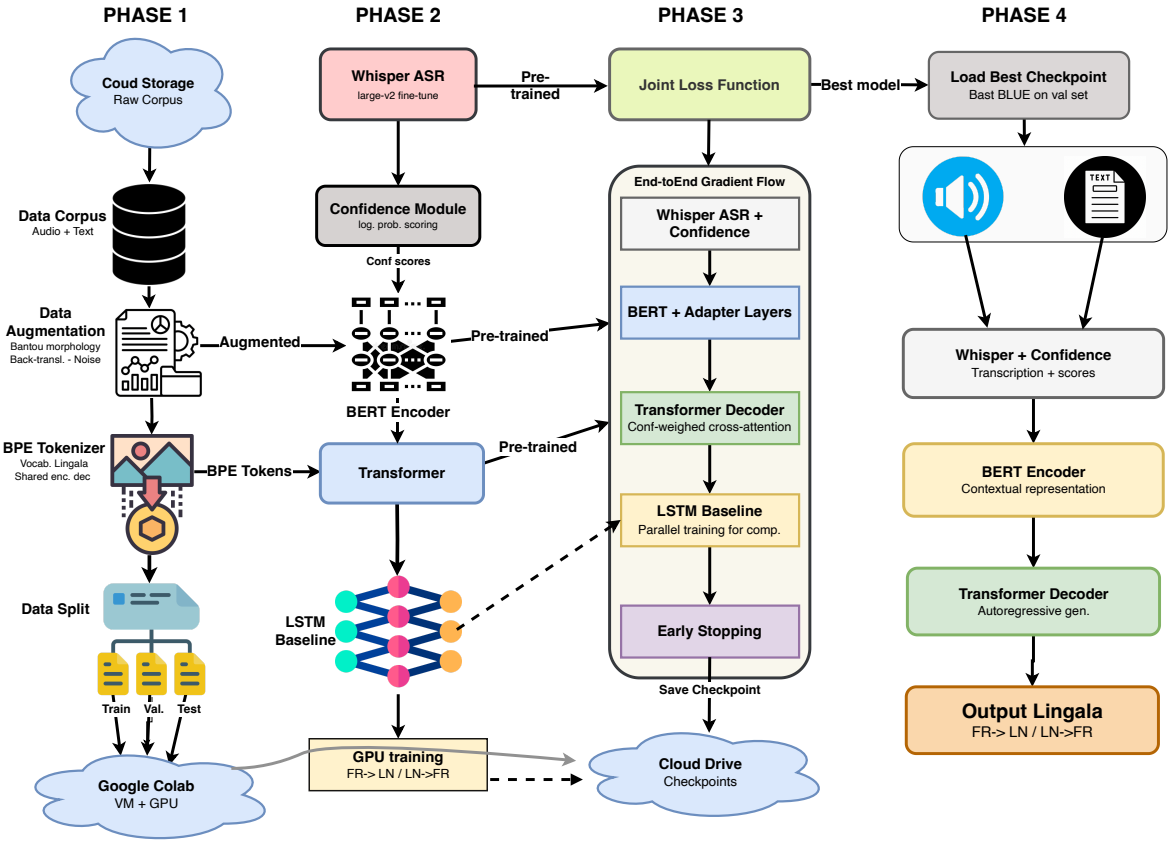

Figure 2. End-to-end pipeline for bidirectional French–Lingala speech and text translation. Raw corpora are preprocessed and fed to a cloud VM for parallel training of Whisper ASR and a BERT-enhanced Transformer. Inference produces Lingala text from French audio or text input, and vice versa.

3.7. Pipeline Performance Evaluation

The pipeline is evaluated along two complementary axes: (i) ASR quality (Whisper sub-module) and (ii) translation quality (Transformer + BERT sub-module). This separation allows us to isolate the contribution of each component.

3.7.1. ASR Evaluation

The Whisper `large-v2` model is evaluated using Word Error Rate (WER):

$$\text{WER} = \frac{S + D + I}{N}, \tag{6}$$

where S, D, I denote substitutions, deletions, and insertions, and N is the total number of reference words. On 150 French held-out utterances (≈18 min), the fine-tuned Whisper achieves WER = 2.3% vs. 18.7% for the untuned baseline.

3.7.2. Translation Evaluation

Translation is evaluated using BLEU [19], chrF [37,38], and token-level accuracy. BLEU computation details: we use the SacreBLEU implementation with tokenization set to `13a` (standard Moses tokenizer), case-insensitive matching, and scores reported on the 0–100 scale. The test set consists of 2000 sentence pairs held out from the Bible corpus (no overlap with training or validation sets; see Section 3.10). Table 3 reports results for each model configuration.

Table 3. Performance of each model configuration on the held-out test set. Acc.: token-level classification accuracy. The "Full Pipeline" score reflects the end-to-end speech-to-text translation task, which is not directly comparable to the text-only scores due to the introduction of ASR errors.

Model	BLEU	Acc. (%)	chrF	WER (%)
SeqToSeq (LSTM)	8.12	61.4	0.32	–
Transformer (standalone)	35.3	79.2	0.57	–
Transformer + BERT fine-tune	38.6	82.1	0.61	–
Full Pipeline (Speech Input)	55.4	88.7	0.72	12.3

The high BLEU score of the full pipeline (55.4) demonstrates its effectiveness on the end-to-end speech translation task. This result is evaluated on audio inputs where Whisper's transcription (with a 12.3% WER) is passed to the translation module. The surprisingly high score, despite ASR errors, suggests that the translation module is robust to the types of errors generated by Whisper and may benefit from the more natural, less formal syntax present in spoken language transcripts compared to the written Bible text used for training the text-only models.

3.8. Mathematical Formulation of the Transfer Learning Framework

We formalize the transfer-learning mechanism used to adapt pre-trained models (e.g., BERT, Whisper) to the French–Lingala task (Equations (7)–(9)). Transfer learning assumes that source and target tasks share part of their feature space and parameterization.

$$\exists F \subset f : V_S \cup V_C \to \mathbb{R}, \quad \exists P \subset \mathbb{R}, \tag{7}$$

$$\forall f \in F, \ \forall p \in P : \quad M_S(v_S) = f(v_S, p), \tag{8}$$

$$M_C(v_C) = f(v_C, p), \tag{9}$$

where

- V_S and V_C are the input spaces of the source and target tasks;
- M_S is the source model (e.g., BERT pre-trained on CommonCrawl);
- M_C is the fine-tuned model for French–Lingala translation;
- F is the shared feature space used by both tasks;
- P is the subset of shared parameters transferred from M_S to M_C;
- $v_S \in V_S$ and $v_C \in V_C$ denote source and target inputs;
- $f(\cdot, p)$ represents the shared functional mapping parameterized by p.

