

Computational Preservation of Endangered Pamiri Languages through NLP: Insights from Bartangi and a Roadmap for Digital Inclusion

Warda Tariq

HSE University

Faculty of Computer Science

Moscow, Russia

varda@hse.ru

Abstract

Endangered languages remain significantly underrepresented in modern Natural Language Processing (NLP) due to the scarcity of foundational computational resources. Although recent advances in multilingual and low-resource NLP have improved language technologies for many underrepresented languages, evaluation commonly emphasizes technical measures such as corpus quality, representation quality, and downstream task performance. While these metrics are essential for assessing computational quality, they provide limited insight into the broader societal impact of NLP resource development, particularly for endangered languages where digital inclusion, cultural preservation, and long-term sustainability are central objectives. In this paper, we propose a conceptual framework for assessing the societal impact of NLP resource development for endangered languages. The framework complements conventional technical evaluation by introducing six dimensions: technical readiness, accessibility, community participation, educational and cultural value, research reusability, and sustainability. We illustrate the application of the framework through the development of computational resources for Bartangi, an endangered Pamiri language, demonstrating how it can be used to identify strengths, limitations, and future priorities beyond technical performance alone. Rather than replacing conventional evaluation, the proposed framework extends it by providing a structured perspective for understanding how foundational NLP resources contribute to responsible AI, digital inclusion, and the long-term computational preservation of endangered languages.

Keywords: Endangered Languages; Low-Resource NLP; Computational Language Preservation; Digital Inclusion; Societal Impact Assessment; Responsible AI

1 Introduction

Language is a fundamental component of cultural identity, knowledge transmission, and social inclusion (Harrison, 2007). However, thousands of the world’s languages remain digitally underrepresented, limiting their participation in modern artificial intelligence (AI) and natural language processing (NLP) technologies (Crystal, 2000; Austin and Sallabank, 2011). Although multilingual language models have substantially advanced NLP for widely spoken languages (Devlin et al., 2019; Conneau et al., 2020; Costa-jussà et al., 2022), many endangered and low-resource languages remain digitally excluded due to the absence of foundational computational resources such as corpora, lexicons, annotated datasets, and language-specific preprocessing tools (Hedderich et al., 2021). Consequently, the benefits of modern AI remain unevenly distributed across the world’s linguistic communities.

Among the most underrepresented are the Pamiri languages, spoken primarily in Tajikistan, Afghanistan, Pakistan, and China. Despite their linguistic and cultural significance, they remain largely absent from contemporary NLP research, restricting the development of applications such as machine translation, information retrieval, educational technologies, and digital archives (Bird, 2009). Bartangi represents a particularly challenging example because of its extremely limited digital resources, lack of standardized computational tools, and scarcity of publicly available corpora (Tariq et al., 2025). Developing NLP resources for such languages therefore requires not only technical solutions but also sustainable approaches that support long-term language preservation and digital inclusion.

Recent research on low-resource NLP has commonly assessed progress through resource- and task-oriented measures, including corpus character-

istics, representation quality, and downstream task performance (Hedderich et al., 2021). While these metrics remain essential, they provide limited insight into a broader question: how do foundational NLP resources contribute to societal outcomes such as accessibility, cultural preservation, education, and long-term digital inclusion? This limitation is particularly important for endangered languages, where resource development often represents the primary contribution rather than the final technological outcome.

In this paper, we argue that computational resource development should be viewed not merely as a technical exercise but as a key component of responsible and socially impactful NLP (Bird, 2020a). Using Bartangi as an illustrative case study, we introduce a conceptual framework for assessing the societal impact of NLP resource development for endangered languages. The proposed framework complements conventional technical evaluation by incorporating six dimensions: technical readiness, accessibility, community participation, educational and cultural value, research reusability, and sustainability. Together, these dimensions provide a structured perspective for understanding how foundational NLP resources contribute to responsible AI, digital inclusion, and the long-term computational preservation of endangered languages.

