

A Mobile-assisted Language Learning and Data Collection Pipeline for Endangered Language Communities

Hannah M. Claus, Anna Korhonen

University of Cambridge

{hmc78, alk23}@cam.ac.uk

Abstract

We present a pipeline for mobile-assisted language learning and data collection for endangered language communities with a focus on primarily oral languages. The language learning app has been created utilising various features to teach users how to read and write their own language. With growing interest from such communities to have more representation on the internet and partake in collecting data for their own usage, we create an open-source pipeline that can be used by any community. We are now at the stage of evaluating the app and the connected pipeline to cater for a more community-based design.

1 Introduction

Language technology for endangered languages must centre the communities it serves (Bird and Yibarbuk, 2024). While even the existence of a digital tool has proven symbolically significant for oral and endangered language communities, empirical evidence of educational and cultural impact is essential for sustainable development. The previously developed application integrates vocabulary learning, literacy support through a newly standardised romanised orthography for the supported languages, cultural heritage content, a keyboard with autocorrect and predictive text, and a function to collect data. This abstract reports the upcoming evaluation phase with the language communities.

2 Evaluation Design

Our evaluation addresses five main questions: (1) Does the app improve reading and writing in the standardised orthography? (2) Does the app’s interaction design suit users who are orally proficient but have little to no (digital) literacy? (3) When given a data-collection function, do community members contribute data, and what is its quality/coverage? (4) How adaptable is the overall framework for future communities?

Participants are recruited from local speakers for Dzardzongke (spoken in Nepal) and Puno Quechua (spoken in Peru), spanning age groups from teenagers to elders. Usability is assessed via task completion and an adapted System Usability Scale. Semi-structured interviews and focus groups elicit qualitative feedback on content, design, and community priorities.

3 Preliminary Findings

Initial feedback from prototype demonstrations shows strong engagement with the cultural heritage section. The orthography-based activities require further scaffolding for participants with no prior Latin-script literacy, pointing to the need for improved onboarding in the next iteration. Evaluation insights can be incorporated rapidly into the pipeline, and the framework¹ can be adapted for other endangered language communities without specialist app development expertise.

4 Conclusion

This work contributes empirical evidence for community-centred language technology. Our evaluation provides the basis for iterative co-design with the language communities, and we discuss how the methodology and open-source pipeline can serve as a model for similar revitalisation efforts targeting other underserved languages.

References

Steven Bird and Dean Yibarbuk. 2024. [Centering the Speech Community](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian’s, Malta. Association for Computational Linguistics.

¹The work-in-progress framework can be found here: <https://github.com/cambridge11/CELT-Framework>.