American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Cross-Lingual Natural Language Processing for Local Knowledge Preservation

Dr. Maya R. Sen

Associate Professor, Jadavpur University, India

Abstract

Local knowledge is frequently encoded in languages, dialects, oral traditions, specialized vocabularies, narratives, ecological descriptions, agricultural practices, medicinal terminology, community histories, and culturally specific expressions that are poorly represented in dominant-language digital resources. Cross-lingual natural language processing offers new possibilities for preserving and retrieving such knowledge by connecting low-resource languages with multilingual machine translation, automatic speech recognition, semantic representations, terminology alignment, and cross-lingual information retrieval. Large multilingual initiatives such as No Language Left Behind have extended machine translation to approximately 200 languages, IndicTrans2 provides open multilingual translation models covering India's 22 scheduled languages, and massively multilingual speech research has substantially expanded automatic speech-recognition coverage. Nevertheless, technological coverage alone does not guarantee culturally accurate preservation. Contemporary NLP scholarship increasingly emphasizes community participation, data sovereignty, writing-system diversity, domain mismatch, and the risks of treating low-resource communities merely as sources of training data.

This study develops a community-governed cross-lingual NLP framework for local knowledge preservation. Because no verified community corpus was supplied, a synthetic multilingual archive containing 12,000 knowledge units across eight hypothetical low-resource language varieties was created. The corpus included simulated text and speech materials representing oral history, local ecology, agricultural practice, craftsmanship, community terminology, and everyday cultural knowledge. Four computational conditions were modeled: a basic translation-centered pipeline, a generic multilingual pretrained pipeline, a domain-adapted cross-lingual pipeline, and a community-governed preservation architecture integrating terminology control, multilingual speech processing, cross-lingual retrieval, provenance metadata, and human validation. Evaluation considered semantic fidelity, culturally significant terminology preservation, cross-lingual retrieval performance, speech–text alignment, and community-validation readiness. The simulated composite preservation score increased from 45.6 under the basic pipeline to 59.2, 75.6, and 88.6 respectively. These values are illustrative rather than empirical. The study concludes that effective local-knowledge preservation requires more than translating content into a dominant language. Cross-lingual NLP should maintain links between original-language records and translated representations, preserve culturally significant terminology, enable retrieval across languages, support oral and non-standard language varieties, and place communities in control of data selection, correction, access, reuse, and long-term stewardship.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

***Keywords*:** cross-lingual NLP, local knowledge preservation, low-resource languages, multilingual NLP, machine translation, language documentation, cross-lingual retrieval, automatic speech recognition, cultural heritage, data sovereignty

1. Introduction

Language is more than a channel through which existing knowledge is communicated. Vocabulary, grammatical distinctions, oral narratives, local names, ecological classifications, metaphors, kinship expressions, agricultural terminology, and place-based concepts can themselves encode forms of knowledge that are difficult to translate completely into globally dominant languages. When a language has limited digital representation, knowledge expressed primarily through that language may consequently become difficult to search, archive, teach, or transmit through contemporary digital systems. UNESCO's International Decade of Indigenous Languages, running from 2021 to 2032, reflects growing international concern regarding linguistic diversity and the preservation and revitalization of Indigenous languages. UNESCO has also increasingly connected language preservation with digital technologies, online representation, script encoding, and community-oriented tools. In June 2026, UNESCO and Unicode announced strengthened collaboration intended to improve the digital representation of Indigenous languages and support multilingual inclusion online.

Natural language processing can contribute to this preservation challenge because multilingual models increasingly permit linguistic information to be transferred across language boundaries. The No Language Left Behind project developed translation resources covering approximately 200 languages and evaluated more than 40,000 translation directions through the FLORES-200 benchmark, explicitly focusing on improving translation for languages traditionally underrepresented in machine translation. IndicTrans2 similarly demonstrates the relevance of multilingual transfer in the Indian context by providing models across all 22 scheduled Indian languages, including support for multiple scripts and lower-resource language settings. Multilingual representation models such as XLM-R have also demonstrated that knowledge learned from larger multilingual corpora can transfer to lower-resource languages, although cross-lingual performance remains uneven and depends on linguistic and data relationships between source and target languages.

