Transliteration corpora for low-resource scripts

Supervisor: Silvia Severini
Examiner: Prof. Schuetze
BSc, MSc, Open: BSc
General Topic Area: Transliteration, Multilinguality.
Summary: A transliteration pair contains two words which are translated preserving the sound [1]. Examples of such pairs are John/Zan (Eng/Crs), Alex/Алекс (Eng/Rus), and Paris/பாரிஸ் (Eng/Tam). Such pairs are mostly made of named entities and rare words. However, corpora with these types of pairs are missing for low-resource languages.
In this thesis, you will extract parallel corpora of named-entities and rare words for low-resource languages crawling the web (e.g., Wikipedia).
The goal is to create a corpus made of English words paired to low-resource language words, either with Latin scripts (e.g., Zulu) or non-Latin scripts (e.g, Russian) [2].
Finally, you will test the data with transliteration models (e.g, g2p [3], seq2seq [4], …).
Prerequisites: programming experience

[1] Prabhakar, D. K., & Pal, S. (2018). Machine transliteration and transliterated text retrieval: a survey. Sādhanā, 43(6), 1-25.
[2] Menezes, D., Milidiu, R., & Savarese, P. (2019, October). Building a massive corpus for named entity recognition using free open data sources. In 2019 8th Brazilian Conference on Intelligent Systems (BRACIS) (pp. 6-11). IEEE.
[3] Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. Speech communication, 50(5), 434-451.
[4] Rosca, M., & Breuel, T. (2016). Sequence-to-sequence neural network models for transliteration. arXiv preprint arXiv:1610.09565.