Stylometry with R: Computer-Assisted Analysis of Literary Texts

Jan Rybicki
Joanna Byszuk
Welcome to DHSI 2018!

Thanks for joining the DHSI community!

In this booklet, you will find essential course materials prefaced by some useful information about getting settled initially at UVic, finding your way around, getting logged in to our network (after you’ve registered the day before our courses begin), and so on.

Given our community’s focus on things computational, it will be a surprise to no one that we might expect additional information online for some of the classes - your instructors will let you know - or that the most current version of all DHSI-related information may be found on our website at dhsi.org.

To access the DHSI wifi network, simply go into your wireless settings and connect to the “DHSI” network and enter the password “dhsi2018”.

And please don’t hesitate to be in touch with us at institut@uvic.ca or via Twitter at @AlyssaA_DHSI or @DHInstitute if we can be of any help ....
The 2018 schedule is just about ready! A very few things to confirm, add, etc, but this is the place to be to find out what is happening when / where ...

Psst: Some Suggested Outings

If you're here a day or two before we begin, or staying a day or two afterwards, here are a few ideas of things you might consider doing ....

▼ Suggested Outing 1, Botanical Beach (self-organised; car needed)

A self-guided visit to the wet, wild west coast tidal shelf (and historically-significant former research site) at Botanical Beach; we recommend departing early (around 8.00 am) to catch low tide for a better view of the wonderful undersea life! Consider bringing a packed lunch to nibble-on while looking at the crashing waves when there, and then have an afternoon drink enjoying the view from the deck of the Port Renfrew Hotel.

▼ Suggested Outing 2, Butchart Gardens (self-organised)

A shorter journey to the resplendently beautiful Butchart Gardens and, if you like, followed by (ahem) a few minutes at the nearby Church and State Winery, in the Saanich Peninsula. About an hour there by public bus from UVic, or 30 minutes by car.

▼ Suggested Outing 3, Saltspring Island (self-organised; a full day, car/bus + ferry combo)

Why not take a day to explore and celebrate the funky, laid back, Canadian gulf island lifestyle on Saltspring Island. Ferry departs regularly from the Schwartz Bay ferry terminal, which is about one hour by bus / 30 minutes by car from UVic. You may decide to stay on forever ....

▼ Suggested Outing 4, Paddling Victoria’s Inner Harbour (self-organised)

A shorter time, seeing Victoria’s beautiful city centre from the waterways that initially inspired its foundation. A great choice if the day is sunny and warm. Canoes, kayaks, and paddle boards are readily rented from Ocean River Adventures and conveniently launched from right behind the store. Very chill.

And more!

Self-organised High Tea at the Empress Hotel, scooter rentals, visit to the Royal BC Museum, darts at Christies Carriage House, a hangry breakfast at a local diner, whale watching, kayaking, brew pub sampling (at Spinnaker’s, Swans, Moon Under Water, and beyond!), paddle-boarding, a tour of used bookstores, and more have also been suggested!

Sunday, 3 June 2018 [DHSI Registration + Suggested Outings]

9:00 to 4:00

▼ Early Class Meeting: 4. [Foundations] DH For Department Chairs and Deans (Hickman 120, Classroom)
Further details are available from instructors in mid May to those registered in the class. Registration materials will be available in the classroom.

3:00 to 5:00

DHSI Registration (MacLaurin Building, Room A100)

After registration, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a nice meal.

Monday, 4 June 2018

Your hosts for the week are Alyssa Arbuckle, Ray Siemens, and Dan Sondheim.

7:45 to 8:15

Last-minute Registration (MacLaurin Building, Room A100)

8:30 to 10:00

Welcome, Orientation, and Instructor Overview (MacLaurin A144)
Classes in Session (click for details and locations)

- 1. [Foundations] Text Encoding Fundamentals and their Application (Cornett A128, Classroom)
- 3. [Foundations] Making Choices About Your Data (MacLaurin D109, Classroom)
- 4. [Foundations] DH For Department Chairs and Deans (Hickman 120, Classroom)
- 5. [Foundations] Introduction to Javascript and Data Visualization (Clearihue D132, Classroom)
- 6. [Foundations] Introduction to Computation for Literary Criticism (Clearihue A195, Lab)
- 7. Out-of-the-Box Text Analysis for the Digital Humanities (Human and Social Development A160, Lab)
- 8. Sounds and Digital Humanities (MacLaurin D111, Classroom)
- 9. Digital Humanities Pedagogy: Integration in the Curriculum (MacLaurin D016, Classroom)
- 10. Text Processing - Techniques & Traditions (McPherson Library A003, Classroom)
- 11. 3D Modelling for the Digital Humanities and Social Sciences (MacLaurin D010, Classroom)
- 12. Conceptualizing and Creating a Digital Edition (MacLaurin D103, Classroom)
- 13. Visualizing Information: Where Data Meets Design (MacLaurin D107, Classroom)
- 14. Introduction to Electronic Literature in DH: Research and Practice (MacLaurin D115, Classroom)
- 15. Race, Social Justice, and DH: Applied Theories and Methods (MacLaurin D105, Classroom)
- 16. XML Applications for Historical and Literary Research (Clearihue A103, Lab)
- 17. Process Humanities Multimedia (Human and Social Development A150, Lab)
- 18. Digital Games as Tools for Scholarly Research, Communication and Pedagogy (MacLaurin D110, Classroom)
- 19. Web APIs with Python (Human and Social Development A170, Lab)
- 20. Ethical Data Visualization: Taming Treacherous Data (MacLaurin D101, Classroom)
- 21. Digital Publishing in the Humanities (Clearihue D131, Classroom)
- 22. Linked Open Data and the Semantic Web (Clearihue D130, Classroom)
- 23. Introduction to IIIF: Sharing, Consuming, and Annotating the World’s Images (MacLaurin D114, Classroom)
- 24. Visualizing Information: Where Data Meets Design (MacLaurin D107, Classroom)
- 25. Feminist Digital Humanities: Theoretical, Social, and Material Engagements (Cornett A229, Classroom)
- 26. The Frontend: Modern JavaScript & CSS Development (Clearihue A030, Classroom)

10:15 to Noon

Lunch break / Unconference Coordination Session (MacLaurin A144)
(Grab a sandwich and come on down!)

Undergraduate Meet-up, Brown-Bag (details via email)

1:30 to 4:00

Classes in Session

Institute Panel: Perspectives on DH (or, #myDHis ...)
Chair: Alyssa Arbuckle (U Victoria)
(MacLaurin A144)

- Milena Radzikowska (Mt Royal C): “Release the Kraken: Story-Driven Prototyping for the Digital Humanities.”
  Abstract: I have spent the last 15 years of my career designing text analysis tools for use by humanities scholars. In this brief presentation, I propose to share a concept-based approach to interface design for DH.

- Emily Murphy (U Victoria): “#MyDHis Edgy.”
  Abstract: I will build upon—or, possibly, perform a misprision of—a tweet by Polina Vinogradova; “#myDHis messy, dusty, edgy, and radically inclusive!” Vinogradova evokes the mess and dust of the archives, the edges that connect nodes of a network, and the political impetus to think of cultural history and community together. I argue that these aspects of DH have a renewed importance as we head into a moment of feminist historiography.

- Margaret Konkol (Old Dominion U): “Prototyping Mina Loy’s Alphabet with a 3D Printer.”
  Abstract: This talk discusses the interpretive and methodological implications of using 3D printing technologies to prototype the archival diagrams of a proposed but never constructed plastic segmental alphabet letter kit—a game designed by modernist poet Mina Loy for F.A.O Schwarz. Although intended as a toy for young children, “The Alphabet that Builds Itself,” as a work of “object typography” articulates a theory of language as kinetic, geometric, recombinant, and open to mutation. Alphabetic segments extend into the x, y, and z coordinates in exponential iterations and conjoin with magnets. Combining elements of contemporaneous typefaces like Futura and Gill Sans, which represented modernity’s functional ideals and democratic principles of simplicity, these recombinant letters represent, as this talk argues, Loy’s unpublished modernist poem, an articulation of Loy’s concept of language as a physical fact in which substance, not just form, is semantic.

4:10 to 5:00

Lee Zickel (Case Western Reserve U): “Comfortably Trepid.”
Abstract: #myDHis found outside the well-established, DH-friendly institutions, at an institution that is devoted predominantly to Medicine and Engineering. I, and with increasing frequency other DH practitioners and instructors, am not positioned in a DH Lab or Humanities Center, but in ITS. Part teacher, part technologist, part translator, I will briefly discuss my work supporting humanists and social scientists, particularly those who are new to or less comfortable with computational methodologies.

- Dorothy Kim (Vassar C): “#MyDHis Antifascist.”
  Abstract: I’ve spent a lot of time in the last 12 months thinking about fascism, digital humanities, its long histories, and what it means to do DH work that centers social justice particularly in this global rise of late fascism. I will speak briefly about DH’s history, including the medieval history related to Busa but how that history really connects to data systems that created the Holocaust and also participated in the Cold War nuclear military complex.
Randa El Khatib (U Victoria): “Learning from the Iterative Process.”
Abstract: #MyDHIs Iterative. In addition to the improvements that come with iterative projects, the iterative process itself is a fruitful area for scholarly inquiry. Within this iterative context, the various teams that I work with and I have been reflecting on and rethinking central DH practices, such as what it means to collaborate, prototype, remix, and implement DH values in our work. In this talk, I will present the various lessons learnt along the way.

Sarah Melton (Boston C): “#MyDHIs...People.”
Abstract: Taking seriously Miriam Posner’s exhortation to “commit to DH people, not DH projects,” I invite us to reflect on how people are the core of DH. In this brief talk, I will explore the intersections between DH, labor, and infrastructure.

5:00 to 6:00
Opening Reception (University Club)
We are grateful to Gale Cengage for its sponsorship.

Tuesday, 5 June 2018

9:00 to Noon
Classes in Session

12:15 to 1:15
Mystery Lunches
DHSI Lunchtime Workshop Session (click for workshop details and free registration for DHSI participants)
- 73. Introduction to ORCID (Digital Scholarship Commons, Classroom).

1:30 to 4:00
Classes in Session

4:15 to 5:15
DHSI Colloquium Lightning Talk Session 1 (MacLaurin A144)
Chair: James O'Sullivan
- New Modes of DH and Archival Skills Acquisition in a Graduate Public History Course. Paulina Rousseau (Ryerson U)
- Walking a Transect: Exploring a Soundscape. John Barber (Washington State U)
- Centering the Edge Case: Designing Services for Humanities Data Research. Grace Afsari-Mamagani (New York U)
- Orwellian Vocabulary and the 21st-Century Politics. Ilgin Kizilgunesler (U Manitoba)
- Making Open Data from a Gray Archive. Sara Palmer (Emory U)

6:00 to 8:00
DHSI Newcomer's Beer-B-Q (Felicitas, Student Union Building)

Wednesday, 6 June 2018

9:00 to Noon
Classes in Session

12:15 to 1:15
Mystery Lunches
Brown Bag Lecture: Alexandra Branzan Albu (U Victoria): “Visual Recognition of Symbolic and Natural Patterns” (Digital Scholarship Commons, 3rd Floor McPherson Library)
Abstract: Image-based object recognition is a visual pattern recognition problem; one may characterize visual patterns as either symbolic or natural. Symbolic patterns evolved for human communication; they include but are not limited to text, forms, tables, graphics, engineering drawings etc. Symbolic patterns vary widely in terms of size, style, language, alphabet and fonts; however, literate humans can easily compensate for this variability and instantly recognize most symbolic patterns. On the other hand, natural patterns characterize images of physical structures; they often lack the intrinsic discriminability and structure of symbolic patterns, and vary widely in terms of pose, perspective, and lighting.
This lecture will explore similarities and differences in approaches designed for recognizing visual and symbolic patterns, and will address the following questions via examples.
- What are the distinctive characteristics of natural patterns? What dimensions of variability can we infer?
- What are the distinctive characteristics of symbolic patterns? What dimensions of variability can we infer?

Alexandra Branzan Albu is an Associate Professor with the Department of Electrical and Computer Engineering and cross-listed with Computer Science. Her research interests are related to image analysis, computer vision, and visual computing. She is actively pursuing outreach activities dedicated to increasing the women's presence in electrical engineering and computer science.