2 Related Work

2.1 Endangered Languages and Digital Preservation

Language documentation has traditionally preserved endangered languages through grammars, dictionaries, text collections, and audio recordings (Bird, 2009). However, the growing importance of digital technologies has expanded preservation beyond documentation toward computational preservation, where reusable language resources support future NLP applications (Bird, 2020b). Digital preservation also depends on the availability of reusable resources such as digital corpora, lexical databases, annotated datasets, and open repositories that can support education, cultural preservation, and community engagement (Bird and Simons, 2003; Simons and Bird, 2003). Nevertheless, such foundational resources remain unavailable for many endangered languages, limiting their participation in the digital ecosystem (Austin and Sallabank, 2011).

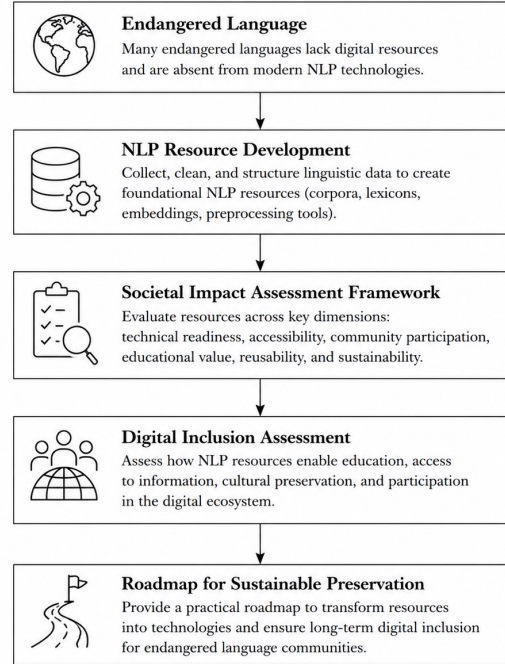


Figure 1: Overview of the proposed approach for computational preservation and digital inclusion of endangered languages.

2.2 NLP for Low-Resource Languages

Recent advances in multilingual language models, cross-lingual transfer learning, data augmentation, and community-driven resource development have significantly improved NLP for low-resource languages (Joshi et al., 2020). Despite this progress, high-quality corpora, standardized preprocessing pipelines, lexical resources, and evaluation datasets remain essential prerequisites for reliable language technologies (Hedderich et al., 2021). Consequently, languages lacking these foundational resources continue to face substantial barriers to modern NLP applications.

2.3 NLP for Positive Societal Impact

Recent research increasingly argues that NLP technologies, particularly those developed for endangered and under-resourced languages, should be considered not only in terms of technical performance but also in relation to community needs, accessibility, and broader social outcomes (Bird, 2020a; Liu et al., 2022). Applications in healthcare, education, accessibility, crisis response, and cultural heritage demonstrate the potential of language technologies to promote social inclusion. Within this context, endangered-language preservation remains a relatively underexplored application area where computational resources can support both

future NLP research and long-term linguistic sustainability (Krauss, 1992).

2.4 Computational Resources for Pamiri Languages

Pamiri languages have received limited attention from the NLP community, with much of the available work focusing on descriptive linguistics and language documentation. For Bartangi in particular, computational resources remain scarce, with only limited corpus and language-modeling resources currently available (Tariq et al., 2025). Consequently, foundational resource development remains a prerequisite for enabling future NLP research on Bartangi and related Pamiri languages.

2.5 Measuring the Societal Impact of NLP

Although NLP has increasingly considered socially beneficial and human-centered applications, widely used evaluation measures continue to include task- and performance-oriented metrics such as accuracy, F1-score, BLEU, and benchmark performance (Papineni et al., 2002; Hedderich et al., 2021). These measures are important for assessing computational performance, but they do not by themselves capture broader considerations such as accessibility, community participation, educational and cultural value, research reusability, or long-term sustainability. This limitation is particularly important for endangered-language technology, where community needs and the social context of resource development can be central to determining whether computational resources provide meaningful benefits (Bird, 2020a; Liu et al., 2022).