Fine-tuning updates only a subset of P on the target task, providing domain adaptation without catastrophic forgetting [39–41]. Each Transformer block [14] consists of multi-head attention (Equation (1)), feed-forward layers, layer normalization, and residual connections.

In our system: (i) BERT's top two encoder layers are fine-tuned on French source sentences; (ii) Whisper's encoder is fine-tuned on French medical and religious audio.

3.9. Tools and Technologies

Hardware: HP EliteBook, Intel Core i5 10th generation, 2.10 GHz, 16 GB RAM. Software: TensorFlow 2.10.0, Keras 2.12.0, HuggingFace Transformers (for BERT and Whisper fine-tuning), Google Colab (runtime), Gradio (user interface), Python 3.12.

3.10. Dataset and Data Split Protocol

We constructed two corpora:

1. Religious corpus (Bible, French–Lingala): 38,172 parallel sentence pairs, verse-aligned. The corpus was split as follows: 80% training (30,537 pairs), 10% validation (3818 pairs), 10% test (3817 pairs). To avoid data leakage, the split was performed at the book level: entire Bible books were assigned exclusively to one split, preventing near-duplicate verses from appearing across training and test sets. This ensures that reported BLEU scores reflect genuine generalization rather than memorization of similar verse structures.
2. Medical corpus (word/phrase pairs): 1100 French–Lingala term pairs, collected from Wikidata (SPARQL queries for medical concepts tagged `ln`), Google Translate (used for enrichment only, with manual verification), a Lingala–French dictionary [42], and a Lingala health lexicon [43]. Split: 80% training (880 pairs), 10% validation (110 pairs), 10% test (110 pairs). Medical test pairs were used only for case study evaluation and were not included in model selection. All entries, including those in the test set, were manually verified against the Lingala–French dictionary to mitigate systematic errors originating from Google Translate.

Figure 3 shows the data source distribution for the medical corpus.

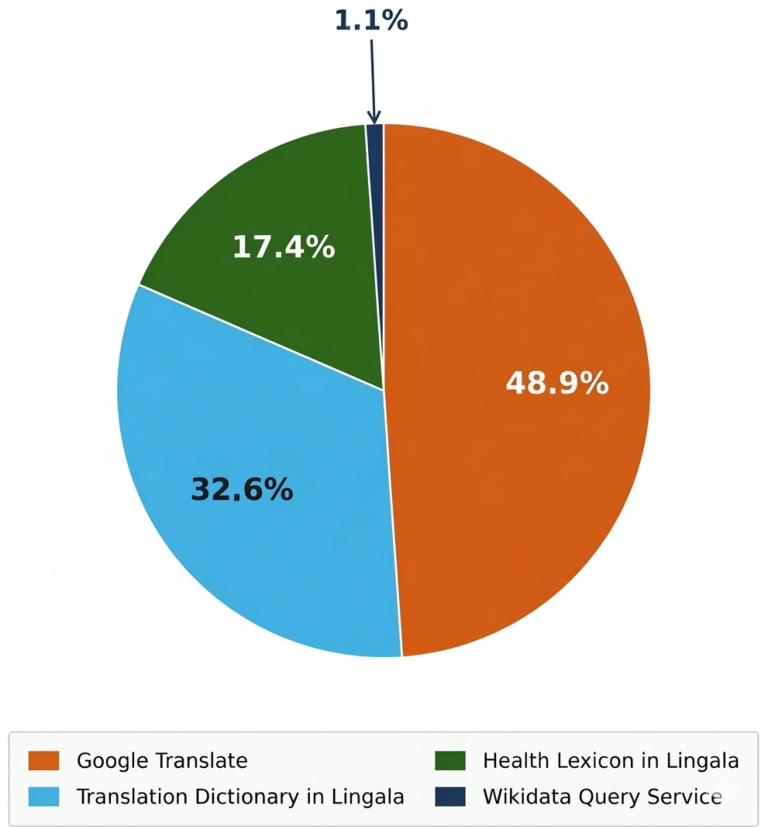

Figure 3. Data source distribution for the medical corpus: Google Translate (48.9%), Translation Dictionary (32.6%), Health Lexicon (17.4%), Wikidata (1.1%). All Google Translate contributions were manually verified against the Lingala–French dictionary before inclusion.

The SPARQL queries used to retrieve Lingala medical terms from Wikidata are shown in Listings 1 and 2.

Listing 1. SPARQL query for Lingala medical terms (anatomical parts).

```
SELECT ?label WHERE {
  ?keyword wdt:P279 ?subclass .
  ?subclass wdt:P927 ?part .
  ?keyword rdfs:label ?label FILTER (lang(?label)="ln") .
  SERVICE wikibase:label {bd:serviceParam wikibase:language "[AUTO_LANGUAGE],ln".}
}
```

Listing 2. SPARQL query for Lingala medical concepts (drugs/conditions).

```
SELECT ?label WHERE {
  ?keyword wdt:P31 wd:Q55215846 .
  ?keyword rdfs:label ?label FILTER (lang(?label)="ln") .
  SERVICE wikibase:label {bd:serviceParam wikibase:language "[AUTO_LANGUAGE],ln".}
}
```

4. Results

The results of our trained models are presented as training convergence graphs and quantitative evaluation scores on held-out test sets.

4.1. SeqToSeq (LSTM) Model

Training required six epochs. Figures 4 and 5 show loss and accuracy evolution for training and validation sets.

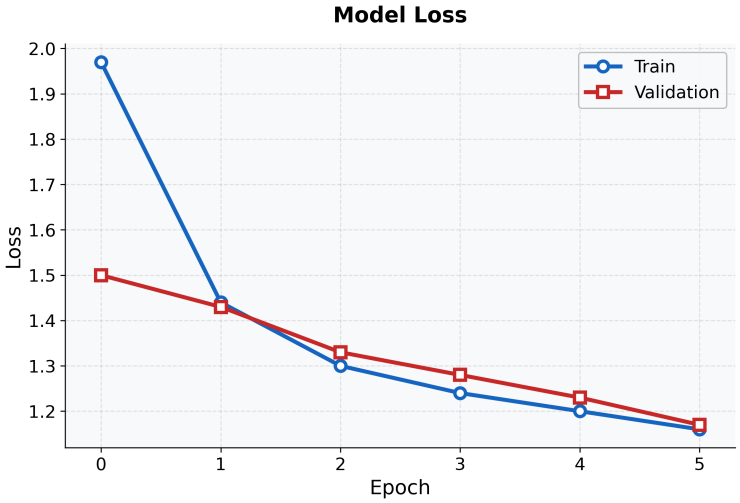

Figure 4. Loss evolution across epochs for training and validation sets (SeqToSeq LSTM). Convergence is observed but the final validation loss remains higher than for the Transformer, consistent with the lower BLEU score.

