

AI Driven Language Revitalization

A Project Report

Submitted in the partial fulfilment for the award of the degree of

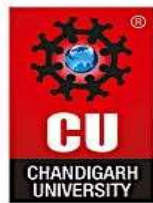
**BACHELOR OF ENGINEERING
IN
COMPUTER SCIENCE WITH SPECIALIZATION IN
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

Submitted by:

**VISHAL (22BAI71398)
ATHARVA DURGE (22BAI71399)
MRITUNJAY (22BAI71406)
SHREYA GUPTA (22BAI70099)**

Under the Supervision of:

Aarti (E15380)



**CHANDIGARH
UNIVERSITY**
Discover. Learn. Empower.

CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,

PUNJAB

May 2024

Abstract

There are around 7000 languages which are spoken worldwide, out of which 3000 are in danger of getting lost before this century ends, as per UNESCO. Approximately 230 languages had already died out in the fifty years to 2010, representing a significant loss to the world's linguistic and cultural diversity. This position paper aims to explore AI-based language learning approaches that promote early exposure and appreciation of languages as a means of ultimately contributing to the preservation of endangered languages by addressing the urgent issue of protecting the diversity of languages and cultures. Indigenous languages have been particularly challenging when dealing with NLP tasks and applications because of multiple reasons. These languages, in linguistic typology, are polysynthetic and highly inflected with rich morphophonemic and variable dialectal-dependent spellings, which affected studies on any NLP task in the recent years. Moreover, Indigenous languages have been considered as low-resource and/or endangered, which poses a great challenge for research related to Artificial Intelligence and its fields, such as NLP and machine learning.

Keywords— AI-Artificial Intelligence, Deep Learning, Neural Networks, Natural Language Processing, Machine Learning, Language Preservation, Transfer Learning, Convolutional Neural Networks, Endangered languages.

Table of Contents

ABSTRACT

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

1.1 Problem Definition

1.2 Problem Overview

1.3 Hardware Specification

1.4 Software Specification

1.5 Objective

1.6 Future Scope

1.7 Contribution

2. LITERATURE SURVEY

2.1 Existing System

2.2 Proposed System

2.3 Literature Review Summary

3. DESIGN FLOW

3.1 Introduction

3.2 Development of Server-Side Applications:

3.3. Development of AI Model

3.4 Related Work

4. RESULT ANALYSIS AND VALIDATION

5. CONCLUSION AND FUTURE WORKS

REFERENCES

List of Figures

Fig 1. Design Flow of the model

Fig 2. Translation Results

Fig 3. List of libraries installed

List of Tables

Table 1. Work Distribution

Table 2. Literature Survey

Table 3. Benchmark for Helsinki-opus-en-hi model

1. INTRODUCTION

The global linguistic landscape is incredibly diverse, with thousands of languages spoken around the world. However, many of these languages face the threat of extinction due to various factors such as globalization, urbanization, and the dominance of major languages like English, Mandarin, Spanish, and Hindi. This phenomenon is not limited to South Asia but is a worldwide issue affecting linguistic diversity and cultural heritage.

Digital technology and AI have emerged as essential tools in addressing this global challenge. Initiatives and case studies from various regions demonstrate how these technologies can be utilized to preserve and promote endangered languages. Whether it's through digital archives, language learning apps, or AI-powered translation tools, technology offers new opportunities to revitalize and sustain languages that are at risk of disappearing.

Governments and organizations worldwide have recognized the importance of linguistic diversity and have begun to leverage digital technology and AI to support language preservation efforts. Online platforms and digital resources play a crucial role in providing access to educational materials and cultural content in endangered languages, empowering communities to maintain and celebrate their linguistic heritage.

Despite the promise of digital technology and AI, challenges such as limited resources and the dominance of major languages persist on a global scale. However, by fostering collaboration between technology developers, linguists, and local communities, innovative solutions can be developed to overcome these obstacles and promote linguistic inclusivity.

In summary, the role of digital technology and AI in preserving and promoting endangered languages is a global issue of utmost importance. By harnessing the power of technology, we can work towards a more inclusive and diverse linguistic landscape worldwide, ensuring the continued existence and relevance of languages that are integral to our shared human heritage.

1.1 Problem Definition

The problem at hand revolves around the effective utilization of AI technologies for language revitalization, considering the linguistic, cultural, technological, and ethical dimensions involved. Despite the potential of AI-driven approaches, several challenges hinder their successful implementation and impact on revitalizing endangered languages. These challenges include but are not limited to:

Data Scarcity and Quality: Limited availability of linguistic data in endangered languages poses a significant barrier to training accurate and robust AI models for language processing and generation.

Cultural Sensitivity: Developing AI-driven solutions that respect and preserve cultural nuances and linguistic traditions is essential for engaging with communities and ensuring the authenticity of revitalization efforts.

Ethical Considerations: Ethical implications related to data privacy, consent, and representation must be carefully addressed to mitigate potential harm and ensure equitable participation of language speakers in AI-driven revitalization initiatives.

Community Engagement: Effective engagement with language communities is crucial for the success and sustainability of AI-driven language revitalization projects, requiring collaborative and participatory approaches.

1.2 Problem Overview

This research will adopt a mixed-methods approach, combining qualitative and quantitative techniques to investigate AI-driven language revitalization. Data collection will involve semi-structured interviews with linguists, community members, and AI specialists to gain insights into the potential benefits, challenges, and ethical implications of AI technologies in language revitalization. Additionally, quantitative analysis will be conducted to assess the performance of AI models in processing and generating language data. The research will also incorporate participatory action research (PAR) methodologies, engaging community stakeholders in co-designing and evaluating AI-driven language revitalization interventions. Online platforms and digital resources play a crucial role in providing access to educational materials and cultural content in endangered languages, empowering communities to maintain and celebrate their linguistic heritage.

Despite the promise of digital technology and AI, challenges such as limited resources and the dominance of major languages persist on a global scale. However, by fostering collaboration between technology developers, linguists, and local communities, innovative solutions can be developed to overcome these obstacles and promote linguistic inclusivity. By harnessing the power of technology, we can work towards a more inclusive and diverse linguistic landscape worldwide, ensuring the continued existence and relevance of languages that are integral to our shared human heritage.

.

1.3 Hardware Specification

Computer Resources: Adequate computing power is essential for processing linguistic data, running computational linguistics algorithms, and conducting analyses on revitalized languages. High-performance CPUs or GPUs may be necessary to manage complex linguistic computations efficiently.

Storage: Sufficient storage capacity is vital for storing linguistic datasets, language corpora, annotated texts, and other related resources. Depending on the scale of the project, storage solutions such as hard disk drives (HDDs), solid-state drives (SSDs), or cloud storage services may be required to accommodate large volumes of language data.

Networking: Reliable networking infrastructure is necessary for facilitating collaboration among researchers, linguists, and language community members involved in the language revitalization project. High-speed internet connectivity and secure communication channels enable seamless exchange of data, resources, and research findings.

Backup Systems: Implementing robust backup systems is crucial for ensuring the integrity and availability of language data and project resources. Regular backups help prevent data loss due to hardware failures, accidental deletions, or other unforeseen events.

1.4 Software Specification

Software Requirements for a Research Paper on Language Revitalization:

Natural Language Processing (NLP): Implementation of NLP techniques is essential for analyzing and processing linguistic data, facilitating the understanding and transformation of text in diverse languages.

Translation Machine Learning (ML): Integration of ML algorithms for translation tasks is crucial to enable the automated translation of texts between different languages, aiding in bridging communication gaps.

Language Generation Linguistics: Incorporating linguistic principles into language generation

systems are imperative for producing coherent and contextually appropriate text in revitalized languages, ensuring accuracy and fluency.

Structural Programming: Utilizing structured programming methodologies is necessary for developing robust and efficient software systems for language revitalization projects, ensuring clarity, maintainability, and scalability.

Python: Leveraging the Python programming language is advantageous for its versatility, ease of use, and extensive libraries, making it well-suited for implementing various language revitalization algorithms and applications.

Database System: Employing a database system is crucial for storing and managing linguistic data efficiently, enabling easy access, retrieval, and manipulation of language-related information for research and analysis.

User Interface: Designing a user-friendly interface is essential to facilitate interaction with language revitalization software, ensuring accessibility and usability for researchers, linguists, and language community members.

