Hidden Authors and Reading Machines: Investigating 19th-century authorship with 21st-century technologies

Hidden Authors and Reading Machines

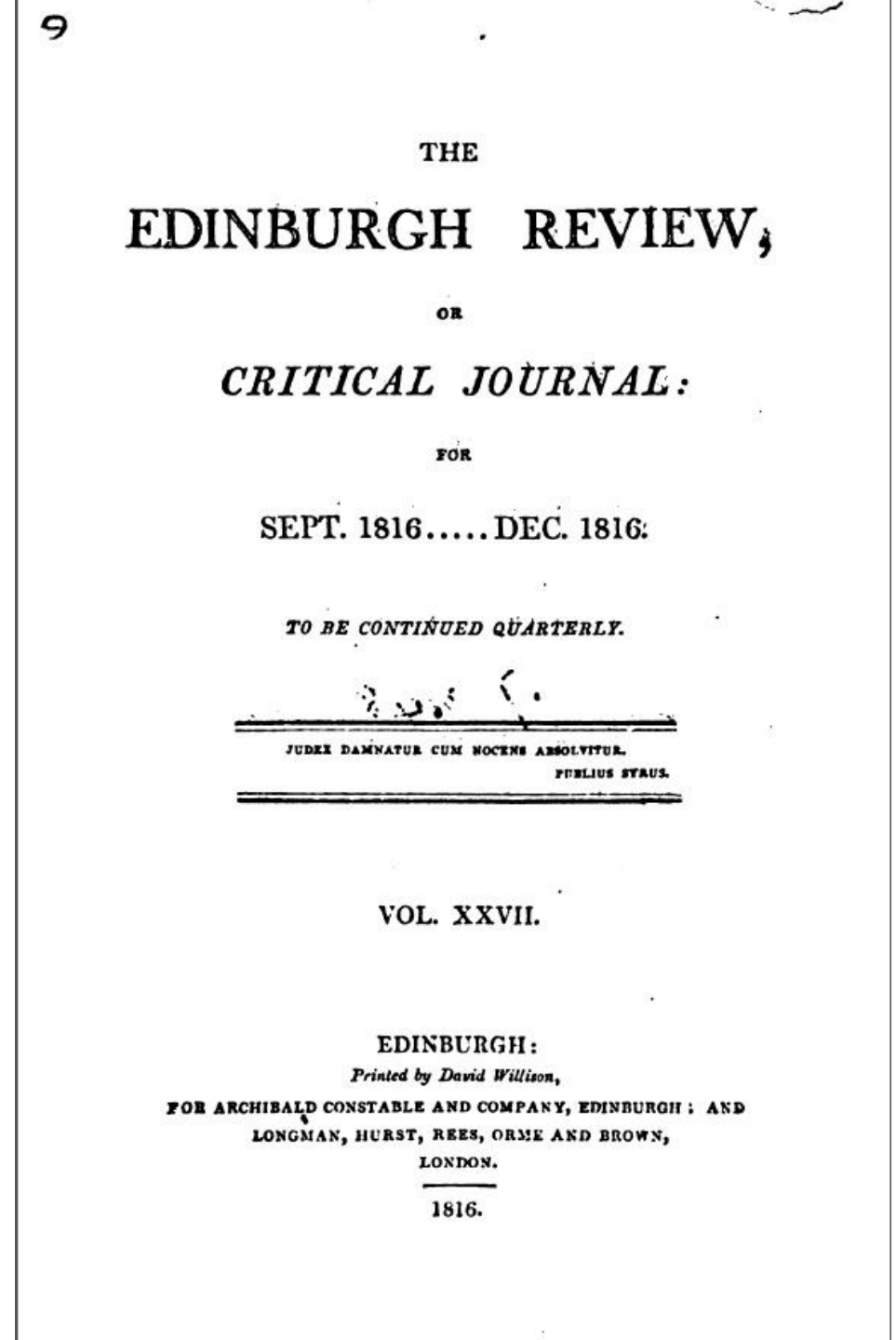
Investigating 19th-century authorship with 21st-century technologies

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)



9

THE

EDINBURGH REVIEW;

OR

CRITICAL JOURNAL:

FOR

SEPT. 1816.....DEC. 1816.

TO BE CONTINUED QUARTERLY.

JUDEX DAMNATUR CUM NOCENS ABSOLVITUR.
PUBLIUS SYRUS.

VOL. XXVII.

EDINBURGH:
Printed by David Willison,
FOR ARCHIBALD CONSTABLE AND COMPANY, EDINBURGH; AND
LONGMAN, HURST, REES, ORME AND BROWN,
LONDON.

1816.