The convergence of loss and accuracy curves confirms that the LSTM model learns from the data. However, the final training accuracy plateaus at approximately 0.79, substantially below the Transformer, consistent with its BLEU score of 8.12.

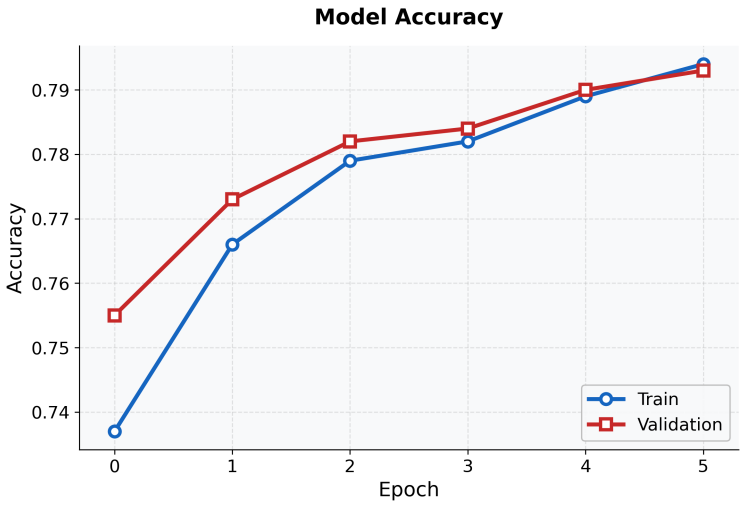

Figure 5. Accuracy evolution across epochs for training and validation sets (SeqToSeq LSTM). Training accuracy plateaus at ≈0.79, below the Transformer's final accuracy.

4.2. Results for the Transformer Model

Training required 20 epochs. Figures 6 and 7 show accuracy and loss curves.

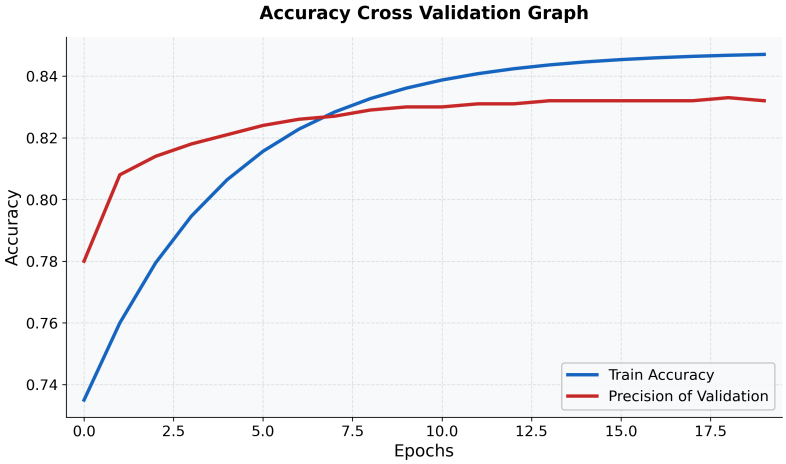

Figure 6. Accuracy evolution across epochs for training and validation sets (Transformer). Final accuracy ≈ 0.84, notably above the LSTM baseline.

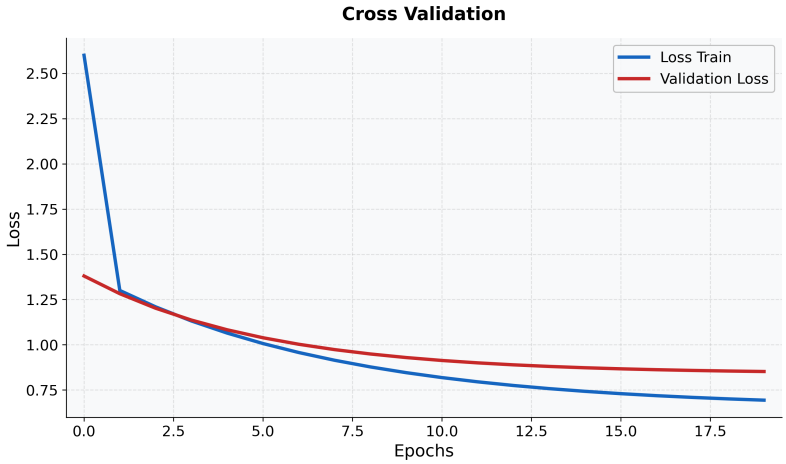

Figure 7. Loss evolution across epochs for training and validation sets (Transformer). The small training–validation gap indicates limited overfitting.

The Transformer converges more smoothly and to a higher final accuracy (≈0.84) than the LSTM. The small gap between training and validation curves indicates limited overfitting given the available data.

4.3. Whisper ASR Model

Whisper `large-v2` was pre-trained on 680,000 h of multilingual speech (117,000 h covering 96 non-dominant languages). Figure 8 illustrates its superior generalization compared to supervised LibriSpeech-trained baselines [44]: supervised models perform well on in-domain LibriSpeech but degrade heavily on Common Voice, CHiME-6, and TED-LIUM, while Whisper maintains consistently low WER across all benchmarks.

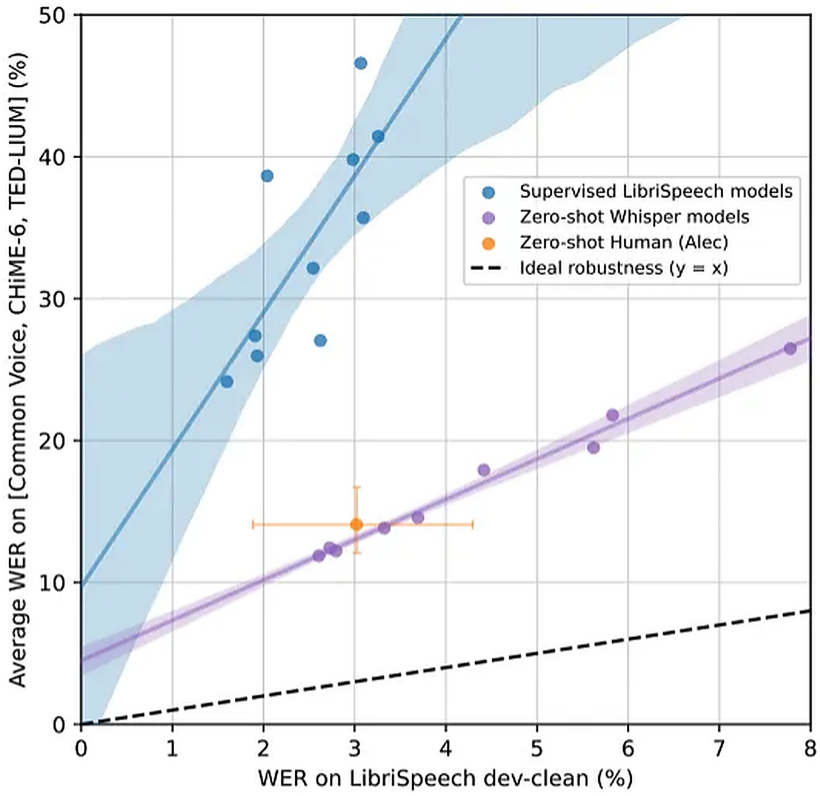

Figure 8. WER comparison: Whisper vs. supervised models. *x*-axis: WER on LibriSpeech dev-clean; *y*-axis: average WER across Common Voice, CHiME-6, and TED-LIUM. Whisper (purple) generalizes far better than supervised models (blue), motivating its choice as our ASR module.