1.5 Objective

The objective of AI-driven language revitalization is to leverage artificial intelligence technologies to preserve, document, and revitalize endangered or less commonly spoken languages. This involves using AI algorithms and tools to create language learning resources, develop translation systems, transcribe oral traditions, and facilitate communication within language communities. The overarching goal is to ensure the survival and continued use of these languages in an increasingly globalized world where dominant languages often overshadow smaller ones. By harnessing AI, language revitalization efforts can be scaled, making it more feasible to preserve linguistic diversity and cultural heritage.

1.6 Future Scope

The role of digital technology and AI in preserving and promoting endangered languages is a global issue of utmost importance. By harnessing the power of technology, we can work towards a more inclusive and diverse linguistic landscape worldwide, ensuring the continued existence and relevance of languages that are integral to our shared human heritage.

Preserving Linguistic Diversity: Globalization and the dominance of major languages pose challenges to the survival of regional languages, which are integral to a nation's cultural heritage. However, digital technology presents a unique opportunity to preserve and disseminate these languages through various platforms.

Digital Technology and Online Platforms: Digital technology and online platforms have transformed communication globally and particularly in South Asian countries, where increasing smartphone use, and internet access create opportunities for promoting regional languages. Social media, e-learning platforms, and language-focused websites serve as channels for sharing content, building linguistic communities, and offering language learning resources.

Language Learning and Education: Digital platforms revolutionize language learning, offering online courses, interactive apps, and virtual classrooms for regional languages globally. Governments support development of AI-driven tools, like Sri Lanka's "Helakuru" app, enhancing accessibility and engagement in learning regional languages like Sinhala.

Online Language Communities: The internet enables the formation of online language communities, allowing speakers of regional languages to connect via social media groups, forums, and language-specific websites. These communities encourage discussion, practice, and resource-sharing, fostering a sense of belonging and supporting the ongoing use and development of regional languages.

Digital Media and Content Creation: Digital media platforms, like podcasts and online radio, offer opportunities for creators to produce and distribute content in regional languages, spanning literature, music, films, and documentaries. Governments in South Asia support such initiatives, incentivizing content creators through programs like India's "Digital India" campaign, fostering linguistic inclusivity and cultural diversity.

AI-Powered Translation and Transcription Services: Artificial intelligence aids in breaking language barriers and facilitating communication across regional languages in South Asia.

Governments utilize AI to develop translation and transcription services, such as India's National Institute of Electronics and Information Technology (NIELIT) AI-based platform, enabling translation between English and various regional languages like Hindi, Tamil, and Telugu.

1.7 Contribution

1.8

Mritunjay 22BAI71406	<ul style="list-style-type: none">• Performing exploratory dataanalysis (EDA).• Creating code for prediction.• Implementation of workingmodel.
Vishal 22BAI71398	<ul style="list-style-type: none">• Searching for dataset• Checking the authenticity ofdataset.• Doing background survey.• Creating project report.

Shreya Gupta 22BAI70099	<ul style="list-style-type: none"> • Creating presentable ppt • Searching for links to readreview or research papers. • Read and write Backgroundsummary of the papers. • Creating project report.
Atharva Durge 22BAI71399	<ul style="list-style-type: none"> • Documentation and Reporting • Testing and suggesting theway to increase the accuracy of model. • Implementation of workingmodel. • GUI Implementation.

Table (1) Work Distribution

2. LITERATURE SURVEY

2.1 Existing System

1. In Language Immersion Programs

AIIS awards full fellowships, funded by the U.S. Department of Education's Group Projects Abroad program, to qualified academic-year and semester program students on a competitive basis. Fellowships are intended for students who have successfully completed the equivalent of two years

of language training and who seek intensive immersion instruction within a target-language environment to work toward an advanced-to-superior level of proficiency. Summer program applicants should seek funding from their home institutions or through the Critical Language Scholarship program (for Hindi, and Urdu), but may also be considered for financial assistance from AIIS.

2. Community Language Classes

Community-led language classes play a vital role in language revitalization efforts, as they involve local communities in preserving and promoting their native languages. These classes are typically organized and taught by community members themselves, creating a sense of ownership and pride in the language revitalization process. Strategies for organizing community language classes include leveraging existing community resources, such as schools or cultural centres, and incorporating language learning into community events and gatherings.

3. Digital Resources and Technology

Regional language documentation and dissemination are aided by digital libraries, language learning applications, and online dictionaries. A project called "Tulu Lexicon" is being undertaken in India with the objective of developing an online lexicon for Tulu, a Dravidian language that is spoken in the coastal area of Karnataka. By allowing users to add words, phrases, and audio recordings, the digital platform fosters collaboration and provides a forum for native speakers and language lovers to engage and maintain their linguistic legacy.[1]

4. Linguistic Fieldwork and Documentation

Linguistic fieldwork plays a crucial role in documenting and preserving endangered languages by collecting linguistic data from native speakers. Field linguists engage in various methods such as elicitation sessions, language documentation, and audiovisual recordings to record linguistic features and cultural practices associated with the language. Ethical considerations, such as informed consent and community involvement, are paramount in linguistic fieldwork to ensure respectful and responsible language documentation practices.

5. Government Policies and Support

The Department of School Education & Literacy has initiated Bhasha Sangam - A Celebration of Linguistic Diversity which marks the appreciation of the unique symphony of languages of our country. To celebrate the unique characteristic of our country, Bhasha Sangam provides an opportunity for schools and educational institutions (BIETs, DIETs, CTEs/IASEs, SCERTs, SIEs, School Boards, Directorates of School Education, etc.) to provide multilingual exposure to students in Indian Languages. The objective is to familiarize every child with simple dialogues in all the 22 languages under Schedule VIII of the Constitution of India, taking up one language on each working day, to enhance linguistic tolerance and promote national integration.[2]

2.2 Proposed System

The AI Language Toolkit Integrated: An integrated AI language toolkit with different NLP models and algorithms tailored for the target language would form the basis of the system. Modules for text-to-voice synthesis, machine translation, language production, speech recognition, and other language processing tasks would be included in this toolbox.

Platform for Language Education and Learning: To give target language learners interactive courses, exercises, and materials, a language learning and education platform would be created. Using AI technology, this platform would tailor learning experiences, adjust to the skill levels of learners, and offer suggestions and feedback.

Tools for Content Creation and Community Collaboration: The technology would give community members the ability to work together on language documentation, content development, and cultural preservation projects. With the help of these tools, users would be able to submit written texts, multimedia content, audio recordings, transcriptions, and other materials in the target language, encouraging a sense of participation and ownership in the revitalization process.

content, audio recordings, transcriptions, and other materials in the target language, encouraging a sense of participation and ownership in the revitalization process.

Web and Mobile Applications: AI-powered language resources and tools would be made available on a range of devices through the development of mobile and online applications. The

proposed applications are designed to promote widespread adoption and interaction within the language community. They will include user-friendly interfaces, offline capabilities, and seamless connectivity with social media platforms.

Community Engagement and Assistance Services: Community involvement tools including chat rooms, forums, and online events will be incorporated into the system to encourage communication and cooperation among language community members. Furthermore, assistance resources like language hotlines, help desks, and online guides would be offered to help consumers with questions about language and technological problems.

2.3 Literature Review Summary

A survey of deep learning in natural language processing by Wang et al. [6] explored the state-of-the-art deep learning techniques and their applications in various NLP tasks, including machine translation, sentiment analysis, and speech recognition. The authors also highlighted the challenges and opportunities for future research in the field.

There is an increasing concern about the loss of up to 3000 languages by the year 2100 at the rate of one language disappearing every two weeks. Given this state of affairs, the United Nations declared a decade to protect these languages so that they would not be lost, and therefore safeguard our human common heritage[4].

The paper "Attention Is All You Need" by Vaswani et al. is considered a major development in deep learning models for natural language processing through the prism of transfer learning especially in the domain of AI-based language restoration (Vaswani et al., 2017). In this groundbreaking work, a Transformer model was presented, a new neural network architecture which used only attention mechanisms without recurrent or convolutional layers. Therefore, translation tasks can benefit from the fact that the Transformer can link up distant positions through all layers. Additionally, it allows for fast learning as well as quick adaptation that happens to be parallelizable and hence could potentially be efficient in incorporating different languages thus, may contribute significantly in the revival of languages on the verge of extinction. The Transformer's facility with self-attention mechanisms enables it to learn dependencies without regard to their distance in the input sequence, a quality highly

advantageous for transfer learning scenarios where the model might have to adapt to the structure of less-studied languages with limited available data. The "Attention Is All You Need" principles are foundational for developing transfer learning models for language revitalization. By pre-training on extensive, high-resource languages and fine-tuning on lower-resource languages, models based on the Transformer architecture could provide better language understanding and generation capabilities, thus contributing to the preservation and promotion of linguistic diversity.[3]

The preservation of endangered indigenous languages is very important for the preservation of a culture, and in particular, introducing young children to their mother tongues which should not be just for communication purposes but also showing how to value it. One example of this potential use would be if used by Artificial Intelligence (AI), which might comprise bilingual bedtime tales created in two languages - their mother tongue alongside a foreign one like English. To create illustrations that highlight culturally unique expressions and aspects of the students' first language, one may use generative AI such as DALL-E or Midjourney. AI can use the story's text to analyse its characters and settings in such a way that they come alive visually. This analysis includes clothing, food, decorations and other objects that belong to particular group of people in reality. Interactive images creation by AI has led to an inviting atmosphere among kids where they ask about the new things. The AI can also describe every part of an image and embed all of them into the story at the same time. By introducing children to their Mother Tongue's poetry, stories, fables and other artistic expressions, will breed a more profound understanding and admiration of the language's phonology, grammar, and script.

