South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com

SSMJ
A Hub of new Knowledge

Online ISSN: 3005-2661

Building a Transformer-Based Neural Machine Translation System for English–Kibajuni Translation: A Low-Resource Deep Learning Approach for Indigenous Language Preservation

Anwar H. Ahmed*
School of Business, Technical University of Mombasa, Kenya
**Corresponding Author*

Wahida M. Bana
School of Business, Technical University of Mombasa, Kenya

Mathew M. Egessa
School of Business, Technical University of Mombasa, Kenya

Collins Kiprotich
School of Business, Technical University of Mombasa, Kenya

Athman L. Omar
AFS – Organisation for Intercultural Education - Kenya

Abstract

Recent progress in artificial intelligence has pushed machine translation to high levels of accuracy for widely resourced languages. Yet for many indigenous and endangered languages, comparable tools remain absent, largely because digitized linguistic data are scarce. Kibajuni, a minimally documented Bantu language spoken along the Kenyan coast, illustrates this gap. Publicly accessible English–Kibajuni machine translation systems are not available, which restricts both everyday digital use and broader language preservation work.This paper reports the design, construction, and assessment of a compact Transformer-based Neural Machine Translation (NMT) system for English–Kibajuni translation. Training relied on a community-produced parallel corpus of roughly 10,000 aligned sentence pairs. A Design Science Research approach guided development of the full translation workflow, beginning with corpus preparation and continuing through Byte Pair Encoding (BPE) tokenization, a custom encoder–decoder Transformer, supervised training in PyTorch, beam-search decoding at inference time, and deployment as a web application. Because data were limited, emphasis was placed on training stability and generalization rather than increasing model size. The system therefore integrated AdamW, OneCycle learning-rate scheduling, dropout, label smoothing, gradient clipping, mixed-precision training, and early stopping driven by validation BLEU. Results indicate that, despite the small dataset, the model learned usable semantic correspondences between English and Kibajuni while remaining computationally light. The final network contains about 6–8 million parameters, occupies roughly 27 MB, and supports real-time translation on modest hardware. In practical terms, the work provides one of the earliest operational English–Kibajuni neural translation platforms. At the methodological level, it offers a reproducible template for developing MT systems for other under-resourced African languages. Taken together, the findings suggest that appropriately scaled Transformer models, paired with subword tokenization and carefully tuned training procedures, can materially advance digital inclusion and language preservation for endangered languages.

Key Words: Artificial Intelligence; Byte Pair Encoding (BPE); Digital Language Preservation; English–Kibajuni Translation; Indigenous Languages; Low-Resource Neural Machine Translation; Natural Language Processing; Transformer Architecture

Doi: 10.61250/ssmj/v1.i4.2

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

1. Introduction

Artificial intelligence (AI) and Natural Language Processing (NLP) have reshaped machine translation (MT), turning it into one of deep learning's most visible successes. Neural Machine Translation (NMT), especially when built on Transformer architectures, has surpassed earlier statistical approaches by learning syntactic and semantic mappings end to end (Bahdanau et al., 2015; Vaswani et al., 2017). Large-scale systems such as Google Translate, DeepL, Meta's No Language Left Behind (NLLB), and MarianMT show strong quality for high-resource language pairs, mainly because they are trained on extremely large parallel corpora, often reaching into the millions or billions of sentence pairs (Costa-jussà et al., 2022; Tiedemann & Thottingal, 2020). For thousands of languages, however, this progress has not translated into practical tools, since digital text and aligned bilingual resources are limited or missing.UNESCO (2021) estimates that close to 40% of the world's roughly 7,000 languages are at risk of disappearing within this century. In Africa, many languages remain poorly represented in digital infrastructures, which reduces their visibility in education, public administration, e-government, and AI-mediated communication (Bird, 2020). The shortage of parallel corpora, standardized writing conventions, annotated datasets, and pretrained models creates a substantial barrier for language technology development (Joshi et al., 2020). The result is persistent digital exclusion for many speech communities even as AI-supported communication expands worldwide.

Kibajuni (Bajuni) is one such language. Spoken by Bajuni communities along the Kenyan and Somali coasts, it is linguistically related to Kiswahili but cannot be treated as a direct substitute for standard Swahili. Kibajuni includes distinct phonological, lexical, and grammatical properties, and its limited written documentation, along with the absence of standard computational resources, has kept it largely outside mainstream NLP research. As a consequence, an accessible English–Kibajuni MT system has not existed, creating a clear need for computational solutions that support practical use, documentation, and preservation.