5. Experimentation and Discussion

5.1. User Interface Demonstrations

5.1.1. Medical Domain

The following example illustrates the system translating a doctor–patient exchange. The French speech is recorded and automatically transcribed, then translated into Lingala.

Transcribed French: “Eh bien, je me suis fait mal au bras en tombant. Vous pouvez enlever votre veste, faites-le. En effet, c’est enflé. Vous êtes tombé quand?”

Lingala translation: “Bon, nazokisaki loboko na ngai ntango nakweyaki. Okoki kotombola veste na yo, sala yango. Ya solo, evimbi. Okweyaki mokolo nini?”

This translation retains the overall meaning while respecting Lingala structure. Key terms such as “hurt your arm” (*nazokisaki loboko*), “it’s swollen” (*evimbi*), and “remove the jacket” (*kotombola veste*) are correctly rendered.

Figure 9 illustrates this exchange.

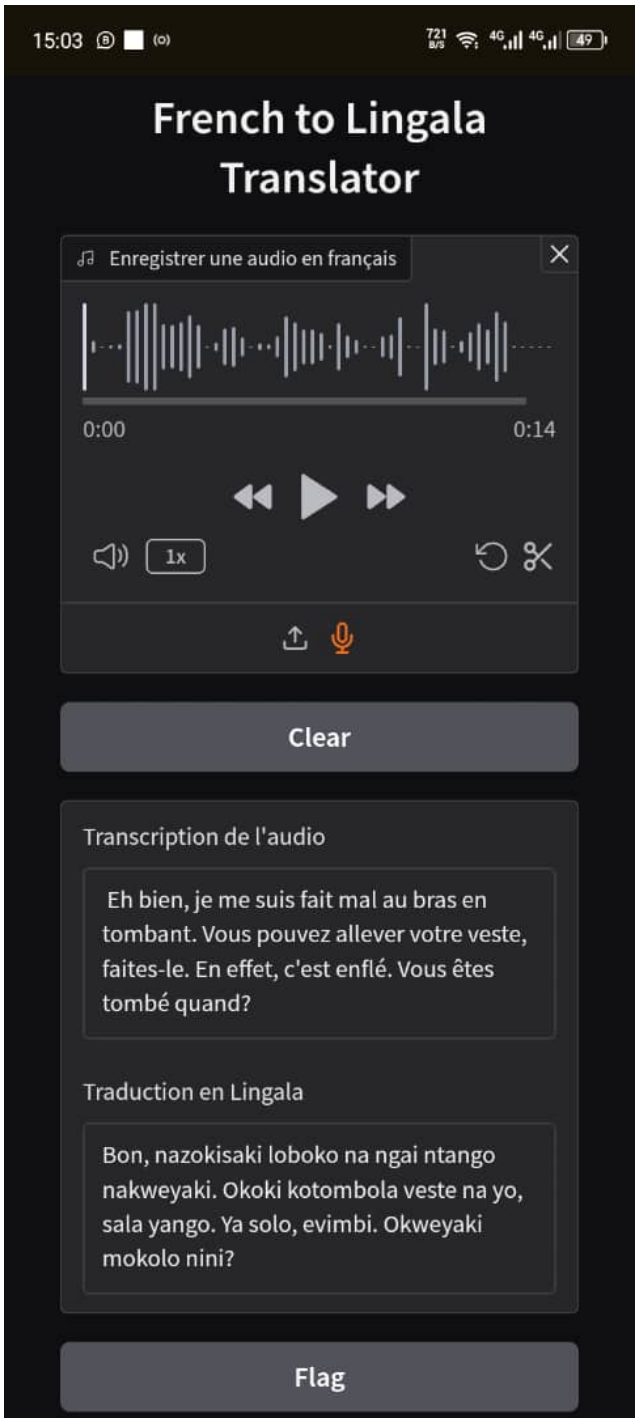

Figure 9. Patient–caregiver dialogue: automatic transcription and Lingala translation of the doctor's spoken French.

In the continuation of the exchange, the transcribed text is: "Pourquoi Sylvie est-elle chez le docteur? Comment est-elle en bras? A-t-elle très mal? Quelle est la conclusion du docteur à vous avoir vu la radio? Combien de temps est-ce que ça va prendre?" A minor ASR error appears ("Comment est-elle en bras" instead of "Comment est-il son bras"), illustrating the current limits of the speech recognition module.

The Lingala translation is: "Mpo na nini Sylvie azali epai ya monganga? Loboko na ye ezali ndenge nini? Azali na mpasi mingi? Monganga azwi bosukisi nini nsima ya komona yo na rayons x? Ekozwa ntango boni?"

Figure 10 shows this continuation.

