Attribution of Authorship for Medieval Persian Quasidas with Stylometry

8th June 2019, Victoria BC, R2L workshop at DHSI19

Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences)

and

Alexey Khismatulin (Institute of Oriental Manuscripts, Russian Academy of Sciences)

Disclaimer of the speaker

https://computationalstylistics.github.io/projects/focs/

10 Computational 01
01 Stylistics 0101000
11 Group 011010110

TT: @jbyszuk

http://joannaby.github.io

PAN IJP Institute of Polish Language Polish Academy of Sciences

Stylometry

What is stylometry?

Stylometry

use of quantitative methods to examine similarities and differences within a group of texts

What is it useful for?

- authorship attribution,
- tracing chronology,
- analysis of cross and inter genre relationships,
- big data analysis,
- style transfer and anonymization,
- and many others.

What is stylometry?

corpus of texts

+

distance measure

+

classification algorithm

+

(visualisation)

Background of the problem

Siyar al-muluk

- first political treatise written in Persian by Nizam al-Mulk(k. 485/1092), the great prime minister of the Saljuq dynasty.
- first publication by Ch.Schefer (d. 1898) in 1891 up to the present time, since then several critical editions

Siyar al-muluk – appended quasida

- anonymous qasida composed in praise of the Sultan Muhammad b. Malikshahand appended to the first redaction of the Siyar al-muluk

- held by the British and Berlin libraries, transcribed in 1032/1623 and1058/1648, from the protocopy made in the city of Urumiya in 564/1168

Our approach

Data

A corpus of normalized texts in UTF-8 encoding consisting of:

- Examined text: the anonymous “...ar”-ending qasida — 837 words
- [and a set of candidate authors]

Data

A corpus of normalized texts in UTF-8 encoding consisting of: [Examined text] and

a set of texts written by candidate authors:

1. 50 “...ar”-ending qasidas by Amir Mu‘izzi — over 35 000 words
2. 12 “...ar”-ending qasidas by Farrukhi Sistani — over 8 000 words
3. 9 “...ar”-ending qasidas by Anwari — over 7 000 words

Methods

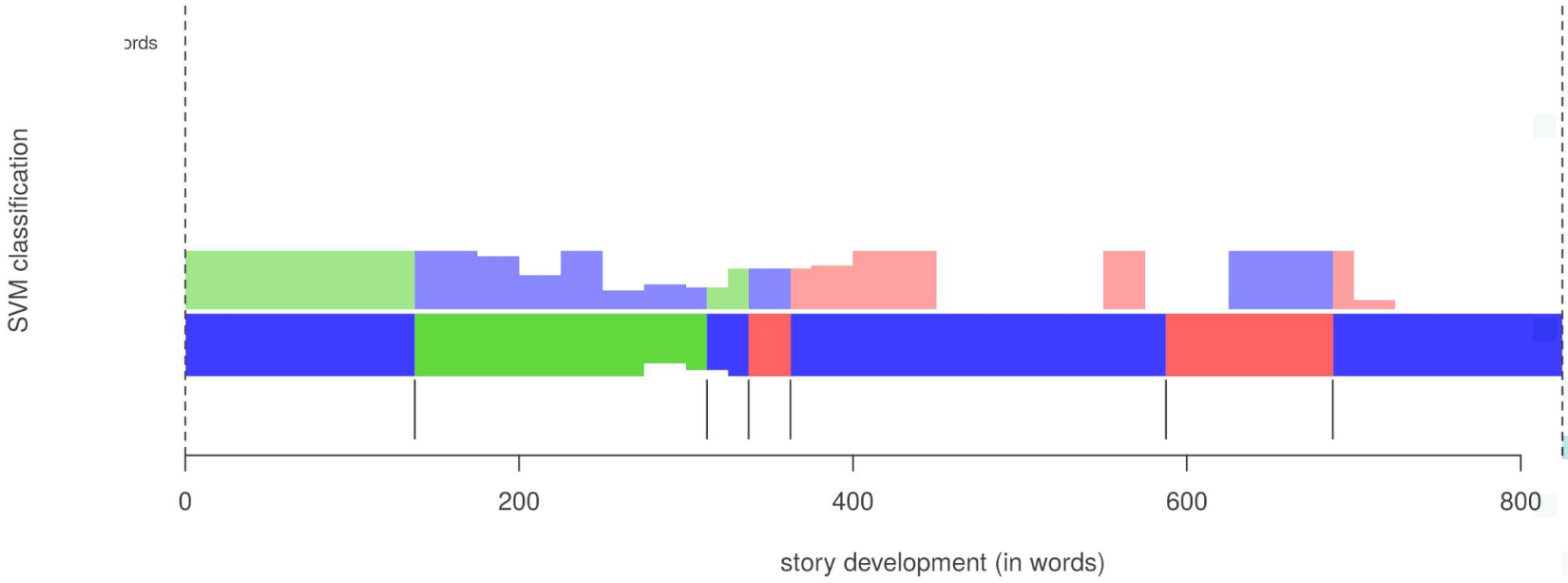
- Cluster analysis
- Classification with Support Vector Machine
- Rolling stylometry method
- Authorship verification with General Imposters (GI) method

All as implemented with stylo R package.

