Assessment 1: Critical Evaluation of 'Participatory Research for Low-resourced Machine Translation'

Watson Ndethi

2025-02-10

Abstract

This paper presents a critical evaluation of Abbott and Martinus's (2020) research on participatory methods for low-resource machine translation, specifically focusing on African languages. The evaluation employs the CRAAP framework to assess the currency, relevance, authority, accuracy, and purpose of the research. Furthermore, it reflects on the note-taking methodology used, provides a summary of the paper, and offers a critical reflection on the strengths and limitations of the participatory approach. The analysis connects the paper's findings to broader African NLP initiatives and the author's research interests in low-resource language preservation.

1 Introduction

My interest in the paper "Participatory Research for Low-resourced Machine Translation" by Abbott and Martinus (2020) stems from a curiosity about how technology specifically Machine Translation could be used to promote uptake of learning and speaking of Kikuyu, my mothertongue and by all measure a low-resourced language, amongst the younger generation of that community while leveraging collaborative creation of annotatable datasets from the older generation who speak it better. The paper's focus on participatory research methods for low-resource machine translation aligns with my research interest in collaboratively developing datasets for low-resourced languages through crowd-sourcing and community engagement. A discussion of broader NLP initiatives particularly in Africa through the Masakhane Community adds to the relevance of the paper in the context of my research goals.

2 Critical Analysis Using CRAAP Framework

2.1 Currency (2020 Publication)

Five years on (2025), post the EMNLP event where the paper was unveiled and even in the wake of major advancements in AI (reasoning Models, Transformer 2 architecture), more recent research *Empirical Methods in Natural Language Processing* (n.d.) shows African Languages still remain massively low-resourced. Even when there have been more recent notable attempts at creation of datasets for LRL, for example the No Language Left Behind (NL2B) initiative by Meta *No Language Left Behind (NL2B)* (2022) (2022), the participatory approach to dataset creation as proposed by Abbott and Martinus (2020) is yet to be fully explored. Additionally, the paper's growing citation counts on ArXiv and Google Scholar indicate the continued popularity of the participatory approach to date.

2.2 Relevance

The continued prevalence of the lack of Machine Translation useable datasets for low-resourced languages, particularly in Africa, underscores the relevance of this work in the current LRL research landscape. In relation to my personal research interests, the participatory approach discussed in the paper aligns with my crowd-sourced dataset creation goal for my own mothertongue.

2.3 Authority

With more than 50 NLP related papers listed under their names on Google Scholar, a credible directory of research articles, Jade Abbott's credibility (*Google Scholar – Jade Abbott*, n.d.) and Laura Martinus' credibility (*Google Scholar – Laura Martinus*, n.d.) in the field of NLP is undeniable. Beyond their scholarly endeavours, Abbott leads tech as CTO at Lelapa AI, an AI research and product lab, based off Johannesburg and Martinus works as a data scientist for a finance outfit. That they are both of South African descent adds to the acceptability of their intention to digitalize African languages. While most of the articles listed are preprints on ArXiv, this particular paper was published in the Association for Computational Linguistics (ACL) Anthology *ACL Anthology* (n.d.), a reputable platform for NLP research affiliated with the Computational Linguistics journal, arguably the leading publication in the field. The annual Empirical Methods in Natural Language Processing (EMNLP) conference, where the paper was presented, is one of the premier venues for NLP research *Empirical Methods in Natural Language Processing* (n.d.), further enhancing the paper's authority.

2.4 Accuracy

The paper clearly outlines the data collection and dataset creation process, associated quality control processes, model development (JoeyMT) and benchmarking and evaluation metrics. Renowned benchmark tests like BLEU, ChrF and TER form part of the eval methodology. A public Github repository (*Masakhane*, n.d.) provides access to NLP research artifacts such as datasets, benchmarks and models code and datasets enabling reproducibility and further research. An extensive bibliography, including citations to foundational work in Machine Translation e.g. (Firat et al., 2016) (Zoph et al., 2016)

2.5 Purpose

The paper sought to address the inadequacy of machine translation for African languages which are mostly low-resourced, by proposing a participatory approach encompassing greater stakeholdership - beyond just the Machine Transalation researchers cohort - in the dataset creation process. The paper's purpose aligns with the broader goal of promoting African language preservation and technological advancement in the NLP field. It sets out on a path to democratize dataset creation for low-resourced languages and manages to do so by creating a well documented and replicable methodology, an actual NLP Model (the citation for JoeyMT has been removed) and tangible community impact in the birthing of the Masakhane community (*Masakhane*, n.d.). The paper, going by its citation counts, has indeed contributed fundamentally to the field of LRL and in that it has achieved its purpose.

3 Note-Taking Methodology

I employed the Cornell note-taking method on paper and via Claude to synthesise the paper's key points, making a summary of the paper's content and crafting critical questions for further reflection. Below is a snapshot of this approach in action via Claude:

← Cornell Notes Page 1: Introduction and Problem Defi... Preview Code

Cues/Questions	Notes
What defines "low-resourced"?	• Complex problem beyond data availability • Symptom of societal problems • Authors oppressed by colonial governments • Low access to tertiary education • Limited PhD candidates from affected regions • Lack of geographic and language diversity in NLP
What is the scale of the problem?	• Africa has 2144 living languages • Small fraction of available resources • Most languages have no annotated data • Limited monolingual resources • Only 5 out of 2695 NLP conference affiliations from African institutions (2018)
What are the research gaps?	• NLP research rarely considers African languages • Existing resources hard to discover • Many resources published in closed journals • Much content not digitized • Standard crowdsourcing pipelines infeasible
What is the paper's approach?	• Proposes participatory research method • Emphasizes value of research partners • Collaborative research process definition • Iterative development • Focus on community involvement

Summary: The paper identifies low-resourced languages as a complex socio-technical challenge requiring community-driven solutions. It proposes participatory research as a method to address both technical and societal aspects of the problem.

4 Paper Summary

The paper outlines a collaborative approach to tackling the challenge of low-resource languages machine translation by incorporating a greater circle of stakeholders beyond the machine translation research community. The novelty in the approach is the inclusion of non-technical stakeholders such as linguists, language activists, and native speakers in the dataset creation process. The implementation was largely successful giving rise to 45 benchmarks for over 30 African languages, having involved 400 + participants across the continent, birthing a vibrant and still thriving African NLP community in the name of Masakhane The project demonstrated the feasibility of a community-driven approach to dataset creation while supporting this achievement with comprehensive evals and benchmarks. The paper is and will continue to be instrumental in furthering the MT for LRL conversation in years to come.

References

ACL anthology. (n.d.). https://www.aclweb.org/anthology/.

Empirical methods in natural language processing. (n.d.). https://2025.emnlp.org/.

Firat, O., Cho, K., Jean, S., Yarats, D., Zhou, Y., & Bengio, Y. (2016). Zero-resource translation with multi-lingual neural machine translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 261–271.

Google scholar – jade abbott. (n.d.). https://scholar.google.com/citations?user=abc123.

Google scholar – laura martinus. (n.d.). https://scholar.google.com/citations?user=xyz789.

Masakhane: A grassroots NLP movement for africa. (n.d.). https://www.masakhane.io/.

No language left behind (NL2B). (2022). https://meta.com/nlb.

Zoph, B., Yuret, D., May, J., Knight, K., & Marcu, D. (2016). Transfer learning for low-resource neural machine translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1568–1575.