

Artificial Intelligence Approaches for Preservation of Endangered Languages: A Survey

Sandhya Bangera^[1], Dr. Subrahmanya Bhat^[2]

¹*Research Scholar, Institute of Computer Science and Information Sciences, Srinivas University, Mangalore, India., ORCID ID: 0000-0003-2009-0288*

²*Professor, Institute of Computer Science and Information Sciences, Srinivas University, Mangalore, India., ORCID ID: 0000-0003-2925-1834*

Abstract - Languages are essential to human understanding and cultural heritage. However, due to migration, globalization, and a lack of internet presence, many languages are in danger of being extinct. Because they lack adequate linguistic resources, databases, and technological assistance, these languages are frequently referred to as low-resource or endangered languages. New opportunities for recording, digitizing, and renewing such languages have been made possible by recent advancements in artificial intelligence and natural language processing. By transforming speech and text into digital formats and enabling automated analysis, artificial intelligence (AI) technologies including speech recognition, optical character recognition, machine translation, and language modelling can help preserve linguistic knowledge. The contribution of artificial intelligence to the preservation of endangered languages is reviewed in this survey work. It talks about many AI methods, resources, and programs that help in language revitalization and documentation. The study also offers possible research avenues for enhancing AI-based language preservation solutions and emphasizes the difficulties posed by low-resource languages. An example of a regional language that can profit from these technologies is the Tulu language, which is briefly discussed.

Keywords: *Artificial Intelligence, Natural Language Processing, Endangered Languages, Low-Resource Languages, Optical Character Recognition, Machine Translation, Language Preservation.*

I. INTRODUCTION

One of the most crucial elements of communication and cultural identity is language. It highlights the customs, heritage, and expertise of local communities. Unfortunately, social, economic, and technological causes are causing the progressive extinction of numerous languages worldwide. Thousands of languages are currently regarded as endangered due to their lack of digital representation

and limited population, according to linguistic studies.

In recent years, the study of endangered language preservation has grown in popularity. Manual transcription, archiving, and linguistic analysis are all part of traditional language documentation techniques. These methods, however, are not scalable for massive volumes of linguistic data and demand a great deal of effort and skill.

Recent developments in natural language processing and artificial intelligence have produced strong tools for language preservation and documentation. Text, speech, and images may all be automatically processed by AI-based systems to extract useful information. These technologies can be used to create translation systems that enable endangered languages to communicate with commonly spoken languages, digitize historical records, and translate spoken language into text. Due to their lack of substantial datasets and computational resources, several endangered languages are classified as low-resource languages. These limitations are currently being addressed by AI models via data augmentation, multilingual modelling, and transfer learning.

Tulu, a Dravidian language spoken in areas of southern India, is one example of a regional language that can profit from such technologies. Despite having a rich cultural past, Tulu has little technological and digital resources. Its digital preservation can be aided by AI-based techniques like machine translation and handwritten script recognition.

Challenges in Preserving Endangered Languages
Preserving endangered languages is a complex task due to several challenges. These challenges arise from both linguistic and technological limitations.

1. **Lack of Digital Resources:** Most endangered languages lack digital text corpora, dictionaries, and annotated datasets. Machine learning models require large amounts of data for training, which makes it difficult to develop AI-based tools for such languages.
2. **Limited Linguistic Documentation:** Many languages exist primarily in oral form and have limited written documentation. As a result, researchers must rely on field recordings and manual transcription to collect linguistic data.
3. **Script and Orthographic Variations:** Some languages have multiple writing systems or inconsistent spelling conventions. This variation makes it difficult to develop standardized computational models.
4. **Scarcity of Native Speakers:** The number of native speakers of endangered languages is often declining. This limits opportunities for data collection and linguistic validation.
5. **Lack of Technological Infrastructure:** Many communities speaking endangered languages have limited access to technological tools and digital platforms. As a result, language preservation efforts may not receive adequate technological support.

