Title: Authorship Attribution of Poetic Texts

Author: Mgr. Petr Plecháč, Ph.D.

Department: Institute of Czech National Corpus

Supervisor: doc. Mgr. Václav Cvrček, Ph.D.

ABSTRACT

Contemporary stylometry offers a number of methods for authorship recognition of poetic texts based on a variety of textual features (e.g. word frequencies, frequencies of character *n*-grams). However, it seems that one important aspect of these texts has been rather left aside – this aspect is versification. The thesis uses four corpora of poetic texts (Czech, German, Spanish, and English) in order to analyze to what extent versification features – such as frequencies of rhythmic patterns or frequencies of various types of rhymes – may be used as an indicator of authorship. We show that (1) versification-based models significantly outperform the *random baseline*, (2) in some cases versification-based models even outperform the traditionally used lexical models, (3) in most of the cases combination of both types of models outperforms the given models alone. Versification features are consequently employed for the purpose of attribution of two texts of doubted authorship: (1) the versified play *The Famous History of the Life of King Henry the Eigth* which was originally published under the name of William Shakespeare, but where many suppose that some parts were actually written by John Fletcher or even other authors, and (2) poems published under the name of Josef Barák where there is a hypothesis that their real author is Jan Neruda.

KEYWORDS

authorship attribution, stylometry, versification, machine learning, corpus linguistics