SORS/WomenInBSC: Charting the Romance Dialectal Continuum: Insights into Language Similarity for Catalan ASR

Abstract

Estimating language similarity is useful for a range of NLP tasks, including the development of multilingual language models, selecting pivot languages in machine translation, and studying cross-lingual transfer. In particular, exploiting cross-lingual transfer is crucial for low- and mid-resource languages, as it can greatly improve model performance without requiring additional data in the target language. However, choosing auxiliary languages for cross-lingual transfer is often a tedious, resource-intensive process based primarily on intuition rather than objective criteria. While some heuristics and models have been proposed for tasks such as machine translation, part-of-speech tagging, and dependency parsing, no established guidelines exist for language selection in ASR (Automatic Speech Recognition).

Estimating language similarity in the ASR context is especially challenging because it involves multiple dimensions of language—phonology, syntax, and lexical overlap among them. In this talk, we explore linguistic similarity within the Romance dialectal continuum, considering both (1) a priori linguistic knowledge and (2) the internal representations of languages in text-based models (BERT) and speech models (Whisper, Wav2Vec2, HuBERT). We investigate whether these multifaceted perspectives can guide auxiliary language selection in Catalan ASR.

Short Bio

Barbara Scalvini was born in Chiari, Italy in 1992. After studying Physics (Bachelor and Master) in Milan, she completed her Master thesis at TU Delft's Delft Center of Systems and Control, focusing on adaptive optics. She then pursued a PhD at Leiden University in Computational Life Sciences, where she developed computational methods for analyzing biomolecular topology. Currently, she lives in the Faroe Islands, where she applies her NLP expertise to develop Faroese language technology as an Assistant Professor at the University of the Faroe Islands, contributing to the Language Technology Centre (Máltøknidepilin). Here, she also teaches Introduction to AI, Machine Learning and Deep Learning in the Master in Data Science. Passionate about the intersection of technology and language preservation, Barbara is deeply committed to revitalizing low-resource languages, believing that language and identity are intertwined. Outside of academia, she loves writing music, drawing and learning languages.

Speakers

Speaker: Barbara Scalvini, PhD. Assistant Professor at The Faculty of Science and Technology of The University of the Faroe Islands
Host: Carlos Daniel Hernandez Mena, Research Engineer. Language Technologies - Life Sciences , BSC
Barcelona Supercomputing Center - Centro Nacional de Supercomputación

---

Source URL (retrieved on *3 Abr 2025 - 14:37*): https://www.bsc.es/es/research-and-development/research-seminars/sorswomeninbsc-charting-the-romance-dialectal-continuum-insights-language-similarity-catalan-asr