1:30 to 4:00
Classes in Session
Thursday, 7 June 2018

9:00 to Noon
- Classes in Session

12:15 to 1:15
- Lunch break / Unconference
- "Mystery" Lunches
- UVIC Library/ETCL lunchtime talk: “A Humanities Application of 3D printing and Machine Translation in the ChessBard and Loss Sets” by Dr. Aaron Tucker
- Digital Scholarship Commons, 3rd floor, Mearns Centre for Learning / McPherson Library
- Bring your lunch and come on up!.
- [Instructor lunch meeting]

1:30 to 4:00
- Classes in Session

4:15 to 5:15
- DHSI Colloquium Lightning Talk Session 3 (MacLaurin A144)
- Chair: James O'Sullivan
- Documenting Deportation: A Collaborative Digital Collection. Paulina Rousseau (Ryerson U)
- Unleashing the Power of Texts as Networks: Visualizing the Scholastic Commentaries and Texts Archive. Jeffrey Witt (Loyola U Maryland) and Drew Winget (Stanford U)
- #haunteDH: Punching holes in the International Busa Machine Narrative. Arun Jacob (McMaster U)
- Text in World: Computational Analysis of Trauma in Genocide Narratives. Nanditha Narayananamoorthy (U York) and Krish Perumal (U Toronto)

7:30 to 9:30
- (Groovy?) Movie Night (MacLaurin A144)

Friday, 8 June 2018 [DHSI; DLFxDHSI Opening]

9:00 to Noon
- DHSI Classes in Session

12:15 to 1:15
- DHSI Lunch Reception / Course E-Exhibits (MacLaurin A100)

1:00 to 2:00
- DLFxDHSI Registration (MacLaurin A100)

1:30 to 1:50
- [DHSI] Remarks, A Week in Review (MacLaurin A144)

2:00 to 3:00
- Joint Institute Lecture (DHSI and DLFxDHSI):
  - Bethany Nowviskie (CLIR DLF and U Virginia): “Reconstitute the World: Machine-reading Archives of Mass Extinction”
    - Chair: Lisa Goddard (U Victoria)
    - Abstract: The basic constitution of our digital collections becomes vastly more important in the face of two understandings: first, that archives of modernity are archives of the sixth great mass extinction of life on our planet; and next, that we no longer steward cultural heritage for human readers alone. In the same way that we people are shaped by what we read, hear, and see, the machine readers that follow us into and perhaps beyond the Anthropocene have begun to learn from "unsupervised" encounters with our digital libraries. What will we preserve for the living generations and artificial intelligences that will come? What do we neglect, or even choose to extinguish? And from an elegiac archive, a library of endings, can we create forward-looking, speculative collections—collections from which to deep-dream new futures? The most extra/ordinary power we possess is the power to make poetry from records of the past. Could it be called on, one day, to reconstitute the world?

2:30 to 3:30
- DHSI Classes in Session
Saturday, 9 June 2018 [DLFxDHSI + DHSI Conference and Colloquium]

8:30 to 9:00  DLFxDHSI Registration (MacLaurin A100)

9:00 to 5:30  DLFxDHSI UnConference Sessions

- DHSI All Day Workshop Session (click for workshop details and free registration for DHSI participants)

9:00 to 4:00

- 53. Building Your Academic Digital Identity (MacLaurin D105, Classroom)

- DHSI Colloquium Day Conference (MacLaurin A144)

Welcome

People I: Documenting Online Lives. Chair: Molly Nebiolo (University of New York)

- Examining Gendered Harassment Online and in Silicon Valley. Andrea Flores (Utica College)
- This is Just to Say I Have <X> the <Y> in your <Z>: Modernist Memes in an Era of Public Apology. Shawna Ross (Texas A&M University)

Break

People II: Documenting Lives Online. Chair: Dheepa Sundaram (College of Wooster)

- Youtube Yoga and Ritual on Demand: The Virtual Economics of Hindu Soteriology. Dheepa Sundaram (College of Wooster)
- The Resemblage Project: Creativity and Digital Health Humanities in Canada. Andrea Charise (University of Toronto) and Stefan Krecsy (University of Toronto)

Lunch

Projects I: Building and Analyzing. Chair: Yannis Rammos (New York University)

- Building the ARTECHNE Database: New directions in Digital Art History. Marieke Hendriksen (Old Dominion University)
- The Ineffective Inquisition: The Holy Office’s Sphere of Influence in Early Modern New Spain. Kira Homo (Pennsylvania State University)

Break

Projects II: Mapping and Visualizing. Chair: Innocent Opara (Quimet Institute)

- Mapping Sarah Sophia Bank’s Numismatic Collection. Erica Hayes (North Carolina State University) and Kacie Wills (University of California, Riverside)
- Text Mining and Visualizing 18th Century American Correspondence. Ashley Sanders Garcia (University of California, Los Angeles)

Break

Practices: Digital Scholarship on Campus and in the Classroom. Chair: Alyssa Arhuckle (University of Victoria)
Concluding Remarks

Sunday, 10 June 2018 [SINM + DHSI Registration, Workshops]

8:30 to 9:00 Symposium on Indigenous New Media Registration (MacLaurin A100)

9:00 to 5:00 DHSI Registration (MacLaurin A100)

9:00 to 4:00

▼ SINM Sessions

63. Symposium on Indigenous New Media: Reading Group (Hickman 105, Classroom)
72. Symposium on Indigenous New Media: Indigitization (Hickman 120, Classroom)
Full details here

▼ DHSI All Day Workshop Sessions (click for workshop details and free registration for DHSI participants)

9:00 to 4:00

53. Building Your Academic Digital Identity (MacLaurin D105, Classroom)
54. An Introduction to the Archaeology of 1980s Computing (MacLaurin D114, Classroom)

▼ DHSI AM Workshop Sessions (click for workshop details and free registration for DHSI participants)

9:00 to Noon

55. Regular Expressions (MacLaurin D111, Classroom)
56. 3D Visualization for the Humanities (MacLaurin D101, Classroom)
58. DH Fieldwork Methods (MacLaurin D016, Classroom)
60. Pedagogy of the Digitally Oppressed: Inculcating De-/Anti-/Post-Colonial Digital Humanities (MacLaurin D107, Classroom)
61. Introduction to #GraphPoem. Digital Tools for Poetry Computational Analysis and Graph Theory Apps in Poetry (MacLaurin D101, Classroom)
62. Creating a CV for Digital Humanities Makers (MacLaurin D115, Classroom)

▼ DHSI PM Workshop Sessions (click for workshop details and free registration for DHSI participants)

1:00 to 4:00

64. Agent-Based Modelling in the Humanities (MacLaurin D111, Classroom)
65. Unleash Linux on MacOS (MacLaurin D010, Classroom)
66. DHSI Knits: History of Textiles and Technology (MacLaurin D016, Classroom)
67. Crowdsourcing as a Tool for Research and Public Engagement (MacLaurin D109, Classroom)
69. Web Annotation as Critical Humanities Practice (MacLaurin D103, Classroom)
70. Dynamic Ontologies for the Humanities (MacLaurin D107, Classroom)
71. Social Media Research in the Humanities (MacLaurin D101, Classroom)

4:10 to 5:00 Joint Institute Lecture (DHSI and SINM):
David Gaertner (U British Columbia): "A Landless Territory?: CyberPowWow and the Politics of Indigenous New Media."
Chair: Deanna Reder (Simon Fraser U) (MacLaurin A144)

Abstract: Following the 1997 launch of Skawennati’s (Mohawk) CyberPowWow, digital space has become a vital new territory for the resurgence of Indigenous storytelling and cultural practice: "We have signed a new treaty," Cree artist Archer Pechawis wrote of this period, "and it is good. We have the right to hunt, fish, dance and make art at www.CyberPowWow.net, .org and .com for as long as the grass grows and the rivers flow." This talk will critically explore the theoretical, cultural, political-economic, and gendered dynamics underwriting the histories and futures of Indigenous new media. Particular attention will be given in examining the ways in which new media and digital storytelling connect to and support key issues in the field of Indigenous studies, such as sovereignty, self-determination, decolonization, and land rights.

After the day, many will wander to Cadboro Bay and the pub at Smuggler's Cove OR the other direction to Shelbourne Plaza and Maude Hunter's Pub OR even into the city for a bite to eat.

Monday, 11 June 2018 [DHSI + SINM]
Your hosts for the week are Ray Siemens and Dan Sondheim.

7:45 to 8:15  DHSI Last-minute Registration (MacLaurin A100)

8:30 to 10:00  DHSI Welcome, Orientation, and Instructor Overview (MacLaurin A144)

9:00 to 4:00  SINM Sessions

10:15 to Noon  DHSI Classes in Session (click for details and locations)

27. [Foundations] Understanding The Predigital Book: Technology and Texts (McPherson Library A003, Classroom)
28. [Foundations] Developing a Digital Project (With Omeka) (Clearihue D132, Classroom)
29. [Foundations] Models for DH at Liberal Arts Colleges (& 4 yr Institutions) (MacLaurin D109, Classroom)
30. [Foundations] Fundamentals of Programming/Coding for Human(s)ists (Clearihue A108, Lab)
32. Stylometry with R: Computer-Assisted Analysis of Literary Texts (Clearihue A102, Lab)
33. Digital Storytelling (MacLaurin D111, Classroom)
34. Text Mapping as Modelling (Clearihue D131, Classroom)
35. Geographical Information Systems in the Digital Humanities (Clearihue A105, Lab)
36. Open Access and Open Social Scholarship (MacLaurin D114, Classroom)
37. Introduction to Machine Learning in the Digital Humanities (Cornett A229, Classroom)
38. Queer Digital Humanities: Intersections, Interrogations, Iterations (MacLaurin D110, Classroom)
41. Using Fedora Commons / Islandora (Human and Social Development A160, Lab)
42. Documenting Born Digital Creative and Scholarly Works for Access and Preservation (MacLaurin D115, Classroom)
43. Games for Digital Humanists (MacLaurin D016, Classroom & Human and Social Development A170, Lab)
44. XPath for Document Archeology and Project Management (Cornett A128, Classroom)
46. Surveillance and the Digital Humanities (MacLaurin D103, Classroom)
47. Text Analysis with Python and the Natural Language ToolKit (Clearihue A103, Lab)
48. Information Security for Digital Researchers (Clearihue D130, Classroom)
49. Wrangling Big Data for DH (Human and Social Development A150, Lab)
50. Accessibility & Digital Environments (MacLaurin D101, Classroom)
51. Critical Pedagogy and Digital Praxis in the Humanities (MacLaurin D105, Classroom)
52. Drupal for Digital Humanities Projects (MacLaurin D107, Classroom)

12:15 to 1:15  Lunch break / Unconference Coordination Session (MacLaurin A144)

DHSI Undergraduate Meet-up, Brown-Bag (details via email)

1:30 to 4:00  DHSI Classes in Session

4:10 to 5:00  Joint Institute Lecture (DHSI and SINM): Jordan Abel (Simon Fraser U): "Indigeneity, Conceptualism, and the Borders of DH."
Chair: Michelle Brown (U Hawaii) (MacLaurin A144)

Abstract: This talk brings together digital humanities discourses in computational textual analysis and Indigenous Literary Studies to analyze a corpus comprised of every book of Indigenous poetry published in Canada, extending from Pauline Johnson's 1895 book The White Wampum to Marilyn Dumont's 2015 book The Pemmican Eaters. While the main goal of this research project initially centered on the topic modeling of a corpus of Indigenous poetry, the project also addresses the systemic barriers that have prevented such work gaining traction, and likewise attempts to address the specific challenges that Indigenous writing (and in particular Indigenous poetry) present to current Digital Humanities methodologies.

5:00 to 6:00  Joint Reception: DHSI and SINM (University Club)

Tuesday, 12 June 2018

9:00 to Noon  Classes in Session

12:15 to 1:15  Lunch break / Unconference

"Mystery" Lunches

DHSI Lunchtime Workshop Session (click for workshop details and free registration for DHSI participants)