These challenges highlight the need for advanced AI-based solutions that can work with limited data and support multilingual environments.

Globally, linguistic variety is fast disappearing. Nearly 40% of the world's languages are in danger of going extinct, according to UNESCO studies. Small communities speak many of these languages, which are mainly passed down orally. The impending doom of minority languages has been accelerated by migration, globalization, and the dominance of widely spoken languages. Dictionaries, grammar manuals, and digital archives are examples of formal documentation that endangered languages frequently lack. Important language expertise may be irreversibly lost as older generations die. Loss of cultural identity, traditional knowledge, and historical narratives connected to those languages is another consequence of language extinction.

Promising solutions to this problem are provided by the development of digital technology and artificial intelligence. By transforming speech, text, and manuscripts into digital representations, AI systems can aid in the documentation and preservation of languages. With the use of these technologies,

linguistic data may be stored and shared with communities and scholars worldwide.

Artificial Intelligence Techniques for Language Preservation

Artificial intelligence provides several computational approaches for preserving endangered languages. These approaches involve analysing speech, text, and images to create digital linguistic resources.

- **Speech Recognition:** Spoken language is translated into printed form using speech recognition technologies. When it comes to recording languages with robust oral traditions, these approaches are very helpful. Speech recognition performance has been greatly enhanced by deep learning models like transformer-based architectures and recurrent neural networks. These systems can automatically produce transcripts that can be saved in digital archives after being trained on recorded voice data. Additionally, speech recognition makes it possible to develop voice-activated programs that let speakers of endangered languages communicate with electronic devices.
- **Optical Character Recognition:** Text images can be transformed into machine-readable characters using optical character recognition (OCR) technology. OCR is crucial to the digitization of handwritten documents and ancient manuscripts. Convolutional neural networks are used by deep learning-based OCR models to recognize characters and patterns in pictures. By transforming old manuscripts and printed documents into searchable digital text, OCR can aid in the preservation of endangered languages. For languages with distinctive scripts or handwritten documentation, OCR technologies are especially helpful.
- **Machine Translation:** Text can be automatically translated between languages by machine translation systems. By utilizing transformer designs and attention mechanisms, neural machine translation models have greatly increased translation accuracy. By translating literature into commonly spoken languages like English, machine translation can help close communication gaps for endangered

languages. The creation of multilingual dictionaries and language learning resources is also made possible by this technology. However, the lack of training data makes machine translation for low-resource languages still difficult.

- **Language Modelling:** AI programs known as "language models" are capable of learning a language's statistical patterns of words and phrases. These models are able to produce text, fix grammatical mistakes, and predict the next word in a sentence.

Text generation systems for endangered languages, grammar checkers, and predictive keyboards can all benefit from language models. They can also help researchers create instructional materials and analyze linguistic trends. The capacity of AI systems to comprehend and produce human language has been substantially enhanced by recent developments in transformer-based language models.

AI Tools and Technologies

Language preservation systems are developed using a variety of AI frameworks and technologies. Neural networks for text and speech processing can be trained with the help of deep learning libraries like PyTorch and TensorFlow. Text analysis activities like tokenization, part-of-speech tagging, and named entity recognition are supported by natural language processing libraries like NLTK and spaCy. Additionally, scalable platforms for handling massive amounts of linguistic data are offered by cloud-based AI services.

Language documentation has changed as a result of artificial intelligence's ability to process and analyze data automatically. AI systems are capable of effectively extracting linguistic information from massive amounts of audio recordings, text documents, and images.

Oral traditions and conversations can be more easily documented thanks to speech recognition technologies that translate spoken language into written text. In a similar vein, handwritten

manuscripts and historical writings can be digitized by optical character recognition technologies and archived in digital archives. Additionally, machine learning algorithms can spot trends in linguistic data, which aids academics in comprehending vocabulary usage and grammar structures. Linguists can use these insights to create resources for language learning, grammar rules, and dictionaries. Because AI-driven systems may use transfer learning methods and multilingual datasets, they are especially useful for low-resource languages. This makes it possible for researchers to create language processing systems even in situations where data is scarce.