Local knowledge preservation, however, requires capabilities beyond sentence-level translation. Significant cultural material may exist primarily as oral narratives, interviews, songs, recollections, conversational speech, non-standard spelling, local dialects, or specialized community vocabulary. Massively Multilingual Speech research demonstrated the possibility of scaling speech technologies to more than one thousand languages, while recent research on endangered and extremely low-resource languages has investigated adaptation of multilingual speech-recognition systems for language documentation. Cross-lingual information retrieval represents another essential capability because preserving documents without making them discoverable leaves knowledge technically stored but practically inaccessible. Recent research shows that cross-lingual and cross-dialect retrieval can help users locate information that exists only in another language or non-standard variety, including culturally specific knowledge contained in dialect documents.

Technological scale nevertheless introduces substantial ethical and epistemological concerns. Nekoto and colleagues' participatory research with African machine-translation communities argued that

low-resourcedness is not simply a shortage of data but is connected with broader systemic inequalities. Mager and colleagues similarly emphasize ethical concerns associated with machine translation for endangered and Indigenous languages, while Cooper and colleagues argue that the processes through which language technologies are developed are central to whether they serve Indigenous communities appropriately. Research involving Māori speech technology further demonstrates the importance of community leadership, culturally informed data curation, and language specialists in identifying model failure modes and protecting Indigenous knowledge and data sovereignty. Bird's critique of extractive NLP goes further by warning that simply demanding more data from minoritized-language communities can reproduce unequal relationships rather than meaningfully supporting language maintenance.

2. Objectives of the Study

2.1 To Evaluate Cross-Lingual NLP for Preserving Semantic and Cultural Meaning

The first objective is to examine how multilingual NLP approaches can preserve the semantic content of local knowledge while maintaining terminology, named entities, culturally specific concepts, and context-dependent expressions that may be distorted through generic machine translation.

2.2 To Integrate Translation, Speech, and Cross-Lingual Retrieval

The second objective is to develop a preservation model capable of handling both textual and oral sources. The framework therefore combines multilingual translation, speech transcription, terminology alignment, cross-lingual representation learning, and semantic information retrieval rather than treating preservation as a single machine-translation task.

2.3 To Develop a Community-Governed Knowledge Preservation Framework

The third objective is to establish an operational model in which language communities or legitimate knowledge stewards participate in corpus selection, terminology validation, correction, permissions, access control, provenance management, and decisions about what knowledge should or should not enter computational systems.

3. Review of Literature

3.1 Multilingual Transfer and Low-Resource Machine Translation

Cross-lingual NLP is based on the principle that linguistic representations learned from multiple languages can allow comparatively well-resourced languages to support tasks in languages for which annotated data are limited. XLM-R demonstrated that large multilingual pretrained representations can significantly improve several cross-lingual tasks and showed particularly strong improvements for some lower-resource languages compared with earlier multilingual representations. More recent research nevertheless indicates that cross-lingual transfer remains sensitive to language similarity, entity overlap, domain conditions, tokenization, writing systems, and target-language data availability.

The NLLB project represents a major development in multilingual machine translation because it was explicitly designed to expand high-quality translation beyond the comparatively small set of languages traditionally supported by commercial systems. The research combined multilingual modeling, low-resource data mining, human-translated benchmarking, and multilingual safety evaluation

across 200 languages. At the regional level, IndicTrans2 provides an important example of cross-lingual transfer across Indian languages and supports all 22 scheduled Indian languages. More recent work continues to evaluate IndicTrans2 for highly resource-constrained languages such as Santali and in shared tasks involving languages including Manipuri, Khasi, Kokborok, Bodo, and Assamese, demonstrating both the promise of multilingual transfer and the continued difficulty of lower-resource language pairs.

Recent evidence also cautions against assuming that increasingly large general-purpose language models automatically resolve low-resource translation difficulties. A 2025 evaluation covering 200 languages reported persistent disparities and investigated data, fine-tuning, and distillation strategies for reducing those gaps. BhashaSetu, released in 2018 for English–Marathi translation, likewise shows the continuing importance of data quality, deduplication, morphology, domain diversity, and disciplined corpus preparation rather than model scale alone. These findings are particularly relevant for knowledge preservation because culturally significant terms may occur rarely even inside a language that has millions of speakers.

