Algorithmic criticism, Distant Reading and the *Edinburgh Review*

Conference or Workshop Item

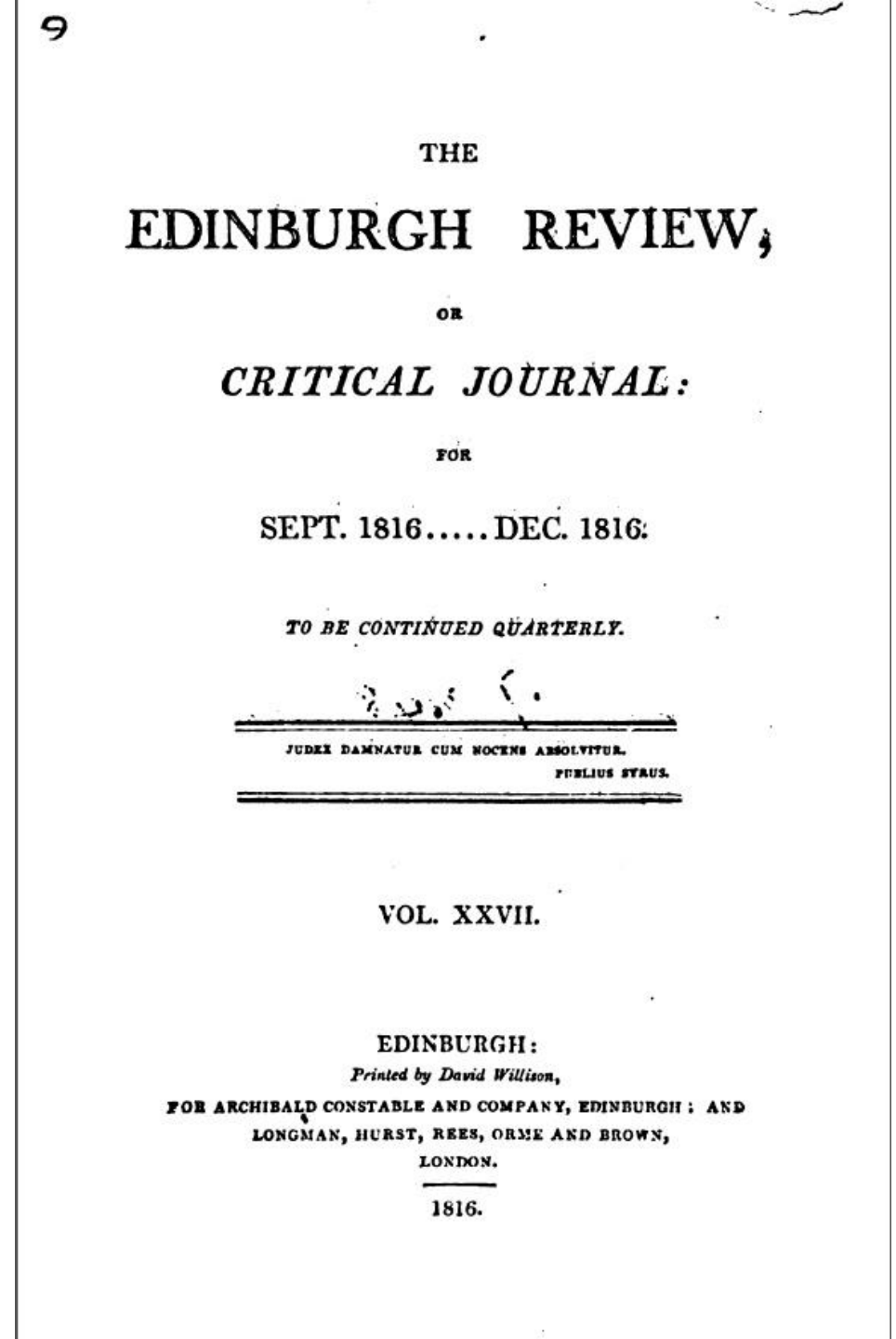
Algorithmic criticism, Distant Reading and the *Edinburgh Review*

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)



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.