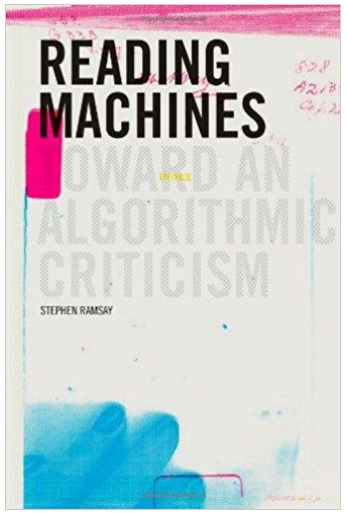
Research questions

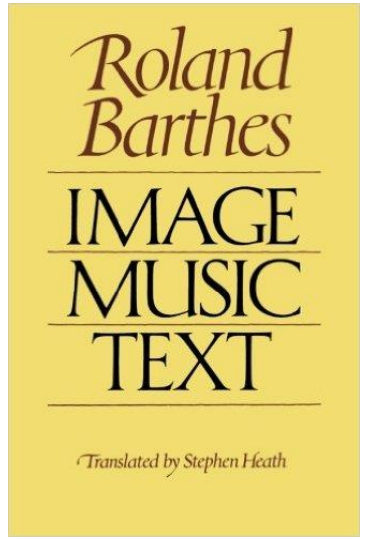
- Did a 19th-century periodical like the *Edinburgh Review* create a "transauthorial discourse" (Klancher 1987) that hid individual authors behind a unified corporate voice?

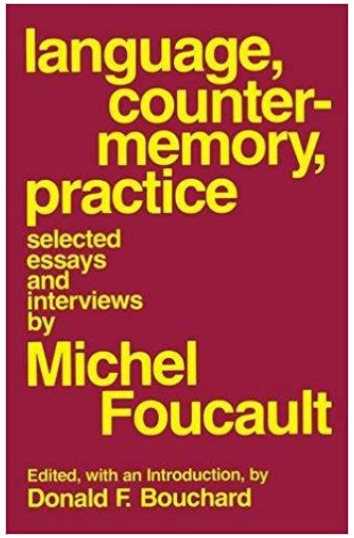
Death of the author, birth of the reader

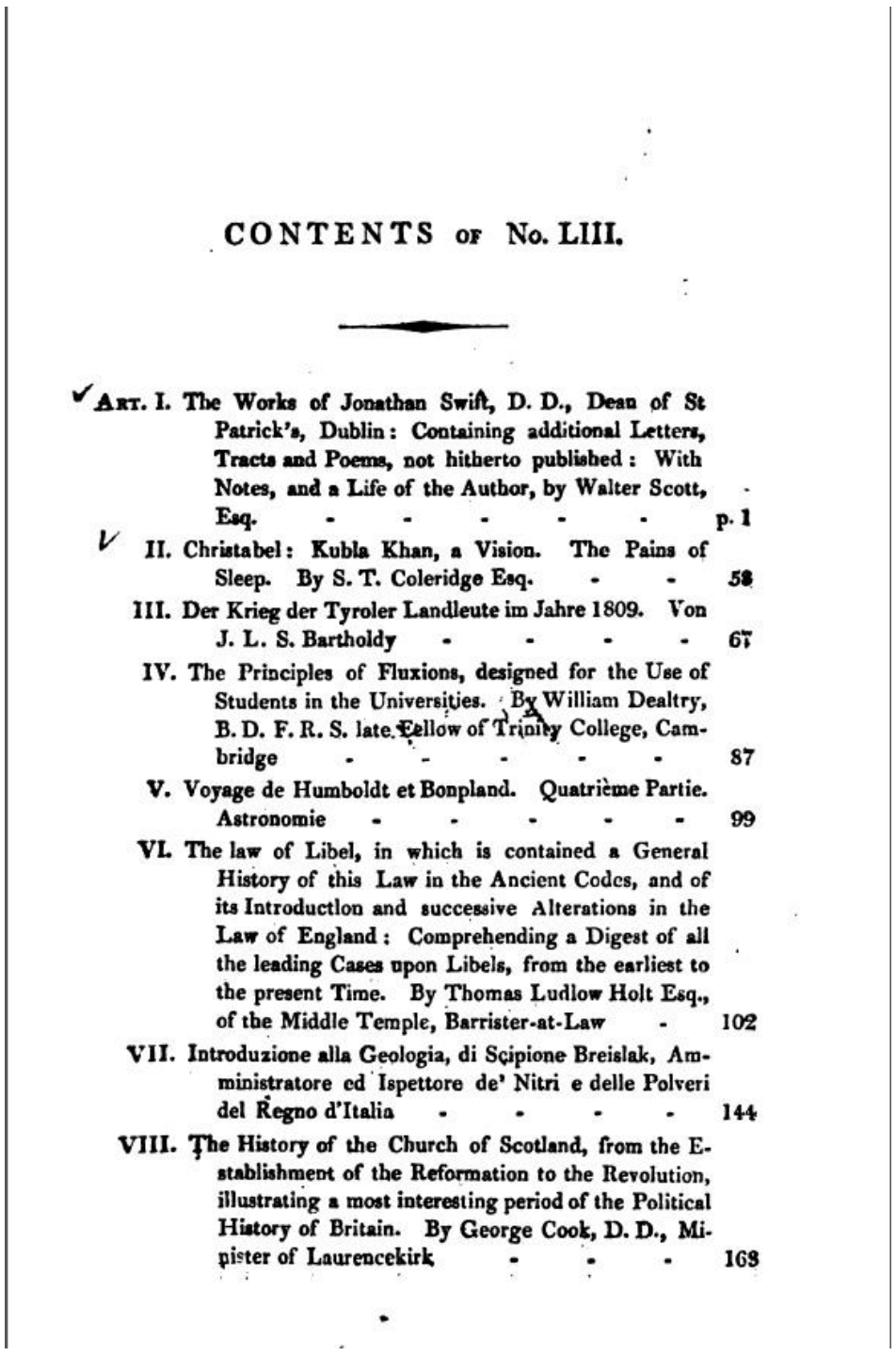
- We study the reception of human readers (e.g. UK Reading Experience Database)…
- … and now of machine readers also

