The bachelor thesis focuses on basic pre-processing (tokenization and segmentation) of Czech texts, mainly for purposes of Czech internet corpus. The texts for this corpus will be automatically obtained from the world wide web, therefore the segmentation is preceeded by character encoding recognition, cleaning and language identification. We performed experiments with two methods of language identification and present their results. The first method is based on comparison of the most frequent n-grams (substrings of length n) extracted from an unknown document and a large Czech corpus. The second one employs a model estimating word probabilities by conditional probabilities of trigrams estimated on the same corpus. For wider usage, we developed a module for tokenization and identification of sentences boundaries by a decision tree analysis of the nearest context of potential sentence boundaries and utilizing extensive lists of Czech abbreviations. The decision tree was trained on a set of manually processed data. Its evaluation was based on independent human judgements and results are presented in the work.