In this thesis, we explore the differences between tokenization methods for multilingual neural language models and investigate their impact on language model representation quality.

We propose a set of metrics to evaluate the quality of tokenizations. We show that the metrics capture the differences between tokenizers and that they correlate with the downstream performance of multilingual language models.

Then, using our metrics, we assess why is the standard tokenizer training on a multilingual corpus reported to be ineffective for multilingual models. We investigate design choices such as data size, implementation or alphabet size. We identify that the issue might be caused by data imbalance and to solve it we propose to sample tokenizer training data uniformly.

We compare the standard tokenizer training with three proposed methods we replicate, that aim to mitigate the same reported issues. We show that the principle behind the improvements of the proposed methods is the same as with the uniform sampling.

Our findings offer a deeper understanding of tokenization methods for multilingual models. We propose a methodology and guidelines for training multilingual tokenizers. Lastly, we show how to achieve improvements in tokenization without the need for more complex tokenization methods.