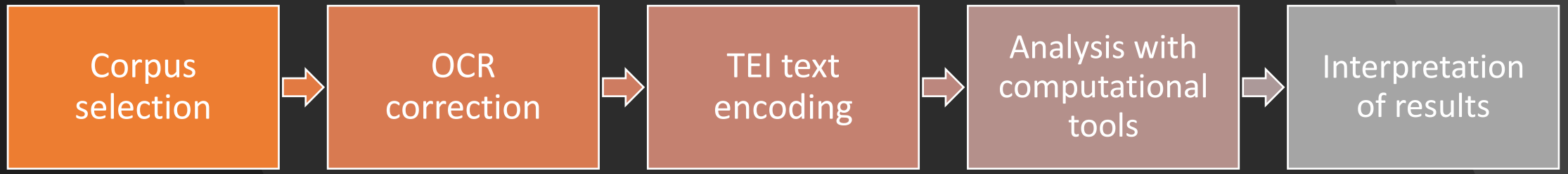
Research question

- 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

- 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 “Shakespear”
  - 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 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**

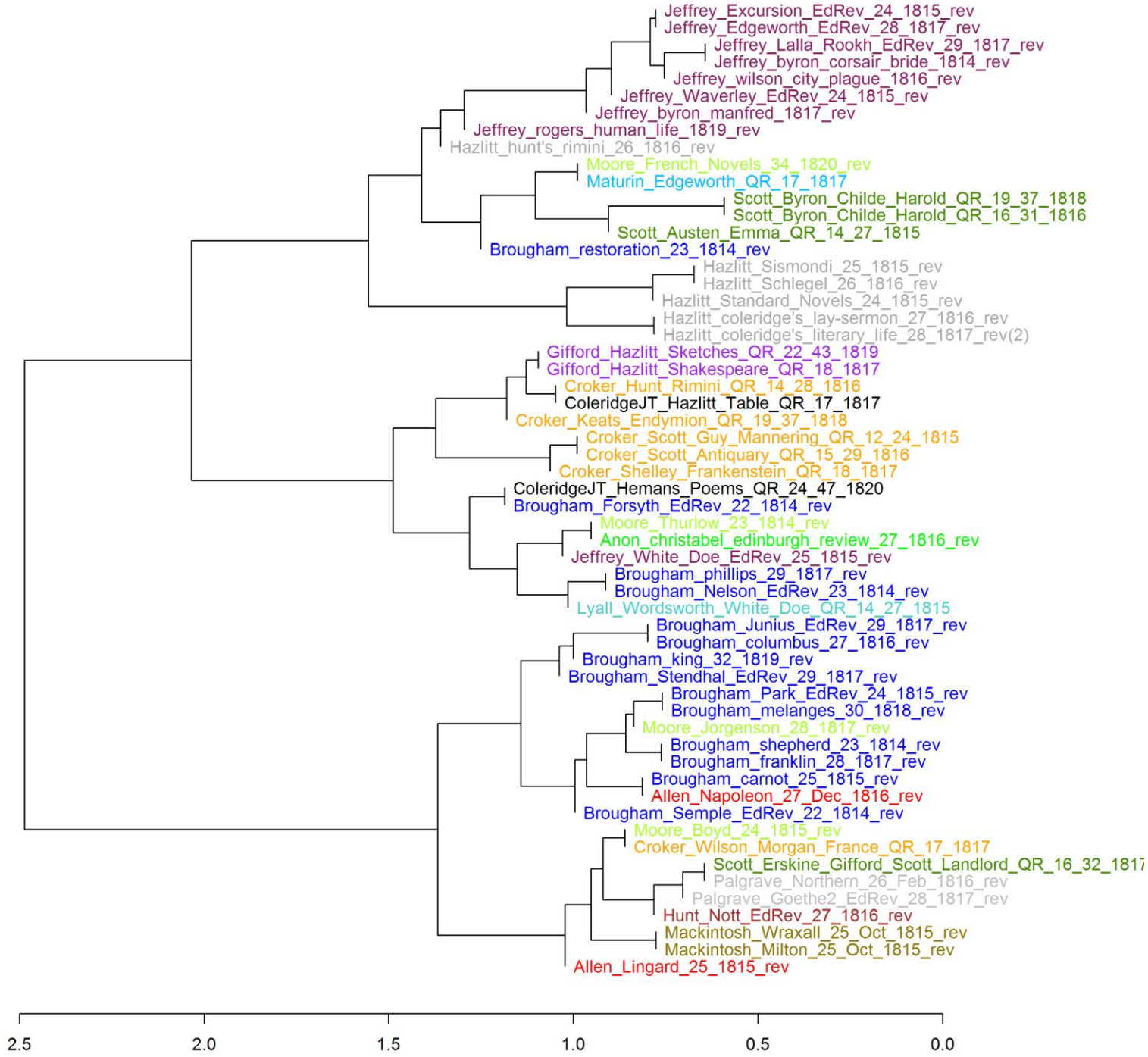
Example: “the”

