Computational stylistics: tools and methods

Maciej Eder

Pedagogical University in Krakow & Polish Academy of Sciences

29/10/2012
1. Stylometry: measuring word frequencies (counting words)

2. Multidimensional approaches to capturing authorial style

3. Grammatical structures: do they improve stylometric performance?
Some research questions of stylometry

- What is *common* in the language and what is related to cultural contexts and/or writer’s *individuality*?

- What elements of style are affected by literary period, genre, topic?

- What is unconsciously incorporated by the author and reflects his/her education, gender, religious background, social or historical conditions?

- Which features of a written text can betray the person who wrote it *despite* his/her aesthetic, social, or historical conditions?
chapter 1

Intuitively,

style is determined by meaningful words
abbot Hugo de Saint-Cher: first concordance (ca 1230)

Lorenzo Valla: authorship attribution of *The Donation of Constantine*

Erasmus Roterodamus: attribution of St. Paul’s and Seneca’s letters

Concordances, attributions, ...

- abbot Hugo de Saint-Cher: first concordance (ca 1230)
- Lorenzo Valla: authorship attribution of *The Donation of Constantine*
- Erasmus Roterodamus: attribution of St. Paul’s and Seneca’s letters

Mostly friars and monks! (stylometry requires patience...)

Maciej Eder  Computational stylistics
Cruden’s Concordance to the *King James Bible* (1737)
How to read 5 million books? (Google ngram viewer)

129 million books published

15 million books scanned

5 million books analyzed

Frequency of the word "apple"
Evolution of language: burnt vs. burned

Maciej Eder | Computational stylistics
Decline of classical education

Maciej Eder
Computational stylistics
Hamlet: a word cloud (www.wordle.net)
Hamlet: a word cloud (function words excluded)

Maciej Eder

Computational stylistics
(1) Sterne, *Sentimental*, (2) Hor. *Ars*, (3) *Bartachom*.

<table>
<thead>
<tr>
<th></th>
<th>the</th>
<th>et</th>
<th>δ’</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.7384</td>
<td>4.69104</td>
<td>4.3154</td>
</tr>
<tr>
<td>2</td>
<td>3.2138</td>
<td>1.26173</td>
<td>3.09004</td>
</tr>
<tr>
<td>3</td>
<td>3.1292</td>
<td>1.22938</td>
<td>2.18434</td>
</tr>
<tr>
<td>4</td>
<td>3.0518</td>
<td>1.13232</td>
<td>1.97123</td>
</tr>
<tr>
<td>5</td>
<td>2.7640</td>
<td>0.905856</td>
<td>1.86468</td>
</tr>
<tr>
<td>6</td>
<td>2.4061</td>
<td>0.841152</td>
<td>1.17208</td>
</tr>
<tr>
<td>7</td>
<td>1.9104</td>
<td>0.776448</td>
<td>1.06553</td>
</tr>
<tr>
<td>8</td>
<td>1.8838</td>
<td>0.744096</td>
<td>1.01225</td>
</tr>
<tr>
<td>9</td>
<td>1.1583</td>
<td>0.711744</td>
<td>0.745871</td>
</tr>
<tr>
<td>10</td>
<td>1.1462</td>
<td>0.711744</td>
<td>0.745871</td>
</tr>
<tr>
<td>11</td>
<td>1.1100</td>
<td>0.420576</td>
<td>0.692595</td>
</tr>
<tr>
<td>12</td>
<td>1.0834</td>
<td>0.420576</td>
<td>0.639318</td>
</tr>
<tr>
<td>13</td>
<td>1.0108</td>
<td>0.388224</td>
<td>0.639318</td>
</tr>
<tr>
<td>14</td>
<td>0.9673</td>
<td>0.388224</td>
<td>0.586042</td>
</tr>
</tbody>
</table>

...
Zipf’s law

Rank/frequency dependence (Zipf’s law)
chapter 2

*the, in, of, or, I, is, ...*

or the mystery of bare function words
Wincenty Lutosławski (1863–1954)
Non-traditional authorship attribution

Given:

- a text of uncertain or anonymous authorship, and
- a comparison corpus of texts by known authors,

is it possible to find the ‘nearest neighbor’ among the available candidates?
What should be analyzed? What is the style?

- An idea of stylistic fingerprint
  - indiscernible with the naked eye
  - beyond authorial control
  - resistant to imitation, plagiarism and parody
  - popular solution: usage of function words

- Is the style determined by an individual? (cf. human DNA code, fingerprint, or patterns in one's iris)
‘If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting’

(Lutosławski 1897: 66)
‘If handwriting can be so exactly determined as to afford certainty as to its identity, so also with style, since style is more personal and characteristic than handwriting’

(Lutosławski 1897: 66)

However, nowadays we rather seek for some statistically significant regularities rather than for a determined uniqueness in style
How to measure differences between texts?

- **One-dimensional methods**
  - indexes of lexical density
  - mean word length, mean sentence length
  - degree of rhythmicity
  - ...

