In Search of the Voice of the Edinburgh Review


For guidance on citations see FAQs.
In Search of the Voice of the Edinburgh Review

Francesca Benatti and David King
The Open University
Introduction

Did the *Edinburgh Review* create a “transauthorial discourse” (Klancher 1987) that hid the voices of individual contributors behind a corporate style?
Corpus

*Edinburgh Review:*
- 45 articles
- 10 authors and one anonymous article
- 269,622 ‘words’

Preparation:
1. OCR with manual curation
2. TEI manual mark-up
3. Attention to quotations
Stylometry

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
“Many interesting things cannot be counted, but many others can.”

John Burrows
Why use stylometry?

Perception of authorial “voice” is quite subjective
• e.g. Duncan Wu (2007)

Computer-aided analysis can supplement humanistic research
Delta

“Delta is the mean of the absolute differences between the z-scores for a set of word-variables in a given text-group and the z-scores for the same set of word-variables in a target text.”

John Burrows
Delta continued

Delta works on the Most Frequent Words present in a given set of texts

We all use Most Frequent Words differently

Underpinned by solid mathematical and linguistic foundations
Data exploration with multidimensional scaling — spot the cluster
False clusters

Female pronouns

- Moore_French_Novels_34_1820_corr 36%
- Jeffrey_Edgeworth_28_1817 33%
- anon_christabel_edinburgh_review_27_1816 32%
- Jeffrey_Lalla_Rookh_29_1817 23%
- Brougham_melanges_30_1818 21%

…and 10 texts contained no female pronouns at all
Increasing rigour

With clustering techniques that
• rely on random seeding, the results depend too heavily on the random starting point
• have parameters, the results depend too heavily on those parameters

Therefore, applied
• both *agglomerative* (hierarchy) and *partition* (kmeans) clustering techniques
• drilled down through two feature sets initially (lexical, POS), and later a third (tf:idf)
Two weak clusters emerge
## MFW vs TF:IDF

Both attempt to remove the influence of content over style in the analysis.

<table>
<thead>
<tr>
<th><strong>MFW</strong></th>
<th><strong>TF:IDF</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequent words</td>
<td>Significant words</td>
</tr>
<tr>
<td>Choose what to <em>include</em> in the analysis</td>
<td>Choose what to <em>exclude</em> from the analysis</td>
</tr>
</tbody>
</table>
Evaluation

Distinctions between authors weak but present

How noticeable are they to a human reader?

Do Jeffrey’s editorial interventions erase individual voices?

Spot patterns that deserve close reading
Future work

Extend corpus:
- Quarterly Review
- CLMET

AntConc:
- keywords (words statistically more frequent)
- methods from corpus linguistics

Funding from RSVP to extend research (Jan-Oct 2017)
Digital Humanities at the Open University
The Open University
Walton Hall
Milton Keynes
MK7 6AA

Arts-digital-humanities@open.ac.uk
www.open.ac.uk