Transformer models changed sequence modeling by replacing recurrence with multi-head self-attention, improving the handling of long-range dependencies and enabling more parallel computation (Vaswani et al., 2017). This architectural shift underlies much of modern NLP, including BERT (Devlin et al., 2019), GPT-style models, T5 (Raffel et al., 2020), mBART (Liu et al., 2020), and multilingual NMT frameworks such as NLLB (Costa-jussà et al., 2022). Still, these systems typically assume large training corpora and significant computing capacity, assumptions that are rarely met for endangered languages where only a few thousand translated sentences may exist.

Given these constraints, low-resource NMT research has focused on techniques that reduce data requirements or extract more value from limited corpora. Subword tokenization through Byte Pair Encoding (BPE) is widely used to lower vocabulary sparsity and handle previously unseen words (Sennrich et al., 2016). Other strategies include multilingual transfer learning (Aharoni et al., 2019), back-translation (Edunov et al., 2018), unsupervised pretraining (Conneau & Lample, 2019), and multilingual denoising objectives (Liu et al., 2020). While such approaches help for moderately resourced settings, they are harder to apply when suitable pretrained models, large monolingual corpora, or closely related training data are not available, conditions that often describe Kibajuni.

For genuinely low-resource cases, model scale and training discipline become central. Evidence indicates that reducing model capacity to fit the available data, and pairing that design with strong regularization, subword modeling, careful learning-rate scheduling, label smoothing, and early stopping can yield workable performance even on small corpora (Ott et al., 2019; Popel & Bojar, 2018).

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

Over-sized models tend to memorize rather than generalize in such settings, whereas appropriately sized Transformers can retain enough representational power without excessive overfitting.

Beyond technical performance, language technologies can support cultural continuity. Digital preservation programs increasingly treat computational tools as part of language documentation, multilingual education, cultural identity work, and broader participation in digital life (Bird, 2020; UNESCO, 2021). Machine translation, in particular, can improve access to information and encourage language use in educational, governmental, and commercial domains.This study therefore develops and evaluates a custom Transformer-based NMT system translating English into Kibajuni, trained on about 10,000 community-generated parallel sentence pairs. The system is built without relying on pretrained multilingual models, instead following a full pipeline constructed from first principles: corpus preparation, BPE tokenizer creation, encoder–decoder Transformer design, supervised training in PyTorch, beam-search decoding, and deployment through a lightweight web interface. The work contributes in four ways: (1) it introduces one of the earliest end-to-end Transformer English–Kibajuni translation systems targeted at an endangered African language, (2) it shows that a compact Transformer can support practical translation with a modest corpus, (3) it provides a reproducible workflow that can be adapted to other under-documented African languages, and (4) it advances digital language preservation by delivering usable computational infrastructure for Kibajuni.

The rest of the paper is structured as follows. Section 2 surveys research on NMT and low-resource language technologies. Section 3 details the methodology, corpus development, tokenization, and model design. Section 4 describes implementation and experimental configuration. Section 5 reports results and system performance. Section 6 discusses limitations and future work, and Section 7 concludes.

2. Literature Review

2.1 Evolution of Machine Translation

Machine Translation has moved through three broad approaches: rule-based MT (RBMT), statistical MT (SMT), and neural MT (NMT). RBMT relied on handcrafted linguistic rules and bilingual dictionaries, which made development expensive and scaling difficult (Hutchins & Somers, 1992). SMT introduced probabilistic modeling and alignment based on bilingual corpora, improving quality relative to purely rule-driven systems (Brown et al., 1993; Koehn et al., 2003). Phrase-based SMT dominated much of the early 2000s, learning correspondences between phrases from large parallel datasets. However, SMT often struggled with sparsity, brittle phrase alignments, and limited capacity to model long-distance context (Koehn, 2010).

Deep learning shifted MT toward end-to-end sequence modeling. The attention mechanism proposed by Bahdanau et al. (2015) allowed neural translators to focus dynamically on relevant source content. Soon after, Vaswani et al. (2017) introduced the Transformer, removing recurrence and relying on self-attention. The result was both improved accuracy and faster training through parallelization. Most widely used systems, including Google Translate, MarianMT, DeepL, and NLLB, now build on Transformers (Tiedemann & Thottingal, 2020; Costa-jussà et al., 2022).