73. Introduction to ORCID (Digital Scholarship Commons, Classroom).
<table>
<thead>
<tr>
<th>Time</th>
<th>Events</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
<tr>
<td></td>
<td>▼ DHSI Colloquium Lightning Talk Session 4 (MacLaurin A144)</td>
</tr>
<tr>
<td></td>
<td>Chair: Lindsey Seatter</td>
</tr>
</tbody>
</table>
| 4:15 to 5:15  | - Mapping Indigenous and Chicana/o Environmental Imaginaries using GIS. Stevie Ruiz (California State U, Northridge), Quetzalli Enrique (California State U, Northridge), Enrique Ramirez (California State U, Northridge), and Tomas Figueroa (California State U, Northridge)  
|               | - "But is it any good?": A quantitative approach to the popularity of digital fanfiction. Suzanne Black (U Edinburgh)  
|               | - The American Prison Writing Archive (APWA). Doran Larson (Hamilton C), Janet Simons (Digital Humanities Initiative, Hamilton C), and William Rasenberger (Hamilton C)  |
| 6:00 to 8:00  | DHSI Newcomer's Beer-B-Q (Felicitas, Student Union Building)           |
| 9:00 to Noon  | Classes in Session                                                     |
| 12:15 to 1:15 | Lunch break / Unconference                                            |
|               | "Mystery" Lunches                                                      |
| 1:30 to 4:00  | Classes in Session                                                     |
|               | ▼ DHSI Colloquium Lightning Talk Session 5 (MacLaurin A144)            |
|               | Chair: Lindsey Seatter                                                 |
| 4:15 to 5:15  | - Faraway, so close: Has the political environment really changed in Ecuador?. Luis Meneses (Electronic Textual Cultures Lab, U Victoria)  
|               | - Re-mixing Melville's Reading: Text Analysis of Marginalia with R and XSLT. Christopher Ohge (U London, School of Advanced Study) and Steven Olsen-Smith (Boise State U)  
|               | - Developing Interactive and Open-Source OER: Inquiry-Based Music Theory. Evan Williamson (U Idaho)  
|               | - Spatial Humanities and the Web of Everywhere. Ken Cooper (SUNY Geneseo)  |
| 6:00 to 7:00  | "Half Way There (yet again)!" [An Informal, Self-Organized Birds of a Feather Get-Together] (Felicitas, Student Union Building)  
|               | Bring your DHSI nametag and enjoy your first tipple on us!             |
|               | **Wednesday, 13 June 2018**                                           |
| 9:00 to Noon  | Classes in Session                                                     |
| 12:15 to 1:15 | Lunch break / Unconference                                            |
|               | "Mystery" Lunches                                                      |
| 1:30 to 4:00  | Classes in Session                                                     |
|               | ▼ DHSI Colloquium Lightning Talk Session 6 (MacLaurin A144)            |
|               | Chair: Lindsey Seatter                                                 |
| 4:15 to 5:15  | - Composition not Inheritance: Imagining a Functional Digital Humanities. Andrew Pilsch (Texas A&M U)  
|               | - Plotting Our Trajectories: Navigating, Situating, and Re-Inventing Research Topoi with R. Sean McCullough (Texas Christian University) and Jongkeyong Kim (Texas Christian U)  
|               | - Herb Simon and His Books. Avery Wiscomb (Carnegie Mellon U) and Daniel Evans (Carnegie Mellon U)  
|               | - (De/Re)Defining "The Digital": A Decolonial Approach to Digital Humanities. Ashley Caranto Morford (U Toronto) and Arun Jacob (McMaster U)  |
| 7:30 to 9:30  | (Groovier?) Movie(r) Night (MacLaurin A144)                            |
|               | **Thursday, 14 June 2018**                                             |
| 9:00 to Noon  | Classes in Session                                                     |
| 12:15 to 1:15 | Lunch break / Unconference                                            |
|               | "Mystery" Lunches                                                      |
| 1:30 to 4:00  | Classes in Session                                                     |
|               | ▼ DHSI Colloquium Lightning Talk Session 6 (MacLaurin A144)            |
|               | Chair: Lindsey Seatter                                                 |
| 4:15 to 5:15  | - Composition not Inheritance: Imagining a Functional Digital Humanities. Andrew Pilsch (Texas A&M U)  
|               | - Plotting Our Trajectories: Navigating, Situating, and Re-Inventing Research Topoi with R. Sean McCullough (Texas Christian University) and Jongkeyong Kim (Texas Christian U)  
|               | - Herb Simon and His Books. Avery Wiscomb (Carnegie Mellon U) and Daniel Evans (Carnegie Mellon U)  
|               | - (De/Re)Defining "The Digital": A Decolonial Approach to Digital Humanities. Ashley Caranto Morford (U Toronto) and Arun Jacob (McMaster U)  |
|               | **Friday, 15 June 2018**                                               |
| 9:00 to Noon  | Classes in Session                                                     |
| 12:15 to 1:15 | Lunch Reception / Course E-Exhibits (MacLaurin A100)                  |
Institute Lecture: William Bowen (U Toronto Scarborough): “Discovery, Collaboration and Dissemination: Lessons Learned and Plans for the Future” (MacLaurin A144)

Abstract: Much has changed and continues to change in digital humanities since the formal establishment of Iter in the Fall of 1997. However, the mandate of the not-for-profit partnership to support “the advancement of learning in the study and teaching of Middle Ages and Renaissance (400–1700) through the development and distribution of online resources” continues to have relevance. This presentation explores the striking challenges faced by Iter and presents our current thinking on the realization of this mandate for the future through a platform with a focus on facilitating the discovery of the academic resources necessary to our work; creating an environment for collaboration, sharing and developing projects; and on enabling the distribution and publication of our scholarship.

2:40 to 3:00

Awards and Bursaries Recognition
Closing, DHSI in Review (MacLaurin A144)

Contact info:
institut@uvic.ca  P: 250-472-5401  F: 250-472-5681
Stylometry with R: Computer-Assisted Analysis of Literary Texts

Jan Rybicki and Joanna Byszuk

Schedule

Monday, 11 June 2018
10:15–12:00 Visualizing Literature: Trees and Networks
1:30–4:00 stylo: Introduction to a Stylometric R Package

Tuesday, 12 June 2018
9:00–12:00 stylo: Grow Your Own Tree
1:30–4:00 Graph Theory: Knit Your Own Network

Wednesday, 13 June 2018
9:00–12:00 oppose: Words That Matter
1:30 to 4:00 classify: Introduction to Supervised Analysis

Thursday, 14 June 2018
9:00–12:00 Now That we Know So Much: Own Project Time
1:30 to 4:00 Sequential Stylometry and Other Clever Tricks

Friday, 15 June 2018
9:30–12:00 What Have We Learned?
Useful links

R project, http://cran.r-project.org
Computational Stylistics Group: https://sites.google.com/site/computationalstylistics
Gephi, https://gephi.org

Reading (in the order of discussing in the class)


If you feel ambitious, you can also look up:


Computational Stylistics and Text Analysis

1. Introduction

Computational stylistics and text analysis have a long, rich history. In retrospect, because of the nature of texts and the capabilities of computers, it seems quite predictable that they would be among the first applications of computers to the humanities. Many religious, literary, and historical texts are highly valued cultural achievements, and some of them have been analyzed for hundreds or even thousands of years. They also contain large numbers of highly significant and meaningful words and other textual features. Thus it seems natural that scholars should have moved quickly to enhance and augment their own description, characterization, and analysis of these rich cultural documents by harnessing the power of computers to store, search, count, and compare textual features. The rapid growth of the power of computers and the rapid increase in the availability of electronic versions of texts have revolutionized the scope and the kinds of analysis that can be performed. At their core, however, computational stylistics and text analysis have remained true to their origins, and continue to use the power of the computer to improve our understanding of texts.

“Computational Stylistics” seems a relatively transparent phrase, but it may be useful to pick it apart a bit. First, “computational” obviously and correctly implies the use of computers, but it leaves unexpressed the rather wide range of ways they can be used. Simple text searches, concordances, and textual manipulation and selection could be counted as computational, but
most practitioners would reserve the term in this context for some kind of statistical analysis, ranging from t-tests, to Principal Components analysis and cluster analysis, to Delta analysis, data mining of various kinds, support vector machines, topic modeling, and even neural networks. Most of these analytic methods have their origins in the closely related field of authorship attribution, and, given that many of them focus on stylistic differences, the two fields are sometimes difficult to distinguish (Craig 1999). The process of distinguishing authors emphasizes the importance of differences and similarities, and almost all of the analytic methods applied to texts are focused on detecting these differences and similarities.

1.1 Style

Style, the subject matter of “Stylistics” is, most broadly, simply a way of doing something. In the simplest case, it is an author’s way of writing. In practice, however, the focus is on the effects of an author’s style on his or her texts. In one recent formulation, “Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively” (Herrmann, van Dalen-Oskam, and Schöch 2015: 16). It is widely assumed, though unprovable, that the features that constitute each author’s style form a unique stylistic fingerprint, so that, if the correct features are chosen, any two authors can be distinguished from each other. Considered as a property of texts, style can also be extended to apply to what can loosely be called genres (the Gothic novel, epic poetry, satire, narrative, drama), literary-historical periods (Victorian, Romantic), chronological divisions within an author’s career (early and late Henry James), or to variations within a single text (the “voices” of different characters or narrators, for example), among many other possibilities.
Style is chiefly linguistic, though in some cases graphological features, the layout or arrangement of text, and even the physical characteristics of a text may contribute to a style. In addition to the obviously linguistic elements of style, such as vocabulary, grammar, morphology, phonology, and figures of speech, most practitioners would include broader characteristics, such as world view, theme, and tone, as potential elements for analysis (for an excellent checklist of stylistic features, see Leech and Short 2007: 61ff). Style is also patterned and distributed. Local and unique stylistic features can be important, but a recognizable style normally involves some kind of repetition, consistency, or pattern.

1.2 Stylistics

Stylistics is essentially comparative, even if the comparison is not always explicit. Almost all statements about a style imply a comparison; for example, even the seemingly simple statement that Faulkner’s style is marked by long sentences implies a contrast with the lengths of the sentences of other authors. Long compared to what? Although the question of what norm is appropriate for the comparison remains vexed, the widespread availability of electronic texts and corpora of texts has made defensible choices and the creation of specialized corpora to use as norms easier to make. Pattern, distribution, and comparison obviously invite a computational approach. Indeed, computation analysis is the only practical way to analyze extremely frequent textual characteristics, or to study unreadably large collections of texts.

Despite the variety of stylistic features, it is fair to say that the overwhelming majority of computational stylistic analyses have involved words, though word n-grams (sequences of words) have recently become increasingly popular features to analyze. Not only are words (and
n-grams) easily identifiable and countable, compared to figures of speech, themes, or syntactic patterns, they are also much more frequent than most other textual characteristics and are obviously, though not unproblematically, meaningful (unlike, for example, sequences of letters or parts of speech). A “word” seems an intuitively simpler concept than it is in practice, and various decisions about how to identify and count the words of a text are defensible under different circumstances. For the purposes of computational stylistics, a word (type) is normally defined as any unique sequence of alphanumerical characters that is not interrupted by a space, or by any punctuation mark except the apostrophe or hyphen. (A type is a unique form, while a token is an individual occurrence of a type: the previous sentence contains two tokens of the type “or”.) Unfortunately, this definition does not distinguish homographic forms like the noun and verb meanings of desert, but experience has shown that computational stylistics is robust enough that the resulting errors in counting do not seriously distort analysis.

1.3 Text Analysis

Text Analysis is a close relative of computational stylistics, but with a wider range and a heavier emphasis on analysis. While computational stylistics has focused almost exclusively on literary texts, text analysis has been applied to many other kinds of texts, from political speeches to blogs, from historical documents to tweets, from legal documents to the sacred texts of religions, from letters to philosophical treatises, from poetry to programming. Perhaps the most obvious further difference between computational stylistics and text analysis is that the latter is more likely to focus on meaning and content. Nevertheless, almost all of the methods of text analysis
have also been applied to questions of literature, authorship, and style.

It would be an exercise of folly to attempt an introduction here to data mining, topic modeling, sentiment analysis, semantic analysis, neural networks, part-of-speech analysis, word-frequency analysis, and of the wide range of statistical analysis techniques that have been applied to the dozens of different textual features that have been analyzed. Instead, it seems more useful to approach computational stylistics and text analysis by focusing on problems at three different scales: microanalysis, middle-distance analysis, and macroanalysis or distant reading.

The first analysis on the micro scale is an authorship problem involving a collaboratively written text, *The World’s Desire* by H. Rider Haggard and Andrew Lang, using a popular recent technique called Rolling Delta. This is followed by an analysis of the voices of the six narrators in Virginia Woolf’s *The Waves*. Middle-distance analysis is represented by a modification of John Burrows’s Zeta (Burrows 2006) that examines the vocabulary of more than 350 high-stakes exit essays written by American high school students. Macroanalysis is demonstrated by turning to the chronological signal visible in much larger corpora: in this case, in 1000 English novels from Swift (Jonathan) to James (E. L.).

### 2. Microanalysis, or Zooming into a Single Text

Empirical investigations in the field of computational stylistics and text analysis are, as noted above, almost exclusively focused on comparison: to reliably describe a given text’s statistical characteristics, in a vast majority of cases one compares the text to other texts collected in a comparison corpus. From this perspective, a single text, perceived in a context of similar or not-
so-similar texts, becomes a monadic entity *per se*. Even if such a text is further divided into smaller samples (see e.g. Kestemont, Moens, and Deploige 2015), the main goal of finding relations between discrete textual entities (works) continues to be the main focus.

This approach assumes that a (literary) work is a monolith, which is not always true: an epistolary novel might consist of multiple stylistic registers, a Menippean satire might combine sections of epic poetry, tragedy, and philosophical prose, a collaboratively written work might contain two or more independent authorial voices, and so forth. In such cases, capturing an average stylistic profile from the text in its entirety is certainly not the optimal scenario. Arguably, much more can be observed when such a text is divided into segments and treated independently. One of the possible applications of this approach is discussed below, where Virginia Woolf’s *The Waves* is dissected according to particular characters’ voices; another application involves chunking the input text into consecutive samples, or equal-size blocks of *n* words (tokens), that are then measured as independent, yet sequentially ordered, samples.

