Textual similarities
stylometry
authorship attribution

a short introduction

Maciej Eder

Institute of Polish Language
(Polish Academy of Sciences)

Quantitative methods in lit studies: why?

- They uncover stylistic (or linguistic) patterns invisible for a naked eye.
- They can generalize from large amounts of data (text collections), which is beyond capabilities of any single scholar.
- They reduce complex (stylistic, linguistic, etc.) phenomena into a number quantifiable features that can be modeled using statistical apparatus.

Stylometry

- Statistical analysis of written style, ...

- ... aimed at finding some order in "randomness"

- Or better: extracting a signal from noise

- Usual application: authorship attribution

Authorship attribution
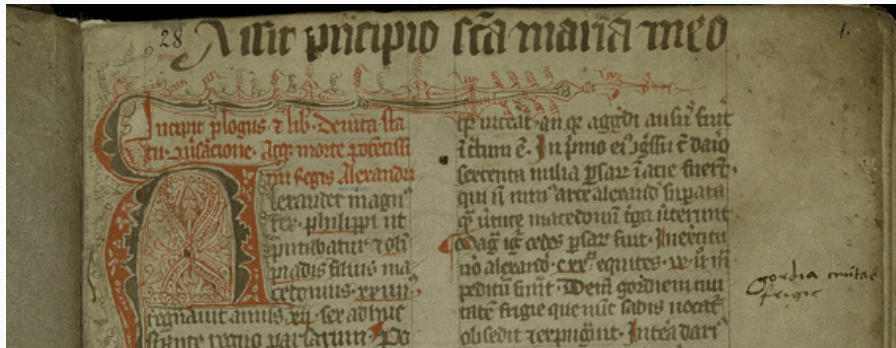
Counting the words

'Ni vestra auctoritate suffultus, patres pretitulati, vestraque opitulatione fretus fierem, meis viribus in vanum tanti ponderis onus subirem, et cum fragili lembo periculose tantam equoris immensitatem introirem.'
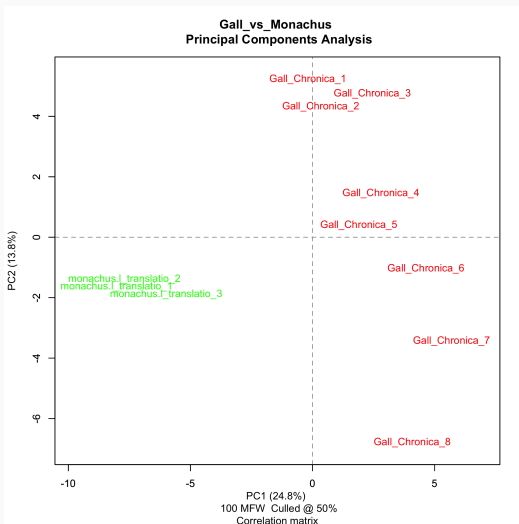
(Gallus Anonymous, *Chronica*)

"et"	3.646 %
"in"	1.840 %
"non"	1.840 %
"est"	0.475 %
"ut"	0.298 %
"ad"	0.852 %
. . .	. . .

Table of frequencies: the evidence

	Gallus *Chronica*	Monachus *Translatio*	Benedictus *Regula*	Alanus *Planctus*	Abelardus *Consolatio*	...
"et"	3.646	4.756	4.159	0.778	3.540	...
"in"	1.840	2.144	2.445	2.849	2.189	...
"non"	1.840	2.144	2.445	2.849	2.189	...
"est"	0.475	0.392	1.057	0.242	0.772	...
"ut"	0.298	0.325	1.185	0.684	1.065	...
"ad"	0.852	1.176	0.898	0.572	1.418	...
"cum"	1.150	1.235	0.755	0.543	1.040	...
"quod"	0.532	0.668	0.906	0.100	0.856	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

6

Gall_vs_Monachus
Principal Components Analysis

Gallus_Anonymus
Cluster Analysis

bernardus_nova_militia
bernardus_de_laudibus_mariae
bernardus_vita_malachiae
hist_alex_magni
translatio_agathae
chronica_monasterii_michaelis
monachus.l_translatio
Gall_Chronica
alex.telesinus_de_rogerio

