One Year After. A Question of Style: individual voices and corporate identity in the Edinburgh Review, 1814-1820

Conference or Workshop Item

How to cite:


For guidance on citations see FAQs.

© [not recorded]

Version: Accepted Manuscript

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
(One Year After) A Question of Style: individual voices and corporate identity in the *Edinburgh Review, 1814-1820*

Dr Francesca Benatti and Dr David King, The Open University
Introduction

A Question of Style in brief
INTRODUCTION
A Question of Style

• Winner, 2016 RSVP Field Development Grant
  • (thank you!!!)

• Research question:
  • Did the *Edinburgh Review* create a “transauthorial discourse” (Klancher) that hid individual authorial voices behind an impersonal corporate style?
When last we saw our heroes in July 2017…

David presented at RSVP 2017 in Freiburg on
  • corpus selection
  • OCR correction

Francesca presented at SHARP 2017 (here in Victoria) and BARS 2017 in York on
  • authorship
  • computational criticism

Where are we now?
Corpus creation

The composition and rationale of our corpus
CORPUS
Size, composition, rationale

- Driven by research goal:
  - Analyse the style of ER
  - Compare it to style of QR

- Driven by prior research on *Christabel* review:
  - Focus on literature, history, travel

- Emergent research questions:
  - Does multiple authorship hide style? (6 multi-authored articles)
  - Are reviews of the same text similar to one another? (10 pairs of reviews)
  - Does the genre of the text being reviewed influence the style of the review? (3 genres)
OCR correction

Post- Optical Character Recognition processing
Post-OCR processing

- **Aim:** Develop semi-automated process for OCR correction
- **Problems:** OCR errors too inconsistent for automation
- **Individual spelling choices**
  - Publick
- **Regional identities**
  - Perswaded
- **Political perspectives**
  - Breslaw/Breslau
- **Language transformation**
  - Shakspear, Shakspeare, Shakespear, Shakespeare
- **Solution:** David reviewed all automated corrections and “spelling mistakes” against the digitised source image
Quotations

Or, what is in a review?
QUOTATIONS
Or, what is in a review?

Chart Title

ERQRCorp

ERCorp

QRCorp

Non-quote %

Quote %
Quotations % Max and Min

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Quote % Max</th>
<th>Quote % Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERCorp</td>
<td>79%</td>
<td>1%</td>
</tr>
<tr>
<td>QRCorp</td>
<td>65%</td>
<td>11%</td>
</tr>
</tbody>
</table>
QUOTATIONS
Or, what is in a review?

• The presence of quotations is a problem for our chosen analytical methods

• e.g. we want to analyse Francis Jeffrey’s style (reviewer), not Walter Scott’s (reviewed)

• Solution:
  • David marked quotations using TEI XML <quote> element
  • Then we removed quotations using XSL transformation

• This reduces the total size of the corpora to **512,702 words**
Analysis

Stylometry and corpus stylistics
Two interpretations of style*

Style as fingerprint

Unconscious elements in the way we write
(e.g. Van Halteren et al. "Existence of a human stylome." (2005))
Reflected by use of Most Frequent Words (MFW)

Style as signature

Conscious choice of words, sentences, tone
(e.g. Van Dalen-Oskam Riddle of Literary Quality project)
Still working out how to identify with computational methods

* as defined by Sarah Allison at DH2016, Stylistics workshop, 12 July 2016
• Style as Fingerprint:
  • Stylometry

• Style as signature:
  • Corpus stylistics
    • Keywords
    • N-grams
    • Clusters
ANALYSIS

Stylometry in brief

• The study of how hidden stylistic traits can be measured through statistical methods to trace an author's voice

• Made better known by John Burrows in his 2001 Busa Award lectures and beyond

• Generally concerned with authorship attribution but increasingly used to study authorship more broadly

• Burrows’ s Delta method implemented by Eder, Rybicki and Kestemont’s Stylo software package

• Improved method Cosine Delta developed by University of Würzburg

• Based on analysis of most frequent words
Stylometry: by journal using Stylo (Eder, Rybicki, Kestemont); Cosine Delta; 300 MFW
Stylometry: by genre (using Stylo (Eder, Rybicki, Kestemont); Cosine Delta; 300 MFW)
Stylometry: by journal (using Stylo (Eder, Rybicki, Kestemont); Cosine Delta; 900 MFW)
Stylometry using Stylo (Eder, Rybicki, Kestemont); Cosine Delta; 300 MFW
Keywords

- Keywords are words that are significantly more frequent in the target text/corpus than in a chosen reference corpus

**Target corpus:**
- ERCorp
- QRCorp

**Reference corpus:**
- 1780-1850 part of CLMET (Corpus of Late Modern English Texts; created by Hendrik de Smet)
  - 99 texts
  - 11 million words

- Used to represent (written) language norm for the period
- How do ERCorp and QRCorp differ from this baseline?
### ANALYSIS
Keywords: top 10 keywords, ERCorp and QRCorp