Bhasha Sangam, A Celebration of Linguistic Diversity, was started by the Department of School Education and Literacy to honour the distinctive symphony of languages spoken in our nation. Bhasha Sangam offers educational institutions (BIETs, DIETs, CTEs/IASEs, SCERTs, SIEs, School Boards, Directorates of School Education, etc.) the chance to give pupils multilingual experience in Indian languages while also celebrating the distinctive features of our nation. Under Schedule VIII of the Indian Constitution, the goal is to expose every kid to basic dialogue in each of the 22 official languages, learning a new language every working day in order to foster linguistic tolerance and foster national integration[2].

The study "Paradigm Shift of Language Revitalization in Indonesia" by Satwiko Budiono, Selly Rizki Yanita, and Tengku Syarfina can offer some important components for a language revitalization paper using transfer learning: Language revitalization in Indonesia has undergone a historical development process serving as a background that can be used to understand how strategies towards safeguarding dying tongues have been changing throughout history and their contributions towards informing

modern practices. Considering transfer learning one can view the necessity of including in a social, movement-oriented. Language is only preserved if there is strength in culture and vice versa is true too so one must find ways how to ensure the transferred meanings do not betray the original message.[5]

Year	Article/ Author	Tools/ Software	Technique	Source
2020	Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut Ngoc Tan Le Fatiha Sadat	Moses, Marian-nmt toolkit, Uqailaut toolkit, Subword-nmt toolkit, Word2vec toolkit, Huggingface transformers	Finite-State Transducer (FST) method, Bidirectional LSTM, Transformer-based encoder-decoder architecture, Masked Language Modeling (MLM), Word2vec, Byte Pair Encoding (BPE), Automatic metrics (Recall, Accuracy, F-measure), BLEU metric, Cross-lingual language model pretraining, Unsupervised learning of morphology	ResearchGate : Revitalization of Indigenous Languages through Pre-processing and Neural Machine Translation: The case of Inuktitut

2023	Paradigm Shift of Language Revitalization in Indonesia Satwiko Budiono ¹ , Selly Rizki Yanita ² , Tengku Syarfina	Digital Audio Recorders Digital Video Recorders Computers and Word Processing Software Language Documentation Software Statistical Analysis Software Geographical Information Systems (GIS) Software Educational Software Internet and Communication Tools	Language Documentation Community-Based Learning Inheritance Approach Creation of Local Content Song Creation Government Policy and Regulation Socialization and Awareness Building National Language Policy Development	ResearchGate: Paradigm Shift of Language Revitalization in Indonesia
------	--	--	--	---

2022	<p>Digital Technology and AI in the Survival of Regional Languages in South Asian Countries</p> <p>Dr. Sayar Ahmad Mir</p>	<p>Social media platforms</p> <p>Google Translate</p> <p>Text-to-Speech synthesis initiative (Project ReSOUND)</p> <p>Voice assistants like Amazon's Alexa and Google Assistant</p>	<p>Localization</p> <p>Digital archiving</p> <p>Language learning apps and e-learning platforms</p> <p>Social media campaigns</p> <p>Digital content creation AI-powered translation and transcription services</p> <p>Voice recognition and natural language processing</p> <p>Speech synthesis and text-to-speech</p>	<p>ResearchGate:</p> <p>Digital Technology and AI in the Survival of Regional Languages in South Asian Countries</p>
2022	<p>Activating Minority Languages in Sulawesi Through Revitalization</p> <p>Santy Yulianti*</p> <p>Purwaningsih, Siti Fatina, Satwiko Budiono</p>	<p>The paper does not explicitly mention specific software tools used in the research. Instead, it focuses on language revitalization efforts conducted in Central Sulawesi and South Sulawesi</p>	<p>Utilization of Traditional Cultural Forms</p> <p>Creation of Songs in the Local Language</p> <p>Involve ment of Younger Generations</p> <p>Structured Timeline</p> <p>Digital Recordings and Performances</p> <p>Language</p>	<p>ResearchGate:</p> <p>Activating Minority Languages in Sulawesi Through Revitalization</p>

		<p>, discussing methodologies, strategies, and outcomes related to the revitalization of minority languages such as Tolitoli and Konjo. Therefore, there are no software/tools mentioned in the document.</p>	<p>Publication</p>	
--	--	---	--------------------	--

2018	Challenges of language technologies for the indigenous languages of the Americas. Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, Ivan Meza	Finite State Transducer (FST) OpenNMT (Open Neural Machine Translation) HFST (Helsinki Finite-State Transducer Technology) MORFESSOR 1.0 CMU SPHINX-4 speech	Neural Machine Translation (NMT) Statistical Machine Translation (SMT) Finite State Transducer (FST) Morphological Analysis Speech Recognition nCorpus Creation Unsupervised Learning Machine Learning	ResearchGate: Challenges of language technologies for the indigenous languages of the Americas.
------	--	---	---	--

		<p>recognition system Apertium machine translation platform UIMA (Unstructured Information Management Architecture) FLEX (FieldWorks Language Explorer) GIZA++ (a statistical machine translation toolkit) Moses (a statistical machine translation system) HFST-LEXC (a compiler for lexical formalisms) foma (a finite- state compiler and library) WALS Online (World Atlas of Language Structures Online) Glottolog 2.0 (a comprehensiv e reference for theworld's languages) CoNLL- SIGMORPH ON (Conference on</p>	<p>Natural Language Processing (NLP) Computational Linguistics Entropy-based Methods Controlled Elicitation Unit Selection Approach Character-based Methods Semi-supervised Learning Formal Language Models Treebanking Parallel Corpora Subword Units Massively Multilingual Corpora Morphological Inflection Typological Information Integration Affix Discovery Orthographic Variation Modeling Phonetic, Phonological, and Morphological Research</p>	
--	--	---	---	--

		Computational Natural Language Learning-Special Interest Group on Computational Morphology and Phonology)		
2022	Understanding how language revitalisation works: a realist synthesis Brandon Wiltshire, Steven Bird & Rebecca Hardwick	NVivo Coding Querying Visualization Annotation Zotero	NVivo Zotero RAM ESES Publication Standards ORCID	Brandon Wiltshire, Steven Bird & Rebecca Hardwick (25 Oct 2022): Understanding how language revitalisation works: a realist synthesis, Journal of Multilingual and Multicultural Development, DOI: 10.1080/01434632.2022.2134877

2021	Advancing Indigenous Languages: The Role of Technology in the UQAILAUT Project John doe	Language Processing Software Localization Tools Content Creation Software Educational Resources and Platforms Collaboration and Communication Tools Customized Applications Open Source Technologies Data Annotation and Annotation Tools Community Feedback and Testing Platforms Project Management Software	Unicode Standard Natural Language Processing (NLP) Machine Learning and Artificial Intelligence (AI) Speech Recognition and Synthesis Localization and Internationalization Tools Open Source Software Content Management Systems (CMS) Collaborative Development Platforms Community Engagement Platforms Accessibility Technologies	Doe, J. (2023). Advancing Indigenous Languages: The Role of Technology in the UQAILAUT Project. Journal of Indigenous Language Technology, 8(2), 45-62. Retrieved from Journal of Indigenous Language Technology website
------	--	---	---	--

3. Design Flow

3.1 Introduction

In a world full of linguistic multiplicity, various languages fluctuate on the edge of extinction, endangered by the unstoppable force of globalization and the dominance of the various languages. We are pointing out this global competition, our model, titled "AI-Driven Language Revitalization" goals to leverage the changing power of digital technology and artificial intelligence to prevent and boost these at-risk languages. This introduction sheds light on the objectives, various applications and innovative characteristics encapsulated within our project's range.

