Studying Endangered / Under-Documented Language Corpora with Large Language Model

The deep learning revolution has not only facilitated the development of mass-market applications with undeniable visibility and impact but has also opened up new possibilities for the documentation, analysis, and modeling of languages. Leveraging neural networks that autonomously construct representations of language, both spoken and written, encoding numerous linguistic properties, offers significant advancements in language analysis and automatic processing.

Neural language models play a pivotal role in reducing the annotation effort required by linguists. By enabling the development of systems capable of automatically annotating data, these models prove invaluable. After explaining how neural networks can uncover language representations from raw data only, we will explain how one such model allowed us to create a system for phonemic transcriptions with only a few annotated data, particularly for rare languages currently under documentation. This effort aligns with recent developments in Natural Language Processing, aiming to equip field linguists with computational tools to help them in their documentation effort. We will delve into how these models provide essential building blocks for such tools and the implications of these advancements.

But we also think that neural language models can assist linguists in automatically extracting typological information, such as phoneme inventories and morphosyntactic complexity indices, from audio recordings. In the third part of our presentation, we will share our preliminary work in this direction, illustrating how it becomes possible to identify phonetically similar languages. We will also address the challenges associated with interpreting such measurements, particularly concerning the metadata typically collected in linguistic fieldwork.