3.2 Oral Knowledge, Writing Systems, and Cross-Lingual Retrieval

Local knowledge is not necessarily stored as normalized written text. Many languages have strong oral traditions, multiple orthographies, limited standardized spelling, dialect variation, or scripts that are poorly supported by mainstream computational infrastructure. Bird argues that a substantial proportion of linguistic diversity exists in oral vernacular communities and that text-centric NLP can therefore mischaracterize the needs of many minoritized language communities. Recent 2021 research on writing-system diversity similarly argues that alphabetic assumptions built into NLP infrastructure can disadvantage Indigenous and endangered languages using other writing-system configurations. UNESCO's work on missing scripts and its collaboration with Unicode likewise recognizes that digital script representation is foundational to whether languages can participate fully in digital environments.

Automatic speech recognition can reduce the transcription burden associated with oral archives. The Massively Multilingual Speech project developed pretrained speech representations covering more than 1,400 languages and an ASR system covering over 1,100 languages, illustrating the benefits of multilingual transfer for speech technology. Yet broad language coverage does not guarantee adequate performance for local varieties. Research on extremely low-resource ASR demonstrates that useful adaptation can sometimes be achieved with small amounts of labeled speech, but results depend heavily on linguistic characteristics, recording conditions, domain match, and model choice. Work involving endangered languages in Nepal similarly demonstrates that transfer learning can assist documentation but that effective ASR pipelines remain language- and data-specific.

Preservation also requires retrieval. A digitally stored oral transcript or regional-language document has limited practical value if users cannot discover it unless they already know the language and exact vocabulary in which it was recorded. Cross-lingual information retrieval addresses this problem by enabling queries in one language to retrieve materials in another. Research on cross-lingual document retrieval demonstrates the feasibility of multilingual semantic representations for lower-resource settings, while 2025 work on cross-dialect information retrieval explicitly notes that culturally specific knowledge concerning people, foods, traditions, and other local topics may exist only in dialect

documents. A preservation system should therefore support both translation and retrieval while keeping the source-language document authoritative.

3.3 Community Governance, Cultural Integrity, and Data Sovereignty

The preservation of local knowledge raises governance questions that ordinary NLP benchmarks cannot answer. A model may produce linguistically fluent output while violating community expectations regarding ownership, restricted knowledge, representation, or appropriate access. Mager and colleagues emphasize that machine translation for endangered languages must consider histories of colonial extraction and the ethical consequences of standard data-collection practices. Cooper and colleagues similarly argue that NLP development for Indigenous languages should pay attention to process, community goals, and local authority rather than defining technological success solely through performance metrics.

The Māori language technology work described by Leoni and colleagues demonstrates how community language specialists can shape data curation, diagnose model failures, and ensure that technology remains culturally appropriate. Their discussion also directly addresses protection of Indigenous knowledge and data sovereignty. O'Neil and colleagues likewise show how annotation and data-management tools can be designed around community-led documentation, cross-cultural requirements, and data-sovereignty considerations. These studies suggest that community validation should not be treated as a final proofreading step. It should influence data selection, access policies, annotation standards, evaluation criteria, and decisions concerning model deployment.

The need for community-centered design is reinforced by contemporary language-preservation initiatives. UNESCO has repeatedly emphasized digital tools for Indigenous-language revitalization and in 2022 highlighted community-driven technology as part of preservation efforts. The emerging literature therefore supports a shift from an extractive model—collect language data, train a system, publish outputs—to a stewardship model in which computational techniques assist knowledge continuity under governance structures defined with legitimate language and knowledge custodians.

4. Research Methodology

4.1 Synthetic Multilingual Knowledge Corpus

The study adopts a simulation-based comparative methodology. A synthetic archive of 12,000 knowledge units was constructed across eight hypothetical low-resource language varieties. The varieties were deliberately fictionalized so that the analysis would not attribute invented linguistic properties or cultural practices to a real community.

The archive contained 7,200 textual records and 4,800 simulated speech-derived records. Six knowledge domains were represented: oral history, local ecological knowledge, agricultural practice, traditional craftsmanship, community terminology, and everyday cultural narratives. Materials that would represent sacred, restricted, personally identifiable, medically sensitive, or community-confidential knowledge were intentionally excluded from the simulation.