### 2.1 The World’s Desire

Pioneering work in sequential stylometry was presented in a study on the authorship of *Walewein* (van Dalen-Oskam and van Zundert 2007), in a comparison of three disputed English prose texts (Burrows 2010), and in a study of *The Tutor’s Story*, written collaboratively by Kingsley and Malet (Hoover 2012). The sequential methodology evolved into the Rolling Delta method (Ryacki, Hoover, and Kestemont 2014), later extended and generalized as Rolling Classify (Eder 2015a). This method will be used here to assess the nature of collaboration between Henry Rider Haggard and Andrew Lang on *The World’s Desire*, first published in 1890.
2.1.1 Background

Henry Rider Haggard (1856-1925) is the author of several adventure novels, among which the bestsellers *She* (1887) and *King Solomon’s Mines* (1885) attracted a good deal of attention. Andrew Lang (1844-1912), a poet, novelist, literary critic, and folklore scholar, earned his fame as a translator of Homeric poems. *The World’s Desire*, a classic fantasy novel written collaboratively by the duo, not particularly long (*ca.* 85,000 words), is a story of the hero Odysseus, who returns home to Ithaca after his journey: instead of finding his home at peace, however, he is involved in several new adventures. The plot of the novel as well as its mythological background was set by Lang, while Haggard contributed his imagination and style. From the correspondence of the two writers, we know that Haggard had written a first draft, entitled *The Song of the Bow*, that was later reworked by Lang. Haggard then took over and wrote a great share of the text. In Haggard’s own words:

Roughly the history of this tale […] is that Lang and I discussed it. Then I wrote a part of it, which part he altered or rewrote. Next in his casual manner he lost the whole MS. for a year or so; then it was unexpectedly found, and encouraged thereby I went on and wrote the rest. […] The MS. contains fifty-three sheets at the beginning written or re-written by Lang, and about 130 sheets in my writing, together with various addenda. (Haggard 1926)

It is assumed that Haggard actually wrote most of the novel except the first four chapters, which were written entirely by Lang. Working on the first drafts of the novels, the two authors were
aware of stylistic differences between them. Haggard quotes Lang’s letter, which confirms his habit of depreciating his own work:

Nov. 27th. The typewritten “Song of the Bow” has come. The Prologue I wrote is better out. It is very odd to see how your part (though not your *chef d’oeuvre*) is readable, and how mine—isn’t. (Haggard 1926)

2.1.2 The Two Authors in *The World’s Desire*

The work by Haggard and Lang seems to be a perfect case study of mixed authorship of a single text. To tell its authorial voices apart, the Rolling Classify was applied. First, the goal was to compile a reference corpus containing authorial profiles of both Haggard and Lang. Out of an extensive list of Haggard’s works, 10 novels and 2 collections of short stories were selected to train a Haggardian profile: *Cetywayo and His White Neighbours* (1882), *Allan’s Wife and Other Tales* (1887), *Allan Quatermain* (1888), *Colonel Quaritch, V.C.* (1888), *Cleopatra* (1889), *Beatrice* (1893), *Black Heart and White Heart* (1900), *Ayesha: The Return of She* (1905), *Benita* (1906), *The Yellow God* (1908), *Child of Storm* (1913), *Allan and the Holy Flower* (1915). When it comes to Lang, a similar selection of 12 novels and short stories collections was compiled: *Much Darker Days* (1884), *In the Wrong Paradise and Other Stories* (1886), *He* (1887), *The Gold of Fairmilee* (1888), *Prince Prigio* (1889), *The Green Fairy Book* (1892), *Prince Ricardo of Pantouflia* (1893), *The Disentanglers* (1902), *The Crimson Fairy Book* (1903), *The Olive Fairy Book* (1906), *The Brown Fairy Book* (1904), *The Lilac Fairy Book* (1910).

The above representative text samples are used to train a model using one of the supervised classification techniques, namely, Support Vector Machines. The testing procedure
starts with chunking *The World’s Desire* into consecutive samples, or equal-size blocks of 5,000 words, with an overlap of 4,500 words, to achieve a dense sampling rate. Next, the support vector machine classifier is applied sequentially to the particular samples, which are checked against the training set, in order to identify the most similar authorial profile. The final stage of the analysis involves a graphical representation of stylistic changes throughout the chunked text. To this end, horizontal stripes are used, which are colored according to the assigned class.

In Fig. 1, the results of the Rolling Classify technique applied to *The World’s Desire*, using 100 MFWs, are shown. One can easily observe a stylistic takeover in the first part of the text. The break point takes place in the middle of the sixth chapter. Also, some sections by Lang seem to appear in the central part of the novel. However, these evaporate when a different MFW stratum is tested.

*Figure 1. Sequential analysis of The World’s Desire by Haggard and Lang: Rolling SVM and 100 MFWs*
In Fig. 2, one can observe the behavior of *The World’s Desire* when 500 MFWs are analyzed. This time, Haggard’s signal shows up for a short moment in the first chapters of the novel. The picture is once more slightly different when 1000 MFWs are taken into consideration (Fig. 3). It is quite clear that for very long vectors of frequent words, the distinction between two authorial voices in the sixth chapter is the only takeover that can be observed.

*Figure 2. Sequential analysis of The World’s Desire by Haggard and Lang: Rolling SVM and 500 MFWs*

*Figure 3. Sequential analysis of The World’s Desire by Haggard and Lang: Rolling SVM and 1000 MFWs*
Comparison of Figures 1-3 leads to the conclusion that the mixed authorship has a form of a sudden takeover rather than a mixture of interwoven authorial voices. However, it is much more difficult to explain the clutter that appears in different segments of the novel depending on the input parameters of the model. This observation is confirmed by a series of similar tests using different classifiers and different style-markers, such as the most frequent word 2-grams (word pairs). At this point, one of the most difficult problems of text classification arises, namely, the distinction between an actual signal and wrong decisions of the classifier (also referred to as false positives). Since this problem goes far beyond the scope of this chapter, it will not be discussed in detail. However, an intuitive and relatively simple way of filtering out the false positives is to perform a series of similar yet not identical tests, in which the parameters of the model (e.g., the number of MFWs) are modified. Patterns that appear despite differences in MFWs tested seem to suggest the existence of a signal, while any ephemeral clutter in the results might simply mean false positives. If this rule-of-thumb is true, *The World’s Desire* has one takeover only, which falls in the middle of the sixth chapter.

2.2 Virginia Woolf’s Character Voices

*The Waves*, Virginia Woolf’s most experimental novel, consists of alternating soliloquies or monologues by three male and three female characters, from childhood through middle age, each clearly indicated by a simple speech-reporting phrase like “said Bernard” or “said Jinny.”
Woolf’s technique has invited considerable critical comment about what axes of difference or unity characterize the novel, as Stephen Ramsay has noted:

Are Woolf’s individuated characters to be understood as six sides of an individual consciousness (six modalities of an idealized Modernist self?), or are we meant to read against the fiction of unity that Woolf has created by having each of these modalities assume the same stylistic voice? (2011: 10).

Ramsay’s claim that it would be a mistake to treat the question of whether the six voices are the same or different as one that can be answered has been disputed (Hoover, forthcoming; see also Hoover 2014 and Plasek and Hoover 2014). However, it seems worthwhile to leave the polemic aside, here, and look a bit more closely at the voices. (For a thorough recent discussion of the various views about the similarities and differences among the voices in The Waves, see Balossi, 2014: Chapters 1-2.)

2.2.1 Distinguishing the Six Voices

Testing the similarities and differences among the six voices is not as simple as it might seem. Because they are so obviously different from the monologues, it seems prudent first to eliminate the sections of third-person narration that begin each chapter and to remove all quotations from other characters from the monologues, so as to analyze only each character’s voice (as does Burrows 1987: 191, 205-07). More problematically, the lengths of the six monologues vary from Susan’s 6,067 words to Bernard’s 32,664. The final chapter of the novel, which is all in Bernard’s voice, begins “Now to sum up,” showing that it is likely to be quite different from the rest of the novel. This chapter has been excluded from the analysis, as it was in three previous
analyses of the novel (Burrows 1987: 206; Ramsay 2011; Balossi 2014: 84). Unfortunately, this still leaves the numbers of words by each character quite unbalanced: Bernard, 16460; Jinny, 6281; Louis, 8694; Neville, 9958; Rhoda, 8401; Susan, 6067. To give each character the same weight, each monologue has been reduced to the length of Susan’s. Simply taking the first 6,067 words of each, however, would mean that only Susan’s whole life would be represented, so the lines of each monologue were randomized and each was cut to 6,067 words. After creating a word frequency list based on the six equal parts, a series of cluster analyses was performed, based on the 100, 200, . . . 1,000 most frequent words of this list on the six equal parts and all the remaining text in sections of about 3000 words. All the analyses correctly group all of the sections by a single character except the one based on the 500 most frequent words; a representative analysis based on the 900 most frequent words is shown in the cluster analysis in Fig. 4. A similar analysis, along with others based on word 2-grams (sequences of two words) and words selected on the basis of consistent occurrence rather

---

1 Cluster analysis is an exploratory method of analysis that is often used in authorship studies. It compares the similarities and differences among the frequencies of all the words being analyzed in all the texts. The texts that use the words in the most similar way are grouped together in such a way that the more similar two texts are, the closer to the left of the graph they join together. In Fig. 4, the two most similar sections are the top two by Bernard (the vertical proximity of sections like Neville Rand Bal and Jinny Rand 6067 is not meaningful).
than high frequency also clearly distinguish the six voices in *The Waves* (see Hoover, forthcoming, for details). In spite of the fact that these six voices are all obviously versions of Woolf’s voice, they are much easier to distinguish than are sections of texts by some pairs of authors.

### 2.2.2 Age and Gender and the Six Voices

Given how distinctive the six voices are when analyzed with the lines randomly organized, it may be surprising that, at the same time, she also manages to distinguish the young voices from the older ones. In chapter one, the six are young children and in chapter two they go off to boarding schools. By chapter three they are entering college, and at the end of the novel the five surviving characters are middle-aged. A cluster analysis based on the 500MFW of the first and second chapters is shown in Fig. 5. Note that only Bernard’s part, at 1,598 words is longer than 1,000 words, Jinny’s is only 405 words, and Neville’s only 505; most analysts would consider these too short for reliable analysis.

Nonetheless, all six of the first chapters cluster separately from those from the second (sections from the second chapter that are longer than 2,000 words have been divided into one section of 1,000 words and a second section consisting of the remainder). The two sections by Bernard and
Neville (though not Louis) also cluster together, and the sections of chapter two by the three female characters cluster separately from those of the male characters, suggesting that, at least when they are children, Woolf has also created a gender split in their language. Cutting these chapters into still smaller sections (405-799 words) shows just how carefully Woolf has distinguished the voices in her novel. As Fig. 6, based on the 700MFW, shows, only the second half of Rhoda’s chapter one fails to cluster with the rest of the chapter one sections at the top of the graph, and the sections of Rhoda, Jinny, and Susan from chapter two cluster together, though those of Bernard, Neville, and Louis cluster only partially. What is extraordinary here is how well these very short sections group together both by age and by character, and, except for two of the sections by Louis, also by gender. (For a more detailed analysis of the younger and older voices, based on very different methods, see Balossi 2014: chapter 6 and Appendix E; for a discussion of the strengths of authorship, genre, gender, and other signals in texts, see Jockers 2013: chapter 6).
One more analysis will show that Woolf also distinguishes the voices of her characters as boarding school students in chapter two from their voices as adults in chapter eight. As Fig. 7 shows, with the exception of Bernard’s monologue from chapter eight, all of the monologues from chapter two cluster separately from those of chapter eight. An examination of the monologues from chapter two suggests that, even at a young age, Bernard’s style is more mature and complex that those of the other characters. Although the gender separation of chapter two naturally reappears here, it disappears entirely in the monologues from chapter eight (looking back at Fig. 4 shows that the randomized parts also fail to group by gender). Obviously, there is no space here for any approach to a full investigation, but these results show convincingly that the methods of computational stylistics can be valuable for exploratory studies, not least by suggesting productive new possibilities for further analysis and discussion.
3. Middle-Distance Analysis: High Stakes Writing Exams

Consider now a middle-sized group of very different texts: high-stakes exit-level writing exams. Although these texts lack any significant intrinsic value of their own, the tools of computational stylistics and text analysis can still produce revealing and worthwhile results. The texts to be analyzed here are a set of 366 essays written by North American high school students in the context of state-wide exit-level writing exams, administered in the final year of high school. These essays were collected and analyzed in a study of the idea of voice in writing assessment (Jeffrey 2010). The essays are not ideal for computational text analysis because they are quite short, averaging only about 430 words and ranging from 128 to 1307 words (only 21 are shorter than 200 or longer than 800 words). They also come from 39 states, and are responses to many different prompts that call for writing in a variety of genres; for example, analytic, narrative, argumentative, explanatory, and informative. If analyzing texts can achieve interesting results under these unfavorable conditions, we can expect excellent performance on longer texts under more favorable conditions.

3.1 Methodology

The method demonstrated here is a variant of Burrows’s Zeta (Burrows 2006; Hoover 2007), as modified by Craig and Kinney (2010) and further modified in Wide-Spectrum analysis (Hoover 2013). The method is especially useful for characterizing the vocabularies of any pair of authors, genres, texts, or indeed any pair of text collections that can be divided unambiguously into two classes. Here, the first comparison will be between low-scoring and high-scoring passing exams (failing exams tend to be very short and defective). Although there are undoubtedly differences of many kinds between the two groups, here we concentrate on what
kinds of consistent vocabulary differences, if any, exist between the low-scoring and high-scoring essays.