“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

Article	“The” as % total words
Anon “Christabel”	6.4%
Jeffrey “Excursion”	6.6%
Moore “Boyd”	7.4%
Hazlitt “Sismondi”	8.6%
Palgrave “Goethe”	5.8%


Jeffrey_Excursion_EdRev_24_1815_rev
Jeffrey_Edgeworth_EdRev_28_1817_rev
Jeffrey_Lalla_Rookh_EdRev_29_1817_rev
Jeffrey_byron_corsair_bride_1814_rev
Jeffrey_wilson_city_plague_1816_rev
Jeffrey_Waverley_EdRev_24_1815_rev
Jeffrey_byron_manfred_1817_rev
Jeffrey_rogers_human_life_1819_rev
Hazlitt_hunt's_rimini_26_1816_rev
Moore_French_Novels_34_1820_rev
Maturin_Edgeworth_QR_17_1817
Scott_Byron_Childe_Harold_QR_19_37_1818
Scott_Byron_Childe_Harold_QR_16_31_1816
Scott_Austen_Emma_QR_14_27_1815
Brougham_restoration_23_1814_rev
Hazlitt_Sismondi_25_1815_rev
Hazlitt_Schlegel_26_1816_rev
Hazlitt_Standard_Novels_24_1815_rev
Hazlitt_coleridge's_lay-sermon_27_1816_rev
Hazlitt_coleridge's_literary_life_28_1817_rev(2)
Gifford_Hazlitt_Sketches_QR_22_43_1819
Gifford_Hazlitt_Shakespeare_QR_18_1817
Croker_Hunt_Rimini_QR_14_28_1816
ColeridgeJT_Hazlitt_Table_QR_17_1817
Croker_Keats_Endymion_QR_19_37_1818
Croker_Scott_Guy_Mannering_QR_12_24_1815
Croker_Scott_Antiquary_QR_15_29_1816
Croker_Shelley_Frankenstein_QR_18_1817
ColeridgeJT_Hemans_Poems_QR_24_47_1820
Brougham_Forsyth_EdRev_22_1814_rev
Moore_Thurlow_23_1814_rev
Anon_christabel_edinburgh_review_27_1816_rev
Jeffrey_White_Doe_EdRev_25_1815_rev
Brougham_phillips_29_1817_rev
Brougham_Nelson_EdRev_23_1814_rev
Lyall_Wordsworth_White_Doe_QR_14_27_1815
Brougham_Junius_EdRev_29_1817_rev
Brougham_columbus_27_1816_rev
Brougham_king_32_1819_rev
Brougham_Stendhal_EdRev_29_1817_rev
Brougham_Park_EdRev_24_1815_rev
Brougham_melanges_30_1818_rev
Moore_Jorgenson_28_1817_rev
Brougham_shepherd_23_1814_rev
Brougham_franklin_28_1817_rev
Brougham_carnot_25_1815_rev
Allen_Napoleon_27_Dec_1816_rev
Brougham_Semple_EdRev_22_1814_rev
Moore_Boyd_24_1815_rev
Croker_Wilson_Morgan_France_QR_17_1817
Scott_Erskine_Gifford_Scott_Landlord_QR_16_32_1817
Palgrave_Northern_26_Feb_1816_rev
Palgrave_Goethe2_EdRev_28_1817_rev
Hunt_Nott_EdRev_27_1816_rev
Mackintosh_Wraxall_25_Oct_1815_rev
Mackintosh_Milton_25_Oct_1815_rev
Allen_Lingard_25_1815_rev
2.5
2.0
1.5
1.0
0.5
0.0

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 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/