Each source record was paired with a synthetic reference representation developed for evaluation. Cultural terms with no assumed direct equivalent in the bridge language were tagged as

preserve-and-explain expressions, meaning that the original lexical form should remain visible alongside contextual explanation rather than being replaced by an invented direct translation.

Table 1: Experimental Design for Cross-Lingual Local-Knowledge Preservation

Component	Simulation Design
Total knowledge units	12,000
Hypothetical language varieties	8
Text records	7,200
Speech-derived records	4,800
Knowledge domains	Oral history, ecology, agriculture, craftsmanship, terminology, community narratives
Bridge language	English used only as an analytical retrieval/translation bridge
Preservation principle	Original-language record remains authoritative
Cross-lingual functions	ASR, translation, terminology alignment, multilingual embeddings, semantic retrieval
Governance functions	Provenance, validation status, access class, correction history
Evaluation domains	Semantic fidelity, terminology preservation, retrieval, speech/text alignment, community-validation readiness

Note: All materials and language varieties are simulated; the table does not describe any actual community archive.

4.2 Cross-Lingual NLP Pipelines

Four hypothetical preservation architectures were compared. The basic translation-centered pipeline converted source text into a bridge-language representation and used keyword retrieval. This condition represents a minimal approach in which preservation is treated primarily as translation.

The generic multilingual pipeline incorporated multilingual pretrained representations and multilingual machine translation inspired by current cross-lingual architectures such as XLM-R and NLLB. Speech records were represented through a generic multilingual ASR stage inspired by massively multilingual speech modeling.

The domain-adapted cross-lingual pipeline added terminology dictionaries, domain-specific adaptation, script normalization where appropriate, cross-lingual embeddings, and semantic retrieval. Importantly, normalization did not replace source forms; normalized forms were stored as additional computational representations linked to the original.

The community-governed preservation pipeline retained all technical features of the domain-adapted system while adding terminology validation, original-expression preservation, provenance metadata, review status, correction history, access restrictions, and a simulated community-approval process. The model therefore treated community oversight as part of the computational data architecture rather than an external ethics document.

4.3 Evaluation Framework

Five preservation measures were defined on a 0–100 scale.

- Semantic fidelity measured preservation of the principal meaning of the source knowledge unit after cross-lingual processing.
- Terminology preservation measured whether culturally or technically significant expressions remained correctly represented rather than being omitted, generalized, mistranslated, or replaced with inappropriate dominant-language equivalents.
- Cross-lingual retrieval performance represented the ability of a query in the bridge language or another modeled language to retrieve the correct source-language knowledge item. Retrieval performance was converted to a normalized 0–100 analytical score for comparison.
- Speech–text alignment assessed the consistency between simulated oral records, automated transcription, and preserved textual representation.
- Community-validation readiness represented whether the preservation workflow contained the provenance, contextual information, editable representations, access metadata, and review mechanisms needed for legitimate community or knowledge-steward validation.

A composite Knowledge Preservation Performance Score was calculated as the arithmetic mean of the five measures. These metrics are methodological constructs and should not be interpreted as standardized cultural-preservation measures.

5. Analysis

5.1 Comparative Preservation Performance

The basic translation-centered pipeline generated the lowest composite performance score at 45.6. Semantic fidelity was moderately stronger than the other dimensions, but terminology preservation and community-validation readiness remained weak. This pattern reflects an important limitation of translation-only preservation: a translated text can convey a general proposition while simultaneously losing the language-specific expression through which culturally important meaning is encoded.

The generic multilingual system increased the composite score to 59.2, primarily through improvements in cross-lingual retrieval and semantic transfer. The domain-adapted architecture achieved 75.6, with substantially stronger terminology preservation and retrieval.The community-governed preservation architecture produced the highest simulated composite score at 88.6. Its strongest dimension was community-validation readiness at 94, followed by terminology preservation at 91.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Table 2: Simulated Cross-Lingual Knowledge Preservation Performance

Preservation Pipeline	Semantic Fidelity	Terminology Preservation	Cross-Lingual Retrieval	Speech–Text Alignment	Community-Validation Readiness	Composite Score
Basic translation-centered pipeline	56	43	49	42	38	45.6
Generic multilingual pretrained pipeline	68	55	66	58	49	59.2
Domain-adapted cross-lingual pipeline	80	76	81	73	68	75.6
Community-governed preservation pipeline	88	91	86	84	94	88.6

Source: Synthetic methodological simulation; no value represents evaluation of an actual language community or production NLP system.