199 MFW Culled @ 100%
Classic Delta distance
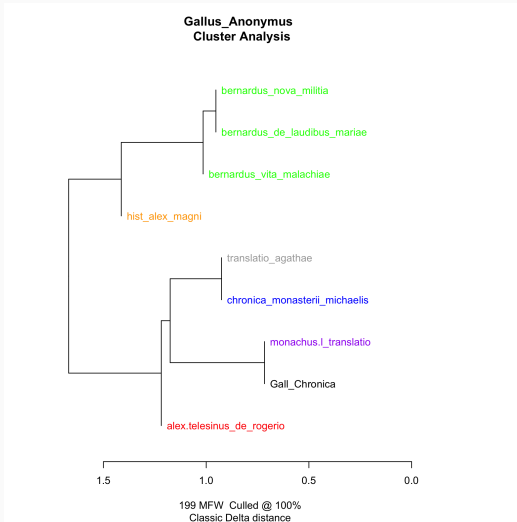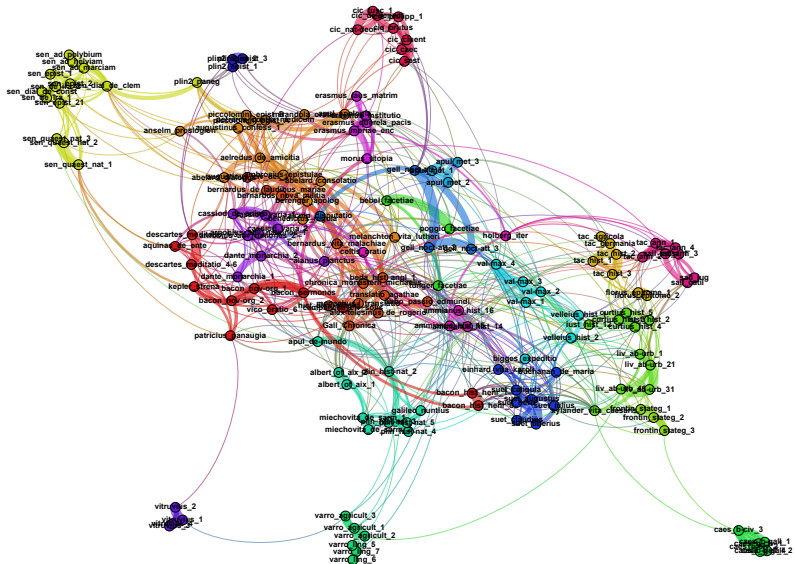
Gallus next to Monachus Littorensis
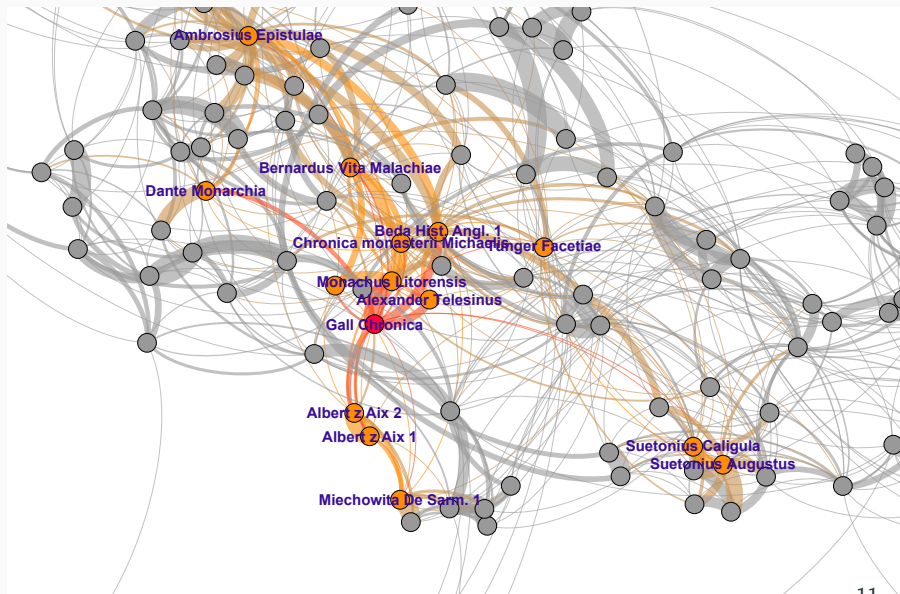
Text similarities beyond attribution