2.2 Neural Machine Translation and Transformer Architectures

NMT typically frames translation as sequence-to-sequence learning. An encoder converts the source sentence into contextual representations, and a decoder generates the target sequence

(Sutskever et al., 2014). Early neural systems used recurrent networks such as LSTMs, but these architectures were constrained by sequential computation and had difficulty capturing long-range dependencies. Transformers addressed these issues through multi-head self-attention, allowing each token to attend to all others regardless of distance (Vaswani et al., 2017). This mechanism, combined with positional encoding, feed-forward sublayers, residual connections, and normalization, supports stable training and strong representational capacity. Transformer-based models also underpin multilingual pretraining systems such as BERT (Devlin et al., 2019), T5 (Raffel et al., 2020), mBART (Liu et al., 2020), and large multilingual NMT toolkits such as NLLB (Costa-jussà et al., 2022). Yet their reliance on large corpora and substantial compute makes direct application difficult for endangered languages.

2.3 Machine Translation for Low-Resource Languages

Despite their strengths, neural MT models degrade sharply in low-resource contexts. Such languages often lack digitized text, parallel corpora, standardized orthographies, and supporting resources (Joshi et al., 2020), a situation common across many African languages.Proposed remedies include multilingual transfer learning (Aharoni et al., 2019), back-translation (Edunov et al., 2018), and unsupervised approaches grounded in cross-lingual modeling (Conneau & Lample, 2019). These methods, however, tend to assume the availability of pretrained multilingual models or large monolingual data, which may not exist for under-documented languages such as Kibajuni. For this reason, recent work also stresses scaling architectures to match dataset size. Popel and Bojar (2018) argue that oversized models overfit small corpora, while Ott et al. (2019) report gains from careful optimization, regularization, and scheduling under limited data.

2.4 Subword Tokenization in Low-Resource NMT

Vocabulary sparsity is a major obstacle in low-resource MT. Word-level vocabularies lead to many out-of-vocabulary items, while character-level modeling increases sequence length and computational cost. BPE addresses this trade-off by segmenting words into frequent subword units (Sennrich et al., 2016). Subword methods lower vocabulary size and allow unseen words to be represented through combinations of learned segments. This is especially helpful for morphologically rich languages, where inflection and derivation generate many surface forms (Kudo & Richardson, 2018). Modern multilingual systems such as MarianMT and NLLB also depend on subword tokenization (Tiedemann & Thottingal, 2020; Costa-jussà et al., 2022).

2.5 Optimization Strategies for Small Transformer Models

Training Transformers on small datasets requires close attention to overfitting and optimization instability. AdamW separates weight decay from gradient-based updates and often improves convergence relative to standard Adam (Loshchilov & Hutter, 2019). Label smoothing reduces overconfident predictions and can improve generalization (Szegedy et al., 2016). Learning-rate scheduling is also central, and OneCycleLR adjusts the learning rate across training to speed convergence and reduce poor minima (Smith & Topin, 2019). Gradient clipping limits unstable updates, dropout acts as stochastic regularization (Srivastava et al., 2014), and early stopping based on validation metrics protects against memorization when data are scarce (Prechelt, 1998).

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

2.6 Evaluation of Neural Machine Translation Systems

Evaluation remains a core requirement in MT studies. BLEU (Papineni et al., 2002) is still the most common automatic metric, estimating quality by comparing system output against human references using modified n-gram precision. BLEU has documented weaknesses, particularly for morphologically rich or low-resource languages where variation can be high (Callison-Burch et al., 2006). For that reason, recent work often recommends supplementing BLEU with metrics such as chrF, COMET, and structured human evaluation (Popović, 2015). Even so, BLEU continues to provide a widely recognized benchmark.

2.7 Digital Preservation of Indigenous Languages

AI-based tools increasingly contribute to language documentation and cultural preservation. UNESCO (2021) identifies digital technologies as central to preserving endangered languages via documentation, educational use, and community participation. Tools such as speech recognition, OCR, lexical resources, and MT can expand access to indigenous knowledge and support language transmission.

Bird (2020) emphasizes that language technologies should be community-centered, supporting speaker needs rather than treating languages only as objects of documentation. MT systems built for endangered languages can therefore serve both communication and identity maintenance. For Kibajuni, an English–Kibajuni NMT system supports computational research and broader aims linked to sustaining linguistic diversity.

