Algorithmic criticism, Distant Reading and the Edinburgh Review

Francesca Benatti
(Open University)

David King
(Open University)
A Question of Style

- Winner of 2016 Research Society for Victorian Periodicals Field Development Grant ($27,000)
- Funded Jan-Oct 2017
- Francesca Benatti (Digital Humanities and Book History)
- David King (Computer Science and Natural Language Processing)
Research question

• Did a 19th-century periodical like the *Edinburgh Review* create a “transauthorial discourse” (Klancher 1987) that hid individual authors behind a unified corporate voice?
Operationalization

• “Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts.” (Franco Moretti)
Operationalization as criticism

- Corpus selection
- OCR correction
- TEI text encoding
- Analysis with computational tools
- Interpretation of results
Corpus selection

- 325,000 words from *Edinburgh Review*
- 175,000 words from *Quarterly Review*
- Literature, history, biography, travel, 1814-1820
- Fall of Napoleon, Congress of Vienna etc.
- *Waverley, The Corsair, The Excursion, Emma, Lord of the Isles, Christabel, Lalla Rookh, Watt Tyler, Childe Harold, Frankenstein* ...
• Poor quality, mass-digitised scans
• David King working on (semi-) automated OCR correction
• But human intervention needed to work with peculiarities of our data e.g.
  • Hazlitt “Shakespear”
  • Brougham “publick”
• Do we normalise or not?
• Extensive quotations within articles
• Up to 20-30% of each article
• Use TEI to mark them in texts
• Should we exclude quotations as non-authorial texts?
• Or keep them to evaluate critical focus of Edinburgh?
• Transform TEI back into plain text with XSL minus quotations
Analysis with computational tools

• Which aspects of authorship do they bring into focus and which do they instead elide, and must be sought through other methods?
Jerome/Foucault’s four criteria for authorship

01  author as standard level of quality

02  author as conceptual or theoretical coherence

03  author as stylistic uniformity

04  author as definite historical figure in which series of events converge
03 Stylistic uniformity

• Authorial **fingerprint**
• Van Halteren’s "human stylome." (2005)
• Unconscious elements in the way we write
• Reflected by use of Most Frequent Words
• Sought by machine reader through **stylometry**
Example: “the”
“the” is (almost) always the most frequent word in an English-language text

Yet there are variations in how often it is employed
e.g. “the” as **percentage** of total number of words in five *Edinburgh Review* articles

<table>
<thead>
<tr>
<th>Article</th>
<th>“The” as % total words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anon “Christabel”</td>
<td>6.4%</td>
</tr>
<tr>
<td>Jeffrey “Excursion”</td>
<td>6.6%</td>
</tr>
<tr>
<td>Moore “Boyd”</td>
<td>7.4%</td>
</tr>
<tr>
<td>Hazlitt “Sismondi”</td>
<td>8.6%</td>
</tr>
<tr>
<td>Palgrave “Goethe”</td>
<td>5.8%</td>
</tr>
</tbody>
</table>
• One possibility: Keywords

• “A keyword is a word that is more frequent in a text or corpus under study than it is in some (larger) reference corpus.” (McEnery)

• Comparing ER corpus with corpus of Romantic Nonfiction texts, 1770-1830:
  • 5.7 million words
  • 42 texts
  • 29 authors
Positive Keywords

- First person plural: we, us, our
- Present tense verbs: is, has, seems
- Third person pronouns: he, she, his, her etc.
We: Top collocates

- Confess
- Apprehend
- Suspect
- Venture
- Presume
- Shall
- Think
- Inclined
- Help
- Conceive
- Believe
01 Quality

- Conscious choice of tone
- e.g. Van Dalen-Oskam *Riddle of Literary Quality* project
- Authorial *signature*
Quality?

- Van Dalen-Oskam
  - vocabulary richness?
  - word length?
  - sentence length?

- Allison
  - medium-frequency words?
  - words used vs. words avoided?

- Mahlberg
  - word clusters
• Finally, can we successfully combine the use of computational methods with literary interpretation in a process of “algorithmic criticism” (Ramsay)?
• Are Digital Humanities methods an improvement compared to traditional Humanities research?
Stylometry evaluation

• Some authorial fingerprints are visible
• But others are less clear
• Could this be due to:
  • Editorial intervention?
  • Multiple authorship?
  • Not enough data/bad data?
Keyword analysis

• “We” and collocates suggest
• Corporate identity?
• “Imagined community” with readers?
• Construction of shared values and shared canon?
Next steps

01 Enhance scripts
02 Include more texts
03 Expand reference corpora
04 Share scripts, TEI texts
05 Evaluate and critique
Conclusion

• Digital analysis can improve our understanding of Romantic authorship by focusing on elements of style and authorship that escape the naked brain
• “Algorithmic criticism” can complement close reading, not replace it
  • Good at finding patterns
  • Not at finding meaning
Thank you!

Francesca Benatti
David King

Faculty of Arts and Social Sciences
Faculty of Science, Technology, Engineering and Mathematics
The Open University
Milton Keynes

Project blog:
http://www.open.ac.uk/blogs/styleproject/

Project outputs (in 2018):
https://ou.figshare.com/