Methods – data selection

- A series of experiments:
  - Random selection of 9 sample texts per candidate author
  - How many most frequent words? 50-350
  - Specialized wordlist

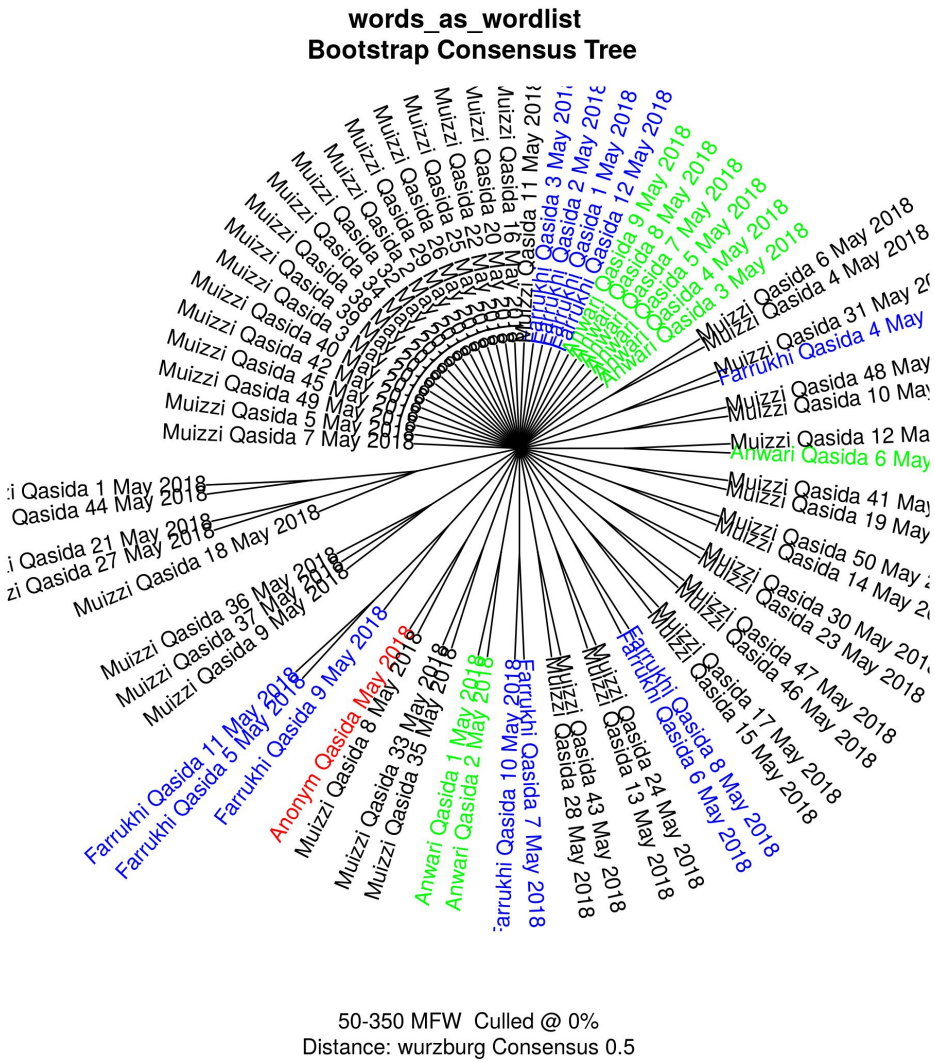
Methods


ords
Anwari
Farrukhi
Muizzi
SVM classification
0
200
400
600
800
story development (in words)

Problem – inconsistent authorial signal

Solution – using rhyming *-ar* words as features

Result – almost perfect classification



Bootstrap consensus

50-350 most frequent words, Cosine Delta as a distance measure

Conclusions

Amir Mu‘izzi likely the author

- Most often pointed by computational methods
- Supported by literary and textual arguments

questions, questions, questions

Confusing attribution with existing methods

Feature of

- only this set?
- historical texts (editions, redactions)?
- texts written in R2L systems?

Confusing attribution with existing methods

Attribution with specialized wordlist(s)

–

a solution?

Further problems

- Can we (how to?) use ngrams of words / characters?
- How to prepare data?
- Lemmatize?

Thank you!

joanna.byszuk@ijp.pan.pl
khism@mail.ru