Wide-spectrum analysis, unlike most methods of text analysis, is based on consistency of use rather than frequency, and its calculation is simple and straightforward. For example, assume there are 200 texts, approximately the same length, by two authors, 100 by each author. Assume further that the word “eyes” is present in 62 (.62) of the texts by the first author and absent from 92 (.92) of the texts by the second author. These percentages are added together to yield a distinctiveness score for “eyes” of .62 + .92 = 1.54. Although distinctiveness scores can range from two (100% presence plus 100% avoidance) to zero (0% presence plus 0% absence), in practice scores above 1.5 are strongly characteristic of the first author and those with distinctiveness scores below .5 strongly characteristic of the second author. Note that quite different distributions can produce similar distinctiveness scores. For example, a word that is present in 77% of the texts by one author and absent from 77% of texts by the other would also have a distinctiveness score of 1.54.

Because of the wide range of sizes in the essays to be analyzed and the variety in prompts and genres, all the low-scoring essays have been combined into one text and all the high-scoring essays into another text and then the lines of each combined text have been sorted in random order. The final 14,000 words of each randomized text have been reserved for testing, leaving the rest of the texts for training purposes. Finally, to avoid basing the analysis on topical words, or words that are frequent because of specific prompts used in states with large numbers of essays, the word list has been manually culled by removing proper names (names of states and character names from text-based prompts, for example) and other topical words. The randomized training
and testing texts were then cut into blocks of 2000 words and analyzed in the Wide-Spectrum spreadsheet, which automates the process of comparing the two main sets of texts, calculates a distinctiveness score for all the words, and sorts those above a neutral score of one from high to low and those below one from low to high. The sheet also collects the most distinctive words for each group in order of distinctiveness, graphs the results, and prepares them for further graphing.

3.2 Vocabulary and the Evaluation of High School Writing

The results of the analysis described above are presented in Fig. 8.

*Figure 8. High and Low Scoring Essays*
The horizontal axis in Fig. 8 indicates the percentage of the word types in each section that are characteristic of the training sections of the high-scoring exams and the vertical axis shows the percentage of the word types in each section that are characteristic of the training sections of the low-scoring exams. (A word type is defined here as a unique spelling; each 2,000-word section typically contains only about 700-800 types because many common words are repeated.) For example, for the high-scoring training section at the bottom right of the graph, only about 23% of the 820 different types are characteristic of low-scoring exams, while almost 53% are characteristic of high-scoring exams. It is easy to see that this method does an excellent job of categorizing the sections that were held out for testing, in spite of the fact that the test sections had no part in creating the word lists on which Fig. 8 is based. All fourteen of the high-scoring and low-scoring sections are much closer to the appropriate training texts than to the opposite ones. (Tested sections do not usually fall within the clusters of training texts because they contain many words that are not in the training sets, so that the percentages of types that are characteristic of each group are lower than for the training texts.)

Although Wide-spectrum analysis does an excellent job of categorizing the test texts here, its real value is in characterizing the texts themselves. Below are the fifty most distinctive words for the two groups, unlabeled so that readers can try their hands at identifying which list is characteristic of the high-scoring essays and which is characteristic of the low-scoring ones.

big, lot, might, makes, stop, maybe, talking, major, example, picture, conclusion, older, sometimes, I'm, try, bad, tell, trouble, hurt, harder, anywhere, got, end, give, remember, problems, affect, car, understand, times, goes, cars, younger, basketball, nice, anything,
wouldn’t, saying, public, everybody, kids, decision, weather, doing, mom, teen, extending, normally, technique, someone

completely, must, eyes, desire, physical, constantly, individual, far, stress, mind, precious, human, light, involved, simply, rather, understanding, fit, build, led, ever, reader, air, changes, beginning, red, responsibility, order, nothing, began, increase, eventually, future, views, clock, merely, which, college, true, difficult, actually, science, once, learning, one’s, ability, walk, although, various, aware

Most readers correctly identify the first list as words characteristic of the low-scoring essays and the second list as words characteristic of the high-scoring essays. There is no space for a full analysis of the differences in the lists, but two important dimensions are formality and specificity (see Jeffrey 2010 and Jeffrey, Hoover, and Han 2013 for more details). The low-scoring essays use a much more casual and informal vocabulary, most noticeable in big, lot, I’ll, bad, got, wouldn’t, kids, and mom, while the high-scoring essays use more formal words like desire, physical, individual, precious, responsibility, science, one’s, and various. The low-scoring essays also tend to use vague and unspecific vocabulary like lot, sometimes, bad, anywhere, affect, nice, anything, everybody, doing, someone, while the high-scoring essays tend to be more specific, using words like completely, constantly, stress, human, build, increase, difficult, and learning.

An examination of the 1,000 most distinctive words for each group confirms these trends. For example, the first list above contains two contractions and the second none, and there are just
three contractions among the 1,000 most distinctive high-scoring words, compared to twenty-two among the 1,000 most distinctive low-scoring words. Along with anywhere, anything, sometimes, and someone in the list above are somebody, someday, someone’s, something, sometime, whenever, and whatever among the 1,000 most distinctive. The nouns in the two lists are also revealing:

Low-scoring nouns:

- lot, example, picture, conclusion, trouble, end, problems, car, times, cars, basketball,
- kids, decision, weather, mom, teen, technique

High-scoring nouns:

- eyes, desire, individual, stress, mind, light, understanding, reader, air, changes,
- beginning, responsibility, order, future, views, clock, college, science, ability

It isn’t hard to use a lot of nice low-scoring examples to give a picture of kids who at times have problems, athletes who really love cars, watching television, and playing sports like basketball, and want to talk about the weather (words that are among the 1,000 most distinctive low-scoring words are in italics).

However, the high-scoring words present a well-rounded individual, a reader with a passion for learning and a desire for future responsibility who is admired by her teachers, almost
all of whom acknowledge her leadership potential and her deep understanding of literature.
(words that are among the 1,000 most distinctive high-scoring words are in italics).

This kind of analysis provides a great deal of interesting evidence about what the exam
graders value and do not value, though it would be dangerous and irresponsible to imagine that
the two passages above fairly characterize the two groups of students. While there is no space
here to pursue the analysis further, it should be clear that such evidence can be very profitably
used in discussing and evaluating the process of high-stakes writing tests and their implications
for education.

4. Massive Text Analysis of 1000 Novels

It seems that any quantitative analysis worth its salt should yield significant chronological
differences in a 1000-novel corpus that extends from the times of Jonathan Swift to those of E.
L. James, especially if it is based on such simple linguistic features as word frequencies. There is
an obvious expectation of linguistic difference between the former’s and the latter’s images of,
say, bondage in, respectively, *Gulliver’s Travels into Several Remote Nations of the World* and
*Fifty Shades of Grey*:

I had the fortune to break the strings, and wrench out the pegs that fastened my left
arm to the ground, for, by lifting it up to my face, I discovered the methods they had
taken to bind me, and at the same time with a violent pull, which gave me excessive
pain, I a little loosened the strings that tied down my hair on the left side, so that I was just able to turn my head about two inches.

His deft fingers skim my back occasionally as they work down my hair, and each casual touch is like a sweet, electric shock against my skin. He fastens the end with a hair tie, then gently tugs the braid so that I’m forced to step back flush against him. He pulls again to the side so that I angle my head, giving him easier access to my neck.

Or is there? While no machine is needed to learn and discover the difference between the two texts, most of the differences between the two depictions – apart from what was then and what is now acceptable in print – seem to concern syntax rather than lexis: “I a little loosened the strings” sounds 18th-century, not 21st; but the narrator of *Shades*, had it been part of the game, would probably have said “I loosened the strings a little,” and the words and their frequencies would have remained unaffected. And yet Fig. 9 shows the two texts at two different extremities in a network visualization of nearest-neighbor links between texts based on cluster-analyzed Delta distances between them based in turn on most-frequent word frequencies (Eder 2015b). While the texts by the two authors are bound by the strongest signal known to stylometry, that of authorship, the authorial groups for both groups clearly order according to general chronology.

More importantly, not only those two texts behave this way. Whatever is between and around their data points follows suit, and the entire network exhibits a very Morettian evolution of greyscale from black to light grey, from left to right and from early to late. There are departures from a purely linear sequence, of course, and some authors move and/or evolve
vertically rather than horizontally; but the overall phenomenon is unmistakable. But then this has already been noted a long time ago with other methods, different statistical tools, and other corpora (not to reach too far back, cf. Brainerd 1980, Burrows 1994, Opas 1996). It is perhaps more significant that, against the background of this great evolution, local evolutions in chronology can be observed in some authors – those, obviously, who are represented by more texts in this corpus than either J. Swift or E. L. James. The latter’s more august last-name-sake, Henry, has already been shown to evolve in the context of his own work (Hoover 2007); but here, too, the chronological sequence of his works in Fig. 9 is almost perfect. Dickens evolves as well: while less linearly, his evolution seems to follow the general left-to-right trend. By contrast, Joseph Conrad’s evolution seems to go against the grain: it is almost as if the Polish-born writer evolved his English towards a more “traditional” usage of most frequent words.

When the two evolutionary phenomena – that of an English literature more or less smoothly moving in a single direction with time, and of a similar movement within a single author – are put together, the picture becomes, perhaps paradoxically, less clear. When one compares the scale of the evolution of Henry James with that of the entire corpus, the former seems to be blown out of proportion. Obviously, there is no direct reflection of any quantitative distance between the texts, and the distances themselves are the combined results of the balanced forces of the gravitational pull of the networking algorithm; still, if the entire “stylistic drift” were to be blamed on historical-linguistic factors alone, that proportion would probably be better maintained. Perhaps most significantly, whatever it is that a network analysis of this kinds represents, it seems to accord quite well with combined literary and linguistic expectations. In simpler language, this network makes a lot of sense.
Figure 9. Network visualization of 1000 English novels between published between 1704 and 2013. Chronology is indicated by the transition from black (early) to light grey (late).

It would then seem that, in this context, the real interest of a network graph like the one in Fig. 9 is not where it agrees with literary history, but where it does not. The anomaly of counter-chronological Conrad has already been mentioned, but another Slavic ESL user, Nabokov, behaves in exactly the same way. Lawrence Sterne’s *Sentimental Journey* and *Tristram Shandy* have been unseasonably removed to the right; this seems to illustrate quite well the very experimental nature of the latter work that culminated in Shklovsky calling its author “a radical revolutionary as far as form (emphasis added) is concerned” (Shklovsky 1990: 147). By contrast,
Beckett’s trilogy leans left, and it has been called “archaic” by Harold Bloom (1988: 9). Nearby lies the magical land of Narnia, but that is even easier to explain as a throwback to archaized, or mediaevalized, fantasy. The other famous fantasist, Tolkien, departs from the flow altogether as one of the outsiders; hanging on to the main body of the English novel network by the threads of his links to the historical romances of Bulwer-Lytton, he gravitates towards the poetic romances of William Morris – and that, too, seems to make a lot of sense.

5. Conclusions

The presentation of the above examples – the examples being necessarily simple and the presentation necessarily short – is intended to show how the tools and methods of Computational Stylistics can be used in the study of literature by any interested scholar who takes the trouble of learning the comparatively simple and comparatively user-friendly software. The authors have hoped to show that while Computational Stylistics sails into uncharted waters by becoming a partner in what Moretti has called distant reading (2015) and what Jockers has called macroanalysis (2013), there is much it can do to help in the traditional study of texts, whether it sticks to its earliest application in authorship attribution; whether it produces an image of literary history; or whether it helps to tell good student essays from bad.

It is interesting to see how the three levels of computational text analyses, the micro, the medium, and the macro, can produce results that create an added value at different levels of literary study. It is a very traditional task for scholars of literature to discern who wrote which part of a text; here, the computational stylist provides support – or correction – to what has been
established on the basis of such obvious sources as the correspondence and the reminiscences of Haggard and Lang. It should be remembered that the importance of their *The World’s Desire* in the body of popular fiction of their time can probably equal the recent media frenzy about the authorship of Harper Lee’s *Go Set A Watchman* and *To Kill A Mockingbird*.

It is safe to assume, faced with the results of the study of *The Waves*, that Virginia Woolf did not count words she used to create the idiolects of her six “voices”; but the final outcome is as if she did with the sole purpose of diversifying her characters’ voices. The most significant added value of this analysis to the existing body of academic criticism on Woolf is in fact nothing less that a heightened appreciation of the writer’s genius: she has been able to produce character styles so distinctive that the differences between them are discernible through quantitative analysis; and that she has diversified them by age and by gender. The nightingale knows not how to read musical notes and how to measure intervals; but its song is brilliant nevertheless.

In a much more down-to-earth material, that of the high school essays, quantitative textual analysis produces results that are interesting for other reasons. Rather than providing an additional method of dealing with a large body of papers to be graded, it seems like a direct indication to students and teachers alike how to write essays; and this sort of investigation fits well into recent psycho- and sociolinguistic approaches exemplified by James Pennebaker’s *The Secret Life of Pronouns* (2011).

One of the oldest tasks of literary studies, classification and systematics of authors, is well served by macroanalysis of the kind presented by the last example in this short survey. History of literature has been trying to put authors in groups and periods and tendencies without the
possibility to read them all; stylometric software reads books differently than humans, but at least it “reads” more of them than any single human – and it can graph the results of this “reading” according to much less impressionistic criteria than those criticized – and perhaps also those adopted – by such specialists on the function and the task of criticism as T.S Eliot or Terry Eagleton.