Taken together, the literature shows major advances in Transformer-based MT while also pointing to the continuing challenges faced by truly low-resource languages. Many multilingual systems remain dependent on large datasets and pretrained models, leaving numerous indigenous African languages without effective tools. This study addresses that gap by building a lightweight English–Kibajuni Transformer trained on community-generated parallel text, contributing to both low-resource NMT work and digital language preservation.

3. Materials and Methods

3.1 Research Design

A design science research (DSR) methodology was selected to guide the development, implementation, and evaluation of an English–Kibajuni NMT artefact. DSR fits the aims of the study because the central objective is to create and test a working computational system responding to a practical need, specifically the lack of digital translation tools for Kibajuni (Hevner et al., 2004). Development proceeded iteratively, with evaluation conducted through standard MT metrics.The workflow was organized into six phases: (i) corpus preparation, (ii) tokenizer construction, (iii) Transformer design, (iv) training and optimization, (v) inference and decoding, and (vi) deployment through a web interface.

3.2 Data Collection and Corpus Development

The dataset contained approximately 10,000 English–Kibajuni parallel sentence pairs, collected through community elicitation activities. Data were compiled in Microsoft Excel worksheets covering domains such as everyday communication, education, health, culture, administration, and social interaction.Cleaning steps included removing incomplete alignments, filtering duplicates, trimming

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

whitespace, normalizing formatting, consolidating worksheets into a single bilingual corpus, and checking alignment consistency. The final dataset was stored in CSV format with two columns, English source and Kibajuni target. Summary statistics were produced to track completeness, sentence length distributions, and domain coverage.

3.3 Tokenization Strategy

To control vocabulary sparsity, BPE tokenizers were trained separately for English and Kibajuni using the Hugging Face Tokenizers library. Configuration included a 4,000-token vocabulary, minimum frequency of 2, whitespace pre-tokenization, special tokens ([PAD], [UNK], [BOS], [EOS]), and template-based sequence formatting. Subword modeling allowed representation of unseen words while keeping vocabularies small enough to be adequately trained given corpus size.

3.4 Transformer Architecture

A custom encoder–decoder Transformer was implemented in PyTorch. The model used an embedding dimension of 256, four encoder layers, four decoder layers, four attention heads, a feed-forward dimension of 512, dropout of 0.20, and a maximum sequence length of 64 tokens. Learned positional embeddings were used.

Table 1: Transformer Model Architecture

Component	Configuration
Embedding dimension	256
Encoder layers	4
Decoder layers	4
Multi-head attention	4 heads
Feed-forward dimension	512
Dropout	0.20
Maximum sequence length	64 tokens
Positional encoding	Learned embeddings

The encoder produced contextual representations of English inputs. The decoder generated Kibajuni output using masked self-attention and encoder–decoder cross-attention.

3.5 Model Training

Training followed a supervised setup. Key settings were: PyTorch framework, AdamW optimizer, OneCycleLR scheduler, cross-entropy loss with label smoothing (0.1), batch size 32, up to 80 epochs, early stopping with patience of 10 epochs, gradient clipping at 1.0, mixed-precision FP16 training, a 90:10 train/validation split, and a fixed random seed of 42. Model selection was based on validation BLEU.

3.6 Translation Inference

Inference used beam-search decoding rather than greedy decoding. Parameters were beam width 4, length penalty 0.6, and maximum length 64 tokens. Generation ended when an end-of-sequence token was produced or when the length cap was reached.

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

3.7 Model Evaluation

System performance was measured using BLEU (Papineni et al., 2002). BLEU compared model outputs to reference translations using modified n-gram precision. Validation BLEU was monitored during training, and early stopping was applied if BLEU did not improve for ten consecutive epochs.

3.8 System Deployment

Deployment was implemented as a lightweight web application. Components included a Flask backend REST API, an HTML5 frontend, a Tailwind CSS interface, JavaScript for client interactions, and Docker containerization to support platform-independent execution. Users provide English input and receive Kibajuni translations with low latency, consistent with resource-constrained deployment needs.

4. System Architecture

The English–Kibajuni NMT system was built as an end-to-end pipeline that converts bilingual sentence pairs into an interactive web service. The design joins preprocessing, subword tokenization, Transformer translation, beam-search decoding, and web deployment. Each element was chosen to balance translation quality with computational cost, reflecting low-resource constraints.