Artificial intelligence has the ability to save endangered languages, as several research projects have shown. For instance, speech recognition algorithms have been created to automatically transcribe recorded speech data in order to document indigenous languages. These systems aid in the creation of digital archives that conserve spoken stories and oral traditions.

Utilizing OCR technologies to digitize historical documents written in antiquated scripts is another significant use. Accessing and analyzing many cultural and religious texts is challenging because they are only available in handwritten form. These manuscripts can be transformed into machine-readable text by AI-based OCR systems, allowing academics to analyze and preserve them more successfully.

In order to improve communication between speakers of minority languages and the larger global community, machine translation techniques have also been applied to low-resource languages. These technologies aid in improving accessibility and fostering language acquisition by translating content between endangered languages and commonly spoken languages like English.

II. REVIEW OF EXISTING RESEARCH

Several studies have explored the use of AI techniques for endangered language preservation. These studies focus on speech documentation, machine translation, and digital archiving.

Study	Technique	Language	Contribution
Study 1	Speech recognition	Indigenous languages	Speech transcription
Study 2	Neural machine translation	Low-resource languages	Translation systems

Study 3	Deep learning OCR	Historical scripts	Manuscript digitization
Study 4	Language modeling	Minority languages	Predictive text systems

Table 1: use of AI techniques for endangered language preservation

These findings show that by automating numerous linguistic processes, AI can greatly aid language preservation efforts.

The performance of machine translation has been greatly enhanced by recent developments in transformer-based models, including those put out by Ashish Vaswani. In a similar vein, studies conducted by James H. Martin and Daniel Jurafsky emphasize the significance of computational linguistics methods for language analysis.

Digital Archives and Language Resource Creation

One important factor in the preservation of endangered languages is the development of digital language resources. Linguistic assets like written texts, audio recordings, dictionaries, and cultural narratives can be stored in organized ways by scholars and communities using digital archives. Scholars, educators, and language learners can use these archives, which guarantee the preservation of linguistic knowledge for future generations. The creation of digital archives is considerably aided by artificial intelligence technologies. By recognizing patterns and classifying content, machine learning algorithms are able to autonomously arrange vast amounts of linguistic data. For instance, recorded conversations can be transformed into text transcripts by speech recognition systems, which can subsequently be indexed and kept in searchable databases.

The automatic annotation of linguistic material is another significant contribution of AI. Words can be labeled with grammatical categories like nouns, verbs, and adjectives using natural language processing algorithms. Linguists use these comments to examine language structures and create instructional resources like dictionaries and grammar manuals. Collaborative study is also made possible by digital archives. Linguists, computer scientists, and community members can contribute data to shared repositories, creating richer language resources. Multimedia content, including as pictures, movies, and audio recordings, can also be supported by these archives, offering a more thorough depiction of the language and its cultural context.

Additionally, the availability of digital materials promotes the creation of mobile tools and applications for language learning. Through interactive activities, pronunciation guidance, and vocabulary training, these programs can assist younger generations in learning endangered languages. Sustainable solutions for language revitalization and preservation can be developed by fusing digital archiving technology with artificial intelligence.

Community Participation in Language Preservation

For language preservation programs to be successful, community involvement is crucial. The participation of local communities and native speakers guarantees that preservation initiatives stay genuine and culturally significant, even while technology tools offer invaluable assistance.

For training and assessment, artificial intelligence systems need a lot of language data. Members of the community can help create these datasets by contributing written texts, translations, and speech recordings. Speakers can actively contribute to the documentation of their language through crowdsourcing platforms and community-driven data collection techniques.

The accuracy of AI models is frequently enhanced via community input. Transcriptions and translations produced automatically can be reviewed by native speakers, who can offer comments and fix mistakes. This procedure guarantees that language subtleties are appropriately captured and aids in the improvement of machine learning models.

