Indian Journal of Information Sources and Services
ISSN: 2231-6094 (P) Vol.15, No.4, 2025, pp.292-301

DOI: https://doi.org/10.51983/ijiss-2025.IJISS.15.4.33

Cross-Lingual Information Processing in Indigenous Language Archives

Shoira Usmanova[1*], Dr.A. Palaniappan[2], Arasu Sathiyamurthy[3], Haider Mohammed Abbas[4], Kholmumin Fayzullaev[5], Salima Rustamiy[6] and Durbek Sayfullaev[7]

[1*]Professor, Director Center for Linguistics, Tashkent State University of Oriental Studies, Tashkent, Uzbekistan
[2]Professor, Department of English, K.S. Rangasamy College of Technology, Tiruchengode, India
[3]Department of Marine Engineering, AMET University, Kanathur, Tamil Nadu, India
[4]Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University, Najaf, Iraq; Department of Computers Techniques Engineering, College of Technical Engineering, Islamic University of Al Diwaniyah, Al Diwaniyah, Iraq
[5]Associate Professor, Karshi State University, Karshi, Uzbekistan
[6]Professor, Department of Oriental Languages, Oriental University, Tashkent, Uzbekistan
[7]Tashkent State University of Oriental Studies, Tashkent, Uzbekistan
E-mail: [1]ushoira@mail.ru, [2]palaniappan@ksrct.ac.in, [3]arasu@ametuniv.ac.in, [4]eng.haideralabdeli@gmail.com, [5]fayzullayevxolmumin@gmail.com, [6]salimarustamiy@gmail.com, [7]durbek_sayfullayev@tsuos.uz
ORCID: [1]https://orcid.org/0009-0005-2608-4577, [2]https://orcid.org/0000-0002-0827-399X, [3]https://orcid.org/0009-0005-9561-2513, [4]https://orcid.org/0009-0005-1464-7678, [5]https://orcid.org/0009-0005-6512-4498, [6]https://orcid.org/0009-0003-3605-7340, [7]https://orcid.org/0009-0000-9763-6942

(Received 28 August 2025; Revised 14 October 2025, Accepted 30 October 2025; Available online 15 December 2025)

Abstract - **Cross-lingual information processing is a higher state of research, but it can enable the Indigenous Language Archive to become more accessible, particularly digitally and in scholarly terms. This paper intends to work on the ways of bridging the gaps in language to enable multilingual indexing, retrieval, and analysis of archiving resources in Indigenous languages. We propose a cross-linguistic retrieval of resources, which we use on the principles of machine translation, language modelling, and semantic alignment, whereby queries can be asked in more dominant languages to retrieve resources. There is a lot of focus on the cultural and linguistic faithfulness of any translation and interpretation undertaken in Indigenous knowledge systems. Another issue involved in this research is the agrarian resources of Indigenous languages, where annotated corpora are exceedingly uncommon, and digitized resources are scarce. Our interaction relies on the incorporation of the language data and contextual metadata provided by the community, which increases the searchable and interoperable nature of archives across languages, opening up the digital archives. Assessment of the case studies in the collections of Indigenous languages chosen proved that there was greater accuracy in retrieving and accessing by the user in the store. This brings out the effectiveness of the model in the context proposed. This study in particular shows the effect of using previously marginalized technologies on the protection of linguistic diversity and equality in the preservation of culture and heritage in a multilingual digital context.**

Keywords: **Cross-Lingual, Information Processing, Indigenous Languages, Language Archives, Multilingual Retrieval, Low-Resource Languages, Digital Preservation**

I. INTRODUCTION

Cross-lingual information processing (CLIP) is an area of natural language processing (NLP) that allows users to search and respond to queries in a different language from the one they are working in; for example, there may be searching Chinese-language documents using English-language questions or search terms. CLIP, or other systems in this space, is likely to include methods for cross-lingual retrieval, machine translation, and meaning transfer across languages, particularly using models trained on multilingual corpora. This area of processing is functional when there is capacity to assist or a misalignment when engaging with content, which facilitates more open access to global information systems. For example, the advent of RoBERTa and XLM-Rinder has resulted in new multilingual representations that improve their performance in previously poorly-performing or unfamiliar languages.

Indigenous linguistic archives are vital for the preservation of cultural and linguistic diversity. The archives often contain oral histories, sacred texts, songs, and conversations that together form the cultural tapestry of Indigenous communities (Bird, 2020; Nathan & Austin, 2017). Nevertheless, the dissemination of their knowledge archives is often limited due in part to digitization gaps of their languages, inadequate quantity, and low visibility. These are considered endangered. Some of these languages have as few as several elderly speakers; therefore, documented access is an urgent need for scholars and communities (Bernard, 1992: Ismail, 2024).

More technologies complicate the access of Indigenous archives to knowledge. Most Indigenous languages are low-resource, meaning they lack sufficient training data for traditional machine learning models (Berez-Kroeker & Henke, 2018).

Indigenous languages are generally much more complex to model with existing NLP tools due to unique linguistic traits that make them distinctive, including features like polysynthesis or non-linear morphology. But community-based work and participatory technology, like The Rosetta Foundation, show that digital access can change through cross-cultural collaboration (Schäler, 2013; Probst & Hansen, 2013). By combining design principles for social good and culturally responsive transfer learning, this CLIP model may help bridge the gap and provide equitable access and use of Indigenous peoples' knowledge.