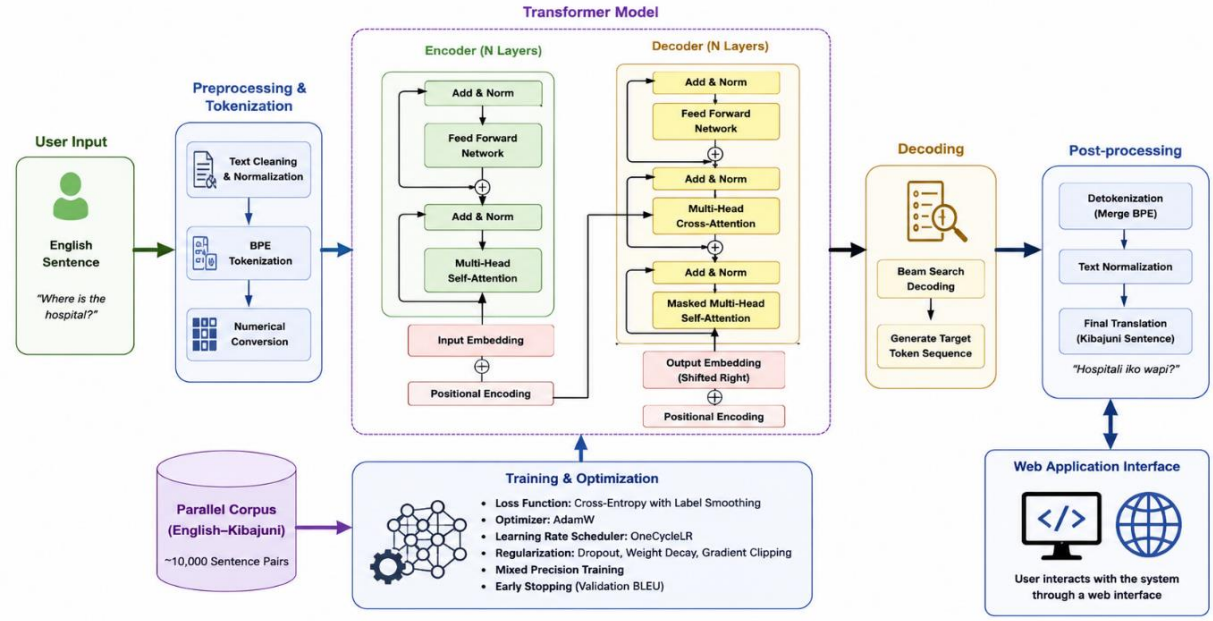

Figure 1. System Architecture of the English–Kibajuni Neural Machine Translation System

Processing begins with corpus preparation. English and Kibajuni sentence pairs collected through community elicitation were cleaned by removing duplicates, incomplete entries, inconsistent formatting, and extraneous whitespace. The resulting aligned corpus was stored in CSV form. For low-resource training in particular, reducing noise is crucial, since each sentence pair has a substantial influence on what the model learns (Koehn, 2010; Sennrich et al., 2016).

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
SSMJ
A Hub of new Knowledge
Online ISSN: 3005-2661

After cleaning, BPE tokenizers were trained separately for the two languages. BPE reduces sparsity by learning frequent subword units, lowering out-of-vocabulary rates and allowing unseen words to be represented through known segments (Sennrich et al., 2016; Kudo & Richardson, 2018). Both tokenizers used a 4,000-token vocabulary and included [PAD], [UNK], [BOS], and [EOS] for padding and sequence generation.

The translation core is a PyTorch-based Transformer encoder–decoder. The encoder maps tokenized English to contextual vectors through four stacked self-attention layers, each combining multi-head attention, residual connections, normalization, and position-wise feed-forward networks. In contrast to recurrent models, attention allows tokens to reference one another directly, supporting long-range dependency modeling and efficient computation (Vaswani et al., 2017).

The decoder contains four corresponding layers and generates Kibajuni output autoregressively. Masked self-attention restricts each step to previously generated tokens, while cross-attention links output generation to encoder representations. Learned positional embeddings preserve order information. Decoder outputs pass through a linear projection and softmax to obtain probabilities over the Kibajuni vocabulary.