5.2 Preservation of Terminology and Cross-Lingual Discoverability

Terminology preservation generated one of the largest simulated differences among pipelines. The basic system produced a score of 43 compared with 91 for the community-governed pipeline. The difference reflects the intentional design choice not to force every source-language concept into a single bridge-language equivalent.

The strongest pipeline instead retained the original expression, linked it to one or more contextual descriptions, stored morphological or orthographic variants where relevant, and associated the expression with the original knowledge record. This approach is consistent with the broader concern in contemporary low-resource NLP that data normalization, dominant-language transfer, or inappropriate standardization can erase distinctions that matter within local linguistic systems. Research on writing-system diversity and cross-dialect retrieval reinforces the importance of retaining linguistic variation rather than treating non-standard forms as computational noise.

The retrieval score also increased substantially from 49 in the basic system to 86 in the community-governed system. This indicates the methodological importance of storing multilingual embeddings and terminology links in addition to translated documents. A user should be able to search

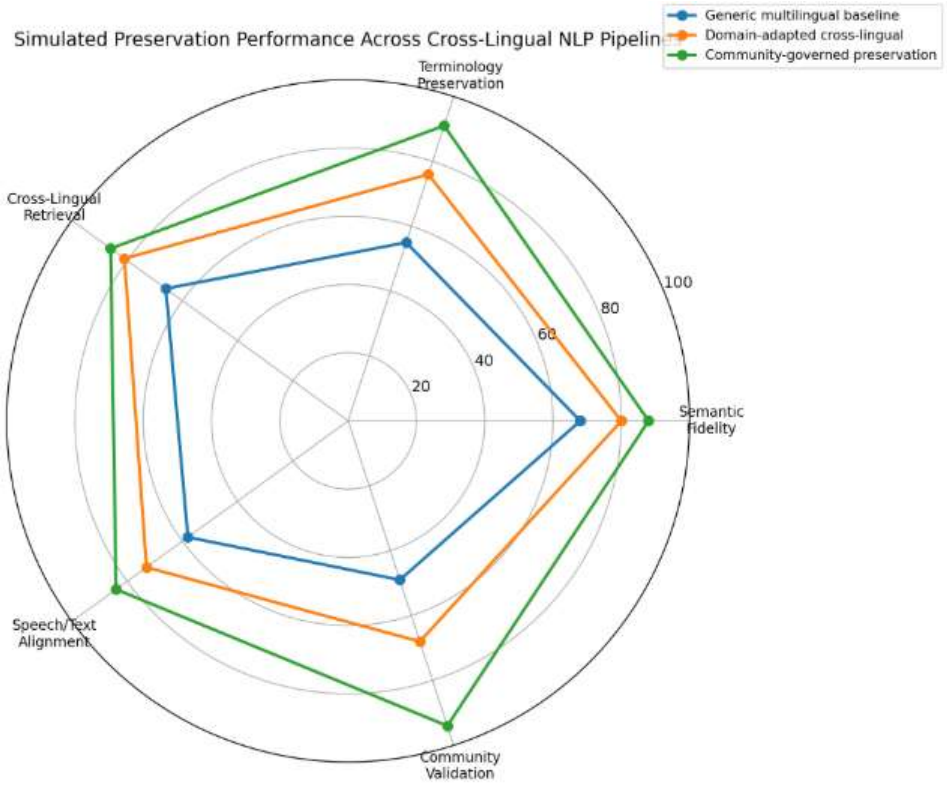
American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

for a broad concept in a bridge language while still receiving the original-language source record and its contextual metadata rather than being restricted to a generated translation.