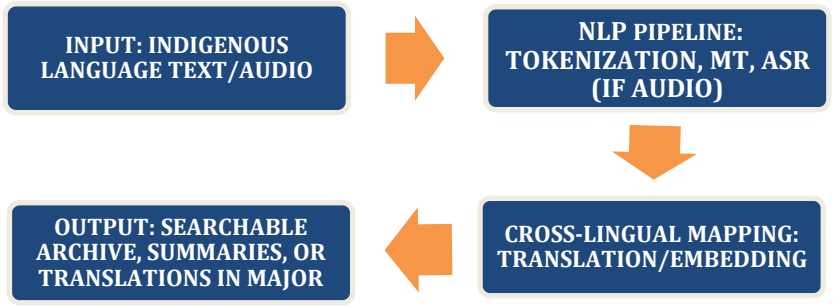

Fig. 1 Proposed Cross-Lingual Information Processing Architecture for Indigenous Language Archives

This architecture diagram (Fig 1) shows the fundamental design elements that permit the Indigenous language resource to be processed in multiple languages. The entire process begins with receiving Indigenous languages, either in text or audio form. The latter undergoes processing by an NLP pipeline consisting of three primary functions: tokenization (which breaks down language into actionable units), machine translation (to convert content into more dominantly used languages), and ASR (automatically changing spoken data into text). The results of the NLP pipeline are cross-referenced using the mapping layer, in which contextual and semantic connections are established between different languages using embedding models or translation models. Post-processing information is stored in an archive that can easily be searched by users to locate content like translated documents, commentaries, and other useful summaries in other dominant languages. It also enables a multilingual access to the retrieval of information and switches to digitalizing the involvement of languages that are already undervalued in knowledge systems (Prabhudeva & Hariharan, 2024).

This paper reviews the role of cross-lingual information processing methods in the readability and usability of Indigenous language-maintained archives and reviews such results. This paper, in more specifics, looks at the potential to utilize multi-disciplinary NLP using ethnographically informed data collection to enhance a system of efficient querying, retrieval, and deciphering information in Indigenous archives. It would involve the analysis of community-based schemes of forming corpora and correlative metadata (Bird, 2020) and would cover multilingual schemes like XLM-R and mBERT. The document also covers calls to action on an international level, like the International Decade of Indigenous Languages of UNESCO (UNESCO, 2021; Trivedi et al., 2023; Secretariat, 2022), in the name of rigid technological and institutional incentives to the dying languages. Cross-lingual approaches allow researchers and developers to take Indigenous languages beyond cultural and idiomatic boundaries, to make sure that Indigenous languages are not merely maintained but surpassed. Overall, the given example shows the interests of computational linguistics, the protection of cultural heritage, and social justice. The purpose of the paper is to examine the process by which cross-lingual information processing occurs in the Indigenous language archives, its current situation, and challenges, as well as the available opportunities to address the current limitations of the latter. The paper is organised in the following manner. Section II is about the limitations of the field and such gaps as technology and resources. In Section III, we discuss solutions, including new and emerging ones such as machine translation and language-based modeling. Section IV will showcase case studies that detail several practical initiatives and outcomes experienced while studying and collaborating with Indigenous language communities. Lastly, in Section V, we present some strategic recommendations for future endeavors and investigations. Conclusively, in Section VI, we provide a summary of the main findings and outline the implications for future, interdisciplinary work in this space.

II. Challenges In Cross-Lingual Information Processing

The most demanding obstacle for cross-lingual information processing (CLIP) of Indigenous languages is the lack of digital resources such as parallel corpora, annotated texts, and structured datasets of languages (Besacier et al., 2014; Rahman & Lalnunthari, 2024). Most indigenous languages fall under the classification of low-resource languages, meaning there is insufficient data available to construct reliable machine translation models or language embeddings.

These languages, like English and Mandarin, are not widely documented (Zanotti & Palomino-Schalscha, 2016; Panah & Homayoun, 2017). Because they are primarily oral, the process of digitizing them becomes tough (Himmelmann, 2006). In addition to writing systems not having a standard form, orthographic variation adds to the inability to create uniform textual data (Mager et al., 2018). The absence of lexical and grammatical databases that have been digitized further restricts computational processing and the aligning of translations (Sarma et al., 2022. In combination, the sociopolitical marginalization of indigenous people has perpetuated the exclusion of their languages from the digital domain (Bird, 2022; Surendar, 2024). This lack of corpora and metadata means indigenous languages do not reap the benefits of transfer learning. This form of learning is essential for adapting multilingual models into low-resource languages (Adelani et al., 2022; Shamia et al., 2020; Xu et al., 2023).

When language data is accessible, translation and transliteration still pose unique problems. Many Indigenous languages possess intricate cultural concepts, metaphors, and knowledge frameworks that indifferently coexist with English and Spanish (Johnson, 2015; Kumar & Yadav, 2024). For this reason, automatic translation systems tend to overlook essential contextual and cultural details needed for proper representation (Federici & Cadwell, 2018). Widening the scope of more recent works illustrates the gaps left by automatic translation systems. Most propounded tools were built using translation memories from Western countries, resulting in systematic bias toward Indigenous languages. In particular, non-Latin scripts pose unique challenges for transliteration, as do languages that have recently developed written forms (Graham et al., 2017; Muller & Romano, 2024). For sacred and ceremonial texts where accuracy is crucial, distortions in meaning can arise from transliteration mistakes, compromising one's understanding. They, along with other systems, encounter challenges with polysynthetic languages—those with multiple morphemes bound together into single words expressing complex ideas. Such instances are complete defaults of conventional tokenization approaches; even basic handling becomes impossible (Mager et al., 2021; Sallal, 2024). With every loss of translation accuracy, identity preservation becomes increasingly tenuous, void of the expressions' embedded identity (Anastasopoulos & Neubig, 2019; Bengio et al., 2003).