<table>
<thead>
<tr>
<th>Rank</th>
<th>ERCorp keyword</th>
<th>QRCorp keyword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>of</td>
<td>the</td>
</tr>
<tr>
<td>2</td>
<td>the</td>
<td>of</td>
</tr>
<tr>
<td>3</td>
<td>cannot</td>
<td>cannot</td>
</tr>
<tr>
<td>4</td>
<td>we</td>
<td>we</td>
</tr>
<tr>
<td>5</td>
<td>author</td>
<td>which</td>
</tr>
<tr>
<td>6</td>
<td>poetry</td>
<td>author</td>
</tr>
<tr>
<td>7</td>
<td>which</td>
<td>river</td>
</tr>
<tr>
<td>8</td>
<td>his</td>
<td>readers</td>
</tr>
<tr>
<td>9</td>
<td>is</td>
<td>poem</td>
</tr>
<tr>
<td>10</td>
<td>poetical</td>
<td>poetry</td>
</tr>
</tbody>
</table>
## Keyword classification

<table>
<thead>
<tr>
<th>Keyword classification</th>
<th>ERCorp</th>
<th>QRCorp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texts, authors, readers</td>
<td>author poetry poetical readers literature</td>
<td>story character poem description poet</td>
</tr>
<tr>
<td>Reviewing</td>
<td>genius style diction antient modern</td>
<td>interesting great original romantic imagination</td>
</tr>
<tr>
<td>Grammatical</td>
<td>of the we ... most</td>
<td>of the we</td>
</tr>
<tr>
<td>Verbs</td>
<td>cannot</td>
<td>cannot</td>
</tr>
</tbody>
</table>
## Genius

<table>
<thead>
<tr>
<th>GENIUS</th>
<th>ER</th>
<th>QR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>189</td>
<td>56</td>
</tr>
<tr>
<td>Range</td>
<td>37/61</td>
<td>14/24</td>
</tr>
<tr>
<td>Authors</td>
<td>14</td>
<td>8</td>
</tr>
<tr>
<td>Highest user</td>
<td>Mackintosh (54)</td>
<td>Scott (33)</td>
</tr>
</tbody>
</table>
ANALYSIS

N-grams

• Repeated sequences of n number of words
• Repeated patterns are repeated because they are associated with the communicative needs of a certain register (Mahlberg)
• 4-grams and above have an associated semantic prosody (Fischer-Starcke)
• Semantic prosody: attitudinal element, motivation
## ANALYSIS

### 4-grams

<table>
<thead>
<tr>
<th>Rank</th>
<th>ERCorp</th>
<th>QR Corp</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>one of the most</td>
<td>the greater part of</td>
</tr>
<tr>
<td>2</td>
<td>the greater part of</td>
<td>one of the most</td>
</tr>
<tr>
<td>3</td>
<td>a great deal of</td>
<td>it is impossible to</td>
</tr>
<tr>
<td>4</td>
<td>as one of the</td>
<td>for the most part</td>
</tr>
<tr>
<td>5</td>
<td>we have no doubt</td>
<td>greater part of the</td>
</tr>
<tr>
<td>6</td>
<td>for the most part</td>
<td>it is plain that</td>
</tr>
<tr>
<td>7</td>
<td>it is impossible to</td>
<td>as much as possible</td>
</tr>
<tr>
<td>8</td>
<td>a great part of</td>
<td>in a great measure</td>
</tr>
<tr>
<td>9</td>
<td>is by no means</td>
<td>it is clear that</td>
</tr>
<tr>
<td>10</td>
<td>a good deal of</td>
<td>it was impossible to</td>
</tr>
</tbody>
</table>
ANALYSIS

4-grams

- Out of the 200 most frequent 4-grams in ERCorp and QRCorp
  - There are more types of 4-grams marking strong judgement in ERCorp
  - 4-gram types marking weak judgement are similar in number
  - ERCorp strong judgement 4-grams occur in multiple authors
- Fewer shared expressions of judgement in QRCorp
- A marker of house style?

<table>
<thead>
<tr>
<th>Type</th>
<th>ERCorp</th>
<th>QRCorp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong judgement</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>Weak judgement</td>
<td>22</td>
<td>19</td>
</tr>
</tbody>
</table>
Clusters

• Certain clusters seem more characteristic of ERCorp e.g.
  • “have no doubt”
  • “great deal”

• N.B. CLMET is 30 times the size of ERCorp

<table>
<thead>
<tr>
<th>Cluster</th>
<th>ERCorp</th>
<th>QRCorp</th>
<th>CLMET</th>
</tr>
</thead>
<tbody>
<tr>
<td>have no doubt</td>
<td>19</td>
<td>1</td>
<td>58</td>
</tr>
<tr>
<td>great deal</td>
<td>35</td>
<td>2</td>
<td>694</td>
</tr>
</tbody>
</table>
Conclusion
And next steps

• Some traces of “house style”
  • Relatively faint through stylometry (“fingerprint” level)
  • More perceptible through corpus stylistics (signature level)

• Influence of genre of text being reviewed on its stylistic fingerprint
CONCLUSION
And next steps

• Corpus stylistics with:
  • Part of speech tags
  • Lemmatisation
  • Author-based corpora
    • Jeffrey, Brougham, Croker, Scott

• Stylometry with:
  • Character n-grams
  • Positive vs. negative reviews

• Assessment of the benefits of curation:
  • Keeping quotations
  • Using “raw” OCR
“Many interesting things cannot be counted, but many others can.”

—John Burrows
THANK YOU!

Download our corpus from ORDO (search for RSVP)
doi: 10.21954/ou.rd.6850865

Follow us on
http://www.open.ac.uk/blogs/styleproject/
@rhymesontheroad
@dhiston