Normalization of Orthographic and Script Variances with the Help of Parallel Corpus

- **Supervisor**:
- **Examiner**: Prof. Hinrich Schütze
- **Open**: BSc
- **Summary**: Multilingual parallel corpora such as the Parallel Bible Corpus (PBC) [1] served as valuable resources for scaling the NLP models to cover a large number of languages. However, one single language sometimes could be written in various conventions and different sources of text could employ different scripts and/or orthographies. For example, the Uyghur language can be written in Arabic, Cyrillic or Latin alphabet; and the Achi language, although only uses Latin alphabet, has two distinct orthographies for two Bible translations. These variations render data from different sources non-comparable and will affect a model's performance when encountering a non-familiar orthography or script. Drawing inspiration from recent work like [2], this project aims to address this problem by extracting correspondences from available parallel corpora as training data, and train a model to unify different orthographies and scripts. The expected outcome is a model that can automatically normalize differences in orthographies and scripts, for a wide range of languages within our parallel corpus.

- **Prerequisites**: enthusiasm, interest for multilingual NLP, good programming background (preferably Python), familiarity with Deep Learning

[1] http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf
[2] https://aclanthology.org/2023.acl-long.809.pdf