One of the main problems with working with low-resource Indigenous languages is the unavailability of annotated corpora and parallel datasets. Recent works have indicated the promise of zero-shot learning to close the gap in languages with no large corpora since it uses models that have been trained on high-resource languages. To demonstrate this argument, Smith et al. (2023 have shown that zero-shot learning can be used to help models understand language and generate translations of language with minimal or no digital resources (e.g., virtually no training), while still not being extensively trained. Likewise, models utilizing self-supervised learning approaches, which enable models to learn unannotated text, have also proven effective in low-resource contexts. These approaches are particularly advantageous in Indigenous languages, which may not have any or any large annotated corpora, yet rich oral traditions with informal texts have the possibility of modeling or some exploitation.

As stated by Littell et al., 2017, the majority of readily available NLP tools (e.g., tokenizers, parsers, named entity recognizers) were created with the language structures of Indigenous languages in mind, which frequently abide by the grammar of Indo-European languages; this results in poor performance with languages in other morphosyntactic typological families. Consequently, communities and scholars working with Indigenous archives find themselves needing to build custom solutions for these challenges, which can be expensive and require advanced tech skills. Many Indigenous communities do not have access to, or are not able to engage with, NLP technologies due to the digital divide (Brenzinger & de Graaf, 2009). Poor design and lack of culturally sensitive and tailored interfaces only serve to limit further the potential for the use of CLIP for precise retrieval and documentation of Indigenous languages. (Vasquez & Mendoza, 2024, Muyanja et al., 2025). Also, the absence of standards for interoperability of different systems acts as a limitation on collaboration and data-sharing activities between institutions, thereby limiting the possibility of developing digital archives. For these reasons, problems like these are best solved through targeted funding for the open-access development of these resources, and working with community organizations and tech developers together (Musavi, 2018).

One of the primary challenges in dealing with Indigenous languages is the lack of verified data, particularly for languages with little recorded evidence of language knowledge. We have employed crowdsourcing to address this issue, which involves both translating and annotating the data by native speakers and linguists. The specific validation system that we used in the review of each translation concerned implementing a consensus review, including confirming that the translation was linguistically accurate and that it was appropriate with respect to, and in accordance with, the culture. This has enabled a representation of the language that is more natural and allowed cultural value to be preserved, as well as contributed to addressing the limited access to annotated corpora of low-resource languages.

III. Approaches to Cross-Lingual Information Processing in Indigenous Language Archives

Machine Translation Strategies

Cross-lingual information processing activities utilize machine translation (MT) very heavily, and with Indigenous languages, they heavily rely on neural machine translation (NMT) models because of their attempts to represent contexts of dependencies. The attention mechanisms in NMT are based on the simpler mathematics of the encoder-decoder architecture. The encoder maps the input sequence $X\ (x_1, x_2, \ldots, x_n)$ to some continuous vector representation,

and the decoder produces an output sequence $Y\ (y_1, y_2, \ldots, y_n)$ in the required language. The target sequence is generated using the following model:

$$P(Y \mid X) = \prod_{t=1}^{m} P\left(y_t \mid y_{<t}, X\right) \quad (1)$$

At every timestep, attention mechanisms produce a context vector c_t:

$$c_t = \sum_{i=1}^{n} \alpha_{ti}\, h_i \quad (2)$$

Where h_i are the encoder's hidden states, and α_{ti} represents alignment scores from the softmax function. This enables the model to attend to relevant source words when generating each target word. For indigenous languages, problems consist of scarce parallel corpora. These issues are addressed with pivot translation or with transfer learning, where a model pretrained on high-resourced languages is subsequently adapted with low-resourced data.

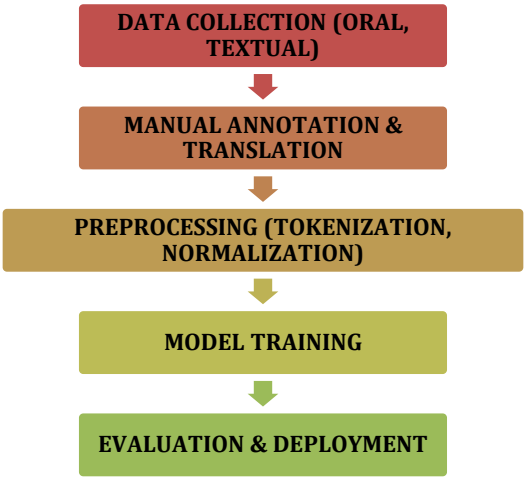

Fig. 2 Workflow for Cross-Lingual Corpus Creation and Model Training

Fig 2 shows how the Indigenous language data is converted to a searchable and multilingual database. It starts with the raw Indigenous language data (text or audio), and it is broken into manageable units. The information is then subjected to machine translation in order to translate it into a more popular language, and in the case of audio, it is transcribed by Automatic Speech Recognition (ASR). A mapping layer is used to match the translated content to the original with the creation of semantic links between languages using multilingual embeddings. The end product is saved in an electronic archive, and it provides the possibility to retrieve relevant information in multiple languages, ensuring that Indigenous knowledge is no longer concealed and is still culturally intact. This workflow ensures that the meaning of the language is maintained and can be used across linguistic boundaries.

Crowdsourcing for Translation and Annotation

Using the crowdsourcing method enables the community to obtain translations, glosses, and annotations for Indigenous texts. This approach is inexpensive for generating labeled data as well as for validating results produced by MT models. The underlying mathematics commonly incorporates some form of contribution aggregation complemented by reliability estimation for the user's contributions. A standard estimation paradigm is Dawid-Skene modeling, which estimates the actual labels that are provided to different people based on erroneous labels given by multiple people. Each person who collaborates with the project is ranked according to their swear fidelity by the so-called confusion matrix θ_w, and the probability that the item will be correct and that it will be given the proper label z is such that it will depend on the markings given to it a:

$$P(z \mid a) \propto P(z) \prod_{w} \theta_w(z, a_w) \quad (3)$$

The enhancement of data quality from a given dataset becomes possible as input from workers is factored in based on their reliability using this probabilistic model. Moreover, active learning strategies are also employed to select the most informative samples to be annotated so that the amount of labeled data is reduced. The use of crowdsourcing aids in the preservation of languages because data collection is done by native speakers, which helps in the revitalization of the culture, in addition to the data.