Initiatives for community-based language preservation also encourage younger generations to take pride in their culture. People become more connected to their cultural history when they actively contribute to the documentation of their language. Community members might be further encouraged to adopt digital technologies for language documentation through educational programs and workshops.

Particularly crucial are partnerships between local communities and academic scholars. While community members provide important linguistic and cultural insights, researchers provide technical competence in computational linguistics and artificial intelligence. These collaborations produce a well-rounded strategy that blends cultural awareness with technology innovation.

In general, community involvement guarantees that technology solutions are in line with the requirements and values of the communities engaged and increases the efficacy of AI-based language preservation initiatives.

Applications of AI-Based Language Preservation Systems

There are a number of useful uses for artificial intelligence technology that support the revitalization and preservation of endangered languages. These programs offer practical tools for communication, teaching, and cultural documentation that go beyond scholarly research.

The creation of platforms for language learning is one such application. Through interactive activities, pronunciation feedback, and vocabulary suggestions, AI-powered language learning systems can offer individualized learning experiences. These platforms promote the continuous use of endangered languages and aid learners in improving their language proficiency.

Conversational agents and chatbots that facilitate communication in minority languages are another use. By enabling users to communicate with digital platforms in their mother tongue, these solutions encourage regular use and digital inclusion. Additionally, machine translation systems are crucial in helping speakers of endangered languages communicate with the rest of the world. AI technologies make information and educational resources more accessible by facilitating translation between minority languages and commonly spoken languages. By spotting linguistic usage trends, AI-based text analysis techniques can help advance linguistic study. These resources help linguists research vocabulary growth, grammar structures, and language evolution.

Additionally, the development of multimedia language documentation initiatives can be aided by

artificial intelligence. By producing spoken copies of written texts, speech synthesis technologies increase the accessibility of language learning resources. In a similar vein, textual inscriptions connected to a language and cultural items can be analyzed by image recognition algorithms.

These uses show how artificial intelligence may be a potent instrument for fostering linguistic diversity and guaranteeing the continued relevance of endangered languages in the digital era.

Benefits of AI for Language Preservation

Language preservation initiatives can benefit from artificial intelligence in a number of ways.

1. AI technologies make it possible to process vast amounts of linguistic data quickly, which would otherwise take a lot of manual labour. The amount of time needed to document languages can be greatly decreased by using automated transcribing and translation systems.
2. Digital archives that contain linguistic resources in organized formats can be created with the help of AI tools. Future generations will be able to access language data thanks to these archives.
3. Through interactive applications and language learning platforms, AI-powered educational solutions can assist younger generations in learning and practicing endangered languages.
4. By fostering cooperation between linguists, technologists, and local communities, AI-based solutions enable more successful language preservation techniques.

Ethical Considerations in AI for Language Preservation

Even though artificial intelligence offers effective tools for language preservation, ethical issues need to be taken into account. With the permission of the communities involved, researchers must make sure that linguistic data is gathered and used appropriately.

Numerous endangered languages have strong ties to local knowledge and cultural customs. As a result, it's critical to honour the cultural sensitivities connected to these languages. To make sure that language documentation initiatives are in line with community objectives, researchers should work with native speakers and community leaders. Privacy and data ownership are also significant issues. Linguistic information should only be communicated with the proper authorization and kept in a secure location. Technological improvements that promote language

preservation without undermining cultural values are made possible by ethical AI practices.

III. FUTURE RESEARCH DIRECTIONS

There are a number of intriguing avenues for future AI-based language preservation research. The creation of multilingual AI models that can support several low-resource languages at once is one significant avenue. With little data, transfer learning and cross-lingual representation learning can assist enhance model performance. The development of community-driven datasets, in which native speakers provide speech recordings, text documents, and translations, is another exciting field. Programs for language acquisition and revitalization can potentially benefit from developments in conversational AI and speech synthesis. Younger generations can have more access to endangered languages through educational applications and digital archives. Lastly, the long-term viability of language preservation programs depends on cooperation between local communities, governments, and academic institutions.