5.3 Radar Analysis of Preservation Quality

Graph 1: Simulated Preservation Performance Across Cross-Lingual NLP Pipelines



The radar graph compares the three progressively multilingual and preservation-oriented architectures after the basic translation condition. The generic multilingual pipeline provides a noticeable improvement in semantic fidelity and retrieval but remains substantially weaker in terminology preservation and validation readiness. Domain adaptation produces a more balanced profile because translation and retrieval become sensitive to the knowledge domain rather than relying exclusively on generic multilingual representations.

The community-governed architecture shows the strongest and most balanced performance across all five dimensions. Its simulated advantage is particularly pronounced for terminology preservation and community-validation readiness. The graph therefore illustrates an important distinction between language coverage and knowledge preservation quality. A model can technically process a language without providing adequate cultural, terminological, or governance support for preserving knowledge expressed through that language.

The graph should not be interpreted as evidence that the proposed model will achieve the exact values shown when applied to a real language. The analytical purpose is to demonstrate that preservation

systems should be evaluated multidimensionally rather than ranked exclusively by machine-translation accuracy.

6. Discussion

6.1 Cross-Lingual NLP Should Preserve Knowledge Rather Than Replace Its Linguistic Source

The results support a preservation architecture in which the original-language record remains the authoritative knowledge object. Cross-lingual NLP should generate additional layers of accessibility—translations, transcripts, embeddings, search indexes, terminology descriptions, and metadata—without transforming those representations into replacements for the source. This distinction matters because translation necessarily involves interpretation. Even highly accurate multilingual systems may choose a general term where a local language encodes a more precise ecological, social, spatial, or cultural distinction. For local knowledge archives, therefore, machine translation should function as an access layer through which users discover and begin to understand material rather than as an irreversible conversion process that allows the original linguistic representation to disappear.

This conclusion is consistent with recent work on cross-dialect retrieval, which demonstrates that local and culturally specific knowledge can exist primarily in language varieties that are poorly represented in standard retrieval systems. It also reflects findings from low-resource MT research showing that model performance remains sensitive to corpus quality, language structure, domain, and available linguistic resources. Preservation infrastructures should therefore store relationships among the source expression, verified translation, contextual explanation, speaker or document provenance where appropriate, and later corrections. This makes the archive capable of accommodating improvements in translation technology without losing the linguistic evidence on which future reinterpretation depends.

6.2 Community Governance Is a Technical Requirement as Well as an Ethical Requirement

The simulation places community-validation readiness inside the preservation performance framework because governance directly influences data quality and the meaning of linguistic outputs. Community knowledge holders may identify incorrect translations that external annotators consider fluent, distinguish ordinary public knowledge from restricted material, clarify terminology that has no direct dominant-language equivalent, and determine whether particular knowledge should be digitally disseminated at all. Research involving Māori language technologies demonstrates that language and data specialists with deep community knowledge can substantially improve model diagnosis and ensure that development remains culturally appropriate. Participatory machine-translation research similarly shows that low-resource NLP can benefit when speakers and local researchers participate in defining the research process rather than functioning only as data providers.

This governance requirement becomes even more important when cross-lingual NLP makes local knowledge discoverable to audiences who previously could not access it linguistically. Improved retrieval is not always equivalent to desirable unrestricted access. Some archives may contain knowledge governed by community-specific rules relating to age, role, ceremony, season, location, family, or other contextual factors. The present simulation therefore treats access metadata as part of preservation rather than assuming that every successfully digitized knowledge object should automatically become public.

The technical system should be capable of representing differentiated access rules established through legitimate governance processes.

6.3 From Multilingual Models to Sustainable Local-Knowledge Infrastructure

Large multilingual models have substantially changed what is technically possible in low-resource NLP. NLLB demonstrates translation at a scale of 200 languages, multilingual speech research has extended speech technology beyond one thousand languages, and cross-lingual representations increasingly permit semantic transfer across language boundaries. These achievements provide useful infrastructure, but sustainable preservation requires a second layer of investment in local corpora, terminology, script support, annotation, evaluation, governance, and long-term storage. A universal multilingual model can provide transfer learning, but it cannot by itself determine whether a culturally specific plant name has been translated appropriately, whether an oral narrative should be publicly searchable, or whether two local orthographic variants should be normalized or preserved separately.