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









CONTENTS OF No. LIII.

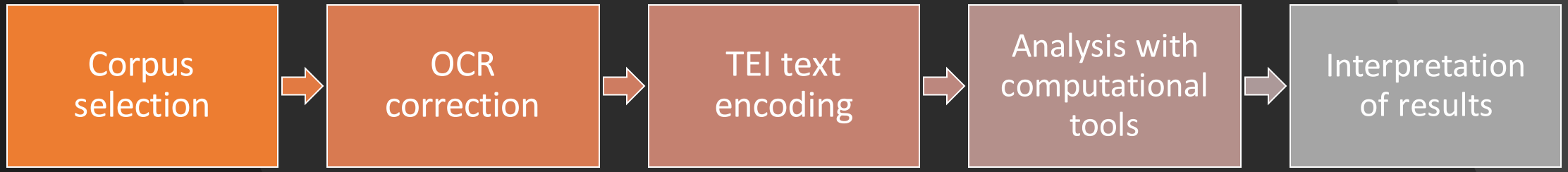
Authorship in the *Edinburgh Review*

- 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"?

Operationalization

- 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



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

Analysis with computational tools

- 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

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**

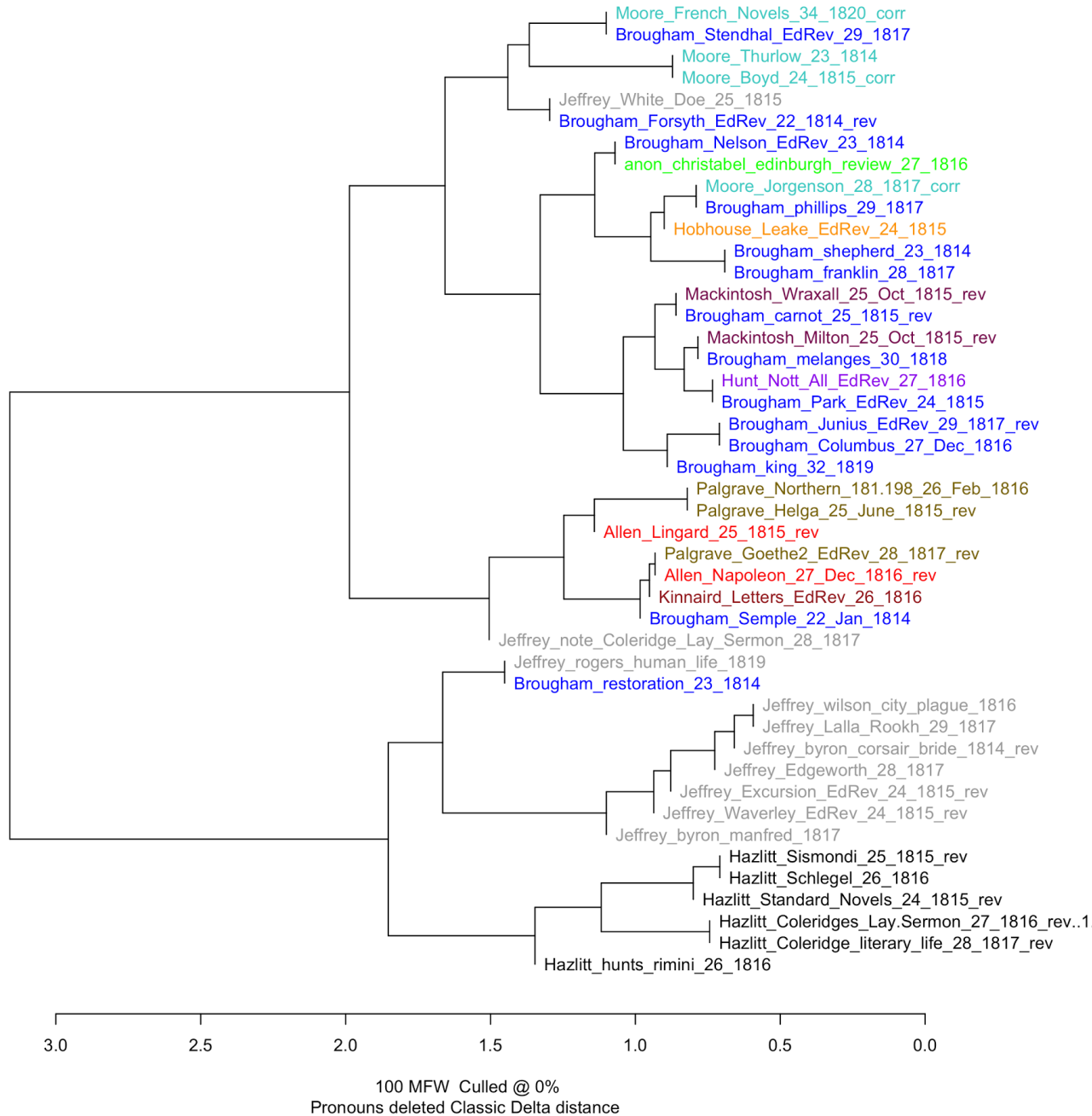
Example: “the”

“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

Anon “Christabel”	6.4%
Jeffrey “Excursion”	6.6%
Moore “Boyd”	7.4%
Hazlitt “Sismondi”	8.6%
Palgrave “Goethe”	5.8%


Moore_French_Novels_34_1820_corr
Brougham_Stendhal_EdRev_29_1817
Moore_Thurlow_23_1814
Moore_Boyd_24_1815_corr
Jeffrey_White_Doe_25_1815
Brougham_Forsyth_EdRev_22_1814_rev
Brougham_Nelson_EdRev_23_1814
anon_christabel_edinburgh_review_27_1816
Moore_Jorgenson_28_1817_corr
Brougham_phillips_29_1817
Hobhouse_Leake_EdRev_24_1815
Brougham_shepherd_23_1814
Brougham_franklin_28_1817
Mackintosh_Wraxall_25_Oct_1815_rev
Brougham_carnot_25_1815_rev