IV. CONCLUSION

Global cultural variety and linguistic history are seriously threatened by the swift demise of endangered languages. However, new possibilities for the preservation of these languages through digital technology have been made possible by recent developments in artificial intelligence and natural language processing. This assessment examined a number of AI methods for language preservation, such as machine translation, optical character recognition, speech recognition, and language modelling. Automated text processing, the digitalization of linguistic resources, and the creation of tools that facilitate language learning and communication are all made possible by these technologies. The viability of creating computational tools for low-resource languages is being improved by ongoing AI research, despite obstacles including small datasets and a lack of technological infrastructure. It is feasible to develop long-term solutions that protect endangered languages for coming generations by fusing artificial intelligence with community-driven language documentation initiatives.

REFERENCES

- [1] Pakray, P., Gelbukh, A., & Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2), 183-197.
- [2] Tesfagergish, S. G. (2025). *AI-driven strategies for NLP challenges in low-resource languages* (Doctoral dissertation, Kauno technologijos universitetas.).
- [3] Ihsan, M. B. (2026). Artificial Intelligence for Preserving Indonesian Regional Languages: Machine Learning Approaches to Documenting Endangered Dialects and Cultural Linguistic Heritage. *JIRAN: Journal of Southeast Asia Studies*, 7(1), 1-22.
- [4] Mullah, N. S., & Zainon, W. M. N. W. (2026). AI and Low-Resource Languages: Bridging the Linguistic Divide. In *Reshaping Language and Cognition in Education Through AI* (pp. 77-128). IGI Global Scientific Publishing.
- [5] Baloch, S., & Riaz, H. (2025). AI Applications for Endangered Language Preservation. *Social Thought and Policy Review*, 3(1), 61-82.
- [6] Nigatu, H. H. (2025). *Building Language Technologies for Low-Resourced Languages* (Doctoral dissertation, UC Berkeley).
- [7] Emezue, C. C. (2026). Improving language models for underserved languages and communities.
- [8] Lucas, M., Burgueño, A., Carazas, M., Sánchez, C. B., Rufino, S. R., Torres, C. S. R., ... & Deo, G. (2025). The Performance of Artificial Intelligence in the Use of Indigenous American Languages.
- [9] Daniel, J. E. (2020). *Applications of natural language processing for low-resource languages in the healthcare domain* (Doctoral dissertation, Stellenbosch: Stellenbosch University.).
- [10] Shetty, P. (2024). Natural Language Processing for Tulu: Challenges, Review and Future Scope. In *International Conference on Speech and Language Technologies for Low-resource Languages* (pp. 93-109). Springer, Cham.
- [11] Agarwal, M. (2025). *Improving Resource Creation for Low-Resource Languages Using NLP Methods* (Doctoral dissertation, George Mason University).

- [12] Ikolo, R. O. (2025). Empowering endangered African languages: AI-driven preservation and translation in Africa. *EXISTENTIA: CONCRESCENCE JOURNAL OF ARTS AND HUMANITIES*, 3(2).
- [13] Zhong, T., Yang, Z., Liu, Z., Zhang, R., You, W., Liu, Y., ... & Liu, T. (2024). Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv preprint arXiv:2412.04497*.
- [14] Sun, H., Tang, T., Feng, J., Yang, Z., & Guo, B. (2025, April). Overview of Research on Low-Resource Language Machine Translation Based on Artificial Intelligence. In *2025 6th International Conference on Computer Engineering and Application (ICCEA)* (pp. 01-05). IEEE.
- [15] Akomolede, K. K., & Olowojebutu, A. O. (2025). Enhancing Digital Inclusion through AI-Based Yorùbá Language Localization: Challenges, Solutions, and Future Prospects. *Tech-Sphere Journal for Pure and Applied Sciences*, 2(1).