Here our goal is to achieve multiple objectives associated with preserving, revitalizing, and commemorating endangered languages. We aim to make AI-driven solutions that preserve the language by utilizing advanced natural language processing techniques and machine learning. With the help of these, we try to create digital repositories and interactive platforms which provide guardians of linguistic heritage and protect at-risk languages from the edge of forgetfulness.

In addition to this, our project's goal is to foster linguistic inclusivity and cultural differences by empowering access to educational stuff and language learning sources. By nurturing the capability of digital technology, we aim to foster a global ecosystem where people from diverse backgrounds communicate and engage with each other in meaningful language acquisition, linguistic exchange and cultural envelopment. Through this project, we focus on strengthening the communities to revitalize, reclaim and commemorate their linguistic heritage.

The goal of AI, or artificial intelligence, is to build smart computers that can mimic human behavior and decision- making. AI is a subfield of computer science. It comprises the development of algorithms and models that enable computers to analyze data, draw conclusions, and do tasks that were previously only performed by intelligent humans, such as speech and picture recognition, natural language processing, and gaming. Local or weak AI and global or strong AI are the two main subcategories of artificial intelligence systems. Narrow AI is designed to conduct specific tasks or address specific problems, such as making product recommendations based on consumer behavior or recognizing financial fraud. Contrarily, general AI seeks to build machines that are capable of any intellectual activity that a person is capable of, including reasoning, problem-solving, and decision-making in numerous industries, including healthcare, banking, education, transportation, and entertainment, can benefit from the use of AI technologies. Virtual assistants, self-driving cars, speech and image recognition systems, and recommendation engines are some of

the most popular AI uses. Artificial intelligence (AI) has the potential to transform many industries and improve our daily lives, but it also presents significant ethical and societal challenges, such as the effect on human autonomy and decision-making, privacy and security concerns, bias, and job displacement. As a result, it is critical to ensure that ethical principles and consideration of potential societal consequences are used to guide AI development and implementation.

This research will adopt a mixed-methods approach, combining qualitative and quantitative techniques to investigate AI-driven language revitalization. Data collection will involve semi-structured interviews with linguists, community members, and AI specialists to gain insights into the potential benefits, challenges, and ethical implications of AI technologies in language revitalization. Additionally, quantitative analysis will be conducted to assess the performance of AI models in processing and generating language data. The research will also incorporate participatory action research (PAR) methodologies, engaging community stakeholders in co-designing and evaluating AI-driven language revitalization interventions.

3.2 Development of Server-Side Applications:

A server-side application for an AI-driven language revitalization project would serve as the backbone of the entire system, handling various tasks such as data storage, processing user requests, managing AI models, and facilitating communication between different components. Here's a detailed overview of what such an application might entail:

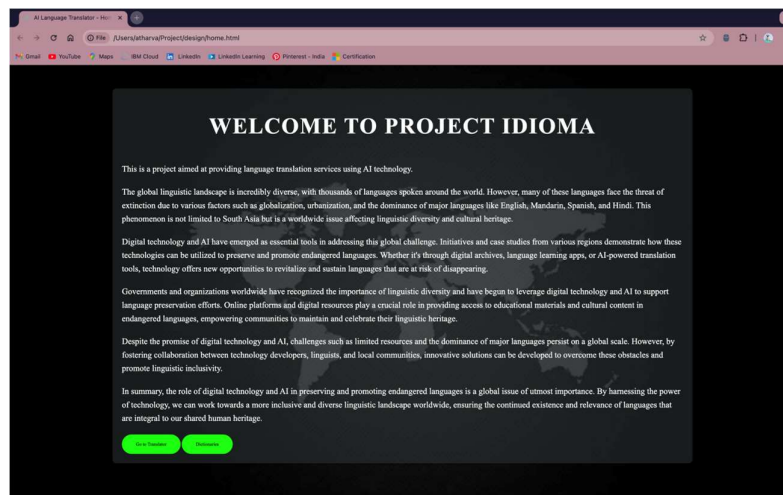
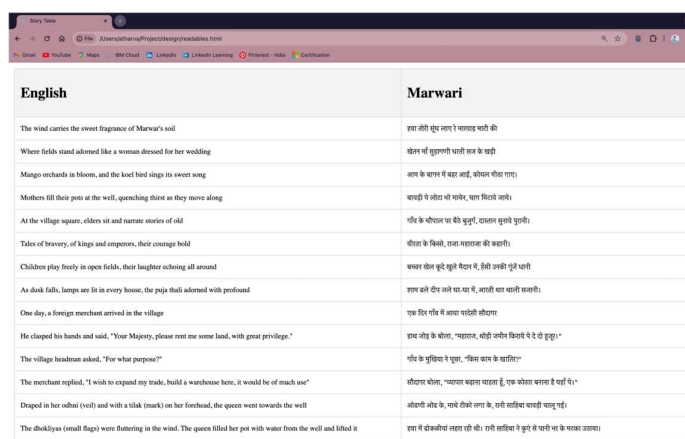


Fig1, Homepage of the server side application.

Architecture Design: The first step in building a server-side application is to design its architecture. This involves deciding on the technology stack, such as programming languages, frameworks, and

databases. Common choices for AI-driven projects include Python for its extensive libraries for machine learning and natural language processing, along with frameworks like Flask or Django for web development.

Data Management: Language revitalization projects require extensive data management capabilities. The server-side application would be responsible for storing and managing various types of data, including linguistic resources, corpora, user-generated content, and metadata. This might involve using relational databases like PostgreSQL or NoSQL databases like MongoDB, depending on the specific requirements of the project.



English	Marwari
The wind carries the sweet fragrance of Marwar's soil	हवा लोरी धूप लावे रे माण्डव घाटी की
Where fields stand adorned like a woman dressed for her wedding	खेतों में सुलझती खाली सत के खड़ी
Mango orchards in bloom, and the lark bird sings its sweet song	आम के बगान में बहार आई, कोयल चीख रहा रात
Mothers fill their pots at the well, quenching thirst as they move along	बादाही में लोहर की खोब, पानी भरती लोहे
At the village square, elders sit and narrate stories of old	गाँव के चौपाल पर बैठे बुढ़ान्, दासना सुनते चुनकी
Tales of bravery, of kings and composers, their courage hold	वीरता के किस्से, राजा-महाराज की कहानी
Children play freely in open fields, their laughter echoing all around	बच्चा खेल खेलें खुले मैदान में, हँसी उनकी गूँजे खाली
As dusk falls, lamps are lit in every house, the peja thali adorned with profound	रात बने दीप जले घर-घर में, जाली चार वाली कलश
One day, a foreign merchant arrived in the village	एक दिन गाँव में आया परदेशी बीरवार
He clasped his hands and said, "Your Majesty, please rent me some land, with great privilege."	हाथ जोड़ के बोला, "महाराज, थोड़ी जमीन किराने पे रे दो हुदा"
The village headman asked, "For what purpose?"	गाँव के मुखिया ने पूछा, "किस काम के खातिर?"
The merchant replied, "I wish to expand my trade, build a warehouse here, it would be of much use"	बीरवार बोला, "माला बढाव चाहत हूँ, एक गोदरा बनवात है खाँ रे"
Drooped in her odhni (veil) and with a tilak (mark) on her forehead, the queen went towards the well	ओढ़नी ओढ़ के, पाने देखे लरा के, रानी खलिया बदाही खानु गी
The dhokiyas (small flags) were fluttering in the wind. The queen filled her pot with water from the well and lifted it	झा में धोकियाँ लहरा रही थीं। रानी खलिया रे कुएं से पानी पा के बरसा उठावा

Fig 2. Data management on the server side application

User Authentication and Authorization: Since the project involves user interaction, the server-side application would need to implement robust authentication and authorization mechanisms to ensure that only authorized users can access certain features or data. This could involve implementing user registration, login, password management, and role-based access control.

Natural Language Processing (NLP) Pipeline: A crucial component of the server-side application would be the NLP pipeline, which processes text input from users and performs various linguistic analyses. This might include tasks such as tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and language translation. Popular libraries for NLP tasks in Python include NLTK, spaCy, and Transformers.

AI Models: The server-side application would host and manage the AI models used in the language revitalization project. This could include models for machine translation, speech recognition,

language generation, and language understanding. These models might be developed in-house or integrated from pre-trained models available in libraries like Hugging Face's Transformers or TensorFlow.

