Instytut Języka Polskiego
Polskiej Akademii Nauk

tel. +48 12 632 56 92
fax. +48 12 632 92 41
www.ijp.pan.pl
ijp@ijp.pan.pl

31-120 Kraków
al. Mickiewicza 31
NIP: 675-00-01-892
REGON: 357534720

Stylometry Expertise Conclusion

1. Task
Assessing authorship of an anonymous Persian qasida

2. Data set
A corpus in UTF-8 encoding consisting of:

a) Examined text: the anonymous "...ar"-ending qasida — 837 words
b) A set of texts written by candidate authors:
   i) 50 "...ar"-ending qasidas by Amir Mu'izzi — over 35 000 words
   ii) 12 "...ar"-ending qasidas by Farrukhi Sistani — over 8 000 words
   iii) 9 "...ar"-ending qasidas by Anwari — over 7 000 words

3. Methods applied
All the methods were used as implemented in 'stylo' R package (by Eder et al. 2016).

1. Authorship verification with General Imposters (GI) method (2014 paper by Koppel and Winter, and 2016 by Kestemont et al.).
2. Cluster analysis with cosine delta measure, using:
   a. frequency tables built on relative frequencies of 50—300 most frequent words (MFW) as the frequencies of MFWs — usually grammatical words — are very strong predictors of authorial uniqueness (first shown by Mosteller and Wallace in 1964);
   b. frequency table built on relative frequencies of "...ar"-ending words.
3. Classification with Support Vector Machine (SVM), a series of 100 tests using randomly selected equal numbers of texts by each author as training data.
4. Rolling stylometry method (supervised machine learning in sequential analysis) using Nearest Shrunken Centroid (NSC) as well as the SVM and Delta classification methods, 100 MFWs and samples of 100 words with 75 word overlap.

4. Conclusions
First of all, it is important to state that authorship attribution and verification never guarantees absolute accuracy. This is even less so in cases of very short (shorter than 2000 words) texts, poetry and works created within one "school" of writing.
All these conditions increase the noise and dilute author's signal. However, successful attempts were made recently (2018 study by Franzini et al.) in examining authorship of similarly short texts.

Having disclaimed that, the results of all conducted tests generally point to Mu'izzi as the most likely author of the examined anonymous qasida, with a margin of error that is however impossible to eliminate with current state of the art methods and knowledge. Even if the quality of the input dataset has been carefully cross-checked and the texts normalized, the inherent nature of any written text involves some noise, and for this very reason one cannot guarantee 100% accuracy of the attribution.

As the results were fairly stable for examined scope of 50 to 350 MFWs, most tests were conducted on 50 and 100 MFWs. Below, I focus on the SVM results rather than those of the NSC because the SVM tends to be less sensitive to feature selection and so more reliable, even if it at times scoring lower accuracy. I do note however, that the NSC results were even more in favour of Mu'izzi's authorship.

To even out the size of training data for candidate authors, the most reliable tests, classification with the SVM and verification with the GI method were conducted on adjusted subsets (100 iterations, in each 8 texts by each of the candidate authors were randomly selected, half of the texts being used as training data, and the rest for classifier's evaluation).

The SVM classification recognized Mu‘izzi as the author in 75 runs, and Anwari and Farrukhi respectively in 21 and 4. While success rate of the SVM 100 MFW classification varied greatly, it is important to stress that Mu‘izzi’s texts were misattributed much less often than Anwari’s and Farrukhi’s. This suggests his signal to be more distinct than those of other authors. Length of the training texts did not seem to be a crucial factor — both shortest and longest texts in the corpus got misattributed with similar likelihood.

The GI method — as a method of authorship verification — considers the problem in the open set settings (that is, assesses whether any of the candidate authors is at all likely to be the author), also recognized Mu‘izzi as the author with higher accuracy, scoring over 0.8 in most (76) runs (1 being the highest possible score for this method). With a few runs’ exceptions, Farrukhi and Anwari were significantly less likely to author the text.

The imperfect attribution in the results of both tests seems likely to be caused by genre and time period noise, which should be verifiable by means of qualitative analysis of the texts.

On the suggestion of Alexey Khismatulin noting the importance of “...ar”-ending words finishing every second verse,
I conducted additional classification, bootstrap consensus tree, that is a result of agreement between a series of cluster analyses with a set threshold (here: 50%), using custom wordlist composed of these words as features for classifier. The method examined similarities in distribution of these words and showed two qasidas by Mu‘izzi as the closest neighbors of the anonymous text (Fig. 1).

Finally, I also used rolling stylometry method to examine and visualize the changes in similarity of the text to candidate authors. While different methods showed various scales of influence of individual authors, Mu‘izzi was always the most pronounced author (SVM on Fig. 2).

All in all, in the light of currently available information, I conclude that Mu‘izzi is the most likely to be the author
of the examined anonymous text.

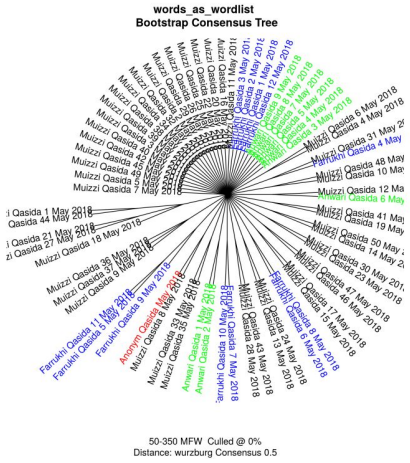

Fig. 1. *Bootstrap consensus cluster analysis classification using “...ar”-ending words as MFW features.*

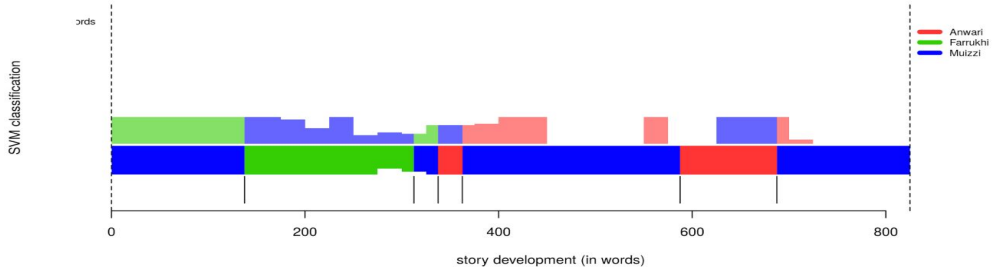

Fig. 2*. Rolling stylometry classification using the SVM method, 100 MFW as features and samples of 100 words with 75 word overlap. The lower of the two lines shows dominating influence, the upper — the second most likely influence. The thickness of each of the lines shows the intensity of the influence.*

Expert
Joanna Byszuk