The emerging research agenda should therefore move from asking whether an NLP model "supports" a language to asking what kinds of support it provides. A language may be technically present in a multilingual model while still receiving weak translation quality, inadequate ASR performance, poor dialect handling, missing script infrastructure, or limited community-governed resources. Recent work explicitly challenges simplistic classifications of languages according to resource levels and highlights continuing inequalities even as multilingual technologies expand. UNESCO's continuing work on digital representation and Indigenous-language technology similarly shows that language inclusion involves infrastructure and community capacity in addition to model coverage.

7. Conclusion

Cross-lingual natural language processing offers significant opportunities for preserving and improving access to local knowledge that has historically remained poorly represented in mainstream digital systems. Multilingual translation, speech recognition, semantic representation, and cross-lingual retrieval can reduce barriers created by limited corpora, linguistic fragmentation, and dependence on dominant-language search systems. Contemporary developments such as NLLB, IndicTrans2, multilingual speech models, and cross-lingual representation learning demonstrate that knowledge encoded in lower-resource languages is becoming increasingly accessible to computational systems.

The present study nevertheless demonstrates conceptually that technological access should not be confused with preservation. A pipeline that simply translates source material into English or another dominant language may improve readability while weakening terminology, provenance, linguistic specificity, and the possibility of later reinterpretation. Preservation requires the original language and, where relevant, original audio or script to remain connected with all computationally generated representations. Machine translation should therefore create an additional access layer rather than become a substitute for the original knowledge record.

The simulation further indicates that domain adaptation and community governance are likely to be essential for high-quality preservation. The composite score increased from 45.6 for the translation-centered baseline to 59.2 under a generic multilingual pipeline, 75.6 following domain adaptation, and

88.6 under the community-governed architecture. These figures are synthetic and do not establish empirical superiority. Their purpose is to demonstrate a testable evaluation framework in which semantic fidelity, terminology, retrieval, speech alignment, and governance readiness are considered simultaneously.

Community participation is particularly important because local knowledge preservation involves rights and responsibilities that cannot be resolved by algorithmic accuracy alone. Communities or legitimate knowledge custodians should have meaningful authority over what is collected, what is translated, how specialized terminology is represented, which corrections are authoritative, which materials may be publicly retrieved, and how digital resources are reused. Contemporary participatory and Indigenous-language NLP scholarship strongly supports this move away from extractive data collection toward community-led technological development.

References

1. Koehn, P. (2005). *Europarl: A parallel corpus for statistical machine translation*. Proceedings of MT Summit X, 79–86.
2. Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, 48–54. https://doi.org/10.3115/1073445.1073462
3. Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics, 19*(2), 263–311.
4. Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51. https://doi.org/10.1162/089120103321337421
5. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems, 27*, 3104–3112.
6. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.
8. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423
9. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of ACL*, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747
10. Artetxe, M., & Schwenk, H. (2019). Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics, 7*, 597–610. https://doi.org/10.1162/tacl_a_00288
11. Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. *International Conference on Learning Representations (ICLR)*.

12. Lample, G., & Conneau, A. (2019). Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems, 32*, 7059–7069.
13. Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research, 65*, 569–631. https://doi.org/10.1613/jair.1.11640
14. Mihalcea, R., & Simard, M. (2005). Creating a multilingual parallel corpus for language preservation. *Proceedings of the 10th International Conference on Intelligent User Interfaces*, 367–368.
15. Bird, S. (2020). Decolonising speech and language technologies. *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519. https://doi.org/10.18653/v1/2020.coling-main.313
16. Mager, M., Oncevay, A., Rios, A., & Gasser, M. (2018). Bilingual dictionaries for endangered languages. *Proceedings of the 27th International Conference on Computational Linguistics*, 1765–1776.
17. Gasser, M. (2009). Building a parser for a language documentation corpus. *Proceedings of the 2009 Workshop on Language Technology and Resources for Indigenous Languages*, 18–25.
18. Zampieri, M., Nakov, P., Rosenthal, S., Ataman, D., Da San Martino, G., Haagsma, H., Pilehvar, M. T., & Tudor, C. (2020). SemEval-2020 Task 12: Multilingual offensive language identification in social media. *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 1425–1447. https://doi.org/10.18653/v1/2020.semeval-1.188
19. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560
20. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.