Stylometry beyond attribution
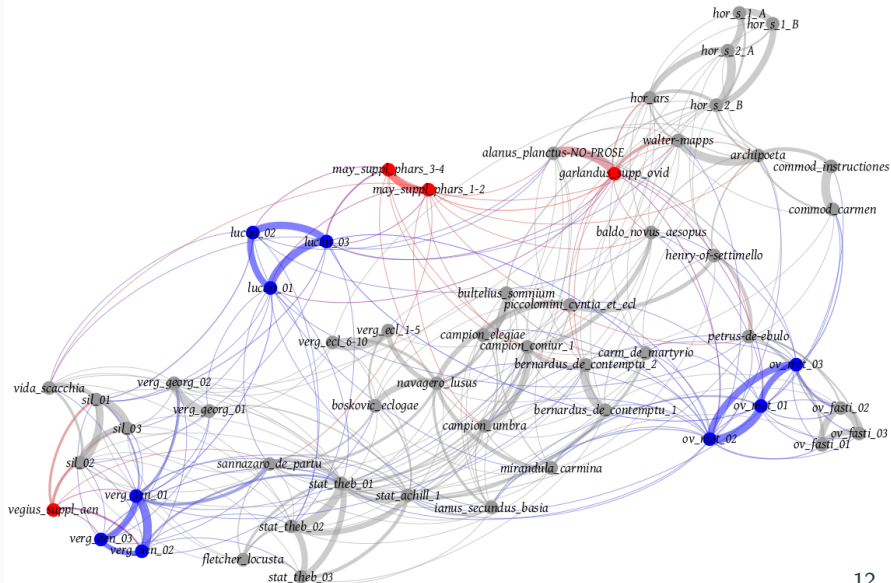
Authorial fingerprint

What **should be analyzed?** *What* **is the style?**

- An idea of stylistic fingerprint

    - undiscoverable with the naked eye
    - beyond authorial control
    - resistant to imitation, plagiarism and parody
    - popular solution: frequencies of the Most Frequent Words

- Is the style is 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)

However, nowadays we rather seek for some statistically significant regularities rather than for a determined uniqueness in style

Function words: English text

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do you not want to know who has taken it?" cried his wife

		(1) Sterne, *Sentimental*,		(2) Hor. *Ars*,		(3) *Bartachom.*
1.	the	5.7384	et	4.69104	δ'	4.3154
2.	i	3.2138	non	1.26173	καὶ	3.09004
3.	and	3.1292	si	1.22938	ἐς	2.18434
4.	of	3.0518	in	1.13232	δὲ	1.97123
5.	to	2.7640	aut	0.905856	ἦν	1.86468
6.	a	2.4061	qui	0.841152	ὁ	1.17208
7.	it	1.9104	ut	0.776448	οὐ	1.06553
8.	in	1.8838	quid	0.744096	ἐπὶ	1.01225
9.	had	1.1583	nec	0.711744	τὸν	0.745871
10.	was	1.1462	est	0.711744	κατὰ	0.745871
11.	as	1.1100	an	0.420576	τε	0.692595
12.	my	1.0834	ad	0.420576	ἄπ'	0.639318
13.	his	1.0108	quae	0.388224	μὲν	0.639318
14.	he	0.9673	ego	0.388224	ἐπ'	0.586042
. . .	. . .	. . .	. . .	. . .	. . .	. . .

16

Table of relative frequencies

	ABronte *Agnes*	Austen *Emma*	CBronte *Jane*	Conrad *Lord*	Dickens *Bleak*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table of relative frequencies

	ABronte	Austen	CBronte	Conrad	Dickens	
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

What should be measured:

17

Table of relative frequencies

	ABronte	Austen	CBronte	Conrad	Dickens	
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

What should be measured: MFWs?

Table of relative frequencies

	ABronte *Agnes*	Austen *Emma*	CBronte *Jane*	Conrad *Lord*	Dickens *Bleak*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

What should be measured: MFWs? Function words?