- **Multidimensional methods**
  - several features (dozens, or hundreds) measured at once
  - they aggregate weak discriminative strength of single features
Dalimil’s Chronicle – stylistic change

Increasing number of syllables in a line

Mean length of lines (in syllables)

Chapters of the chronicle

Maciej Eder  Computational stylistics
Where stylometry (usually) begins: table of frequencies

<table>
<thead>
<tr>
<th></th>
<th>Milton Samson</th>
<th>Milton Paradise</th>
<th>Keats Hyperion</th>
<th>Patmore Eros</th>
<th>Browning Bishop</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the”</td>
<td>4.57</td>
<td>4.24</td>
<td>4.25</td>
<td>4.19</td>
<td>4.47</td>
<td>...</td>
</tr>
<tr>
<td>“to”</td>
<td>3.11</td>
<td>3.29</td>
<td>3.43</td>
<td>3.14</td>
<td>3.71</td>
<td>...</td>
</tr>
<tr>
<td>“and”</td>
<td>3.19</td>
<td>3</td>
<td>3.08</td>
<td>2.85</td>
<td>2.81</td>
<td>...</td>
</tr>
<tr>
<td>“of”</td>
<td>2.6</td>
<td>3</td>
<td>2.63</td>
<td>2.43</td>
<td>2.86</td>
<td>...</td>
</tr>
<tr>
<td>“I”</td>
<td>2.17</td>
<td>2.2</td>
<td>2.13</td>
<td>2.42</td>
<td>2.22</td>
<td>...</td>
</tr>
<tr>
<td>“a”</td>
<td>2.24</td>
<td>1.92</td>
<td>1.92</td>
<td>2.21</td>
<td>1.92</td>
<td>...</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td></td>
</tr>
</tbody>
</table>

Maciej Eder
Computational stylistics
Where stylometry (usually) begins: table of frequencies

<table>
<thead>
<tr>
<th></th>
<th>Milton Samson</th>
<th>Milton Paradise</th>
<th>Keats Hyperion</th>
<th>Patmore Eros</th>
<th>Browning Bishop</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the”</td>
<td>4.57</td>
<td>4.24</td>
<td>4.25</td>
<td>4.19</td>
<td>4.47</td>
</tr>
<tr>
<td>“to”</td>
<td>3.11</td>
<td>3.29</td>
<td>3.43</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>“and”</td>
<td>3.19</td>
<td>3</td>
<td>3.08</td>
<td>2.85</td>
<td>2.81</td>
</tr>
<tr>
<td>“of”</td>
<td>2.6</td>
<td>3</td>
<td>2.63</td>
<td>2.43</td>
<td>2.86</td>
</tr>
<tr>
<td>“I”</td>
<td>2.17</td>
<td>2.2</td>
<td>2.13</td>
<td>2.42</td>
<td>2.22</td>
</tr>
<tr>
<td>“a”</td>
<td>2.24</td>
<td>1.92</td>
<td>1.92</td>
<td>2.21</td>
<td>1.92</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

...?
Where stylometry (usually) begins: table of frequencies

<table>
<thead>
<tr>
<th></th>
<th>Milton Samson</th>
<th>Milton Paradise</th>
<th>Keats Hyperion</th>
<th>Patmore Eros</th>
<th>Browning Bishop</th>
</tr>
</thead>
<tbody>
<tr>
<td>“the”</td>
<td>4.57</td>
<td>4.24</td>
<td>4.25</td>
<td>4.19</td>
<td>4.47</td>
</tr>
<tr>
<td>“to”</td>
<td>3.11</td>
<td>3.29</td>
<td>3.43</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>“and”</td>
<td>3.19</td>
<td>3</td>
<td>3.08</td>
<td>2.85</td>
<td>2.81</td>
</tr>
<tr>
<td>“of”</td>
<td>2.6</td>
<td>3</td>
<td>2.63</td>
<td>2.43</td>
<td>2.86</td>
</tr>
<tr>
<td>“I”</td>
<td>2.17</td>
<td>2.2</td>
<td>2.13</td>
<td>2.42</td>
<td>2.22</td>
</tr>
<tr>
<td>“a”</td>
<td>2.24</td>
<td>1.92</td>
<td>1.92</td>
<td>2.21</td>
<td>1.92</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

. . . . ? . . . . . . . . . ??
Explanatory approach: Cluster Analysis

English_poetry
Cluster Analysis

100 MFWs, Culled @ 0 %
Pronouns deleted: TRUE; Distance: Classic Delta
Greek New Testament and its authors

- **Synoptic Gospels** – strong similarities, numerous identical sentences, c. 50% of common material
  - **Mark**: written c. 68–73
  - **Matthew**: c. 70–100
  - **Luke**: c. 80–100
- **John**: c. 90–100 – gospel developed from a *Johannine circle*
- **Acts**: c. 60–64 – traditionally ascribed to *St. Luke*
- **Epistles** – Paul, James, John, Peter, Jude
- **Revelation** – traditionally ascribed to *St. John*, but also to John the Presbyter (Dionysius of Alexandria, 3rd cent.)
Greek New Testament: textual similarities

Novum Testamentum Graece Unicode
Multidimensional Scaling

30 MFWs, Culled @ 0 %
Classic Delta distance
Different method, similar results

Unrooted consensus tree (p=0.5)
Another approach (less rigorous)

Unrooted consensus tree (p=0.25)
Applications of multidimensional methods

- authorship attribution
- genre recognition
- gender recognition
- chronology
- ...