But it must be understood that computational stylistics has no quarrel with Eliots and Eagletons. It has been all three authors’ experience that, when a computational stylist meets an open-minded traditional literary scholar, the twain come up with a new quality in textual analysis on any of the three scales discussed and exemplified in this paper. Computational Stylistics aims not at replacing Bloom with Burrows; it makes much more sense to bring them together for the benefit of our knowledge of literature, our potential in reading and understanding texts – and our appreciation of both the literary tradition and the individual literary talent.

References


CONTENTS

Acknowledgments ix

PART I. FOUNDATION

1 REVOLUTION 3
2 EVIDENCE 5
3 TRADITION 11
4 MACROANALYSIS 24

PART II. ANALYSIS

5 METADATA 35
6 STYLE 63
7 NATIONALITY 105
8 THEME 118
9 INFLUENCE 154
PART III. PROSPECTS

10 ORPHANS 171

References 177
Index 187
The digital revolution is far more significant than the invention of writing or even of printing.
—Douglas Carl Engelbart

An article in the June 23, 2008, issue of *Wired* declared in its headline “Data Deluge Makes the Scientific Method Obsolete” (Anderson 2008). By 2008 computers, with their capacity for number crunching and processing large-scale data sets, had revolutionized the way that scientific research gets done, so much so that the same article declared an end to theorizing in science. With so much data, we could just run the numbers and reach a conclusion. Now slowly and surely, the same elements that have had such an impact on the sciences are revolutionizing the way that research in the humanities gets done. This emerging field we have come to call “digital humanities”—which was for a good many decades not emerging at all but known as “humanities computing”—has a rich history dating back at least to Father Roberto Busa’s concordance work in the 1940s, if not before.* Only recently, however, has this “discipline,” or “community of practice,” or “field of study/theory/methodology,” and so on, entered into the mainstream discourse of the humanities, and it is even more recently that those who “practice” digital humanities (DH) have begun to grapple with the challenges of big data.† Technology has certainly changed some things about the way literary scholars go about their work, but until recently change has been

---

* Roberto Busa, a Jesuit priest and scholar, is considered by many to be the founding father of humanities computing. He is the author of the *Index Thomisticus*, a lemmatized index of the works of Thomas Aquinas.

† Some have already begun thinking big. In 2008 I served on the inaugural panel reviewing applications for the jointly sponsored National Endowment for the Humanities and National Science Foundation “Digging into Data” grants. The expressed goals of the grant are to promote the development and deployment of innovative research techniques in large-scale data analysis; to foster interdisciplinary collaboration among scholars in
mostly at the level of simple, even anecdotal, search. The humanities computing/digital humanities revolution has now begun, and big data have been a major catalyst. The questions we may now ask were previously inconceivable, and to answer these questions requires a new methodology, a new way of thinking about our object of study.

For whatever reasons, be they practical or theoretical, humanists have tended to resist or avoid computational approaches to the study of literature.* And who could blame them? Until recently, the amount of knowledge that might be gained from a computer-based analysis of a text was generally overwhelmed by the dizzying amount of work involved in preparing (digitizing) and then processing that digital text. Even as digital texts became more readily available, the computational methods for analyzing them remained quite primitive. Word-frequency lists, concordances, and keyword-in-context (KWIC) lists are useful for certain types of analysis, but these staples of the digital humanist’s diet hardly satiate the appetite for more. These tools only scratch the surface in terms of the infinite ways we might read, access, and make meaning of text. Revolutions take time; this one is only just beginning, and it is the existence of digital libraries, of large electronic text collections, that is fomenting the revolution. This was a moment that Rosanne Potter predicted back in the digital dark ages of 1988. In an article titled “Literary Criticism and Literary Computing,” Potter wrote that “until everything has been encoded, or until encoding is a trivial part of the work, the everyday critic will probably not consider computer treatments of texts” (93). Though not “everything” has been digitized, we have reached a tipping point, an event horizon where enough text and literature have been encoded to both allow and, indeed, force us to ask an entirely new set of questions about literature and the literary record.

the humanities, social sciences, computer sciences, information sciences, and other fields around questions of text and data analysis; to promote international collaboration; and to work with data repositories that hold large digital collections to ensure efficient access to these materials for research. See http://www.diggingintodata.org/.

* I suspect that at least a few humanists have been turned off by one or more of the very public failures of computing in the humanities: for example, the Donald Foster Shakespeare kerfuffle.
While still graduate students in the early 1990s, my wife and I invited some friends to share Thanksgiving dinner. One of the friends was, like my wife and me, a graduate student in English. The other, however, was an outsider, a graduate student from geology. The conversation that night ranged over a wine-fueled spectrum of topics, but as three of the four of us were English majors, things eventually came around to literature. There was controversy when we came to discuss the “critical enterprise” and what it means to engage in literary research. The very term research was discussed and debated, with the lone scientist in the group suggesting, asserting, that the “methodology” employed by literary scholars was a rather subjective and highly anecdotal one, one that produced little in terms of “verifiable results” if much in the way of unsupportable speculation.

I recall rising to this challenge, asserting that the literary methodology was in essence no different from the scientific one: I argued that scholars of literature (at least scholars of the idealistic kind that I then saw myself becoming), like their counterparts in the sciences, should and do seek to uncover evidence and discover meaning, perhaps even truth. I dug deeper, arguing that literary scholars employ the same methods of investigation as scientists: we form a hypothesis about a literary work and then engage in a process of gathering evidence to test that hypothesis.

After so many years it is only a slightly embarrassing story. Although I am no longer convinced that the methods employed in literary studies are exactly the same as those employed in the sciences, I remain convinced that there are a good many methods worth sharing and that the similarities of methods exist in concrete ways, not simply as analogous practices.

The goal of science, we hope, is to develop the best possible explanation for some phenomenon. This is done via a careful and exhaustive gathering of evi-
We understand that the conclusions drawn are only as good as the evidence gathered, and we hope that the gathering of evidence is done both ethically and completely. If and when new evidence is discovered, prior conclusions may need to be revised or abandoned—such was the case with the Ptolemaic model of a geocentric universe. Science is flexible in this matter of new evidence and is open to the possibility that new methods of investigation will unearth new, and sometimes contradictory, evidence.

Literary studies should strive for a similar goal, even if we persist in a belief that literary interpretation is a matter of opinion. Frankly, some opinions are better than others: better informed, better derived, or just simply better for being more reasonable, more believable. Science has sought to derive conclusions based on evidence, and in the ideal, science is open to new methodologies. Moreover, to the extent possible, science attempts to be exhaustive in the gathering of the evidence and must therefore welcome new modes of exploration, discovery, and analysis. The same might be said of literary scholars, excepting, of course, that the methods employed for the evidence gathering, for the discovery, are rather different. Literary criticism relies heavily on associations as evidence. Even though the notions of evidence are different, it is reasonable to insist that some associations are better than others.

The study of literature relies upon careful observation, the sustained, concentrated reading of text. This, our primary methodology, is “close reading.” Science has a methodological advantage in the use of experimentation. Experimentation offers a method through which competing observations and conclusions may be tested and ruled out. With a few exceptions, there is no obvious corollary to scientific experimentation in literary studies. The conclusions we reach as literary scholars are rarely “testable” in the way that scientific conclusions are testable. And the conclusions we reach as literary scholars are rarely “repeatable” in the way that scientific experiments are repeatable. We are highly invested in interpretations, and it is very difficult to “rule out” an interpretation. That said, as a way of enriching a reader’s experience of a given text, close reading is obviously fruitful; a scholar’s interpretation of a text may help another reader to “see” or observe in the text elements that might have otherwise remained latent. Even a layman’s interpretations may lead another reader to a more profound, more pleasurable understanding of a text. It would be wasteful and futile to debate the value of interpretation, but interpretation is fueled by observation, and as a method of evidence gathering, observation—both in the sciences and in the humanities—is flawed. Despite all their efforts to repress them, researchers will have irrepressible biases. Even scientists will “interpret” their evidence through a lens of subjectivity. Observation is flawed in the same way that generalization from the specific is flawed: the generalization may be good, it may even explain a total population, but the selection of the sample is always something less than
perfect, and so the observed results are likewise imperfect. In the sciences, a
great deal of time and energy goes into the proper construction of “representa-
tive samples,” but even with good sampling techniques and careful statistical
calculations, there remain problems: outliers, exceptions, and so on. Perfection
in sampling is just not possible.

Today, however, the ubiquity of data, so-called big data, is changing the sam-
pling game. Indeed, big data are fundamentally altering the way that much
science and social science get done. The existence of huge data sets means that
many areas of research are no longer dependent upon controlled, artificial ex-
periments or upon observations derived from data sampling. Instead of con-
ducting controlled experiments on samples and then extrapolating from the
specific to the general or from the close to the distant, these massive data sets
are allowing for investigations at a scale that reaches or approaches a point of
being comprehensive. The once inaccessible “population” has become accessible
and is fast replacing the random and representative sample.

In literary studies, we have the equivalent of this big data in the form of big li-
braries. These massive digital-text collections—from vendors such as Chadwyck-
Healey, from grassroots organizations such as Project Gutenberg, from nonprofit
groups such as the Internet Archive and HathiTrust, and from the elephants in
Mountain View, California, and Seattle, Washington*—are changing how literary
studies get done. Science has welcomed big data and scaled its methods accord-
ingly. With a huge amount of digital-textual data, we must do the same. Close
reading is not only impractical as a means of evidence gathering in the digital
library, but big data render it totally inappropriate as a method of studying liter-
ary history. This is not to imply that scholars have been wholly unsuccessful in
employing close reading to the study of literary history. A careful reader, such as
Ian Watt, argues that elements leading to the rise of the novel could be detected
and teased out of the writings of Defoe, Richardson, and Fielding. Watt’s study is
magnificent; his many observations are reasonable, and there is soundness about
them.† He appears correct on a number of points, but he has observed only a
small space. What are we to do with the other three to five thousand works of

* That is, Google.com and Amazon.com.
† A similar statement could be made of Erich Auerbach’s Mimesis. It is a magnificent
bit of close reading. At the same time, Auerbach was acutely aware of the limitations of
his methodology. In the epilogue to Mimesis, he notes the difficulties of dealing with
“texts ranging over three thousand years” and how the limitations of his library in Istan-
bul made it “probable that [he] overlooked things which [he] ought to have considered.”
Interestingly, however, he says at the same time that “it is quite possible that the book
owes its existence to just this lack of a rich . . . library.” If it had been possible to access
the greater archive, he “might never have reached the point of writing” (1953).
fiction published in the eighteenth century? What of the works that Watt did not observe and account for with his methodology, and how are we to now account for the works not penned by Defoe, by Richardson, or by Fielding? Might other novelists tell a different story? Can we, in good conscience, even believe that Defoe, Richardson, and Fielding are representative writers? Watt’s sampling was not random; it was quite the opposite. But perhaps we only need to believe that these three (male) authors are representative of the trend toward “realism” that flourished in the nineteenth century. Accepting this premise makes Watt’s magnificent synthesis into no more than a self-fulfilling project, a project in which the books are stacked in advance. No matter what we think of the sample, we must question whether in fact realism really did flourish. Even before that, we really ought to define what it means “to flourish” in the first place. Flourishing certainly seems to be the sort of thing that could, and ought, to be measured. Watt had no such yardstick against which to make a measurement. He had only a few hundred texts that he had read. Today, things are different. The larger literary record can no longer be ignored: it is here, and much of it is now accessible.

At the time of my Thanksgiving dinner back in the 1990s, gathering literary evidence meant reading books, noting “things” (a phallic symbol here, a biblical reference there, a stylistic flourish, an allusion, and so on) and then interpreting: making sense and arguments out of those observations.* Today, in the age of digital libraries and large-scale book-digitization projects, the nature of the “evidence” available to us has changed, radically. Which is not to say that we should no longer read books looking for, or noting, random “things,” but rather to emphasize that massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making. The literary scholar of the twenty-first century can no longer be content with anecdotal evidence, with random “things” gathered from a few, even “representative,” texts.† We must strive to understand these things we find interesting in the context of everything else, including a mass of possibly “uninteresting” texts.

* Yes, a simplification, but close enough to serve as a heady foil in this introductory polemic. Along similar lines, Susan Hockey writes of the “somewhat serendipitous noting of interesting features” (2000, 66).

