Hidden Authors and Reading Machines

Investigating 19th-century authorship with 21st-century technologies

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 (Book History and Digital Humanities)
• David King (Computer Science and Natural Language Processing)
Research questions

• Did a 19th-century periodical like the Edinburgh Review create a “transauthorial discourse” (Klancher 1987) that hid individual authors behind a unified corporate voice?
Death of the author, 
birth of the reader

• We study the reception of human readers (e.g. UK Reading Experience Database)...
• ... and now of machine readers also

• Can we work with the 21st-century machine reader to study authorship in the 19th-Edinburgh Review?
Authorship in the Edinburgh Review

• Founded in 1802 by members of Whig intelligentsia
• All articles published anonymously
• Most authors identified by now by Wellesley
  Index to Victorian Periodicals
• How different are these authors from one another? And from those of other periodicals/texts?
• Is there an Edinburgh Review “house style”?
• How can we, in Franco Moretti’s words, “operationalize” the practice of authorship in the *Edinburgh Review*?

• “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.”
Operationalization as criticism

1. Corpus selection
2. OCR correction
3. TEI text encoding
4. Analysis with computational tools
5. 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 ...
OCR correction

- 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 “Shakspeare”
  - Brougham “publick”
- Do we normalise or not?
**TEI Text Encoding**

- 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
• Which aspects of authorship are brought into focus with the help of the machine reader?

• Which aspects of authorship are instead elided through computational analysis, 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
• 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>Author</th>
<th>Percentage</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>
02 Conceptual coherence

- 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
What does it all mean?

• Finally, can we successfully combine the use of computational methods for the empirical measurement of textual features with the synthesis and literary interpretation of these results?

• Can the resulting “algorithmic criticism” (Ramsay 2011) reveal patterns that enable new readings of the complex practice of authorship within the *Edinburgh Review*?
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 Perfect scripts
02 Include more texts
04 Include whole issues
05 Expand reference corpora
06 Share scripts, TEI texts
07 Evaluate and critique
Conclusion

- Machine reader can complement human reader, not replace
- Good at finding patterns
- Not at finding meaning
- But we human readers can work together with it
“Many interesting things cannot be counted, but many others can.”

–John Burrows
Francesca Benatti
David King

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

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

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