Table of relative frequencies

	ABronte	Austen	CBronte	Conrad	Dickens	
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

What should be measured: MFWs? Function words? other features?

Different style-markers: an example

'It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.'

(J. Austen, *Pride and Prejudice*)

Splitting the input text: examples

single words:

'it' 'is' 'a' 'truth' 'universally' 'acknowledged' 'that' 'a' . . .

word bi-grams:

'it-is' 'is-a' 'a-truth' 'truth-universally' 'universally-acknowledged'
. . .

letter bi-grams:

'it' 't-' '-i' 'is' 's-' '-a' 'a-' '-t' 'tr' 'ru' 'ut' 'th' 'h-' '-u' . . .

letter tri-grams:

'it-' 't-i' '-is' 'is-' 's-a' '-a-' 'a-t' '-tr' 'tru' 'rut' 'uth' 'th-' . . .

letter hexa-grams:

Table of relative frequencies (letter tri-grams)

	ABronte	Austen	CBronte	Conrad	Dickens	
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"-th"	1.4901	1.3523	1.5002	1.8030	1.5110	...
"the"	1.2484	1.0411	1.1717	1.4986	1.1899	...
"he-"	0.9835	1.0204	1.1500	1.4710	1.0677	...
"-an"	0.8949	0.7362	0.8166	0.6324	0.8086	...
"nd-"	0.8195	0.6858	0.8380	0.6497	0.8069	...
"er-"	0.6849	0.7428	0.6683	0.4968	0.6282	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

Table of relative frequencies (letter tri-grams)

	ABronte *Agnes*	Austen *Emma*	CBronte *Jane*	Conrad *Lord*	Dickens *Bleak*	...
"-th"	1.4901	1.3523	1.5002	1.8030	1.5110	...
"the"	1.2484	1.0411	1.1717	1.4986	1.1899	...
"he-"	0.9835	1.0204	1.1500	1.4710	1.0677	...
"-an"	0.8949	0.7362	0.8166	0.6324	0.8086	...
"nd-"	0.8195	0.6858	0.8380	0.6497	0.8069	...
"er-"	0.6849	0.7428	0.6683	0.4968	0.6282	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

A non-trivial question: How many markers should be analyzed?

Table of relative frequencies (letter tri-grams)

	ABronte	Austen	CBronte	Conrad	Dickens	
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"-th"	1.4901	1.3523	1.5002	1.8030	1.5110	...
"the"	1.2484	1.0411	1.1717	1.4986	1.1899	...
"he-"	0.9835	1.0204	1.1500	1.4710	1.0677	...
"-an"	0.8949	0.7362	0.8166	0.6324	0.8086	...
"nd-"	0.8195	0.6858	0.8380	0.6497	0.8069	...
"er-"	0.6849	0.7428	0.6683	0.4968	0.6282	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

A non-trivial question: How many markers should be analyzed?

Table of relative frequencies (letter tri-grams)

| | ABronte | Austen | CBronte | Conrad | Dickens | |
	Agnes	*Emma*	*Jane*	*Lord*	*Bleak*	...
"-th"	1.4901	1.3523	1.5002	1.8030	1.5110	...
"the"	1.2484	1.0411	1.1717	1.4986	1.1899	...
"he-"	0.9835	1.0204	1.1500	1.4710	1.0677	...
"-an"	0.8949	0.7362	0.8166	0.6324	0.8086	...
"nd-"	0.8195	0.6858	0.8380	0.6497	0.8069	...
"er-"	0.6849	0.7428	0.6683	0.4968	0.6282	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

A non-trivial question: How many markers should be analyzed?

In search of an ideal vector of features (markers)

- a small number of function words (Mosteller & Wallace 1964)
- 50 Most Frequent Words (or other markers) (Juola 2008)
- 100 Most Frequent Words (Burrows 2002)
- 300 MFWs (Smith & Aldridge 2011)
- 500 MFWs (Craig & Kinney 2009)
- >1000 MFWs (Hoover 2004)
- the entire word-list (Juola 2008)
- there is no universal solution (Rybicki & Eder 2011)

How do we trace the fingerptint?

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?

