Fine-tuning our Stylometric Tools: Investigating Authorship, Genre, and Form in French Classical Theater

Christof Schöch
University of Würzburg, Germany

Digital Humanities Conference
Lincoln, Nebraska, July 15-20, 2013
corpus: french classical drama

32 / 40 plays
comedies in verse
comedies in prose
tragedies in verse
tragedies in prose

Pierre Corneille
Thomas Corneille
Charles Dufresny
Georges de Scudéry
Jean-Francois Regnard
Molière
problem: genre, form, author
one approach: PCA

Pierre Corneille, comedies
Thomas Corneille, comedies
Pierre Corneille, tragedies
Thomas Corneille, tragedies
hypothesis: parts of the wordlist

- ces
- mort
- frere
setup: moving wordlist onset

first run: onset 1, length 100
words 1-100 from the wordlist

second run: onset 50, length 100
words 51-150

third run: onset 101, length 100
words 101-200

e tc.

cluster analysis for each run
method: check low-level cluster pairs

four possibilities:

- author and genre are identical (AG)
- author, but not genre are identical (Ax)
- genre, but not author are identical (xG)
- neither author nor genre are identical (xx)
results: proportions of pair types

author vs. genre
results: proportions of pair types

author vs. form
conclusion: authorship attribution

What you can do

- include plays heterogeneous as to their genre (tragedy vs. comedy),
- if you do not extend the wordlist beyond 1100 most frequent words.

What you can't do

- do not include plays heterogeneous as to form (verse vs. prose),
- even if you are using only the top of the wordlist.
conclusion: authorship attribution

What you can do
• include plays heterogeneous as to their genre (tragedy vs. comedy),
• if you do not extend the wordlist beyond 1100 most frequent words.

What you can't do
• do not include plays heterogeneous as to form (verse vs. prose),
• even if you are using only the top of the wordlist.

More generally, these results
• could make clustering and attribution more reliable
• particularly important when you cannot balance your corpus

However
• this is not actually the best way to solve this problem
Thank you!

And special thanks to Maciej Eder, Jan Rybicki & Mike Kestemont, authors of the „stylo” package for R

All data available at: http://dragonfly.hypotheses.org/424

Christof Schöch
University of Würzburg, Germany

Digital Humanities Conference
Lincoln, Nebraska, July 15-20, 2013