2.6 Research Gap

Despite substantial advances in multilingual NLP and low-resource language technologies, important gaps remain for endangered Pamiri languages. Computational resources for these languages remain limited, and existing technical evaluation approaches do not by themselves capture considerations such as community needs, accessibility, and long-term resource stewardship (Hedderich et al., 2021; Bird, 2020a; Liu et al., 2022). To address this gap, we use Bartangi as an illustrative case study and introduce a conceptual framework that complements technical evaluation with broader societal dimensions relevant to endangered-language NLP resource development.

3 Bartangi as an Illustrative Case Study

3.1 Bartangi in the Context of Computational Preservation

Bartangi is an endangered Eastern Iranian language belonging to the Pamiri language family and spoken primarily in the Bartang Valley of the Gorno-Badakhshan Autonomous Region of Tajikistan. Like many endangered languages, it has limited digital resources and minimal representation in modern NLP, restricting the development of computational tools for language documentation, education, and digital communication.

Unlike high-resource languages with extensive corpora and mature NLP ecosystems, Bartangi has only limited foundational resources for computational language processing, with substantial gaps in standardized preprocessing tools, annotated datasets, and benchmark resources. These characteristics make Bartangi an appropriate illustrative case for examining how foundational NLP resources can support computational preservation.

Rather than presenting Bartangi as a standalone resource-development project, this paper uses it as an illustrative case study for demonstrating how the proposed framework can assess the broader societal impact of foundational NLP resource development in endangered-language settings.

3.2 Resource Development for Bartangi

Existing computational resources for Bartangi remain extremely limited. An earlier publicly available Bartangi corpus (Novokshanov, 2020) provided a foundation for computational work on the language. Building on these available materials, Tariq et al. (2025) developed a structured resource-development pipeline involving corpus cleaning, normalization, morphological processing, lemmatization, and word embedding training. The present study uses these documented resources as an illustrative case for applying the proposed societal-impact framework.

The resulting resources provide an initial computational foundation for future NLP research, supporting applications such as semantic analysis, language modeling, information retrieval, and educational technologies (Tariq et al., 2025). More importantly, they demonstrate that meaningful computational resources can be developed even under severe data limitations when computational methods are combined with appropriate linguistic expertise. However, their broader societal value depends

on how effectively they enable future research, educational initiatives, and community-oriented language technologies. This distinction between technical resource development and broader societal impact motivates the framework proposed in the following section.

4 A Framework for Assessing the Societal Impact of NLP Resource Development

Conventional NLP evaluation often emphasizes technical outputs such as corpus characteristics, representation quality, and downstream task performance (Hedderich et al., 2021). While these measures remain essential, they do not by themselves capture broader dimensions of endangered-language resource development, including accessibility, community participation, educational and cultural value, research reusability, and long-term sustainability (Bird, 2020a; Liu et al., 2022). To complement technical evaluation, we propose a conceptual framework incorporating six dimensions for assessing the broader societal implications of NLP resource development for endangered languages.

Although motivated by the Bartangi case study, the framework is intended to be applicable to a broader range of endangered-language projects facing similar challenges of resource scarcity, digital underrepresentation, and long-term sustainability. The six dimensions should not be interpreted as sequential stages of development; rather, they provide complementary perspectives for evaluating both the technical foundations and the broader societal outcomes of computational preservation. Figure 2 illustrates the relationships among these dimensions, while Table 1 summarizes the corresponding evaluation questions and representative indicators.

4.1 Technical Readiness

Technical readiness forms the foundation of the proposed framework because all subsequent language technologies depend upon the availability of reliable computational resources. This dimension evaluates whether sufficient technical infrastructure has been established to support future NLP research and application development. Key components include cleaned and normalized corpora, standardized preprocessing pipelines, lexical resources, word embeddings, and evaluation datasets. For the Bartangi project, technical readiness was

