MONITIETEINEN IHMISEN MONIMUOTOISUUS ESITELMÄSARJA
INTERDISCIPLINARY HUMAN DIVERSITY SEMINAR SERIES

Novel Approaches to Investigating Linguistic Diversity Using Speech Embeddings

Monday 27 April 2026 by

Tuukka Törö, Antti Suni, and Juraj Simko
Department of Digital Humanities, University of Helsinki

Time: 11:45–12:15 coffee and tea served; 12:15–13:00 presentation
Location: XIV, 3rd floor, Natura, Zoom

Abstract
Traditional research on linguistic diversity requires substantial effort from collecting and processing data to annotating and describing linguistic features. The analytical choices researchers make can influence outcomes and potentially introduce theoretical or observer bias. Moreover, many of the world's languages and varieties remain under documented, limiting the applicability of conventional linguistic methods. Recent advances in deep learning have opened new avenues for exploring language variation. State-of-the-art self-supervised speech models can reliably classify languages from short audio samples of varying quality. This enables us to study language relationships in their latent spaces in a relatively theory independent way, based on how languages sound. These models also allow us to examine relationships involving low resource languages and varieties, including those with minimal available data, no writing system, or prior linguistic description. In our talk, we will highlight both the strengths and limitations of these methods for conducting scientifically sound research on linguistic diversity across multiple levels of variation, from dialects to global language relationships.

Earn credits: HuDi0001 course (2 ECTS)
More information: humandiversity.fi

UNIVERSITY OF TURKU
CENTRES OF EXCELLENCE IN RESEARCH
Suomen Akatemia
Finlands Akademi
Research Council of Finland