Maciej Eder  Computational stylistics
A break in style after World War II

Maciej Eder
Computational stylistics
How far can we go without any words?
My father was a clergyman.

- PRP$: (poss. pron.)
- NN: (noun)
- VBD: (verb)
- DT: (determiner)
- NN: (noun)
Different output formats, similar results

Stanford POS Tagger

My_PRP$ father_NN was_VBD a_DT clergyman_NN ...

TreeTagger

My PP$ my
father NN father
was VBD be
a DT a
clergyman NN clergyman
::
It was by your desire that I first thought of such a composition. So many years have since past, that you may have forgotten this circumstance: but your desires are to me in the nature of commands: and the impression of them is never to be erased from my memory.

Again, Sir, without your assistance this history had never been completed. Be not startled at the assertion. I do not intend to draw on you the suspicion of being a romance writer. I mean no more than that...
What about getting rid of original words?

It was by your desire that I first thought of such a composition. So many years have since past, that you may have forgotten this circumstance: but your desires are to me in the nature of commands: and the impression of them is never to be erased from my memory.

Again, Sir, without your assistance this history had never been completed. Be not startled at the assertion. I do not intend to draw on you the suspicion of being a romance writer. I mean no more than that...
It was by your desire that I first thought of such a composition. So many years have since past, that you may have, perhaps, forgotten this circumstance: but your desires are to me in the nature of commands; and the impression of them is never to be erased from my memory. Again, Sir, without your assistance this history had never been completed. Be not startled at the assertion. I do not intend to draw on you the suspicion of being a romance writer. I mean no more than that I partly owe to you my existence during great part of the time which I have employed in composing it: another matter which it may be necessary to remind you of; since there are certain actions of which you are apt to be extremely forgetful; but of these I hope I shall always have a better memory than yourself. Lastly, It is owing to you that the history appears what it now is. If there be in this work, as some have been pleased to say, a stronger picture of a truly benevolent mind than is to be found in any other, who that knows you, and a particular . . .
... we can use tags as a ‘normal’ running text

Maciej Eder
Computational stylistics
Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una, pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. . . .
... and its grammatical representation.

POS-tags analysis: disappointing

Multidimensional Scaling

tagged_small_collection_ST

55 MFW 1-grams Culled @ 0%
Classic Delta distance
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN ...
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN
n-grams of POS tags

```plaintext
PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ
NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC
PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN
PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$
NN DT NN VBD RB VBN VBN . .

PRP_VBD_IN VBD_IN_PRP$
```
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN  VBD_IN_PRP$  IN_PRP$_NN  PRP$_NN_IN
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN NN_IN_PRP
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN VBD_IN_PRP$ IN_PRP$_NN PRP$_NN_IN NN_IN_PRP IN_PRP_RB
n-grams of POS tags

PRP VBD IN PRP$ NN IN PRP RB VBD IN JJ DT NN . RB JJ NNS VBP IN NN , IN PRP MD VB , RB , VBN DT NN : CC PRP$ NNS VBP TO PRP IN DT NN IN NNS : CC DT NN IN PRP VBZ RB TO VB VBN IN PRP$ NN . RB , NNP , IN PRP$ NN DT NN VBD RB VBN VBN . . .

PRP_VBD_IN  VBD_IN_PRP$  IN_PRP$_NN  PRP$_NN_IN NN_IN_PRP  IN_PRP_RB  . . .
3-grams of POS tags = no words, just labels!

100-5000 MFW 3-grams Culled @ 0%
Classic Delta distance Consensus 0.5

Maciej Eder  Computational stylistics
did Latin philosophers develop their own style?

do POS-tags depend on genre?

is there a difference in performance between MFWs and POS tags?
How did they know which tags were appropriate?!

n-acc v-ind sent
n-acc v-ind pun
n-abl v-ind sent
v-ind sent adv
v-ind pun cs
n-abl cc n-abl
adv v-ind pun
n-acc cc n-acc
n-abl v-ind pun
v-ind sent cc
v-ind pun adv
prep n-abl n-gen
v-ind sent pun
v-ind pun rel
prep adj-abl n-abl
...
Conclusions

- Stylometry can retrieve much information from content (i.e. ‘normal’) words, ...

- ... but bare function words more precisely distinguish authorial uniqueness.

- Best performance, however, is achieved when the words are thrown away!

- Thus, grammatical labels (tags) are worth to be thoroughly tested in future experiments, ...

- ... even if it looks pretty heretic.
Run your own experiment!
Thank you!