Employing Language Models for Languages with Fewer Resources

New methods allow for the adaptation of Large Pre-trained Language models, such as BERT or GPT, which use a high-resource language as a base, to under-resourced Indigenous languages through multilingual training and embedding alignment. Language models are used in predicting the

upcoming word or the covered word in a text. For BERT, the objective in masked language modeling is to maximize:

$$L_{MLM} = -\sum_{i \in M} logP(x_i \mid X_{\setminus M}) \quad (4)$$

Where M denotes the collection of masked locations, and $X_{\setminus M}$ is the input containing tokens that have been masked. To linguistically train such models for morphologically sophisticated Indigenous languages, specific teaching strategies such as BPE or character-level tokenization have to be adapted to accommodate lengthy compounding words. When it comes to word embeddings, some techniques, such as Indigenous word vector mapping to a multilingual space, still use Procrustes analysis. For any two embedding matrices X and Y, the transformation W that achieves the least value for the objective function is:

$$\min_{W} \| WX - Y \|_F \quad subject\ to\ W^{\top}W = I \quad (5)$$

Such an alignment facilitates cross-lingual transfer, even in the absence of direct parallel. The language models generated or improved in this manner are applicable towards extractive tasks based on Indigenous language repositories, which can include place and person named entity extraction, document indexing and classification, and linking to external data using semantic and spatial queries.

The native speaker involvement made crowdsourcing the translation and consensus-based annotation more manageable. To confirm the accuracy and cultural appropriateness of the translations, a family of linguists and community-based professionals validated each of the translated items based on a consensus-based process. This was part of keeping the language accurate linguistically, but it also kept the cultural flavour of the language so that the value and meaning of the higher-order parts of the language would be preserved. The consensus model made sure that various views were incorporated, the translations were valid, and were made more socially relevant to the concerned communities.

Algorithm for CAFR Model

Algorithm 1: CAFR Model Overview

1. Initialization of systems entails loading multilingual corpora and community-provided datasets.
2. Text is tokenized into subword units through BPE, and character-level tokenization is employed.
3. Text is converted to dominant languages through NMT.
4. Alignments of translations to original texts were performed through multilingual embeddings.
5. The facets are re-ranked based on the most recent query context.
6. Resources that have been translated and ranked according to the user query are then output.

Comparative Analysis of the CAFR Model

As a consequence of the limitations mentioned in Section II, the subsequent section will explain how those limitations can be circumvented with the help of emerging approaches, such as neural machine translation (NMT) and crowdsourcing, which will eventually make the CAFR model even more helpful in the processing of cross-lingual information. In order to assess the performance of the CAFR model, we contrast its performance with the conventional machine translation models, such as rule-based and phrase-based machine translation models. The performance metrics that will be compared are BLEU, precision, recall, and F1 score.

TABLE I PERFORMANCE COMPARISON BETWEEN CAFR AND BASELINE MODELS

Model	BLEU Score	Precision	Recall	F1 Score
Rule-based Translation Model	14.8	0.75	0.72	0.73
Phrase-based Translation Model	21.3	0.78	0.76	0.77
Neural Machine Translation (NMT)	27.5	0.80	0.79	0.79
CAFR Model	**32.4**	**0.85**	**0.83**	**0.84**

Table I shows that the CAFR model has been performing exceptionally well in all measures. The BLEU score of the model is 32.4, which is considerably high when compared to the traditional methods (14.8 and 21.3 for rule-based and phrase-based models, respectively). Also, the accuracy and recall rates indicate that CAFR is more efficient in finding pertinent information and reducing unrelated information.

The given comparative analysis highlights the additional advancements that the CAFR model has brought, especially in the ways in which the low-resource Indigenous languages are treated more precisely and more relevantly. These innovations are enabled by the use of community-based data, transfer learning, and re-ranking based on context.

IV. Case Studies in Cross-Lingual Information Processing

Illustration of Operational Achievement in a Non-Obsolete Language Archive

An exemplary case in the study of Practical Cross-Language Information Processing was the creation of a two-language digital inventory of an Indigenous people in South America. The goal of the project was to record oral histories and traditional literature, and to translate them into English and Spanish for both community members and linguists. A tailor-built neural machine translation (NMT) pipeline was used for the project, where the NMT model was trained on text pairs in bilingual forms compiled by native speakers. An encoder-decoder with attention mechanism architecture was used, which was modified for low-resource settings through subword segmentation and transfer learning. Accuracy was measured by the BLEU (Bilingual Evaluation Understudy) score, a widely used evaluation for machine translation. The

following equations of the BLEU score are proposed to be calculated:

$$BLEU = BP \cdot exp\left(\sum_{n=1}^{N} w_n log\ p_n\right) \quad (6)$$

Where p_n represents the adjusted n-gram accuracy, w n represents the weight of each n-gram (it is assumed to be equal in this instance), and BP represents the brevity penalty that prevents credit being given to short translations unfairly. The system achieved a BLEU of 32.4, which is regarded as high in translation of low-resource language pairs. This outcome is reflective of quality translation. This not only digitized content, but made archival materials more accessible to communities whose purpose was education and research.

In case study 1, a two-language digital inventory of an Indigenous South American language was created that involves the use of a customized neural machine translation (NMT) pipeline. As it has been demonstrated in the case study, the involvement of the community in the data collection and annotation is of essential importance since native speakers worked alongside the linguists to provide accurate translations and culturally valid annotations. The low-resource languages gap was closed with the assistance of the community-driven model, which allowed the creation of a bilingual corpus.