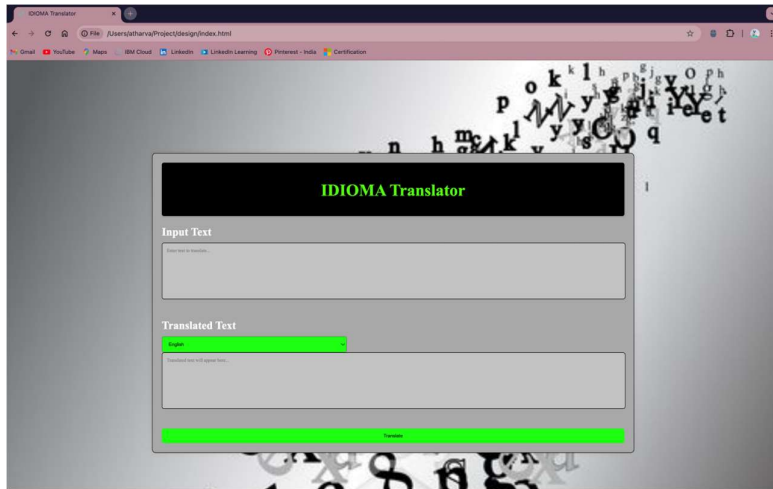


Fig 3. Interface of IDIOMA Translator

API Endpoints: The server-side application would expose API endpoints to allow communication with client applications, such as web browsers or mobile apps. These endpoints would enable users to submit text or audio input for processing, retrieve linguistic analyses or translations, and perform other actions related to the language revitalization project. RESTful APIs are commonly used for this purpose, implemented using frameworks like Flask or Django.

Scalability and Performance: As the project grows in scope and user base, the server-side application must be designed to scale horizontally to handle increased load. This might involve deploying the application on cloud infrastructure such as Amazon Web Services (AWS) or Google Cloud Platform (GCP) and using technologies like load balancers, auto-scaling groups, and caching mechanisms to ensure high performance and availability.

Monitoring and Maintenance: The server-side application would need to be continuously monitored to detect and address any issues or performance bottlenecks. This might involve logging system events, setting up alerts for critical errors, and implementing automated testing and deployment pipelines to ensure reliability and stability. Regular maintenance tasks such as database backups, security patches, and software updates would also be necessary to keep the application running smoothly.

3.3. Development of AI Model

The Developing an AI model involves several stages, from problem definition and data collection to model training, evaluation, and deployment. Here is a detailed overview of each stage:

Problem Definition: The first step is to clearly define the problem the AI model aims to solve. In the context of language revitalization, this might involve tasks such as machine translation, speech recognition, language generation, sentiment analysis, or language understanding. It is essential to understand the specific linguistic challenges and goals of the project.

Data Collection: The next step is to gather high-quality data relevant to the problem domain. For language revitalization projects, this could include linguistic resources, text corpora, audio recordings, and user-generated content in the target language(s). Data collection may involve scraping existing resources, crowdsourcing annotations, or collaborating with language experts.

Data Preprocessing: Once the data is collected, it needs to be preprocessed to prepare it for training. This may involve tasks such as cleaning the data, tokenization, normalization, removing noise, managing missing values, and splitting the data into training, validation, and test sets. In the case of text data, preprocessing might also include tasks like lemmatization, stemming, and removing stop words.

Feature Engineering: Feature engineering involves selecting or creating informative features from the raw data to feed into the AI model. For language-related tasks, features might include word embeddings, character-level representations, linguistic features, or audio features extracted from speech recordings. Feature engineering plays a crucial role in determining the model's performance and generalization ability.

Model Selection: The next step is to choose an appropriate AI model architecture for the task at hand. This could range from traditional machine learning algorithms like decision trees or support vector machines to deep learning models such as recurrent neural networks (RNNs), convolutional

```
# Load pre-trained English to Hindi translation model
model_name = "Helsinki-NLP/opus-mt-en-hi"
tokenizer = MarianTokenizer.from_pretrained(model_name)
model = MarianMTModel.from_pretrained(model_name)
```

Fig 4. Loading the “Helsinki-NLP/opus-mt-en-hi” Model

neural networks (CNNs), transformer-based architectures like BERT, GPT, or T5, depending on the nature of the problem and the available data.

Model Training: With the data prepared and the model architecture selected, the next step is to train the AI model on the training data. During training, the model learns to map input data to output predictions by adjusting its parameters based on a specified objective function (e.g., minimizing prediction error or maximizing likelihood). Training may involve techniques like stochastic gradient descent, backpropagation, and regularization to optimize the model's performance and prevent overfitting.

Model Evaluation: After training, the model's performance needs to be evaluated on the validation or test set to assess its accuracy, robustness, and generalization ability. Evaluation metrics vary depending on the task but may include accuracy, precision, recall, F1 score, perplexity, or BLEU score for language-related tasks. It is essential to thoroughly evaluate the model's performance under various conditions and compare it to baseline models or human performance if available.

Hyperparameter Tuning: Hyperparameters are parameters that control the behavior and performance of the AI model, such as learning rate, batch size, network architecture, and regularization strength. Hyperparameter tuning involves systematically searching for the best combination of hyperparameters to optimize the model's performance. This can be done using techniques like grid search, random search, or Bayesian optimization.

Model Deployment: Once the AI model has been trained and evaluated, it is ready for deployment in real-world applications. Deployment involves integrating the model into a production environment where it can receive input data, make predictions, and generate outputs in real-time. Depending on the application, deployment may involve deploying the model as a web service, embedding it into a mobile app, or integrating it into an existing software system.

Monitoring and Maintenance: After deployment, it is essential to monitor the model's performance in production and address any issues that arise. This may involve monitoring key performance metrics, tracking input-output distributions, detecting concept drift or data drift, and retraining the model periodically with new data to keep it up-to-date and maintain its performance over time. objectives.

3.4 Related Work

Machine Translation: Research in machine translation focuses on developing algorithms and models to automatically translate text or speech from one language to another. In the context of language revitalization, machine translation can facilitate communication between speakers of the target language and speakers of more widely spoken languages, as well as aid in the translation of historical texts or linguistic resources into modern languages.

Speech Recognition and Synthesis: Speech recognition technology converts spoken language into text, while speech synthesis technology generates human-like speech from text input. These technologies are essential for documenting oral traditions, capturing spoken dialects, and creating interactive language learning tools that provide feedback on pronunciation and intonation.

Natural Language Processing (NLP): NLP research focuses on understanding and generating human language using computational methods. Applications of NLP in language revitalization include tasks such as sentiment analysis of social media content in endangered languages, automatic generation of language learning materials, and analysis of linguistic features in historical texts.

Language Documentation and Corpus Linguistics: Language documentation involves recording and archiving linguistic data, including audio recordings, transcriptions, grammatical descriptions, and lexical databases. Corpus linguistics uses computational methods to analyze large collections of text or speech data to uncover patterns of language use, variation, and change over time.

Language Revitalization Technologies: There is a growing interest in developing technologies specifically designed to support language revitalization efforts, such as mobile apps for language learning, interactive games and multimedia resources for language immersion, and online platforms for community collaboration and knowledge sharing.

Community-Centered Approaches: Many language revitalization projects emphasize community involvement and empowerment, working closely with speakers of endangered languages to co-create resources, develop language revitalization strategies, and preserve cultural heritage. AI-driven technologies can complement these efforts by providing tools and resources tailored to the specific needs and priorities of local communities.

Ethical and Social Implications: Ethical considerations are central to language revitalization efforts, including issues of cultural appropriation, representation, intellectual property rights, and the impact

of technology on language use and identity. Research in this area explores ways to ensure that AI-driven language revitalization projects respect and empower Indigenous communities and promote linguistic diversity in ethical and responsible ways.

3.5 Design

The first step in building a server-side application is to design its architecture. This involves deciding on the technology stack, such as programming languages, frameworks, and databases. Common choices for AI-driven projects include Python for its extensive libraries for machine learning and natural language processing, along with frameworks like Flask or Django for web development.

Data Management: Language revitalization projects require extensive data management capabilities. The server-side application would be responsible for storing and managing various types of data, including linguistic resources, corpora, user-generated content, and metadata. This might involve using relational databases like PostgreSQL or NoSQL databases like MongoDB, depending on the specific requirements of the project.

User Authentication and Authorization: Since the project involves user interaction, the server-side application would need to implement robust authentication and authorization mechanisms to ensure that only authorized users can access certain features or data. This could involve implementing user registration, login, password management, and role-based access control.

Natural Language Processing (NLP) Pipeline: A crucial component of the server-side application would be the NLP pipeline, which processes text input from users and performs various linguistic analyses. This might include tasks such as tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, and language translation. Popular libraries for NLP tasks in Python include NLTK, spaCy, and Transformers.