† When writing of “anecdotal” here, I am not thinking of the use made of anecdote in the new historical tradition that we find expressed in, for example, Greenblatt’s “cultural poetics.” Rather, I use the word in the sense of “anecdotal evidence”: that is, evidence that is atypical, informally gathered, speculative, or purely interpretive, which is to say not empirical. On this point, the type of literary data I am exploring allows me to adopt a fundamentally empirical position. Having said that, there is a place for anecdotal evidence in literary study, and I do not intend here a critique of anecdotalism per se, but rather to simply make a distinction and separation between two types of evidence.
“Strictly speaking,” wrote Russian formalist Juri Tynjanov in 1927, “one cannot study literary phenomena outside of their interrelationships” (1978, 71). Unfortunately for Tynjanov, the multitude of interrelationships far exceeded his ability to study them, especially with close and careful reading as his primary tools. Like it or not, today’s literary-historical scholar can no longer risk being just a close reader: the sheer quantity of available data makes the traditional practice of close reading untenable as an exhaustive or definitive method of evidence gathering. Something important will inevitably be missed. The same argument, however, may be leveled against the macroscale; from thirty thousand feet, something important will inevitably be missed. The two scales of analysis, therefore, should and need to coexist. For this to happen, the literary researcher must embrace new, and largely computational, ways of gathering evidence. Just as we would not expect an economist to generate sound theories about the economy by studying a few consumers or a few businesses, literary scholars cannot be content to read literary history from a canon of a few authors or even several hundred texts. Today’s student of literature must be adept at reading and gathering evidence from individual texts and equally adept at accessing and mining digital-text repositories. And mining here really is the key word in context. Literary scholars must learn to go beyond search. In search we go after a single nugget, carefully panning in the river of prose. At the risk of giving offense to the environmentalists, what is needed now is the literary equivalent of open-pit mining or hydraulicking. We are proficient at electronic search and comfortable searching digital collections for some piece of evidence to support an argument, but the sheer amount of data now available makes search ineffectual as a means of evidence gathering. Close reading, digital searching, will continue to reveal nuggets, while the deeper veins lie buried beneath the mass of gravel layered above. What are required are methods for aggregating and making sense out of both the nuggets and the tailings. Take the case of a scholar conducting research for a hypothetical paper about Melville’s metaphysics. A query for whale in the Google Books library produces 33,338 hits—way too broad. Narrowing the search by entering whale and god results in a more manageable 3,715 hits, including such promising titles as American Literature in Context and Melville’s Quarrel with God. Even if the scholar could further narrow the list to 1,000 books, this is still far too many to read in any practical way. Unless one knows what to look for—say, a quotation only partially remembered—searching for research purposes, as a means of evidence gathering, is not terribly practical.* More interesting, more exciting, than panning for nuggets in digital archives

* In revising this section before publication, I went back to Google Books and discovered that the number of hits for this particular search had grown significantly since I first tested. No doubt readers will find even higher numbers today.
is the ability to go beyond the pan and exploit the trommel of computation to process, condense, deform, and analyze the deeper strata from which these nuggets were born, to unearth, for the first time, what these corpora really contain. In practical terms, this means that we must evolve to embrace new approaches and new methodologies designed for accessing and leveraging the electronic texts that make up the twenty-first-century digital library.

This is a book about evidence gathering. It is a book about how new methods of analysis allow us to extract new forms of evidence from the digital library. Nevertheless, this is also a book about literature. What matter the methods, so long as the results of employing them lead us to a deeper knowledge of our subject? A methodology is important and useful if it opens new doorways of discovery, if it teaches us something new about literary history, about individual creativity, and about the seeming inevitability of influence.
3 TRADITION

Talents imitate, geniuses steal.
—[Oscar Wilde?]  

As noted previously, there is a significant tradition of researchers employing computational approaches to the study of literature and an even longer tradition of scholars employing quantitative and statistical methods for the analysis of text. The specifically computational tradition dates back to the work of Father Roberto Busa, and since that time momentum has been building, exponentially, so that now, somewhat suddenly, the trend line has rocketed upward and the “digital humanities” have burst upon the scene and become a ubiquitous topic of discussion in humanities programs across the globe.* Notwithstanding the fact that there is no general agreement as to what exactly the term digital humanities defines, the sudden popularity of this thing called digital humanities has occurred with such rapidity that even we who consider ourselves natives of the tribe have been taken by surprise. Some have suggested that the reason stock in digital humanities is skyrocketing is because literary studies are in a general state of crisis and that we are yearning for a new theoretical construct that would ground our inquiries in science (see, for example, Gottschall 2008). This may be the case, for some, but I am not a member of that club. As someone who has studied diasporas, I understand that there can be pushes and pulls to any migration. For the Irish, British oppression made for an imposing stick and the promise of opportunity in America an enticing carrot. Here, however, the migration to digital humanities appears to be mostly about opportunity. In fact, the sudden motivation for scholars to engage in digital humanities is more than likely a direct by-product of having such a wealth of digital material with which to engage. With apologies to the indigenous, I must acknowledge here

* The tradition may stretch even further if we broaden our definition of computation to include substrates beyond silicon.
that the streets of this “new” world are paved with gold and the colonizers have arrived. A large part of this change in scholarly thinking about the digital has been brought about because of the very simple fact that digital objects, digital data stores, and digital libraries in particular have become both large and easily accessible. We have built it, and they are coming. Despite the success of this “thing called digital humanities,” as William Deresiewicz derided it in 2008, there remains no general agreement or even general understanding of what the term means or describes. Some, including Matthew Kirschenbaum (2010), think that this ambiguity is a good thing. I am not as certain. Do video-game analysis and stylometry really make good bedfellows? Probably not; these are entirely different threads.* Understanding how we got to this point of free-loving digital humanities is useful not simply as a matter of disciplinary history but as a way of contextualizing and understanding the methods and results presented in this book. So, a few words are in order about the traditions informing my macroanalytic approach to digital literary studies.

In 2012 we stand upon the shoulders of giants, and the view from the top is breathtaking. The skies were not always this clear. Susan Hockey summarized the period of the 1980s as one in which “we were still at a stage where academic respectability for computer-based work in the humanities was questionable” (2004, 10). Mark Olsen noted in 1993 that despite advances in text processing, “computerized textual research has not had a significant influence on research in the humanistic disciplines” (309). A decade later, Thomas Rommel argued that “the majority of literary critics still seem reluctant to embrace electronic media as a means of scholarly analysis . . . [and] literary computing has, right from the very beginning, never really made an impact on mainstream scholarship” (2004, 92). Stephen Ramsay wrote in 2007, “The digital revolution, for all its wonders, has not penetrated the core activity of literary studies, which, despite numerous revolutions of a more epistemological nature, remains mostly concerned with the interpretive analysis of written cultural artifacts. Texts are browsed, searched, and disseminated by all but the most hardened Luddites in literary study, but seldom are they transformed algorithmically as a means of gaining entry to the deliberately and self-consciously subjective act of critical interpretation” (478).

* Just so it is clear, I am a big fan of the “big tent,” or the “big umbrella,” if you will. In 2011 Glen Worthey and I cohosted the annual Digital Humanities Conference at Stanford, where our conference theme was “Big Tent Digital Humanities.” In the spirit of the Summer of Love, we donned tie-dyed shirts and let a thousand DH flowers bloom. We love our DH colleagues one and all. This book, however, stands at one side of the tent. We do different things in DH; we are vast.
Others from outside the scholarly community of computing humanists, writers such as Sven Birkerts (1994) and Nicholson Baker (2001), have warned of the dangers inherent in the digitization of books, and Emory English professor Mark Bauerlein has offered a sustained, if unspecific, critique of the digital age in general (2008). Even as recently as 2008, the ever-adversarial William Deresiewicz wrote in the *Nation* about the digital humanities, poking fun at something he imagined to be just another fad of scholarship.* But things change.

Despite some early concerns and several contemporary detractors, today—some few years after the most recent lamentations—the scholarly presses and the mainstream media are buzzing with news of this thing called “digital humanities.”† Humanities computing, or, more popularly, “digital humanities,” is alive and well. The field is healthy: participation in the primary professional organization, the Alliance of Digital Humanities Organizations (ADHO), is vibrant, and attendance at the annual Digital Humanities Conference is at an all-time high.‡ So large have we grown, in fact, that the number of rejected papers now far exceeds the number accepted, and many of the panels and papers that are not rejected draw standing-room crowds and lively discussion. Meanwhile, new degree programs specifically geared toward digital humanities are now offered at universities across the globe.§ Academic jobs for candidates with expertise in the intersection between the humanities and technology are becoming more and


† Matthew Kirschenbaum provides a succinct, six-page overview of the field in his *ADE Bulletin* article titled “What Is Digital Humanities and What’s It Doing in English Departments?” (2010). Other examples include Fischman 2008a, 2008b; Goodall 2008; Howard 2008a, 2008b, 2008c; Pannapacker 2011; Parry 2010; Shea 2008; and Young 2009.

‡ The Alliance of Digital Humanities Organizations is a consortium including the Association for Computing and the Humanities, the Association of Literary and Linguistic Computing, the Society for Digital Humanities, and CenterNet.

§ In terms of numbers of institutions per capita and dollars per capita, Canada is the obvious front runner here, but several universities in the UK, Ireland, and the United States have recently begun programs or “tracks” in digital humanities. Stanford began offering an undergraduate emphasis in “digital humanities” through its Interdisciplinary Studies in the Humanities Program back in 2006. In October 2006, Kings College of London announced a Ph.D. in digital humanities. In 2010 the National University of Ireland, Maynooth, began offering a master’s of arts in digital humanities (http://www.learndigitalhumanities.ie/), and University College London began offering a master of arts and science in digital humanities (http://www.ucl.ac.uk/dh-blog/2010/07/30/announcing-the-new-mamscc-in-digital-humanities-at-ucl/). In 2011 Trinity College Dublin began a master of philosophy program in digital humanities under the direction of Susan Schreibman.
more common, and a younger constituent of digital natives is quickly overtaking the aging elders of the tribe.* By one measure, the number of young scholars and graduate students attending the annual digital humanities conference in 2009 was three times the number of those attending one year earlier.† To my 2006 query to the members of the Humanist List about the health of the field, I received a number of encouraging replies that included remarks about the recent “groundswell of research interest” in digitally oriented projects, the development of new “centers” for computing in the humanities, and institutional support for the hiring of computing humanists.‡ Especially impressive has been the news from Canada. Almost all of the “G 10” (that is, the top thirteen research institutions of Canada) have institutionalized digital humanities activities in the form of degrees such as Alberta’s master’s in digital humanities, programs such as McMaster’s in digital media, centers such as the University of Victoria’s Humanities Computing Centre, or through institutes such as Victoria’s Digital Humanities Summer Institute. Noteworthy too is that the prestigious Canada Research Chair has been appointed to a number of computing humanists.§ Not the least important, the program for the 2011 Modern Language Association conference in Seattle included, by one scholar’s count, at least fifty-seven panels in the “digital humanities,” up from forty-four the previous year when the panel session titled “The History and Future of the Digital Humanities” had standing-room crowds (Pannapacker 2011).¶ All signs indicate that the digital

* A search, conducted in October 2006, of jobs listed in the Chronicle of Higher Education including both the words digital and humanities resulted in thirty-four hits. Recent searches have contained even more, including opportunities in senior-level posts such as that advertised in July 2010 for a director of Texas A&M’s new Digital Humanities Institute. On September 25, 2011, Desmond Schmidt posted the following summary of digital humanities jobs on the Humanist List: “There have been a lot of advertisements for jobs lately on Humanist. So I used the Humanist archive to do a survey of the last 10 years. I counted jobs that had both a digital and a humanities component, were full time, lasted at least 12 months and were at PostDoc level or higher. 2002: 11, 2003: 6, 2004: 15, 2005: 15, 2006: 18, 2007: 24, 2008: 27 (incomplete - 1/2 year), 2009: 36, 2010: 58, 2011: 65 so far.”

† In 2009 I was chair of the ADHO Bursary Awards committee. The prize is designed to encourage new scholars in the discipline. From 2008 to 2009, the number of candidates for the Bursary Award jumped from seven to more than thirty.

‡ Humanist, now in its twenty-second year of operation, is, by general consensus, the Listserv of record for all matters related to computing in the humanities.

§ See http://tapor.ualberta.ca/taporwiki/index.php/Canada_Research_Chairs_and_Award_Winners.

Trad ition

humanities have arrived, even while the fields of study sheltering beneath the
umbrella remain a somewhat ambiguous and amorphous amalgamation of lit-
erary formalists, new media theorists, tool builders, coders, and linguists.