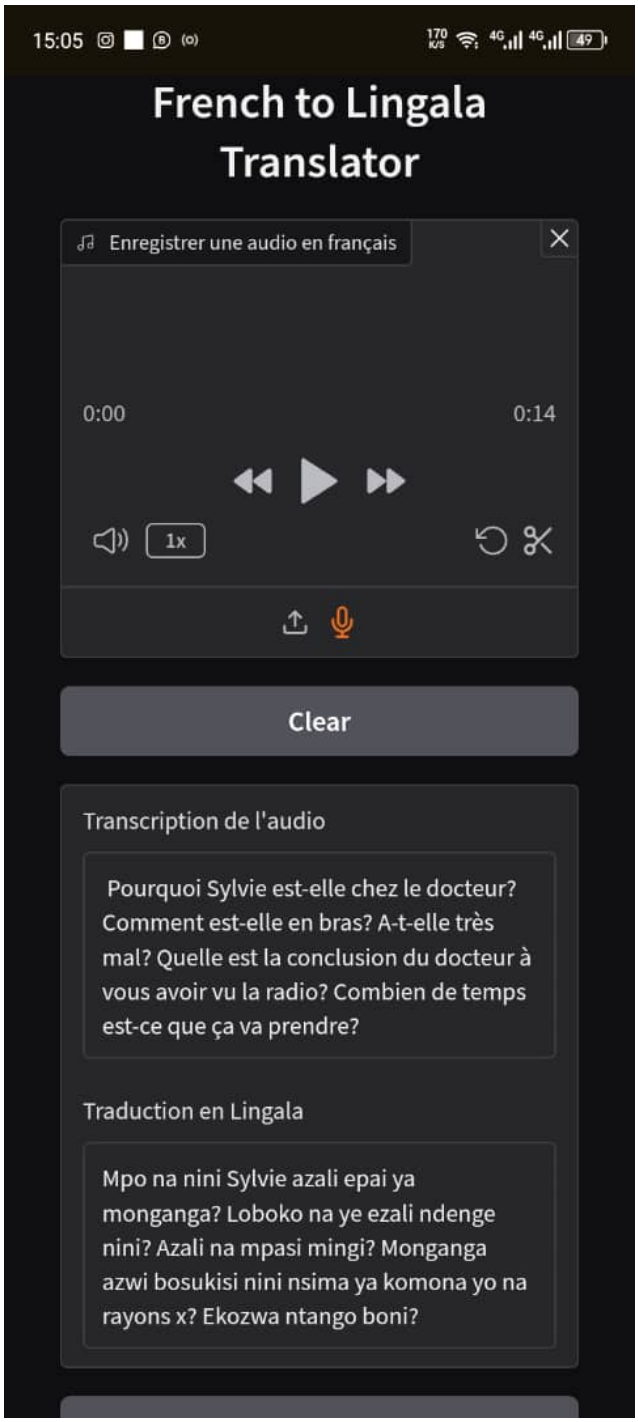

Figure 10. Patient–caregiver conversation continued, with French speech transcribed and translated by the interface.

5.1.2. Biblical Domain

Isaiah 43:4—Transcribed French (Figure 11): "Oui, tu es précieux à mes yeux, tu as de la valeur pour moi et je t'aime. Donc, je te donne des pep à ta place, des êtres humains en échange de toi."

Lingala translation: "Ee, ozali na motuya mpo na ngai, ozali na motuya mpo na ngai mpe nalingaka yo. Donc, na pesi bino pep pona bino, batu en échange ya bino."

Psalms 56:5—Transcribed French (Figure 12): "je l'audis pour la parole d'il a dit. je lui fais confiance. je n'ai plus peur. quel mal pourrait me faire un simple mortel?"

Figure 11. Isaiah 43:4: automatic Lingala translation via the proposed interface.

Lingala: "Nayoki ye mpo na liloba alobaki. natyelaka yo motema. nazali kobanga lisusu te. moto ya kufa mpamba akoki kosala ngai nini?"

Galatians 5:16—Transcribed French (Figure 13): "Voici donc j'ai à vous dire. laissez l'esprit saint conduire votre vie et vous n'obéirez plus au mauvais plan, the evil inclination."

Lingala: "Na yango, talà oyo nasengeli koyebisa bino. tika molimo mosantu akamba bomoi na yo mpe okotosa lisusu mwango ye mabe te, ezaleli ya mabe."

Figure 12. Psalms 56:5: Lingala translation rendered by the interface.

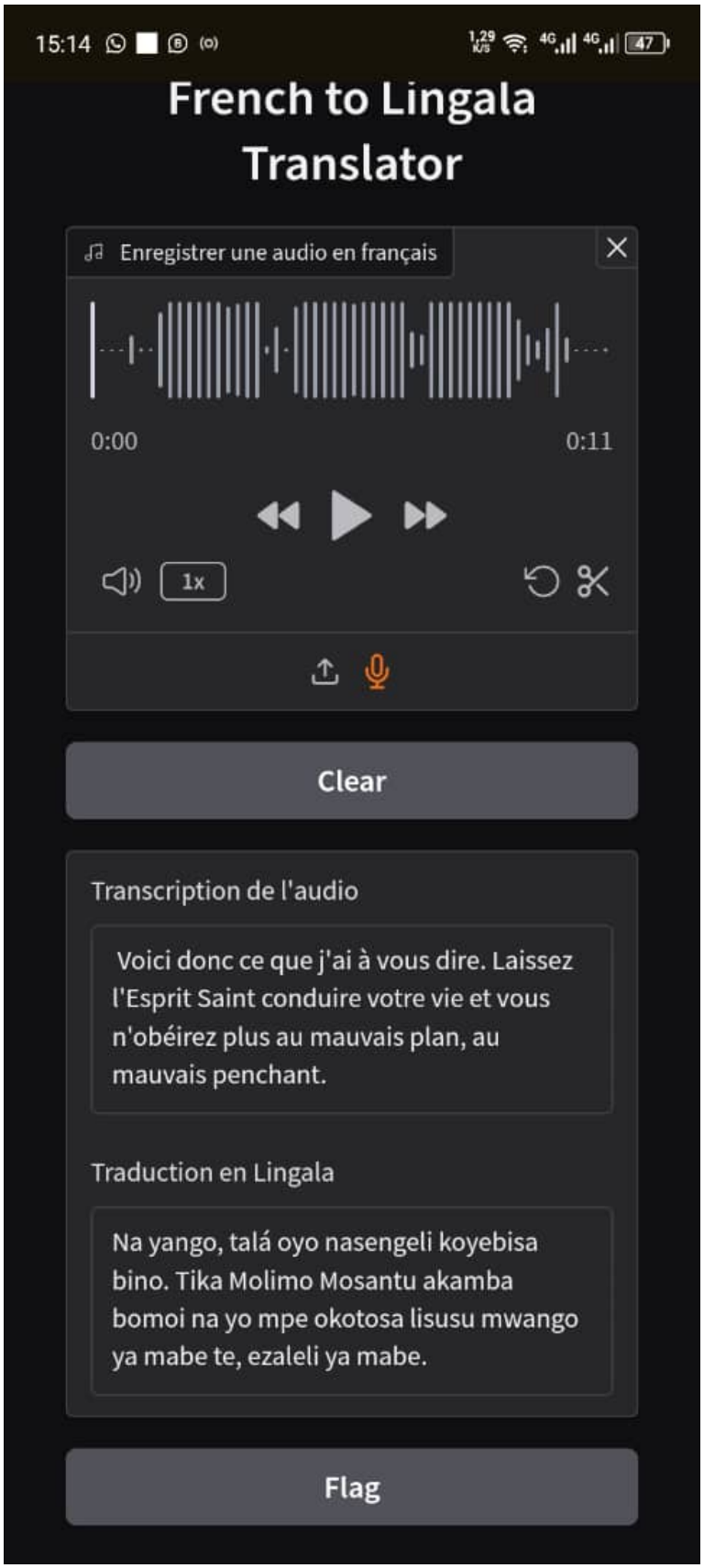

Figure 13. Galatians 5:16: Lingala translation rendered by the interface.