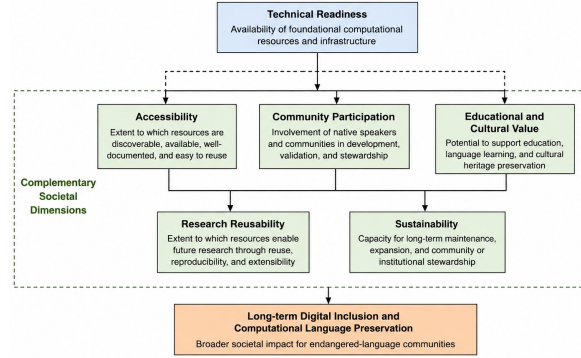


Figure 2: Proposed framework for assessing the societal impact of NLP resource development for endangered languages. Technical readiness provides the foundation for five complementary societal dimensions that collectively support long-term digital inclusion and computational language preservation.

established through the construction of a normalized corpus together with language-specific pre-processing procedures and foundational semantic representations. These resources constitute the initial computational infrastructure required for future language technology development.

4.2 Accessibility

Accessibility concerns whether developed resources can be effectively accessed and reused by researchers, educators, and language communities. Key considerations include documentation, metadata, licensing, public availability, and ease of reuse (Bird and Simons, 2003; Bender and Friedman, 2018). These practices support transparent and responsible resource dissemination.

4.3 Community Participation

Community participation concerns the meaningful involvement of native speakers and community stakeholders in resource development, validation, and stewardship. Such participation helps ensure linguistic authenticity and alignment with community needs and priorities (Bird, 2020a; Liu et al., 2022).

4.4 Educational and Cultural Value

Educational and cultural value concerns whether computational resources can support language learning, cultural preservation, and knowledge transmission. Potential applications include digital dictionaries, language-learning tools, and searchable cultural archives (Crystal, 2000; Jones and Ogilvie, 2013).

4.5 Research Reusability

Research reusability concerns whether computational resources can support subsequent research and technological development. Publicly available datasets, documented methods, and reusable code can reduce duplication of effort and facilitate cumulative progress in low-resource NLP (Geburu et al., 2021; Bender and Friedman, 2018).

4.6 Sustainability

Sustainability concerns whether computational resources can remain available, maintained, and useful beyond the lifetime of an individual project. Long-term preservation requires continued maintenance, institutional or community stewardship, and mechanisms for future resource updates (Bird and Simons, 2003; Grenoble and Whaley, 2006).

4.7 Framework Rationale

The six dimensions synthesize recurring concerns in prior work on endangered-language documentation, language revitalization, computational resource development, data accessibility, and responsible NLP. Technical readiness and research reusability reflect the importance of well-documented and reusable computational resources (Hedderich et al., 2021; Bender and Friedman, 2018; Geburu et al., 2021). Accessibility and sustainability draw on principles of resource portability, preservation, and long-term stewardship (Bird and Simons, 2003; Simons and Bird, 2003). Community participation and educational and cultural value reflect community-centered approaches to endangered-language technology and language revitalization (Grenoble and Whaley, 2006; Bird, 2020a; Liu et al., 2022). The framework integrates these complementary concerns into a single qualitative assessment approach rather than treating them as independent or exhaustive evaluation criteria.

5 Applying the Framework to Bartangi

To illustrate the practical applicability of the proposed framework, we apply it to the Bartangi resource-development project. The objective is not to claim that the project has already achieved comprehensive societal impact, but rather to demonstrate how the proposed framework can be used to assess the current state of computational preservation. The assessment is based on documented outputs and development practices reported for