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1|$

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1| + |a_2 - b_2|$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1| \; + |a_2 - b_2| \; + |a_3 - b_3|$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots$

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	. . .
"the"	4.57	4.24	4.25	4.19	4.47	. . .
"to"	3.11	3.29	3.43	3.14	3.71	. . .
"and"	3.19	3	3.08	2.85	2.81	. . .
"of"	2.6	3	2.63	2.43	2.86	. . .
"I"	2.17	2.2	2.13	2.42	2.22	. . .
"a"	2.24	1.92	1.92	2.21	1.92	. . .
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots + |a_n - b_n|$

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$|a_1 - b_1| + |a_2 - b_2| + |a_3 - b_3| + \ldots + |a_n - b_n| =$$
$$\sum_{i=1}^{n} |a_i - b_i|$$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)}$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)} \quad \delta_{(c,d)}$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)} \quad \delta_{(c,d)} \quad \delta_{(d,e)}$

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)} \quad \delta_{(c,d)} \quad \delta_{(d,e)} \quad \delta_{(a,c)}$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)} \quad \delta_{(c,d)} \quad \delta_{(d,e)} \quad \delta_{(a,c)} \quad \delta_{(b,d)}$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)}$ $\delta_{(c,d)}$ $\delta_{(d,e)}$ $\delta_{(a,c)}$ $\delta_{(b,d)}$ $\delta_{(c,e)}$

Computing distances using a table of frequencies

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)}$ $\delta_{(c,d)}$ $\delta_{(d,e)}$ $\delta_{(a,c)}$ $\delta_{(b,d)}$ $\delta_{(c,e)}$ $\delta_{(a,d)}$

Computing distances using a table of frequencies

	Milton Samson	Milton Paradise	Keats Hyperion	Patmore Eros	Browning Bishop	...
"the"	4.57	4.24	4.25	4.19	4.47	...
"to"	3.11	3.29	3.43	3.14	3.71	...
"and"	3.19	3	3.08	2.85	2.81	...
"of"	2.6	3	2.63	2.43	2.86	...
"I"	2.17	2.2	2.13	2.42	2.22	...
"a"	2.24	1.92	1.92	2.21	1.92	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

$$\sum_{i=1}^{n} |a_i - b_i|$$

$\delta_{(b,c)}$ $\delta_{(c,d)}$ $\delta_{(d,e)}$ $\delta_{(a,c)}$ $\delta_{(b,d)}$ $\delta_{(c,e)}$ $\delta_{(a,d)}$ ...

23

Table of computed distances

	Milton *Samson*	Milton *Paradise*	Keats *Hyperion*	Patmore *Eros*	Browning *Bishop*	...
Milton *Samson*	0	0.9839	1.12	1.0493	1.0864	...
Milton *Paradise*	0.9839	0	1.0891	1.089	1.1047	...
Keats *Hyperion*	1.12	1.0891	0	1.128	1.11	...
Patmore *Eros*	1.0493	1.089	1.128	0	1.1128	...
Browning *Bishop*	1.0864	1.1047	1.11	1.1128	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋱

Unsupervised method: cluster analysis

English_poetry
Cluster Analysis

200 MFW Culled @ 0%
Classic Delta distance
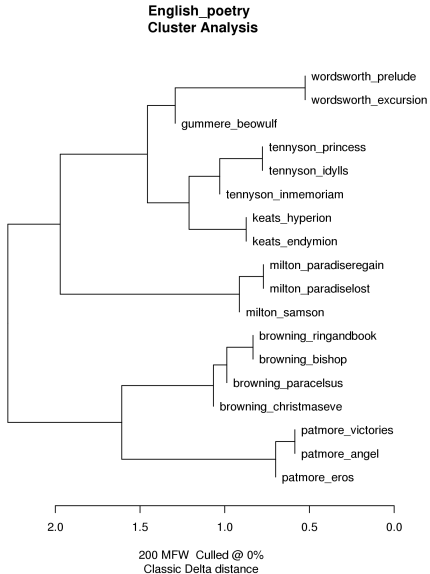
Thank you!

email: maciejeder@gmail.com

Twitter: @MaciejEder

GitHub: https://github.com/computationalstylistics