In this thesis, we address the preprocessing approaches that improve the robustness of the subword tokenization for two types of noising.

We are focusing on the inline approaches to casing and diacritics in the texts, that is, allocating the casing and diacritics information to the special tokens that are separate from the words. In the field of casing noise, we compare the performance of our inline casing algorithm, InCa, and the existing solutions for inline case handling. We show that in some noising scenarios, our algorithm shows the best performance, and in the cases where it performs on par with the alternative solutions, the intrinsic parameters of the tokenizer trained on our data are more stable.

For the task of diacritics encoding, we are providing two solutions of inline diacritization, InDia, and show its improvement on robustness against the de-diacritized texts.

Since the final application that we plan to use our tokenizer is Czech-Ukrainian machine translation, we make a thorough comparison of the intrinsic and extrinsic performance of the inline approaches, and show that they have a correlation, although its scope is limited.