At inference time, beam search replaces greedy decoding. Rather than committing to the single best token at each step, beam search maintains multiple hypotheses, expanding the top four candidates at each generation step and selecting the highest-scoring completed sequence using length normalization. A beam width of 4 and a length penalty of 0.6 were chosen to balance fluency improvements against computational expense (Koehn, 2010).

Deployment uses a Flask REST API that loads the model and tokenizers once at startup to reduce response latency. A browser-based interface built with HTML5, Tailwind CSS, and JavaScript sends asynchronous requests to the server. The server performs tokenization, translation, and detokenization before returning Kibajuni output. Docker packaging includes model artifacts and dependencies, supporting reproducible deployment across operating systems. The resulting design shows that a complete Transformer translation environment can be developed for an under-documented language without relying on high-end infrastructure, provided that architectural scale and training strategy are aligned with available data.

5. Experimental Setup

Experiments were designed to test whether a Transformer-based English–Kibajuni NMT system could be trained effectively on a small bilingual corpus. Given Kibajuni's limited digital resources, the experimental focus was on optimization and regularization choices that improve generalization while limiting overfitting. The workflow covered corpus preparation, tokenizer training, model training and validation, decoding, and deployment.The dataset contained roughly 10,000 manually aligned sentence pairs obtained through community elicitation. Before training, the corpus was processed to remove incomplete entries, duplicates, and formatting inconsistencies. Data were split randomly into training and validation subsets using a 90:10 ratio with a fixed seed (42) to support reproducibility. Training data were used for parameter updates, while validation data served only for monitoring generalization and selecting the best checkpoint.

BPE tokenizers were trained independently for English and Kibajuni with 4,000-token vocabularies. This small vocabulary choice aimed to avoid large numbers of under-trained tokens while still covering the core lexical space. Sequence boundary markers were inserted automatically to support supervised training and generation.

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

The Transformer was implemented in PyTorch using a 256-dimensional embedding, four encoder layers, four decoder layers, four attention heads, a 512-dimensional feed-forward block, and dropout at 0.20. Inputs were limited to 64 tokens, covering almost all sentences in the corpus. The model contained about 6–8 million parameters, a deliberate attempt to match capacity to training data rather than adopting large-scale architectures.

Optimization used AdamW with OneCycleLR scheduling. Cross-entropy loss with label smoothing (0.1) was applied, and padding tokens were excluded from the loss. Gradient clipping at 1.0 was used to stabilize updates. Mixed-precision training reduced memory use and improved efficiency on GPU hardware. Training ran for up to 80 epochs but was halted by early stopping if validation BLEU did not improve for ten epochs.

BLEU served as the primary evaluation metric (Papineni et al., 2002). Validation BLEU was computed after each epoch, and the checkpoint with the highest BLEU was retained as the final model. Beam-search decoding with width 4 and length penalty 0.6 was applied at inference time. The trained model was integrated into a Flask web application, enabling near real-time translation from a browser interface, demonstrating feasibility for community-oriented deployment in resource-limited settings.

6. Results and Discussions

6.1 Results

The study aimed to assess whether a carefully scaled Transformer could produce practical English–Kibajuni translations from a small bilingual corpus. Results indicate that the proposed model learned useful semantic mappings between English and Kibajuni even though it was trained on only about 10,000 sentence pairs. Training remained stable under the chosen optimization settings. AdamW, OneCycle scheduling, dropout, label smoothing, gradient clipping, early stopping, and mixed-precision training together supported smooth convergence. Validation BLEU increased during early epochs and later stabilized, suggesting that the model learned translation patterns without severe overfitting. Early stopping prevented extended training after performance plateaued, helping preserve the strongest generalization checkpoint.

BPE played a central role in reducing sparsity. Subword segmentation allowed the system to represent rare or unseen words through combinations of frequent units. Keeping vocabulary size at 4,000 tokens also meant that subword units were trained more thoroughly, limiting excessive fragmentation while retaining adequate coverage.

Qualitative inspection showed that the model produced coherent Kibajuni output for sentence types and lexical content that were well represented in training. Beam search improved fluency relative to greedy decoding by considering multiple competing hypotheses rather than committing early to locally optimal choices. Overall, translations preserved much of the source meaning and produced word ordering consistent with Kibajuni.

In computational terms, the system remained efficient. The trained model occupies about 27 MB, far smaller than many multilingual translation models that often require hundreds of megabytes or more. This compact footprint supports rapid loading and low-latency inference on ordinary hardware, a key consideration for educational institutions and community settings with limited compute resources.