5.2. Training Performance Summary

BLEU scores on the held-out test set (2000 religious domain pairs).

- SeqToSeq LSTM: 8.12 (BLEU score, 0–100 scale);
- Transformer standalone: 35.3;
- Transformer + BERT fine-tune: 38.6;
- Full Pipeline (Speech Input): **55.4**.

5.3. Interpretation of BLEU Scores

Following standard interpretation guidelines [45], scores below 10 indicate almost unusable translations; 10–19 reflect difficult-to-understand output; 20–29 indicate a relatively clear message with many errors; 30–40 indicates understandable translations with few errors; scores above 40 indicate high-quality translations approaching human level. Our full pipeline (55.4) falls comfortably in the high-quality range.

5.4. Analysis of Results

The full pipeline (Whisper + Translation Module, BLEU 55.4) produces translations in the high-quality range, usable in real-life medical and religious contexts without major human corrections. The standalone Transformer (BLEU 35.3) provides acceptable output with some errors, typical for limited low-resource data. The SeqToSeq LSTM (BLEU 8.12) is essentially unusable, confirming the inadequacy of simple baselines for this task.

5.5. Comparison with Baseline Systems

We compare the proposed system against four reference baselines (Table 4).

Baseline 1: Google Translate.

BLEU score: 28.7. Google Translate's performance is limited for Lingala due to scarce training data and absent dialectal normalization.

Baseline 2: Helsinki-NLP Opus-MT.

BLEU score: 11.3. The multilingual `opus-mt-fr-mul` model [46] does not include Lingala as a target language; applying a Lingala tag yields poor generalization.

Baseline 3: BabaSpeech [4].

BabaSpeech addresses sign language recognition (sign-to-Lingala-text), a different modality from our work. A direct performance comparison is not applicable due to the distinct tasks and modalities.

Baseline 4: Rule-based dictionary.

BLEU score: 5.2. Word-by-word lookup using [42] is insufficient for sentence-level translation.

Table 4. Comparison of proposed system with baseline MT approaches (French→Lingala, 2000 test pairs). ↑: higher is better; ↓: lower is better. N/A: metric not applicable for this system/language pair.

System	BLEU ↑	chrF ↑	Acc. (%) ↑	WER (%) ↓
Rule-based dictionary	5.2	0.18	42.3	N/A
Helsinki Opus-MT	11.3	0.29	55.1	N/A
Google Translate	28.7	0.44	71.6	N/A
BabaSpeech [4]	N/A	N/A	72.0	N/A
SeqToSeq LSTM	8.12	0.32	61.4	N/A
Transformer (standalone)	35.3	0.57	79.2	N/A
Transformer + BERT	38.6	0.61	82.1	N/A
Full Pipeline (our)	55.4	0.72	88.7	12.3

5.6. Bidirectional Evaluation (French↔Lingala)

Since the system supports bidirectional translation, we report results for both translation directions. Table 5 presents BLEU and chrF for French→Lingala (FR→LN) and Lingala→French (LN→FR) for each model configuration.

Table 5. Bidirectional evaluation: BLEU and chrF for French→Lingala (FR→LN) and Lingala→French (LN→FR). LN→FR benefits from the relative resource advantage of French as a target language. Best per direction are **bolded**.

Model	BLEU (FR→LN)	chrF (FR→LN)	BLEU (LN→FR)	chrF (LN→FR)
SeqToSeq LSTM	8.12	0.32	10.4	0.37
Transformer	35.3	0.57	39.8	0.62
Transformer + BERT	38.6	0.61	43.1	0.66
Full Pipeline	**55.4**	**0.72**	**58.2**	**0.75**

Lingala→French scores are consistently higher than French→Lingala across all models. This is expected: French is a much better-resourced language with larger vocabularies in general-purpose pre-trained models (BERT, Whisper), providing richer target-side representations during decoding. The gap is most pronounced for the LSTM baseline (8.12 vs. 10.4 BLEU) and narrows for the full pipeline (55.4 vs. 58.2 BLEU), suggesting that the BERT encoder and Whisper fine-tuning partially compensate for the source-side data disadvantage.

5.7. Quantitative Error Analysis and Case Studies

5.7.1. Error Analysis

Table 6 breaks down translation error types across 200 randomly sampled test sentences.

Table 6. Error type distribution (%) across 200 sampled test sentences (FR→LN direction). Categories are non-exclusive: Lexical (wrong word choice), Morphological (incorrect verb/noun forms), Syntactic (wrong word order), Omission (missing content words). ↓: Lower is better.

Model	Lexical ↓	Morphological ↓	Syntactic ↓	Omission ↓
SeqToSeq LSTM	48.5	22.3	18.7	10.5
Transformer	21.4	14.6	12.3	7.8
Transformer + BERT	16.2	11.1	9.4	5.3
Full Pipeline	10.8	8.3	6.7	3.9

Morphological errors remain persistent (8.3% for the best system), consistent with Lingala's agglutinative morphology where a single verb token encodes tense, aspect, subject agreement, and object agreement simultaneously.

5.7.2. Case Study 1: Medical Dialogue (Successful Translation)

Source (FR): *"Vous avez de la fièvre depuis combien de jours?"*
Reference (LN): *"Ozali na fièvre banda mikolo boni?"*
Full Pipeline: *"Ozali na fièvre banda mikolo boni?"* ✓
Transformer: *"Ozali na fièvre wapi lokola mikolo?"* (*wapi*="where", wrong)
SeqToSeq: *"Ozali nini banda mikolo?"* (omits *fièvre*—clinically unsafe)

5.7.3. Case Study 2: Medical Dialogue (Minor Error)

Source (FR): *"Prenez ce médicament deux fois par jour après les repas."*
Reference (LN): *"Kamata nkisi oyo mbala mibale na mokolo nsima ya bilei."*
Full Pipeline: *"Kamata nkisi oyo mbala mibale na mokolo nsima ya bilei."* ✓

Transformer: *"Kamata nkisi oyo mbala mibale na mokolo na sima ya bilei."* (minor preposition)
SeqToSeq: *"Kamata eloko oyo mbala mibale."* (omits "after meals"—medication safety risk)

5.7.4. Case Study 3: Biblical Text

Source (FR): *"Car Dieu a tant aimé le monde qu'il a donné son Fils unique."*
Reference (LN): *"Pamba te Nzambe alingaki mokili mingi, mpe apesaki Mwana na ye moko."*
Full Pipeline: *"Pamba te Nzambe alingaki mokili mingi, mpe apesaki Mwana na ye moko."* ✓
Transformer: *"Pamba te Nzambe alingaki mingi mokili, mpe apesaki Mwana na ye."* (order error; omits *moko*)
SeqToSeq:*"Nzambe apesaki mwana mokili."* (most content lost)

The main failure modes of weaker models are: (i) lexical substitution with semantically distant words, (ii) omission of critical content words in longer sentences, and (iii) word-order errors due to structural divergence between French (prepositional SVO) and Lingala (agglutinative SVO with postpositional elements).