Mackintosh_Milton_25_Oct_1815_rev
Brougham_melanges_30_1818
Hunt_Nott_All_EdRev_27_1816
Brougham_Park_EdRev_24_1815
Brougham_Junius_EdRev_29_1817_rev
Brougham_Columbus_27_Dec_1816
Brougham_king_32_1819
Palgrave_Northern_181.198_26_Feb_1816
Palgrave_Helga_25_June_1815_rev
Allen_Lingard_25_1815_rev
Palgrave_Goethe2_EdRev_28_1817_rev
Allen_Napoleon_27_Dec_1816_rev
Kinnaird_Letters_EdRev_26_1816
Brougham_Semple_22_Jan_1814
Jeffrey_note_Coleridge_Lay_Sermon_28_1817
Jeffrey_rogers_human_life_1819
Brougham_restoration_23_1814
Jeffrey_wilson_city_plague_1816
Jeffrey_Lalla_Rookh_29_1817
Jeffrey_byron_corsair_bride_1814_rev
Jeffrey_Edgeworth_28_1817
Jeffrey_Excursion_EdRev_24_1815_rev
Jeffrey_Waverley_EdRev_24_1815_rev
Jeffrey_byron_manfred_1817
Hazlitt_Sismondi_25_1815_rev
Hazlitt_Schlegel_26_1816
Hazlitt_Standard_Novels_24_1815_rev
Hazlitt_Coleridges_Lay.Sermon_27_1816_rev..1.
Hazlitt_Coleridge_literary_life_28_1817_rev
Hazlitt_hunts_rimini_26_1816
3.0
2.5
2.0
1.5
1.0
0.5
0.0
100 MFW Culled @ 0%
Pronouns deleted Classic Delta distance

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

Thank you!

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/