Deployment as a web-based translation platform confirmed the system's operational viability. Users could submit English text through a browser and receive Kibajuni translations in real time,

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

demonstrating that an end-to-end NMT pipeline for an under-documented language can be implemented without reliance on large pretrained multilingual models.

While translation quality is constrained by the small and domain-limited corpus, the results show that meaningful English–Kibajuni neural translation is feasible using a carefully tuned, lightweight Transformer. This supports the broader view that compact, well-regularized architectures can provide practical MT capabilities for other low-resource and endangered languages.

6.2 Discussion

This work examined whether a compact Transformer-based NMT system could deliver usable English–Kibajuni translation despite limited bilingual training data. Findings show that a model trained on approximately 10,000 community-generated sentence pairs can learn systematic correspondences between the two languages and produce translations that are adequate for many everyday contexts. The results align with the view that architectural scaling and training strategy can strongly influence outcomes in low-resource settings, not only dataset size (Popel & Bojar, 2018; Ott et al., 2019).

A key contribution is the demonstration that Transformer-based translation does not always require large multilingual corpora to become operational. Dominant MT systems depend on massive resources that do not exist for Kibajuni (Costa-jussà et al., 2022; Tiedemann & Thottingal, 2020). Here, model capacity was reduced to about 6–8 million parameters, and regularization was strengthened. This balance helped avoid the typical overfitting problems associated with training large Transformers on small datasets, consistent with recommendations from Popel and Bojar (2018).

BPE tokenization was also decisive. Whole-word vocabularies are inefficient in low-resource contexts, especially for languages with rich morphology or non-uniform spelling practices. Subword segmentation improved lexical coverage and made it possible to form unseen words from known units, consistent with earlier evidence (Sennrich et al., 2016; Kudo & Richardson, 2018).

The optimization configuration contributed noticeably to stability and generalization. AdamW, OneCycleLR, dropout, label smoothing, gradient clipping, mixed precision, and early stopping worked together to produce reliable training dynamics and to reduce memorization. Prior research has reported similar benefits for constrained data scenarios (Loshchilov & Hutter, 2019; Smith & Topin, 2019; Szegedy et al., 2016).

From a deployment standpoint, the model's small size, around 27 MB, supports real-time translation without high-end infrastructure. This matters in many African contexts where computing capacity and network connectivity are limited, and where practical language tools should run on modest hardware.

The work also contributes to digital language preservation. UNESCO (2021) notes that global pressures are accelerating language loss, and computational tools can support documentation, education, and use in digital settings. By delivering one of the first end-to-end English–Kibajuni neural translation systems, this study provides a base that can support later language revitalization efforts and encourage Kibajuni use within emerging digital environments.

At the same time, persistent low-resource challenges remain. Performance depends strongly on corpus coverage, and quality declines for structures and domains that are underrepresented in training, a pattern observed in many low-resource NMT studies (Joshi et al., 2020; Costa-jussà et al., 2022). The system is therefore best understood as an enabling tool and a foundation rather than a substitute for human translation.

South Sahara Academic Publishing
South Sahara Multidisciplinary Journal
Volume 4||Issue 1||Page26-39||July||2026|
Email: info@ssapub.com
Online ISSN: 3005-2661

7. Conclusions and Recommendations

7.1 Conclusion

This paper presented the design, implementation, and evaluation of a lightweight Transformer-based NMT system for English–Kibajuni translation. To address the limited computational resources available for Kibajuni, the study developed an end-to-end pipeline that spans corpus preparation, BPE tokenization, custom Transformer modeling, supervised training, beam-search decoding, and web deployment. In contrast to multilingual systems that depend on pretrained models and massive parallel corpora, the system was built from first principles using about 10,000 community-generated sentence pairs.Findings show that meaningful NMT for a genuinely low-resource language is achievable when model capacity is matched to data availability and supported by careful optimization. Subword tokenization, dropout, label smoothing, learning-rate scheduling, gradient clipping, and early stopping helped the model learn useful bilingual mappings while limiting overfitting. The resulting model supports practical translation with modest computational requirements, making it suitable for resource-constrained deployment.

Beyond system performance, the work contributes to digital language preservation by delivering one of the first operational English–Kibajuni neural translation platforms. It also offers a reusable methodological blueprint for other under-documented African languages.