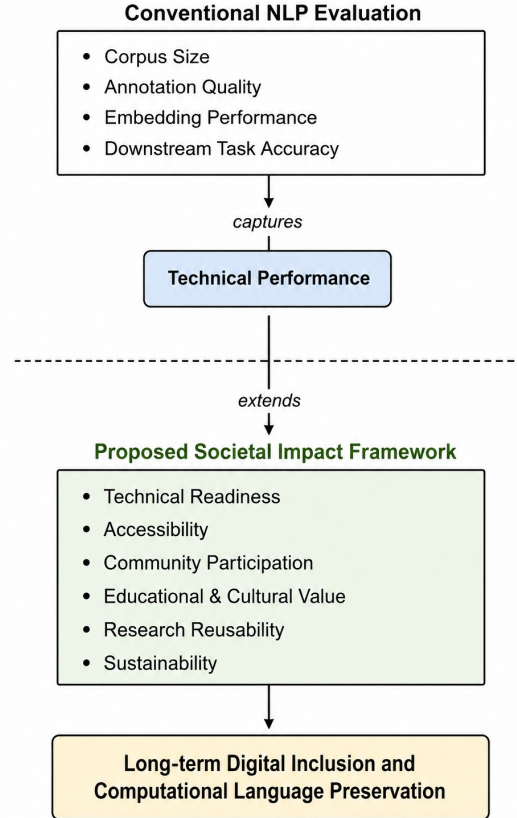


Figure 3: Positioning the proposed societal impact framework relative to conventional NLP evaluation.

the Bartangi resource-development pipeline (Tariq et al., 2025).

To ensure consistency across the six dimensions, we use a three-level qualitative assessment scale: *High*, *Moderate*, and *Limited*. A dimension is rated *High* when most of the indicators defined in Table 1 are demonstrably addressed by the available project evidence; *Moderate* when some indicators are addressed but important components remain incomplete; and *Limited* when few or none of the indicators are currently demonstrated. The ratings should therefore be interpreted as structured qualitative assessments rather than standardized quantitative scores. The assessment was conducted by the author by comparing the documented Bartangi project evidence with the indicators defined in Table 1. As this study presents an illustrative application of the proposed framework rather than a validated annotation protocol, no independent annotators or inter-annotator agreement measures were used.

Table 2 summarizes the assessment across the six framework dimensions.

Dimension	Evaluation Question	Representative Indicators
Technical Readiness	Are the core NLP resources required for future language technologies available?	Cleaned corpus, preprocessing pipeline, lexical resources, word embeddings, evaluation datasets
Accessibility	Can researchers, educators, and language communities readily access and use the developed resources?	Documentation, public availability, licensing, metadata, resource dissemination
Community Participation	Were native speakers or community stakeholders meaningfully involved throughout the resource development process?	Community consultation, linguistic validation, user feedback, collaborative development
Educational & Cultural Value	Can the developed resources support language learning, cultural preservation, or educational initiatives?	Educational materials, digital dictionaries, cultural archives, language-learning resources
Research Reusability	Can other researchers effectively build upon the developed resources?	Public datasets, reusable code, reproducible methodology, benchmark datasets
Sustainability	Can the project continue to evolve and remain useful over time?	Resource maintenance, future updates, institutional support, community stewardship

Table 1: The six dimensions of the proposed framework together with their corresponding evaluation questions and representative indicators.

Figure 3 illustrates how the proposed framework complements conventional technical evaluation. Rather than replacing existing performance metrics, it extends them by incorporating societal dimensions associated with accessibility, community engagement, resource reuse, and long-term sustainability.

The qualitative assessment indicates that the Bartangi project has established a strong technical foundation for future NLP research while also identifying dimensions that require further development. Accessibility and research reusability are assessed as moderate because some relevant indicators are currently addressed, while broader dissemination and reuse mechanisms remain incomplete. Community participation, educational and cultural value, and sustainability are assessed as limited because the corresponding activities and infrastructures have not yet been systematically established. These results illustrate how the proposed framework can identify strengths, limitations, and future priorities that are not captured by technical performance measures alone.

6 Discussion

The primary contribution of this work is the introduction of a framework for assessing the societal impact of NLP resource development for endangered languages, illustrated through the Bartangi case study. Rather than proposing a new NLP model or language resource, the framework provides a complementary perspective for evaluating how foundational computational resources contribute to long-term digital inclusion and computational preservation.

The rapid expansion of NLP has substantially improved language technologies for many low-resource languages; however, evaluating these advances solely through technical performance provides an incomplete understanding of their broader societal value. For endangered languages, the creation of foundational computational resources is often a necessary first step toward digital preservation, yet the existence of such resources does not automatically imply meaningful societal impact. The proposed framework distinguishes between technical outputs and the broader outcomes that resource development seeks to enable, providing a structured perspective for assessing computational preservation beyond conventional evaluation metrics (Fishman, 2001).