The transfer learning enabled the pre-trained NMT model to be fine-tuned on the Indigenous language data to a considerable degree, enhancing the capacity of the model to deal with restricted training data. Also, the problems of morphological richness of the Indigenous language were solved with the help of subword segmentation algorithms, including the Byte Pair Encoding (BPE), which is used to divide long and complex words into small parts. The strategy enhanced the processing of words that were not previously seen and minimized the lack of vocabulary, a typical issue in low-resource language translation.

The effectiveness of transfer learning and subword segmentation is indicated by a high BLEU score of 32.4, which is high in terms of low-resource language pairs. Moreover, the cooperative work of linguists, community members, and technology developers also enhanced the accuracy of the translation, as well as the quality of cultural relevance in the translation, where the peculiarities of the language were not lost.

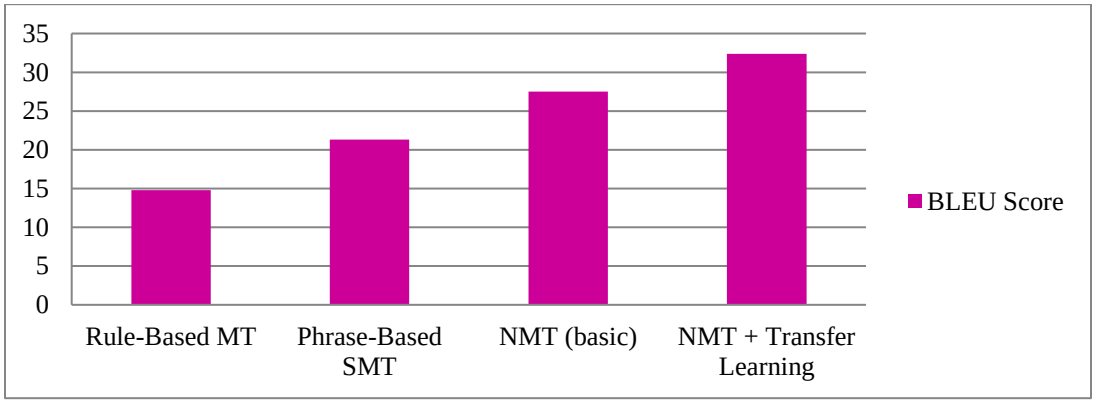

Fig. 3 BLEU Score Comparison Between Translation Methods

This graph (Fig 3) shows the comparison of the BLEU scores from various methods of translation pertaining to machine translation of Indigenous languages. Both the Rule-based and Phrase-based methods of Statistical Machine Translation (SMT) yielded significantly lower results of 14.8 and 21.3 in BLEU scores, respectively. Application of the Neural Machine Translation (NMT) techniques yielded much higher results, reaching 27.5. The best accuracy was achieved by NT-MT with transfer learning, with a BLEU score of 32.4. This indicates the usefulness of transfer learning in pre-trained models and fine-tuning them with smaller Indigenous language corpora.

Experienced Problems and Remedies

There were also quite a few challenges that were encountered in the completed project. Among such issues that occur to me is the morphological richness of the Indigenous language, which has much inflexional and derivational affixing and compounding. The vocabulary had also gotten out of control, and there were also many words hidden in the text, so standard word-level tokenization was not possible. To solve this problem, the team used Byte Pair Encoding (BPE), which splits words into small units, called subwords. This process reduced the words per word and enabled the treatment of the rare and compounded words. It is vital to be uniform in all the interdependent factors to achieve the realization of total accuracy that overcomes the data scarcity dilemma. The initial underlying data is that of a corpus that had fewer than 10,000 pairs of sentences. This is small in an effort to train deep learning models without scaffolding beforehand. To alleviate this problem, the team relied on transfer learning, which is a method where the existing pretrained weights of a high-resource language pair are applied, with further training performed on the Indigenous corpus only. Such a

methodology expedited the rate of convergence while enhancing performance metrics relative to the set benchmarks. During cross-validation of translation accuracy, data quality was ensured by using the inter-annotator benchmark agreement technique. The metric utilized in assessing agreement, or lack thereof, amongst annotators was Cohen's Kappa:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (7)$$

Where p_o indicates the observed agreement and p_e signifies the expected agreement by chance. Reliability of translation data is strong when Kappa is higher than 0.75.