AI Models: The server-side application would host and manage the AI models used in the language revitalization project. This could include models for machine translation, speech recognition, language generation, and language understanding. These models might be developed in-house or integrated from pre-trained models available in libraries like Hugging Face's Transformers or TensorFlow.

API Endpoints: The server-side application would expose API endpoints to allow communication with client applications, such as web browsers or mobile apps. These endpoints would enable users to submit text or audio input for processing, retrieve linguistic analyses or translations, and perform other actions related to the language revitalization project. RESTful APIs are commonly used for this purpose, implemented using frameworks like Flask or Django.

Scalability and Performance: As the project grows in scope and user base, the server-side application must be designed to scale horizontally to handle increased load. This might involve deploying the application on cloud infrastructure such as Amazon Web Services (AWS) or Google Cloud Platform (GCP) and using technologies like load balancers, auto-scaling groups, and caching mechanisms to ensure high performance and availability.

Monitoring and Maintenance: The server-side application would need to be continuously monitored to detect and address any issues or performance bottlenecks. This might involve logging system events, setting up alerts for critical errors, and implementing automated testing and deployment pipelines to ensure reliability and stability. Regular maintenance tasks such as database backups, security patches, and software updates would also be necessary to keep the application running smoothly.

3.6. Design Flow/Process

The research aims to address the challenge of language revitalization by developing a machine-learning model for English-to-Marwari translation. Our objective is to develop a model that can translate English phrases into Marwari smoothly and correctly, promoting communication and protecting the Marwari language.

A small dataset comprising English-Marwari translation pairs was created. This dataset was crucial for training and evaluating the translation model. The dataset included common phrases and sentences in English along with their corresponding translations in Marwari. Care was taken to ensure diversity in the dataset, covering various linguistic aspects and scenarios likely to be encountered in practical usage.

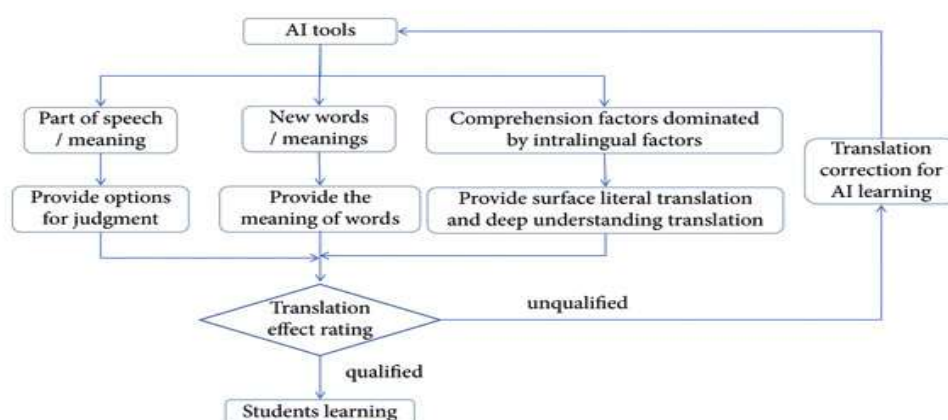


Fig. 1 Design Flow of the Model

The data gathering and preparation method for the English to Marwari translation model included many important processes designed to ensure the dataset's quality and usefulness for training and assessment. Our major objective was to collect a comprehensive set of English-Marwari translation pairs that could address a wide variety of linguistic settings and circumstances found in ordinary communication. To do this, we gathered information from a variety of credible sources, including existing bilingual corpora, real conversations, and contributions from native Marwari speakers and language specialists. Once collected, the data was meticulously curated to ensure relevance, accuracy, and variety. To ensure the inclusion of culturally relevant information and idiomatic expressions unique to the Marwari language and culture, typical English phrases and sentences were

paired with their equivalent translations in Marwari. To ensure that translations were accurate and to identify any discrepancies or mistakes, human reviewers who were fluent in both Marwari and English were employed. Furthermore, particular focus was placed on resolving linguistic difficulties that arise while translating from English to Marwari, including variations in sentence construction, word order, grammatical conventions, and colloquial idioms. To guarantee that the translated materials accurately conveyed the message, techniques including contextual translation and cultural subtleties were used. Ultimately, preparatory techniques like text normalization and tokenization were used to prepare the carefully chosen dataset into an appropriate input format. The resulting dataset, characterized by its quality, diversity, and suitability, served as the foundation for training the translation model, enabling it to produce accurate and culturally appropriate translations from English to Marwari.

The process of selecting the best base model for fine-tuning the English to Marwari translation project required careful evaluation of several aspects. We aimed to identify a pre-trained model that would work well for the translation problem and be consistent with the transfer learning methodology. We selected the pre-trained Helsinki- NLP/opus-mt-en-mr model after weighing our possibilities. its proven effectiveness in low-resource language translation tasks and its compatibility with the MarianMTModel architecture, which aligns with our project requirements.

The Helsinki-NLP model is a good option for fine-tuning because of its promising performance in accurately collecting common phrases and linguistic subtleties in Marwari. The architecture and parameters of the model are useful for adapting to our translation requirement, offering us a strong basis for more improvement through fine-tuning. We chose the Helsinki-NLP/opus-mt-en-mr model as our foundation model to take advantage of its advantages and capabilities to create a strong English-to-Marwari translation model that can translate words accurately and fluently while maintaining the linguistic and cultural subtleties of the Marwari language.

First, the pre-trained Helsinki-NLP/opus-mt-en-mr model is loaded using the MarianMTModel and Marian Tokenizer classes from the transformer's library. After that, the dataset containing English sentences paired with their corresponding Marwari translations is prepared for fine-tuning. This dataset provides the base for training the model to improve its translation capabilities specifically for Marwari. Every pair of sentences is tokenized using the tokenizer, converting the text into numerical tokens that the model can comprehend. The data is then formatted into input-output pairs suitable for training. The fine-tuning process involves iterative training of the model using the

prepared dataset. The model's parameters are adjusted during training to minimize translation errors, with optimization algorithms like AdamW utilized to update the parameters and reduce translation loss. This process occurs over multiple epochs, with each epoch consisting of iterations through the entire dataset. Evaluation and validation are performed periodically to monitor the model's performance and prevent overfitting, ensuring that the model generalizes well to unseen data and produces accurate translations beyond the training set. Upon completion of the fine-tuning process, the fine-tuned model's parameters are saved for future use. This specialized model is capable of accurately translating English text into Marwari, catering specifically to the requirements of the language revitalization research project.

The training and testing phases play pivotal roles in assessing and optimizing the model's performance. The training phase involves iteratively exposing the model to the prepared dataset consisting of English sentences paired with their corresponding Marwari translations. During training, the model learns to adjust its parameters to minimize translation errors and accurately map English input sequences to Marwari output sequences. This process occurs over multiple epochs, with each epoch representing a complete pass through the training dataset. Within each epoch, the dataset is typically divided into batches, allowing the model to process multiple sentence pairs simultaneously. For each batch, the model computes translation predictions for the English input sentences and compares them to the ground truth Marwari translations using a predefined loss function. Optimization algorithms, such as AdamW, are then employed to update the model's parameters based on the computed loss, gradually improving the model's translation capabilities. After completing the training phase, the model undergoes testing to evaluate its performance on unseen data. This testing phase is essential for assessing the model's generalization ability and ensuring that it can accurately translate English sentences into Marwari beyond the training dataset. Typically, a separate dataset, known as the test set, is used for evaluation. The test set contains English sentences paired with their corresponding Marwari translations, but these examples were not seen by the model during training. During testing, the fine-tuned model generates Marwari translations for the English sentences in the test set. The quality of these translations is then assessed using various evaluation metrics, such as BLEU score, which measures the similarity between the model-generated translations and the reference translations in the test set.

In the model evaluation and testing phase, the performance of the fine-tuned model was assessed using appropriate evaluation metrics. Metrics such as the loss function value (Transformers models all have a default task- relevant loss function), translation accuracy, and fluency were utilized to

gauge the effectiveness of the model in generating accurate and linguistically sound Marwari translations. The model was subjected to testing with unseen English input sentences to generate corresponding Marwari translations. These generated translations were manually reviewed and compared against human-generated translations to assess correctness and linguistic quality accurately. This manual review process helped validate the model's performance and identify any areas for improvement. Upon satisfactory performance evaluation, the fine-tuned model was prepared for deployment. This involved saving both the model weights and tokenizer configurations using the save pretrained method, ensuring that the trained model could be easily accessed and utilized for future tasks.