The Bartangi case study illustrates this distinction. The development of a cleaned corpus, preprocessing pipeline, and foundational language resources represents an important achievement in technical readiness. These resources establish the computational infrastructure required for future NLP applications, but they should be viewed as enabling conditions rather than evidence of successful digital inclusion. Whether these resources ultimately support education, cultural preservation, accessibility, or community empowerment depends on their subsequent adoption by researchers, educators, and language communities.

The proposed framework therefore encourages a broader interpretation of success for endangered-language NLP projects by considering both technical quality and the social contexts in which computational resources are developed, shared, and ultimately used (OECD, 2019).

The framework also has practical implications

Framework Dimension	Assessment	Supporting Evidence
Technical Readiness	High	A cleaned corpus, preprocessing pipeline, and word embeddings have been developed, addressing most of the technical-readiness indicators defined in Table 1.
Accessibility	Moderate	The resources have been documented, and an earlier Bartangi corpus is publicly available. However, broader dissemination, licensing information, metadata, and long-term accessibility mechanisms remain incomplete.
Community Participation	Limited	Linguistic expertise informed resource development, but broader community consultation, feedback, and collaborative participation have not yet been systematically incorporated.
Educational & Cultural Value	Limited	The resources provide a foundation for future educational tools, digital dictionaries, and language-preservation initiatives, but these applications have not yet been developed or evaluated.
Research Reusability	Moderate	The corpus and computational resources provide a basis for further NLP research, although broader reuse through publicly documented datasets, code, and benchmark resources remains incomplete.
Sustainability	Limited	Long-term maintenance, resource expansion, institutional support, and community stewardship have not yet been systematically established.

Table 2: Qualitative assessment of the Bartangi resource-development project using the proposed three-level framework.

for future endangered-language initiatives. Rather than treating societal impact as an outcome to be considered only after technical development, the proposed dimensions can guide project design from the outset (Bender et al., 2021). Researchers can identify strengths and limitations across multiple dimensions, prioritize future resource development, and design computational infrastructures that are more closely aligned with the needs of language communities. Although demonstrated using Bartangi, the framework is intended to be adaptable to other endangered-language projects facing similar challenges of resource scarcity and digital underrepresentation.

7 Limitations

The proposed framework has several limitations. First, it is conceptual and is illustrated using a single case study involving the Bartangi language. Although Bartangi provides an appropriate example of an endangered language with extremely limited computational resources, additional case studies are needed to evaluate the framework’s applicability across other endangered-language projects (Chelliah and de Reuse, 2011).

Second, the framework is demonstrated through a qualitative assessment rather than a standardized quantitative methodology. The present assessment was conducted by a single author and therefore does not provide evidence of inter-rater reliability. Future work should involve multiple independent evaluators and develop measurable indi-

cators and evaluation protocols that enable more systematic and comparable assessments across different resource-development initiatives. Finally, several dimensions, particularly community participation, educational and cultural value, and sustainability, require longitudinal evaluation. Moreover, the framework focuses specifically on the societal impact of foundational NLP resource development and does not address broader economic, policy, or governance aspects of AI deployment. Future research should investigate how the proposed framework can be integrated with existing responsible AI and human-centered NLP evaluation methodologies.

8 Ethical Considerations

Computational work on endangered languages requires careful attention to data provenance, community participation, consent, and resource stewardship. Language data may encode culturally sensitive knowledge or originate from communities with limited control over how linguistic materials are subsequently digitized and reused. Researchers should therefore document data sources and collection practices transparently and consider whether subsequent uses are consistent with the interests and expectations of the relevant language communities (Bird, 2020a; Liu et al., 2022).

For the Bartangi case considered in this paper, the analysis relies on previously documented computational resources rather than new data collection from community members (Tariq et al., 2025). Ac-

cordingly, the present study does not claim direct community consultation or consent as part of the framework application. This absence is reflected in the limited community-participation assessment reported in Table 2. Future resource-development efforts should involve native speakers and community stakeholders in decisions concerning linguistic validation, appropriate forms of resource sharing, and long-term stewardship.