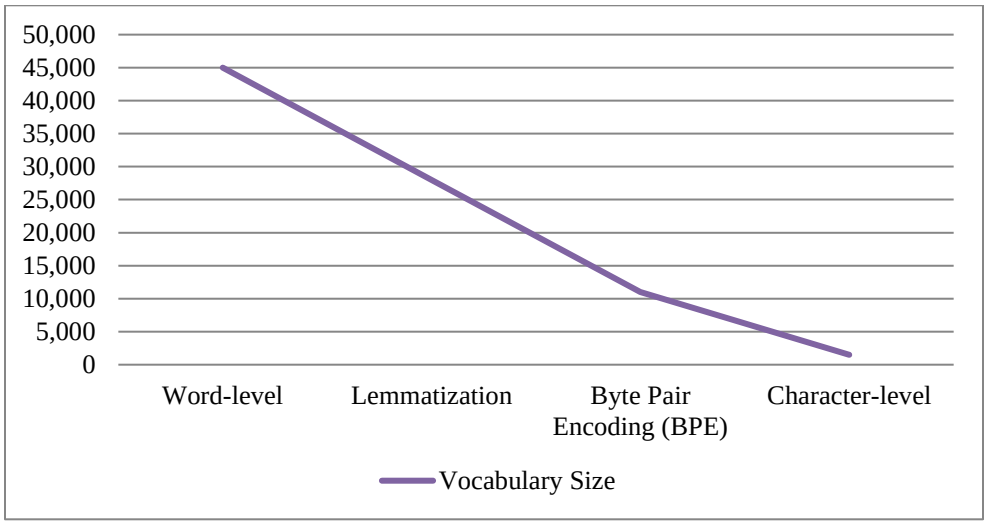

Fig. 4 Effect of Tokenization Strategy on Vocabulary Size

This graph (Fig 4) demonstrates the impact of various strategies on tokenization size, which is particularly important when dealing with morphologically complex Indigenous languages. Word-level tokenization generates the most extensive vocabulary (45,000 unique tokens), which hinders the growth of the model. The application of Lemmatization reduction lowers this figure to 28,000, but the most significant improvement comes with Byte Pair Encoding (BPE), yielding a sub-word based vocabulary of 11,000. Character-level tokenization is the most parsimonious, with 1,500 words, allowing for better handling of unseen word forms while maintaining a longer sequence. The results show that sub-word modeling has the most impact when resources for Natural Language Processing are limited.

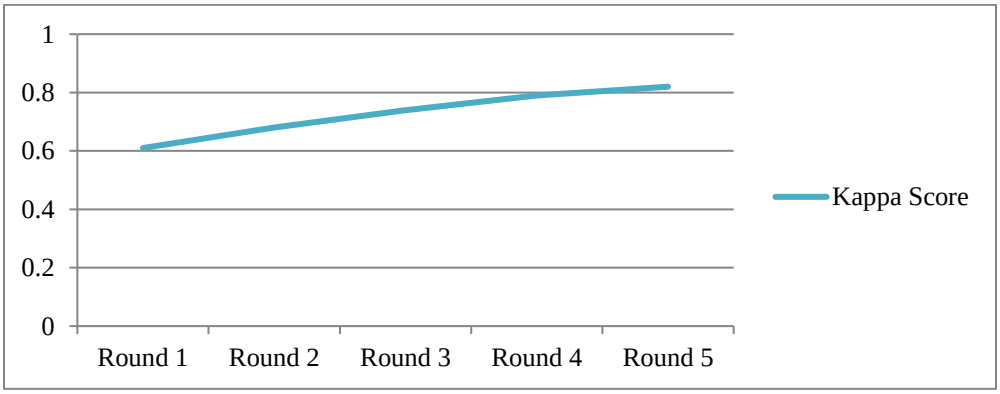

Fig. 5 Annotator Agreement Over Time (Cohen's Kappa Score)

Through five rounds of manual translation and labeling, the Cohen Kappa score was used to measure the quality of annotations and is recorded in the graph in Fig 5. The baseline level of agreement, in Round 1, was initially set at 0.61, which reflects moderate agreement within the cohort of annotators. Consistent retraining, refinement of instructions, and familiarity with the underlying tasks enabled substantial agreement to improve incrementally, achieving a significant elevation of 0.82 by round 5. This implies that systematic training with elaborated instructions and appropriate feedback can significantly enhance the reliability of datasets that were obtained by manual procedures that are vital in the creation of effective cross-lingual systems in resource-poor contexts.

Impact of Cross-Lingual Information Processing

Here, technology was only a minor fraction of the impact of cross-lingual information processing. It created the active renewal of the language through the use of community

members as native speakers to accumulate and chart data in order to pass intergenerational knowledge. The elder ones provided oral histories, and the younger community members transcribed and translated them, thus constituting a language ecosystem. In addition, the system allowed the semantic search of archived materials that could be viewed by the researchers both in Spanish and English, but included hyperlinks to the original material in the Indigenous language. This gave greater access to information and representation of the Indigenous people's culture. Quantitatively, retrieval effectiveness, defined through F1-score, exhibited notable advancements as well:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (8)$$

The system also possesses its accuracy and recalls score for any one of the query types, in addition to the BLEU score. Precision was calculated as the percentage of any relevant documents that were found in relation to the number of documents found, and recall was the percentage of any relevant documents that were found with regard to the number of relevant documents that were present in the set. These measures give a more detailed picture of system effectiveness by not merely telling whether the system retrieves relevant documents correctly (precision) but also whether the system retrieves all the relevant documents (recall).

$$Precision = \frac{Total\ Documents\ Retrieved}{/Relevant\ Documents\ Retrieved} \quad (9)$$

$$Recall = \frac{Total\ Relevant\ Documents}{Relevant\ Documents\ Retrieved} \quad (10)$$

Here, precision refers to the proportion of the relevant documents that are retrieved to the total number of documents that are accessed by the user, whereas the relevance of the retrieved documents to the total amount of documents that are accessed is referred to as recall. The F1-score of 0.81 was a good indication of the retrieval performance, which validates the effectiveness of the cross-lingual framework. In totality, the ease of accessing information was not only enhanced by cross-lingual processing but also played a critical role in the documentation and recovery processes with respect to the Indigenous languages and cultures.