Further advances will depend on expanding the corpus, improving evaluation methods, and exploring transfer learning and augmentation techniques. Still, the present study shows that community-driven data collection, paired with carefully scaled Transformer architectures, can extend modern MT capabilities to highly under-resourced languages, supporting linguistic inclusion and the preservation of cultural heritage.

7.2 Future Work

This study establishes a starting point for further English–Kibajuni NMT research. The most immediate need is corpus expansion, moving beyond 10,000 sentence pairs toward 50,000–100,000 high-quality translations, collected through continued community engagement. Broader domain coverage across education, healthcare, governance, commerce, religion, culture, and daily communication should improve lexical breadth and generalization.

Multilingual transfer learning is another direction. Given Kibajuni's relationship to Kiswahili, experiments that incorporate related Bantu languages, including training with pretrained systems such as mBART or NLLB, might raise quality while reducing the amount of Kibajuni-specific parallel data required (Liu et al., 2020; Costa-jussà et al., 2022).

Data augmentation also remains promising. Back-translation, synthetic parallel generation, active learning, and semi-supervised training could expand effective training data without proportional manual translation effort (Edunov et al., 2018).

Extending the system to bidirectional translation would increase usefulness for education, tourism, healthcare communication, public services, and cultural preservation. Related speech technologies, including speech-to-text and text-to-speech, could support integrated multilingual communication tools.

Evaluation should be strengthened through human assessment with Kibajuni speakers, along with complementary automatic metrics such as chrF and COMET, to better capture adequacy, fluency, and cultural appropriateness.

Finally, deployment research could consider compression, quantization, and distillation so that the model can run on smartphones or offline devices, which is important in areas with unstable connectivity.

References

Aharoni, R., Johnson, M., & Firat, O. (2019). Massively multilingual neural machine translation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 3874–3884. https://doi.org/10.18653/v1/N19-1388

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). https://arxiv.org/abs/1409.0473Bird, S. (2020). Decolonising speech and language technology. Proceedings of the 28th International Conference on Computational Linguistics, 3504–3519. https://doi.org/10.18653/v1/2020.coling-main.313

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2), 263–311.

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 249–256.

Conneau, A., & Lample, G. (2019). Cross-lingual language model pretraining. Advances in Neural Information Processing Systems, 32, 7059–7069.

Costa-jussà, M. R., Hassan, H., Cross, J., Çelebi, O., Schwenk, H., Heafield, K., ... & Barrault, L. (2022). No language left behind: Scaling human-centered machine translation. arXiv Preprint arXiv:2207.04672. https://doi.org/10.48550/arXiv.2207.04672

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 489–500. https://doi.org/10.18653/v1/D18-1045

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Hutchins, W. J., & Somers, H. L. (1992). An introduction to machine translation. Academic Press.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. Proceedings of the 58th Annual Meeting of the

Association for Computational Linguistics, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

Koehn, P. (2010). Statistical machine translation. Cambridge University Press.

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics, 48–54. https://doi.org/10.3115/1073445.1073462

Kudo, T., & Richardson, J. (2018). SentencePiece: A simple and language-independent subword tokenizer and detokenizer for neural text processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 66–71. https://doi.org/10.18653/v1/D18-2012

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., ... & Fan, A. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726–742. https://doi.org/10.1162/tacl_a_00343

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1711.05101

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., ... & Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. Proceedings of NAACL-HLT 2019: Demonstrations, 48–53. https://doi.org/10.18653/v1/N19-4009

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318. https://doi.org/10.3115/1073083.1073135

Popel, M., & Bojar, O. (2018). Training tips for the Transformer model. The Prague Bulletin of Mathematical Linguistics, 110(1), 43–70. https://doi.org/10.2478/pralin-2018-0002Popović, M. (2015). chrF: Character n-gram F-score for automatic MT evaluation. Proceedings of the Tenth Workshop on Statistical Machine Translation, 392–395. https://doi.org/10.18653/v1/W15-3049

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 1715–1725. https://doi.org/10.18653/v1/P16-1162

Smith, L. N., & Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. Proceedings of SPIE, 11006, Article 1100612. https://doi.org/10.1117/12.2520589

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(56), 1929–1958.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in Neural Information Processing Systems, 27, 3104–3112.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT , Building open translation services for the World. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 479–480.UNESCO. (2021). World report on languages: Towards a global assessment framework for language vitality. UNESCO Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008