Abstract

Native Language Identification is the task of identifying an author’s native language based on their productions in a second language. The absolute majority of previous work has focused on English as the second language. In this thesis, we work with 3,715 essays written in Czech by non-native speakers. We use machine learning methods to determine whether an author’s native language belongs to the Slavic language group. By training models with different feature and parameter settings, we were able to reach an accuracy of 78%.