Resource ownership and accessibility also require careful consideration. Public availability should not automatically be treated as an ethical good when language materials contain culturally restricted or sensitive content. Decisions about licensing, dissemination, preservation, and future reuse should balance research accessibility with community preferences and responsible stewardship (UNESCO, 2021; Bird and Simons, 2003). These considerations are particularly important for endangered-language NLP, where long-term computational preservation should support rather than override the interests of the communities whose languages are represented.

9 Conclusion

Endangered languages remain among the least represented within contemporary NLP despite rapid advances in multilingual language technologies. Although substantial progress has been made in developing methods for low-resource NLP, technical evaluation alone does not capture how foundational language resources relate to broader societal outcomes. This paper argues that assessing endangered-language NLP requires moving beyond performance-oriented evaluation toward approaches that also consider digital inclusion, cultural preservation, and long-term sustainability.

Using Bartangi as an illustrative case study, we introduced a conceptual framework that complements conventional technical evaluation with broader societal considerations relevant to endangered-language resource development. The framework provides a structured perspective for identifying both the technical strengths of computational resources and areas requiring further development to support responsible and socially beneficial NLP.

Applying the framework to the Bartangi project demonstrates how technical resource development can be assessed alongside broader societal considerations. While foundational computational

resources establish the basis for future language technologies, their long-term impact ultimately depends on accessibility, adoption, reuse, and sustainable stewardship. Future work should validate the framework across additional endangered languages and develop quantitative and qualitative methodologies for operationalizing each assessment dimension.

By complementing conventional technical evaluation with structured societal impact assessment, the proposed framework provides a practical foundation for future research on responsible, inclusive, and socially impactful NLP for endangered languages.

Acknowledgments

The author thanks the anonymous reviewers for their constructive feedback, which helped improve the clarity and methodological presentation of this work.

References

- Peter K. Austin and Julia Sallabank, editors. 2011. *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM.
- Steven Bird. 2009. [Last words: Natural language processing and linguistic fieldwork](#). *Computational Linguistics*, 35(3):469–474.
- Steven Bird. 2020a. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Steven Bird. 2020b. [Sparse transcription](#). *Computational Linguistics*, 46(4):713–744.
- Steven Bird and Gary Simons. 2003. [Seven dimensions of portability for language documentation and description](#). *Language*, 79(3):557–582.
- Shobhana L. Chelliah and Willem J. de Reuse. 2011. *Handbook of Descriptive Linguistic Fieldwork*. Springer.

- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, et al. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*, pages 8440–8451.
- Marta R. Costa-jussà et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- David Crystal. 2000. *Language Death*. Cambridge University Press.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.
- Joshua A. Fishman, editor. 2001. *Can Threatened Languages Be Saved?* Multilingual Matters.
- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.
- Lenore A. Grenoble and Lindsay J. Whaley. 2006. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press.
- K. David Harrison. 2007. *When Languages Die: The Extinction of the World’s Languages and the Erosion of Human Knowledge*. Oxford University Press.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A survey on recent approaches for natural language processing in low-resource scenarios](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.
- Mari C. Jones and Sarah Ogilvie, editors. 2013. *Keeping Languages Alive: Documentation, Pedagogy, and Revitalization*. Cambridge University Press.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.
- Michael Krauss. 1992. The world’s languages in crisis. *Language*, 68(1):4–10.
- Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud’hommeaux. 2022. [Not always about you: Prioritizing community needs when developing endangered language technology](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3933–3944, Dublin, Ireland. Association for Computational Linguistics.
- Novokshanov. 2020. Tsakorporus bartangi: Bartangi linguistic corpus. GitHub repository. Bartangi linguistic corpus implemented using Tsakorporus.
- OECD. 2019. [Oecd principles on artificial intelligence](#).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.
- Warda Tariq, Victor Popov, and Vasilii Gromov. 2025. [Building a clean bartangi language corpus and training word embeddings for low-resource language modeling](#). In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 1256–1262, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.
- UNESCO. 2021. [Recommendation on the ethics of artificial intelligence](#).