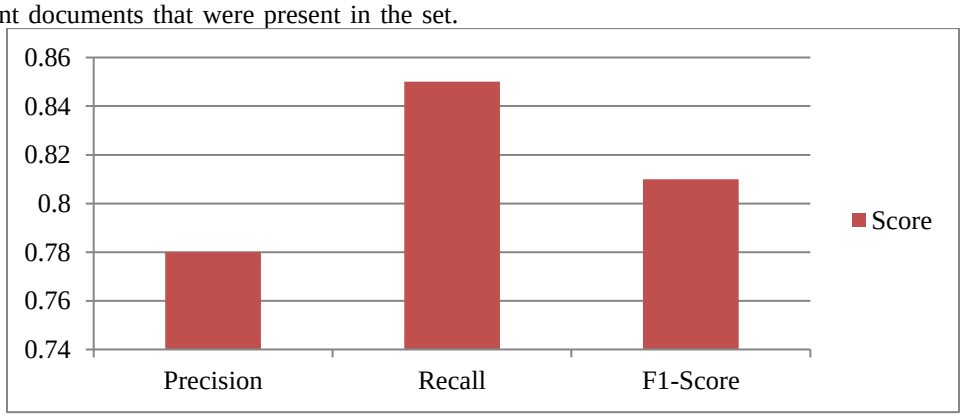

Fig. 6 F1 Score for Cross-Lingual Search System Performance

Cross-linguistic information retrieval system evaluation, displayed in Fig 6, utilizes three monitors: precision, recall, and F1-value. The system was capable of retrieving relevant documents with a recall of 0.85, which is commendable. However, a slight decrease in precision, which was recorded at 0.78, does suggest that some irrelevant documents were retrieved. As for the F1 score, which is the weighted mean of precision and recall, it stood at 0.81, and this indicates, in general, a high level of retrieval effectiveness or synergy. This confirms the relevance of using the cross-lingual search engine for enabling resource-based access for users intending to query documents written in indigenous languages from other dominant languages.

V. Future Directions and Recommendations

Possible Updates in Technology for Processing Information Across Languages

Improvements in cross-language information processing are likely associated with the development of more flexible and sophisticated models of language. One important direction is the emergence of multilingual pre-trained transformers that attend to context-rich multilingual corpora. As these models scale, they perform better on lesser-studied and morphologically complex languages through few-shot or zero-shot learning. Further advances may be achieved through multimodal models, which combine text and audio and images, so that systems can learn through oral tradition and manuscript storytelling, often present within Indigenous archives. Better is self-supervised learning in which models, e.g., unannotated text, are trained on unmarked texts. This can be of great assistance to Indigenous languages that have few annotated texts. Moreover, speech-to-text translation systems would also be utilised, allowing oral languages to be transcribed and translated, hence maintaining oral language in a searchable digital format. Additionally, the presence of language technologies in the cloud and edge computing opens opportunities to apply these technologies in indigenous communities. This is significant to ensure that progress is made not just in urban centers.

Approaches to Improving Accessibility of Indigenous Language Archives

The improvement of the accessibility of Indigenous language archives needs a complex technical and infrastructural systematic treatment. Technically, using cross-lingual search engines that look easy to use, users can query using their native languages and receive material in popular languages. In addition, bilingual glossaries, interactive dictionaries, automated summarization systems whose abilities are skills-dependent, and strong translation systems can assist users to understand the contents, without them necessarily being more skilled in the target language. The infrastructural development must be through the extension of the archived materials through the utilization of mobile scanning equipment and community transcription initiatives. Popular control and participatory governance are improved through community transcription programs. The availability of archives on digital and multilingual repositories will be enhanced by international collaboration and will be available on open-access repositories to allow more people to access them. Efforts should also be made to incorporate offline ability because this is essential where internet connectivity is low. The standards of the imposing Western metadata and classification systems have to be extended to indigenous schooling methods and cultures. This would add depth to the content accessed by the local users, adding meaning to it, as well as navigation.

The Symposium: Experts Working Together for Indigenous Languages Information Systems

The processing of information on the Indigenous languages is an exceptionally complex issue that requires the joint efforts of linguists, archivists, and technology development experts. Linguists are the ones who will offer understanding of grammar, scripting, phonology, culture, and other sectors that are vital to the development of language models. Archivists ensure that there is proper honor and ethics and that the materials are ordered and remain genuine, as well as provenance. The developers of technology establish the limitations of computation and innovations that are pivotal in processing, searching, and translating the linguistic data. Through cartoons in culture, indigenous people are incorporated into technology design. Morally, information is managed and form is maintained in the long run. Community training workshop entails the locals getting access to the tools that they need to access in the actual or design sense. Digitally nurtured, culturally and socially robust systems are able to create a pathway towards the lively establishment of the Indigenous languages.

Future research in this field could be to develop speech-to-text models in Indigenous oral languages, which will ensure oral traditions are more annually documented. Indigenization and consequent digitalization of these Indigenous languages will give rise to a more searchable and accessible database of both cultural preservation and linguistic preservation and linguistic research, as most of these languages are most often oral. Moreover, advocating international collaboration to build shared language corpora could be a key uncertainty in reducing the digital divide of minority languages. This would create a better baseline of machine learning and language processing models that would be more efficient and accessible to Indigenous communities. Intercountry collaboration, international research centers, and Aboriginal people also help in the quality of the information received, and the fact that language diversity is not eroded in the digital era.

VI. CONCLUSION

The findings of this study suggest that cross-lingual information processing is necessary to maintain archives of Indigenous languages, which experience significant shortages in digital infrastructure, language barriers, and technology gaps. The remedy is in a tactical embracing of new technologies, including neural machine translation, crowdsourced labeling, and adaptive language models, and essential ethical and interdisciplinary efforts. To move forward, more effort should be put into the underrepresented languages, the creation of annotated corpora, translation models, and multimodal interfaces. Finally, the most critical intervention is the implementation of a community-based approach to make technological tools not only technically effective but also culturally sensitive and to actively empower the Indigenous peoples, and to not eradicate their cultural identity in the face of globalization.

REFERENCES

[1] Adelani, D. I., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., ... & Klakow, D. (2022, December). Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4488-4508). https://doi.org/10.18653/v1/2022.emnlp-main.298

[2] Anastasopoulos, A., & Neubig, G. (2019). Pushing the limits of low-resource morphological inflection. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 984–996. https://doi.org/10.48550/arXiv.1908.05838

[3] Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, *3*(Feb), 1137-1155.

[4] Berez-Kroeker, A. L., & Henke, R. (2018). Language archiving. *Oxford handbook of endangered languages*, 347-369.