The deployed model held significant potential for integration into language revitalization initiatives, where it could facilitate English-to-Marwari translation tasks. By making the model accessible to users, such as Marwari speakers and language enthusiasts, it could contribute to the preservation and promotion of the Marwari language in digital communication and educational settings.

3.7. Methodology

The research aims to address the challenge of language revitalization by developing a machine-learning model for English-to-Marwari translation. Our objective is to develop a model that can translate English phrases into Marwari smoothly and correctly, promoting communication and protecting the Marwari language.

A small dataset comprising English-Marwari translation pairs was created. This dataset was crucial for training and evaluating the translation model. The dataset included common phrases and sentences in English along with their corresponding translations in Marwari. Care was taken to ensure diversity in the dataset, covering various linguistic aspects and scenarios likely to be encountered in practical usage.

	A	B
1	English	Marwari
2	Hello	राम राम सा
3	Yes	होया
4	you	थे
5	How are you?	थे कीसा हो?
6	Fine, thank you.	म्हे चोखा हां, आभार
7	thank you	आभार
8	your	थारो
9	What is your name?	थारो नाम काई हे?
10	My	म्हारो
11	My name	म्हारो नाम
12	You're welcome	थारो स्वागत है
13	No	कोनी
14	Forgive me	खम्मा घणी
15	Do you speak English?	थाने अंग्रेजी बोलणी आवे हे के?
16	Is there someone here who speaks English?	अठे कोई अंग्रेजी बोलणाळो हे के?
17	I don't understand.	म्हारी समज मांय कोनी आवे
18	I didn't get it.	हुं समज्यो कोनी।
19	where	कठे
20	Where is the toilet?	शौचालय कठे हे?
21	I love you	हुं थाने प्रेम करूं हूं।
22	Do you love me?	थे मन्ने प्रेम करो हो के?
23	where are you going?	थे कठे जा रह्या हो?
24	where do you live?	थे कठे रहवो हो?
25	what are you doing?	थे काई कर रह्या हो?
26	what happened here?	अठे काई होयो हो?

Fig 2. Dataset used in training the model

The data gathering and preparation method for the English to Marwari translation model included many important processes designed to ensure the dataset's quality and usefulness for training and assessment. Our major objective was to collect a comprehensive set of English- Marwari translation pairs that could address a wide variety of linguistic settings and circumstances found in ordinary communication. To do this, we gathered information from a variety of credible sources, including existing bilingual corpora, real conversations, and contributions from native Marwari speakers and language specialists. Once collected, the data was meticulously curated to ensure relevance, accuracy, and variety. To ensure the inclusion of culturally relevant information and idiomatic expressions unique to the Marwari language and culture, typical English phrases and sentences were paired with their equivalent translations in Marwari. To ensure that translations were accurate and to identify any discrepancies or mistakes, human reviewers who were fluent in both Marwari and English were employed. Furthermore, particular focus was placed on resolving linguistic difficulties that arise while translating from English to Marwari, including variations in sentence construction, word order, grammatical conventions, and colloquial idioms. To guarantee that the translated materials accurately conveyed the message, techniques including contextual translation and cultural subtleties were used. Ultimately, preparatory techniques like text normalisation and tokenization were used to prepare the carefully chosen dataset into an appropriate input format. The resulting dataset, characterized by its quality, diversity, and suitability, served as the foundation for training the translation model, enabling it to produce accurate and culturally appropriate translations from English to Marwari.

The process of selecting the best base model for fine- tuning the English to Marwari translation project required careful evaluation of several aspects. We aimed to identify a pre-trained model that would work well for the translation problem and be consistent with the transfer learning methodology. We selected the pre-trained Helsinki- NLP/opus-mt-en-mr model after weighing our possibilities. its proven effectiveness in low-resource language translation tasks and its compatibility with the MarianMTModel architecture, which aligns with our project requirements.

The Helsinki-NLP model is a good option for fine-tuning because of its promising performance in accurately collecting common phrases and linguistic subtleties in Marwari. The architecture and parameters of the model are useful for adapting to our translation requirement, offering us a strong basis for more improvement through fine-tuning. We chose the Helsinki-NLP/opus-mt-en-mr model as our foundation model to take advantage of its advantages and capabilities to create a strong

English-to-Marwari translation model that can translate words accurately and fluently while maintaining the linguistic and cultural subtleties of the Marwari language.

First, the pre-trained Helsinki-NLP/opus-mt-en-mr model is loaded using the MarianMTModel and MarianTokenizer classes from the transformer's library. After that, the dataset containing English sentences paired with their corresponding Marwari translations is prepared for fine-tuning. This dataset provides the base for training the model to improve its translation capabilities specifically for Marwari. Every pair of sentences is tokenized using the tokenizer, converting the text into numerical tokens that the model can comprehend. The data is then formatted into input-output pairs suitable for training. The fine-tuning process involves iterative training of the model using the prepared dataset. The model's parameters are adjusted during training to minimize translation errors, with optimization algorithms like AdamW utilized to update the parameters and reduce translation loss. This process occurs over multiple epochs, with each epoch consisting of iterations through the entire dataset. Evaluation and validation are performed periodically to monitor the model's performance and prevent overfitting, ensuring that the model generalizes well to unseen data and produces accurate translations beyond the training set. Upon completion of the fine-tuning process, the fine-tuned model's parameters are saved for future use. This specialized model is capable of accurately translating English text into Marwari, catering specifically to the requirements of the language revitalization research project.

The training and testing phases play pivotal roles in assessing and optimizing the model's performance. The training phase involves iteratively exposing the model to the prepared dataset consisting of English sentences paired with their corresponding Marwari translations. During training, the model learns to adjust its parameters to minimize translation errors and accurately map English input sequences to Marwari output sequences. This process occurs over multiple epochs, with each epoch representing a complete pass through the training dataset. Within each epoch, the dataset is typically divided into batches, allowing the model to process multiple sentence pairs simultaneously. For each batch, the model computes translation predictions for the English input sentences and compares them to the ground truth Marwari translations using a predefined loss function. Optimization algorithms, such as AdamW, are then employed to update the model's parameters based on the computed loss, gradually improving the model's translation capabilities. After completing the training phase, the model undergoes testing to evaluate its performance on unseen data. This testing phase is essential for assessing the model's generalization ability and ensuring that it can accurately translate English sentences into Marwari beyond the training dataset.

Typically, a separate dataset, known as the test set, is used for evaluation. The test set contains English sentences paired with their corresponding Marwari translations, but these examples were not seen by the model during training. During testing, the fine-tuned model generates Marwari translations for the English sentences in the test set. The quality of these translations is then assessed using various evaluation metrics, such as BLEU score, which measures the similarity between the model-generated translations and the reference translations in the test set.

In the model evaluation and testing phase, the performance of the fine-tuned model was assessed using appropriate evaluation metrics. Metrics such as the loss function value (Transformers models all have a default task- relevant loss function), translation accuracy, and fluency

were utilized to gauge the effectiveness of the model in generating accurate and linguistically sound Marwari translations. The model was subjected to testing with unseen English input sentences to generate corresponding Marwari translations. These generated translations were manually reviewed and compared against human-generated translations to assess correctness and linguistic quality accurately. This manual review process helped validate the model's performance and identify any areas for improvement. Upon satisfactory performance evaluation, the fine-tuned model was prepared for deployment. This involved saving both the model weights and tokenizer configurations using the save pretrained method, ensuring that the trained model could be easily accessed and utilized for future tasks.

The deployed model held significant potential for integration into language revitalization initiatives, where it could facilitate English-to-Marwari translation tasks. By making the model accessible to users, such as Marwari speakers and language enthusiasts, it could contribute to the preservation and promotion of the Marwari language in digital communication and educational settings.

4. Result Analysis and Validation

Our findings reveal that the model, trained on a dataset consisting of 75 translations, achieved an accuracy of 22%. This performance metric underscores the model's effectiveness in translating English sentences to Marwari, albeit with certain limitations due to the relatively small size of the training dataset. Despite these constraints, the results demonstrate the potential of transfer learning in addressing language translation tasks, particularly in low-resource language settings.

test set	BLEU	chr-F
Tatoeba.en.mr	22.0	0.397

Table-3 Benchmarks for Helsinki-NLP/opus-mt-en-mr

A BLEU score of 22.0 suggests that the model's translations exhibit a reasonable level of accuracy and fluency compared to the reference translations. Further research endeavours are warranted to explore strategies for augmenting the training data and refining the model architecture to improve translation accuracy and robustness.

