Crowdsourcing Machine Learning Datasets for Northeastern Neo-Aramaic: Groundwork for Language Revitalization

Matthew Nazari [1,2*]

[1]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Allston, MA, USA
[2]The North-Eastern Neo-Aramaic Database Project, University of Cambridge, Cambridge, UK
matthewnazari@college.harvard.edu

Abstract

This paper presents an online platform for crowdsourcing speech and translation data for the Northeastern Neo-Aramaic (NENA) dialects. By establishing a dedicated space to collect data from individuals regardless of dialect, accent, fluency, gender, or age, the platform facilitates the creation of machine learning datasets. These datasets will enable the introduction of cutting-edge Artificial Intelligence (AI) models to these communities. We demonstrate this by using crowdsourced datasets to train Automatic Speech Recognition (ASR) models on several dialects to assist ongoing language documentation efforts. Beyond its technological implications, this initiative represents a significant cultural moment, prompting researchers and community members to critically consider the role and future of NENA dialects. This work establishes a precedent for community-driven endeavors within NENA-speaking communities, addressing not only the technological advancement of these underrepresented speech communities but also fostering essential dialogue on linguistic identity and cultural preservation.

[*]of the Assyrian community of San Jose, California.