[5] Bernard, H. R. (1992). Preserving language diversity. *Human organization*, *51*(1), 82-89.

[6] Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, *56*, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

[7] Bird, 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

[8] Brenzinger, M., & de Graaf, T. (2009). Documenting endangered languages and maintaining language diversity. *Linguistic Anthropology*, *238*.

[9] Federici, F. M., & Cadwell, P. (2018). Training citizen translators: Design and delivery of bespoke training on the fundamentals of translation for New Zealand Red Cross. *Translation Spaces*, *7*(1), 20-43. https://doi.org/10.1075/ts.00002.fed

[10] Himmelmann, N. P. (2006). Language documentation: What is it and what is it suitable for? *Essentials of Language Documentation*, 1–30.

[11] Ismail, W. S. (2024). Threat Detection and Response Using AI and NLP in Cybersecurity. *Journal of Internet Services and Information Security*, *14*(1), 195-205. http://doi.org/10.58346/JISIS.2024.I1.013

[12] Kumar, A., & Yadav, P. (2024). Experimental Investigation on Analysis of Alkaline Treated Natural Fibers Reinforced Hybrid Composites. *Association Journal of Interdisciplinary Technics in Engineering Mechanics*, *2*(4), 25-31.

[13] Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., & Levin, L. (2017, April). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (pp. 8-14).

[14] Mager, M., Gutierrez-Vasques, X., Sierra, G., & Meza-Ruiz, I. (2018, August). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International conference on computational linguistics* (pp. 55-69).

[15] Muller, H., & Romano, L. (2024). An Exploratory Study of the Relationship Between Population Density and Crime Rates in Urban Areas. *Progression Journal of Human Demography and Anthropology*, *2*(4), 28-33.

[16] Musavi, N. E. (2018). The Relationship between Iranian EFL Learners' Motivation and Attitudes towards English Learning and Their Inference Ability in Listening Comprehension. *International Academic Journal of Humanities*, *5*(2), 1–18.

[17] Muyanja, A., Nabende, P., Okunzi, J., & Kagarura, M. (2025). *Metamaterials for revolutionizing modern applications and metasurfaces. Progress in Electronics and Communication Engineering, 2 (2), 21–30.*

[18] Nathan, D., & Austin, P. K. (2017). Reconceiving metadata: Language documentation through thick and thin. *Language Documentation and Description 2*, *179-188*. https://doi.org/10.25894/ldd299

[19] Panah, Z. S., & Homayoun, M. (2017). The Relationship between Iranian English Teachers' Sense of efficacy and their Classroom Management Strategy Use. *International Academic Journal of Organizational Behavior and Human Resource Management*, *4*(2), 1–9.

[20] Prabhudeva, T., & Hariharan, R. (2024). A Systematic Review and Meta-Analysis of Tuberculosis Patients: Perspectives of Pharmacists Towards Sustainability. *Clinical Journal for Medicine, Health and Pharmacy*, *2*(4), 1-10.

[21] Probst, C. W., & Hansen, R. R. (2013). Reachability-based Impact as a Measure for Insiderness. *Journal of Wireless Mobile Networks, Ubiquitous Computing and Dependable Applications*, *4*(4), 38-48.

[22] Rahman, F., & Lalnunthari. (2024). Development of an image processing system for monitoring water quality parameters. *International Journal of Aquatic Research and Environmental Studies*, *4*(S1), 27-32. https://doi.org/10.70102/IJARES/V4S1/5

[23] Sallal, F. F. J. (2024). The Relationship Between Machiavellianism and the Ethics of the Accounting and Auditing Profession: An Analytical Study of the Opinions of a Sample of Accountants and Auditors in Banks Listed on the Iraq Stock Exchange. (2024). *International Academic Journal of Economics*, *11*(1), 01-10. https://doi.org/10.9756/IAJE/V11I1/IAJE1101

[24] Sarma, N., Singh, R. S., & Goswami, D. (2022). SwitchNet: Learning to switch for word-level language identification in code-mixed social media text. *Natural Language Engineering*, *28*(3), 337-359. https://doi.org/10.1017/S1351324921000115

[25] Secretariat, U. N. (2022). *International Decade of Indigenous Languages, 2022–2032: Global action plan: Note by the Secretariat*. United Nations.

[26] Shamia, D., Chandrika, E., Haripriya, R., & Fathima, S. S. (2020). Cross Spectral and Cross Distance Face Matching System for Military based Camps. *International Journal of Advances in Engineering and Emerging Technology*, *11*(2), 43-51.

[27] Surendar, A. (2024). Internet of medical things (IoMT): Challenges and innovations in embedded system design. *SCCTS Journal of Embedded Systems Design and Applications*, *1*(1), 33-36. https://doi.org/10.31838/ESA/01.01.08

[28] Trivedi, J., Devi, M. S., & Solanki, B. (2023). Step Towards Intelligent Transportation System with Vehicle Classification and Recognition Using Speeded-up Robust Features. *Archives for Technical Sciences*, *1*(28), 39-56. https://doi.org/10.59456/afts.2023.1528.039J

[29] Vasquez, E., & Mendoza, R. (2024). Membrane-Based Separation Methods for Effective Contaminant Removal in Wastewater and Water Systems. *Engineering Perspectives in Filtration and Separation*, *2*(4), 21-27.

[30] Xu, Y., Namazifar, M., Hazarika, D., Padmakumar, A., Liu, Y., & Hakkani-Tur, D. (2023, July). Kilm: Knowledge injection into encoder-decoder language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 5013-5035).

[31] Zanotti, L., & Palomino-Schalscha, M. (2016). Taking different ways of knowing seriously: cross-cultural work as translations and multiplicity. *Sustainability Science*, *11*(1), 139-152.