This project underscores the transformative potential of transfer learning in natural language

English Text: Do you speak English?
Marwari translation: थाने अंग्रेजी बोलणी आवै है के?

English Text: Where are you?
Marwari translation: थे सिध हो?

Fig 3. Resulting translation outputs

processing, particularly in the context of language revitalization. By employing transfer learning techniques, we have leveraged the knowledge encoded in pre-trained language models and adapted it to the specific task of English to Marwari translation. Transfer learning enabled us to overcome the challenge of limited data availability for Marwari by fine-tuning a pre-trained translation model on a small dataset of English-Marwari sentence pairs. This approach not only expedited the training process but also allowed the model to capture the linguistic patterns and nuances of the Marwari language. Moreover, the use of transfer learning has broader implications for language preservation

and promotion efforts. By reducing the dependency on large, annotated datasets and lowering the barriers to entry for language technology development, transfer learning democratizes access to advanced AI capabilities and empowers language communities to reclaim and revitalize their linguistic heritage.

5. Conclusion and Future Works

5.1. Conclusion

In this work, we showed which languages have available digital resources and their related tools. Research has focused on tasks like morphology and machine translation, however, there is still a lot of work to be done since these languages exhibit a wide range of linguistic phenomena while resources are scarce.

Through this work, we discussed some of the challenges that must be taken into account, e.g., small datasets, high dialectal variation, rich morphology, lack of orthographic normalization, and scarcity of linguistic preprocessing tools.

NLP research for these languages can broaden the understanding of human language structures and help to build more general computational models. Moreover, the development of language technologies can have a positive social impact on the speakers of indigenous languages, helping to maintain the living cultural heritage that each language represents.

For several decades, communities, linguists, and language activists have designed programs to maintain and revitalise indigenous languages. However, not much is known about what causes these programs to be successful.

In this paper, we have identified initial theories that might lead to more successful revitalisation efforts by investigating how the involvement of local communities influences language revitalisation outcomes. We note that others have argued that existing language revitalisation methods have not necessarily delivered the desired results (Olawsky 2013; Roche 2020) and that there is a need for a new conceptualisation of language revitalisation (Grenoble and Whaley 2021). Based on our realist synthesis of the literature, we have proposed that language revitalisation efforts need to include a

focus on strengthening communities and promoting commitment. In other words, two necessary contexts for successful revitalisation are a strong community and a committed community.

Two implications flow from this work. First, if strengthening communities and promoting commitment are building blocks to more successful language revitalisation efforts, achieving them is as important as increasing language proficiency. In cases where language programs are not delivering desired results, addressing these factors may improve efforts.

Second, in addition to increasing language proficiency, the commissioning and evaluation of language programs might expand its scope to explicitly include strengthening communities and promoting commitment. Language activities might target an extended range of positive outcomes such as creating opportunities for the community to spend time together. Language programs could recognise and celebrate new milestones in building community strength. This, in turn, offers a formula for sustainability.

In view of this, it is clear that language revitalisation is not a mere reversal of language loss, but instead a new path forward that fosters more resilient Indigenous communities (cf. Budrikis 2021; Leonard 2017). Our work in examining the literature and drawing out these initial theories lays a foundation for new, community-based research to understand the dynamics of language revitalisation. Such research will continue unpacking the mechanisms of language revitalisation efforts, shed light on how mechanisms produce outcomes, and ensure that program design takes local aspirations seriously. The result, we hope, will be more effective programs for revitalising languages and for promoting strong identity, resilience, and well-being.

we have presented how to leverage Inuktitut-English Neural Machine Translation with morphological word segmentation. We intend to apply our proposed approach in other Indigenous language families, especially those related to the Inuit language family, to deal with NLP tasks. With the valuable collaboration of Indigenous communities, we will be able to collect reliable data from the speakers of these Indigenous languages. Moreover, our NLP applications could help preserve ancestral knowledge and revitalize Indigenous languages, heritage and culture with the transfer of knowledge from elders to the youth.

5.2. Future Works

1. Enhanced Data Collection and Annotation:

Crowdsourcing Efforts: Expand data collection efforts through crowdsourcing platforms to gather more diverse linguistic resources, including additional dialects, oral narratives, and cultural artefacts.

Annotation Tools: Develop annotation tools to facilitate the manual annotation of linguistic data by community members, linguists, and volunteers, enabling the creation of more comprehensive and high-quality annotated corpora.

2. Advanced AI Models and Techniques:

Continued Model Development: Explore the use of state-of-the-art AI models and techniques, such as transformer-based architectures and self-supervised learning approaches, to improve the performance of language processing tasks such as machine translation, speech recognition, and sentiment analysis.

Multimodal Learning: Investigate multimodal learning approaches that combine text, speech, and visual information to better capture the richness and complexity of language and cultural expression.

3. Community Engagement and Empowerment:

Capacity Building: Conduct training workshops and educational programs to empower community members with the skills and knowledge needed to actively participate in language documentation, revitalization, and technology development.

User-Centred Design: Involve community members in the co-design and iterative development of language revitalization technologies, ensuring that the tools and resources created are culturally relevant, accessible, and user-friendly.

4. Integration with Educational Initiatives:

Language Learning Platforms: Integrate AI-driven language revitalization tools and resources into formal and informal language learning programs, including schools, community centres, and online platforms, to support language acquisition and proficiency development.

Interactive Learning Experiences: Develop interactive learning experiences, such as language games, virtual reality environments, and augmented reality applications, to engage learners and foster active participation in language revitalization efforts.

5. Long-Term Sustainability and Impact:

Community-Led Sustainability Plans: Collaborate with local stakeholders to develop long-term sustainability plans for the ongoing maintenance, funding, and governance of language revitalization initiatives, ensuring their continued impact and relevance beyond the duration of the project.

Monitoring and Evaluation: Implement robust monitoring and evaluation frameworks to assess the effectiveness, reach, and cultural impact of language revitalization efforts over time, informing future decision-making and resource allocation.

6. Ethical and Responsible Innovation:

Ethical Guidelines: Establish clear ethical guidelines and best practices for conducting research and deploying AI technologies in language revitalization contexts, with a focus on cultural sensitivity, data sovereignty, informed consent, and community ownership.

Community Data Governance: Promote community-driven approaches to data governance and ownership, empowering indigenous communities to control and benefit from the use of their linguistic and cultural heritage data.

7. Cross-Disciplinary Collaboration and Knowledge Sharing:

Interdisciplinary Research Networks: Foster cross-disciplinary collaboration and knowledge exchange among researchers, practitioners, policymakers, and community members working in the fields of linguistics, AI, anthropology, education, and cultural heritage preservation.

Open Access Resources: Create open access repositories and knowledge sharing platforms to facilitate the sharing of linguistic data, tools, methodologies, and best practices within the language revitalization community and beyond.

References

- [1] Languages UNESCO WAL. (n.d.). En.wal.unesco.org. <https://en.wal.unesco.org/languages>
- [2] Bhasha Sangam - Celebrating the Linguistic Diversity of India| National Portal of India. (n.d.). <https://www.india.gov.in/spotlight/bhasha-sangam-celebrating-linguistic-diversity-india>
- [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention Is All You Need. arXiv.org. <https://arxiv.org/abs/1706.03762>
- [4] Claudia Delgado Barrios. (2022, February 21). <https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages/>
- [5] Budiono, Satwiko & Yanita, Selly & Syarfina, Tengku. (2023). Paradigm Shift of Language Revitalization in Indonesia. JURNAL ARBITRER. 10. 338-347. 10.25077/ar.10.4.338-347.2023.
- [6] Y. Zhang, J. Wang, J. Zhao, Y. LeCun, and M. Paluri, "Understanding Deep Learning Techniques for Image Captioning," IEEE Transactions on Multimedia, 2021.
- [7] Howard, J., & Ruder, S. (2018, January 1). Universal Language Model Fine-tuning for Text Classification. <https://doi.org/10.18653/v1/p18-1031>
- [8] Peters, M., Neumann, M., Iyyer, M., & Zettlemoyer, L. (2018, February 14). Deep contextualized word representations. ResearchGate. https://www.researchgate.net/publication/323217640_Deep_contextualized_word_representations
- [9] X. Wang, J. Zhang, Y. Jiang, Z. Dong, and Q. Liu, "A Survey of Deep Learning in Natural Language Processing," IEEE Transactions on Neural Networks and Learning Systems, 2021.
- [10] W. Guo, J. Wang and S. Wang, "Deep Multimodal Representation Learning: A Survey," in IEEE Access, vol. 7, pp. 63373-63394, 2019, doi:10.1109/ACCESS.2019.2916887.