Enhancing Lexical Similarity Calculation through Grapheme-to-Phoneme Conversion

- **Supervisor**:
- **Examiner**: Prof. Hinrich Schütze
- **Open**: MSc
- **Summary**: When calculating lexical similarity, it is fundamental that the calculation should be done in a unified script space. There are transliteration tools such as Uroman [1] to map all texts into the same Latin script, but due to inherent differences of the writing systems, the transliterated words can still be very different from each other even though their pronunciations are the same. For example, while the short vowels in Tajik are expressed in Cyrillic and Latin scripts, they are omitted when written in the Perso-Arabic scripts. To deal with such cases where a simple transliteration is insufficient, this project aims to explore the usage of Grapheme to phoneme (G2P) conversion tools such as [2,3] to map orthographic grapheme sequences into phoneme sequences. In this way, lexical similarity of words written in different scripts can be calculated in a more proper way, which is expected to outperform a simplistic transliteration.

- **Prerequisites**: enthusiasm, good programming background (preferably Python), good knowledge of NLP, familiarity with G2P

[1] https://aclanthology.org/P18-4003/
[2] https://pypi.org/project/epitran/
[3] https://github.com/PasaOpasen/PersianG2P