Language Clustering with Multilingual Language Models

- Supervisor: Peiqin Lin
- Examiner: Prof. Hinrich Schuetze
- BSc, MSc, Open: MSc
- General Topic Area: Multilinguality
- Prerequisites: enthusiasm, Good programming background (preferably python), basic knowledge of NLP, DL, and Pytorch
- Details: Cross-lingual self-supervised models, like mBERT, XLM, and XLM-R for text, XLS-R for speech, have achieved remakable results on a series of downstream tasks. Language clustering, which measures the cross-lingual similarity among languages, can be leveraged to further improve cross-lingual transferablity of these models. Different linguistic distances, including geographic distance, genetic distance, inventory distance, syntactic distance, and phonological distance, are usually adopted to measure similarity among languages and then cluster them. However, how languages are implicitly clustered in multilingual language models, is still unclear.

  In this thesis, the task is to address one or more of the following research questions.

  - Language clustering on text level.
  - Language clustering on speech level.
  - Comparison among clustering results with different methods.