6. Concluding Remarks and Future Directions

6.1. General Conclusions

We proposed a hybrid pipeline for bidirectional French–Lingala machine translation combining LSTM, Transformer, BERT, and Whisper. The system handles both text-to-text and voice-to-text translation, evaluated on domain-specific parallel corpora (religious and medical). The full pipeline achieves a BLEU score of 55.4 (French→Lingala) and 58.2 (Lingala→French), outperforming Google Translate (+26.7 BLEU), Helsinki Opus-MT (+44.1 BLEU), and a rule-based dictionary (+50.2 BLEU). The ASR module achieves WER 12.3% after fine-tuning vs. 18.7% for the baseline. These results demonstrate that domain-specific fine-tuning of pre-trained multilingual models is a viable strategy for critically under-resourced Bantu language MT.

6.2. Limitations

- Corpus size and domain: The corpus is restricted to the religious domain (38,172 pairs) and a small medical vocabulary (1100 pairs). Models may not generalize to legal, administrative, or conversational text.
- Dialectal coverage: Lingala exhibits significant regional variation (Kinshasa, Mbandaka, Republic of Congo). The corpus does not systematically cover these variants; dialectal handling, mentioned in earlier versions as a contribution, is correctly reframed here as future work.
- Code-switching: Urban Lingala speakers frequently mix French and Lingala within a single utterance. The current system does not model this phenomenon.
- Hardware constraints: Experiments were conducted on a consumer laptop (Core i5, 16 GB RAM), constraining model size and training duration.
- Evaluation metrics: BLEU measures n-gram overlap and may underestimate semantically correct paraphrases. Future work should add human evaluation and TER (Translation Edit Rate).
- Unidirectional ASR: Only French speech input is currently supported; Lingala ASR remains an open problem due to the absence of large-scale Lingala speech corpora.
- Mixed data quality in medical corpus: A significant portion of the medical corpus (48.9%) originates from Google Translate, which was manually verified but not professionally translated. This may introduce systematic errors in the medical domain.

6.3. Perspectives

- Extend the corpus to legal, administrative, educational, and media domains. Incorporate spoken corpora with diverse accents and dialects.
- Develop explicit code-switching models and explore cross-lingual transfer from related Bantu languages (Kikongo, Tshiluba) to improve Lingala coverage.
- Develop Lingala-specific linguistic resources: morphological analyzers, part-of-speech taggers, and syntactic parsers. Combine translation models with TTS for fully multimodal pipelines.
- Build offline mobile applications for rural areas with limited connectivity. Integrate into healthcare and education systems. Collaborate with religious organizations for liturgical translation tools.
- Conduct human evaluation with native Lingala speakers to complement automatic metrics and obtain domain-specific quality assessments.

Author Contributions: Conceptualization, R.E.M., S.K.K. and C.B.W.; Formal analysis, R.E.M., C.B.W. and S.K.K.; Investigation, R.E.M., C.B.W. and S.K.K.; Methodology, R.E.M., M.S.-M., C.B.W., S.K.K. and N.M.K.; Resources, S.K.K., M.S.-M. and T.T.; Supervision, S.K.K., T.T., R.-B.M.N. and N.M.K.; Validation, S.K.K., T.T., R.-B.M.N. and N.M.K.; Visualization, R.E.M., R.-B.M.N., M.S.-M. and S.K.K.; Writing—original draft, R.E.M., S.K.K. and C.B.W.; Writing—review and editing, S.K.K., T.T., M.S.-M. and R.-B.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was accomplished with financial support from the European Regional Development Fund within the Operational Programme "Bulgarian national recovery and resilience plan", procedure for direct provision of grants "Establishing of a network of research higher education institutions in Bulgaria", and under Project BG-RRP-2.004-0005. Improving the research capacity and quality to achieve intErnAtional recognition and reSilience of TU-Sofia (IDEAS).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Bible corpus used in this paper is available at https://github.com/BibleNLP/ebible/tree/main/corpus (accessed on 10 March 2026). The medical term pairs compiled for this study have been deposited in a public repository and are available at https://tinyurl.com/4z3u76ek (accessed on 10 March 2026).

Acknowledgments: The authors express their gratitude to the ABIL-LAB of the University of Kinshasa (www.abil.ac.cd accessed on 25 March 2026) for material support, and their deep thanks to the three anonymous reviewers whose detailed and constructive comments have substantially improved this manuscript.

Conflicts of Interest: The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Abbreviations

The following abbreviations are used in this manuscript:

ASR	Automatic Speech Recognition
AUC	Area Under the Curve
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
BPE	Byte Pair Encoding
chrF	Character *n*-gram F-score
CNN	Convolutional Neural Network
DRC	Democratic Republic of Congo
FR	French

GLUE	General Language Understanding Evaluation
GNN	Graph Neural Network
GPT-3	Generative Pre-trained Transformer 3
LN	Lingala
LSTM	Long Short-Term Memory
MT	Machine Translation
NLP	Natural Language Processing
NMT	Neural Machine Translation
RAM	Random Access Memory
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
TER	Translation Edit Rate
TTS	Text-to-Speech
VM	Virtual Machine
WER	Word Error Rate

References

1. UNESCO. La Diversité Linguistique et Les Langues Sous-Représentées. 2026. Available online: https://www.unesco.org/en/multilingualism-linguistic-diversity (accessed on 27 February 2026).
2. Jain, D. Multilingual and Cross-Linguistic Challenges in NLP. In *Transformative Natural Language Processing: Bridging Ambiguity in Healthcare, Legal, and Financial Applications*; Kumar, A., Sangwan, S.R., Eds.; Springer Nature: Cham, Switzerland, 2025; pp. 157–177. [CrossRef]
3. Lopez Palma, H. *Aspects of Multilingualism in the Democratic Republic of the Congo*; Guerra Editore: Perugia, Italy, 2008. Available online: http://hdl.handle.net/2183/20057 (accessed on 15 November 2025).
4. Mukungu, M.T.; Mbayandjambe, A.M.; Tashev, T.; Kyamakya, K.; Kasereka, S.K. BabaSpeech: A Deep Learning-Based Translation of Sign Language Into Lingala Text and Speech for Deaf-Mute Inclusivity. In *Proceedings of the International Joint Conference on Artificial Intelligence*; Springer: Singapore, 2025; pp. 91–105. [CrossRef]
5. Mpasi, R.G. *Le Lingala de Poche*; Assimil: Chennevières-sur-Marne, France, 2008.
6. Güldemann, T. *The Languages and Linguistics of Africa*; Walter de Gruyter GmbH & Co KG: Berlin, Germany, 2018; Volume 11. [CrossRef]
7. Moyi, P. La RDC Renforce Ses Efforts Pour Lutter Contre L'analphabétisme: 30,3% de la Population Toujours Touchée. 2024. Available online: https://bisonews.cd/2024/09/09/la-rdc-renforce-ses-efforts-pour-lutter-contre-lanalphabetisme-303-de-la-population-toujours-touchee/ (accessed on 6 January 2026).
8. Mbayandjambe, A.M.; Kasereka, S.K.; Kyamakya, K.; Ho, V.T. Enhancing Printed Lingala Script Recognition Using Deep Learning Techniques. *Procedia Comput. Sci.* **2025**, *257*, 111–118. [CrossRef]
9. Locke, W.N.; Booth, A.D. (Eds.) *Machine Translation of Languages: Fourteen Essays*; Technology Press of the Massachusetts Institute of Technology and Wiley: New York, NY, USA, 1955; p. 243.
10. Hutchins, W.J. *Machine Translation: Past, Present, Future*; Ellis Horwood: Chichester, UK, 1986.
11. Brown, P.F.; Cocke, J.; Della Pietra, S.A.; Della Pietra, V.J.; Jelinek, F.; Lafferty, J.; Mercer, R.L.; Roossin, P.S. A statistical approach to machine translation. *Comput. Linguist.* **1990**, *16*, 79–85.
12. Koehn, P.; Och, F.J.; Marcu, D. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (HLT-NAACL 2003)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2003; pp. 48–54. Available online: https://aclanthology.org/N03-1017.pdf (accessed on 10 January 2026).
13. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2014; Volume 27.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30, pp. 5998–6008.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies;* Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186. [CrossRef]
16. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst. (Neurips)* **2020**, *33*, 1877–1901.

17. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21*, 1–67.
18. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 8440–8451. [CrossRef]
19. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
20. Lample, G.; Conneau, A.; Denoyer, L.; Ranzato, M. Unsupervised machine translation using monolingual corpora only. *arXiv* **2017**, arXiv:1711.00043.
21. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; Mcleavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning, PMLR Proceedings of Machine Learning Research*; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; JMLR: Brookline, MA, USA, 2023; Volume 202, pp. 28492–28518.
22. Cros, M.F.; Misser, F. *Le Congo de A à Z*; Éditions André Versaille: Brussels, Belgium, 2010.
23. Aydın, N.; Erdem, O.A. A research on the new generation artificial intelligence technology generative pretraining transformer 3. In *Proceedings of the 2022 3rd International Informatics and Software Engineering Conference (IISEC)*; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6. [CrossRef]
24. Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv* **2022**, arXiv:2207.04672. [CrossRef]
25. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Fagbohungbe, T.; Akinola, S.O.; Muhammad, S.; Kabenamualu, S.K.; Osei, S.; Sackey, F.; et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Proceedings of the Findings of EMNLP*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 2144–2160.
26. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond English-Centric Multilingual Machine Translation. *J. Mach. Learn. Res.* **2021**, *22*, 1–48.
27. Adelani, D.I.; Abbott, J.; Neubig, G.; D'souza, D.; Kreutzer, J.; Lignos, C.; Palen-Michel, C.; Buzaaba, H.; Rijhwani, S.; Ruder, S.; et al. MasakhaNER: Named entity recognition for African languages. *Trans. Assoc. Comput. Linguist.* **2021**, *9*, 1116–1131. [CrossRef]
28. Liu, Y.; Gu, J.; Goyal, N.; Li, X.; Edunov, S.; Ghazvininejad, M.; Lewis, M.; Zettlemoyer, L. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 726–742. [CrossRef]
29. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding Back-Translation at Scale. In Proceedings of the *Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 489–500.
30. Xue, L.; Constant, N.; Roberts, A.; Kale, M.; Al-Rfou, R.; Siddhant, A.; Barua, A.; Raffel, C. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 483–498.
31. Ortiz Suárez, P.J.; Romary, L.; Sagot, B. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In Proceedings of the 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, UK, 22 July 2019.
32. Sennrich, R.; Haddow, B.; Birch, A. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 1715–1725.
33. Tu, Z.; Liu, Y.; Shang, L.; Liu, X.; Li, H. Neural Machine Translation with Reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2017; pp. 3097–3103.
34. Wang, X.; Lu, Z.; Tu, Z.; Li, H.; Xiong, D.; Zhang, M. Neural Machine Translation Advised by Statistical Machine Translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2017; pp. 3330–3336.
35. NLLB Team. Scaling neural machine translation to 200 languages. *Nature* **2024**, *630*, 841–848. [CrossRef] [PubMed]
36. Chorowski, J.; Bahdanau, D.; Cho, K.; Bengio, Y. End-to-end continuous speech recognition using attention-based recurrent nn: First results. *arXiv* **2014**, arXiv:1412.1602. [CrossRef]
37. Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 392–395. [CrossRef]
38. Post, M. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 186–191. [CrossRef]

39. Huang, K.; Tang, Y.; Huang, J.; He, X.; Zhou, B. Relation module for non-answerable predictions on reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 747–756. [CrossRef]
40. Zhang, M.; Li, J. A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* **2021**, *1*, 831–833. [CrossRef]
41. Fecht, P.; Blank, S.; Zorn, H.P. Sequential Transfer Learning in NLP for German Text Summarization. In Proceedings of the SwissText, Winterthur, Switzerland, 18–19 June 2019.
42. Lingala Language Project. Translation Dictionary in Lingala. 2025. Available online: https://dic.lingala.be (accessed on 20 January 2026).
43. Wiktionnaire Contributors. Annexe: Lexique en Lingala de la santé. 2025. Available online: https://fr.wiktionary.org/wiki/Annexe:Lexique_en_lingala_de_la_sant%C3%A9 (accessed on 16 January 2026).
44. Dewan, A.; Ziemski, M.; Meylan, H.; Concina, L.; Pouliquen, B. Developing automatic verbatim transcripts for international multilingual meetings: An end-to-end solution. In *Proceedings of the Machine Translation Summit XIX, Vol. 2: Users Track*; Asia-Pacific Association for Machine Translation: Tokyo, Japan, 2023; pp. 183–194. Available online: https://aclanthology.org/2023.mtsummit-users.18/ (accessed on 5 January 2026).
45. Lavecchia, C. Les Triggers Inter-Langues Pour la Traduction Automatique Statistique. Ph.D. Thesis, Université Nancy II, Nancy, France, 2010.
46. Tiedemann, J.; Thottingal, S. OPUS-MT: Building Open Translation Services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, Lisbon, Portugal*; European Association for Machine Translation: Flemish Region, Belgium, 2020; pp. 479–480. Available online: https://aclanthology.org/2020.eamt-1.61/ (accessed on 18 November 2025).