Computational text analysis—by all accounts the foundation of digital hu-
manities and its deepest root—has come a long way since 1949, when Father
Roberto Busa began creation of his word index. These days, humanists routinely
create word indexes and frequency lists using readily available software. With
the spread of broadband and the accessibility of the Internet, many tools that
were once platform dependent and command line in nature have been “rein-
vented” for the web so that scholars may now do small-scale text processing
and analysis on remote web servers using any number of web-based applica-
tions. Keyword-in-context lists can be quickly generated using TactWeb.* Stéfan
Sinclair’s HyperPo and Voyant offer self-serve text-analysis tools for traditional
concurring and co-occurrence alongside more experimental widgets for the
processing and deforming of textual data.† There is a growing number of tools
specifically geared toward the “visualization” of literary materials.‡ A particu-
larly well-conceived, low-entry project is the “Text Analysis Portal” (TAPoR),
which has set itself up as a one-stop shop for basic text analysis. This project,
which began life with a six-million-dollar (CAD) grant from the Canadian
Foundation for Innovation, is distributed across six universities and provides
a centralized and, to some extent, standardized way of accessing a variety of
text-analysis applications. TAPoR serves as a model of collaboration and offers
a foundational, even seminal, approach to future humanities computing work.
Indeed, some in the United States are now attempting to go beyond TAPoR and
develop what Chris Mackey, formerly of the Mellon Foundation, once referred
to as the “mother of all text-analysis applications.”§ These projects, whose names
include “Bamboo” and others with such funky acronyms as MONK, SEASR,
and DARIAH, are all seeking ways to make leveraging computation as easy for
the average literary scholar as finding biblical references in a canonical novel.¶

‡ See, for example, Bradford Paley’s TextArc application (http://www.textarc.org/) or
the word clouds available through Wordle or the Many Eyes project of IBM.
§ The comment was made during a presentation at the Stanford Humanities Center.
The project that eventually emerged from these and other discussions is Project Bam-
¶ MONK stands for “Metadata Offers New Knowledge” (http://www.monkproject
.org/), SEASR for Software Environment for the Advancement of Scholarly Research
(http://seasr.org/), and DARIAH for Digital Research Infrastructure for the Arts
and Humanities (http://www.dariah.eu/). See also Project Bamboo at http://www
.projectbamboo.org.
Computing humanists have made important contributions to humanities scholarship: thanks to them, we have impressive digital archives and critical editions such as the exemplary Women Writers Project of Brown University and Kevin Kerinan’s impressive Electronic Beowulf.* Fellow travelers from linguistics, machine learning, natural language processing, and computer science have developed robust text-analysis programs that can be employed to automatically identify parts of speech, named entities (people, places, and organizations), prominent themes, sentiment, and even poetic meter.† These tools have in turn been deployed for studies in authorship attribution, textual dating, and stylistic analysis.

There are any number of other useful products that have evolved out of collaborations among humanists, linguists, and technologists: the Google search engine performs a type of text analysis when searching for keywords and collocates; using calculations based on vocabulary, sentence length, and syllables, Microsoft Word attempts to determine the grade level of a piece of writing.‡ The XML (extensible markup language) standard that plays such a critical role in data interchange today was heavily influenced by the early work of the Text Encoding Initiative and in particular founding TEI editor Michael Sperberg-McQueen. These have been important and useful contributions, to be sure, and the recent Blackwell publications *A Companion to Digital Humanities* (Schreibman, Siemens, and Unsworth 2004) and *A Companion to Digital Literary Studies* (Siemens and Schreibman 2007) are a testament to the various ways in which technology has established itself in the humanities.

Despite all of this achievement and the overwhelming sense of enthusiasm and collegiality that permeates the DH community, there is much more work to be done. We have in fact only begun to scratch the surface of what is possible. Though the term *digital humanities* has become as omnipresent on our campuses as *multiculturalism* was several years ago, the adoption of “digital” tools and methodologies has been limited, even among those who would self-identify as “digital humanists.” To be sure, literary scholars have taken advantage of digitized textual material, but this use has been primarily in the arena of search, retrieval, and access. We have not yet seen the scaling of our scholarly questions in accordance with the massive scaling of digital content that is now

---

* http://www.wwp.brown.edu/ and http://ebeowulf.uky.edu/.
† Examples include the Stanford Natural Language Processing Group’s Part of Speech Tagger and Named Entity Recognizer, the University of Massachusetts’s Machine Learning for Language Toolkit (MALLET), and many others.
‡ For more on this, just open Microsoft Word’s “Help” and search for “Readability Scores.” MS Word uses both the Flesch Reading Ease score and the Flesch-Kinkaid Grade Level score.
held in twenty-first-century digital libraries. In this Google Books era, we can take for granted that some digital version of the text we need will be available somewhere online, but we have not yet fully articulated or explored the ways in which these massive corpora offer new avenues for research and new ways of thinking about our literary subject.*

To some extent, our thus-far limited use of digital content is a result of a disciplinary habit of thinking small: the traditionally minded scholar recognizes value in digital texts because they are individually searchable, but this same scholar, as a result of a traditional training, often fails to recognize the potentials for analysis that an electronic processing of texts enables. For others, the limitation is more directly technical and relates to the type and availability of software tools that might be deployed in analysis. The range of what existing computer-based tools have provided for the literary scholar is limited, and these tools have tended to conform to a disciplinary habit of closely studying individual texts: that is, close reading. Such tools are designed with the analysis of single texts in mind and do not offer the typical literary scholar much beyond advanced searching capabilities. Arguably, the existing tools have been a determiner in shaping perceptions about what can and cannot be done with digital texts.† The existing tools have kept our focus firmly on the close reading of individual texts and have undoubtedly prevented some scholars from wandering into the realms of what Franco Moretti has termed “distant reading” (2000). Combine a traditional literary training focused on close reading with the most common text-analysis tools focused on the same thing, and what you end up with is enhanced search—electronic finding aids that replicate and expedite human effort but bring little to the table in terms of new knowledge. I do not intend to demean the use of text-analysis tools at the scale of the single text or at the scale of several texts; quite the contrary, there is an incredibly large body of quantitative work in authorship attribution, gender identification, and what is

* My comments here may seem idealistic given the realities of copyright law and contemporary literature in particular. That digital versions of these recent works exist seems a point we can take for granted; that they are or will be readily accessible is a more complicated problem about which I have more to say in chapter 10.

† Duke University historian of science Tim Lenoir has made a similar point in arguing that quarks would not exist were it not for the particle accelerators that were built to discover or produce them. Lenoir has made this comment on multiple occasions, primarily in lectures on pragmatic realism and social construction. He has written about this extensively in his book Instituting Science (1997), particularly the chapter on Haber-Bosch, in which he discusses this issue at length. He derived this line of thinking in part from Ian Hacking’s argument in Representing and Intervening, in which Hacking argues that electrons are real when you can spray them (1983, 23).
more generally referred to as “stylometry” that informs my own work. And even in the less statistically driven realms of computational text analysis, there are tools for visualizing and exploring individual texts that serve as rich platforms for “play,” as Stéfan Sinclair has termed it (2003), or what might more formally be termed “discovery” and “exploration.” Steven Ramsay’s “Algorithmic Criticism” (2007) provides a strong statement regarding the value of text-analysis tools for text “deformation.” Such deformations may lead to new and different interpretations and interpretive strategies.*

Our colleagues in linguistics have long understood the value of working with large corpora and have compiled such valuable resources as the British National Corpus and the Standard Corpus of Everyday English Usage. Linguists employ these resources in order to better understand how language is used, is changing, is evolving. The tools employed for this work are not, generally speaking, web-based widgets or text-analysis portals such as the TAPoR project. Instead, our colleagues in linguistics have learned to be comfortable on the command line using programming languages. They have learned to develop applications that run on servers, and they have developed a willingness to wait for their results. Literary scholars, on the other hand, have generally been content to rely upon the web for access to digital material. Even in the text-analysis community, there is a decided bias in favor of developing web-based tools.† Unfortunately, the web is not yet a great platform upon which to build or deliver tools for doing text analysis “at scale.” Quick queries of indexed content, yes, but not corpus ingestion or complex analysis.‡

Given the training literary scholars receive, their typical skill set, and the challenges associated with large-scale digitalization and computational analysis, it is easy to understand why literary scholars have not asked and probed with computers the same sorts of questions about “literary language” that linguists

* Ramsay’s original article has now been extended into a book-length study. See Ramsay 2011.
† Stéfan Sinclair of McGill University is an accomplished text-analysis tool builder, and his recent offering, Voyant, is the best example I have seen of an online tool that can handle a large amount of text. See http://voyant-tools.org/. Even this exceptional tool is still only capable of fairly basic levels of analysis.
‡ Cloud computing and high-performance computing are certainly beginning to change things, and projects such as SEASR may someday provide the web interface to high-performance text analysis. At least in the near term, the success of web-based macroanalysis will depend in large part upon the users of such tools. They will need to abandon the idea that clicking a link returns an immediate result. The web may become a portal into a complex text-analysis platform, but the web is not likely to evolve as a place for instant access to complex data.
have asked about language in general. On the one hand, literary scholars have not had access, until recently, to large amounts of digital literary content, and, on the other, there is a long-standing disciplinary habit of thinking about literature in a limited way: in terms of “close readings.” Close reading is a methodological approach that can be applied to individual texts or even small subsets of texts but not, for example, to all British fiction of the nineteenth century. A “close reading” of nineteenth-century British fiction would, in fact, be implausible. Consider, for example, the very real limitations of human reading: Franco Moretti has estimated that of the twenty to thirty thousand English novels published in Britain in the nineteenth century, approximately six thousand are now extant. Assuming that a dedicated scholar could find these novels and read one per day, it would take sixteen and a half years of close reading to get through them all. As a rule, literary scholars are great synthesizers of information, but synthesis here is inconceivable.* A computer-based analysis or synthesis of these same materials is not so difficult to imagine. Though the computer cannot perfectly replicate human synthesis and intuition, it can take us a long way down this road and certainly quite a bit further along than what the human mind can process. It is exactly this kind of macroanalytic approach that is the future of computing in the humanities, and, according to some, the future of literary studies (see, for example, Gottschall 2008 and Martindale 1990).

I am not the first, however, to suggest that a bird’s-eye view of literature might prove fruitful. On this point, Franco Moretti has been at the forefront, suggesting “distant reading” as an alternative to “close reading.” In *Graphs, Maps, Trees*, Moretti writes of how a study of national bibliographies made him realize “what a minimal fraction of the literary field we all work on: a canon of two hundred novels, for instance, sounds very large for nineteenth-century Britain (and is much larger than the current one), but is still less than one per cent of the novels that were actually published: twenty thousand, thirty, more, no one really knows—and close reading won’t help here, a novel a day every day of the year would take a century or so” (2005, 3–4). Moretti’s “Graphs” chapter is particularly compelling; it provides a beginning point for the development

* In history and in historical economics, there is a recent tradition of thinking big. The Annales school of historiography developed by the French in the early twentieth century has had the goal of applying quantitative and social-scientific methods in order to study history of the “long-term,” the longue durée. The approach views history in terms of “systems.” Lynn Hunt’s brief and useful overview of the history of the Annales paradigm argues that “in contrast to earlier forms of historical analysis [namely, exemplar and developmental approaches], the Annales school emphasized serial, functional, and structural approaches to understanding society as a total, inter-related organism” (1986, 211).
of a more formal literary time-series analysis methodology. Moretti examines the publication rates for novels (in several countries) over periods of years and decades. Focusing on the peaks and valleys in novel production, he moves from the quantitative facts to speculation and interpretation, posing, for example, that the rise and fall of various novelistic genres in the British corpus can be correlated to twenty-five- to thirty-year cycles or generations of readers. In Moretti’s model, the tastes and preferences of one generation are inevitably replaced by those of the next. He suggests that there are connections between literary cycles and political ones, arguing, for example, that the French Revolution was a critical factor in the fall of the French novel. Although such an argument could certainly be made anecdotally, the accompanying data—and the graph showing the sharp decline in novel production in about 1798—leave little room for debate.

Nor am I original in considering the applications of technology to large textual collections. Already noted are the linguists, and there is, of course, an entire community of computer scientists (many of them at Google) who work in the field of text mining and information retrieval. Along with similar agencies in other nations, the National Security Agency is in this business as well: the NSA is reported to have been employing text-mining technologies since the Cold War, and the “classified” ECHELON surveillance system is purported to capture all manner of electronic information, from satellite communications to email correspondences. These captured materials are then analyzed, mined by machines, in order to sniff out threats to national security. The amount of information devoted to ECHELON online is somewhat staggering—a Google search for this supersecret program along with the keywords text and mining provides 375,000 sites of interest. This figure is trivial next to the Google results for a search for the keyword Area 51 (154 million hits) but does demonstrate the point that text mining, and ECHELON for that matter, is nothing new. Similar to ECHELON is the technology developed by Palantir Technologies in Palo Alto, California. The company’s website describes their software as being “a platform for information analysis . . . designed for environments where the fragments of data that . . . tell a larger story are spread across a vast set of starting material” (Palantir Technologies 2011). Translation: we build technologies for the macroanalysis of large, disparate corpora.

Not quite as spectacular as Palantir and the NSA are projects more specifically aimed at the application of text mining to the humanities. The NORA, MONK, and SEASR projects originally led by John Unsworth at the University of Illinois are three such projects. The expressed goal of the NORA project was to “produce software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries” (NORA 2006). Using software developed by the University of Illinois's
National Center for Supercomputing Applications and the “Data to Knowledge” applications of Michael Welge’s Automated Learning Group, the NORA team successfully deployed a Java-based application for “sniffing” out preidentified “patterns” in large digital collections. An early version of the software allowed an end user to “tag” or “mark” certain works in a collection, and the system then used those works to build a model—what some biologists who work with DNA and gene expression call a “signal.” This signal is then sought throughout the larger collection. The example offered on the NORA website involves marking “erotic” passages in the works of Emily Dickinson. Some 260 individual documents are presented, and the user “marks” or rates a small percentage of these for erotic content.* The human-marked documents constitute a training set, which is used by the software to “predict” which works in the collection are likely to contain erotic content as well. This is essentially an information-retrieval task. MONK and SEASR are more advanced implementations of the NORA technologies. SEASR provides the most deeply abstracted and robust imagining of the early NORA work. SEASR is both a back-end infrastructure and a semifriendly web interface that allows researchers to build text-analysis “flows” that get executed on a server.†

Outside of the humanities, computer scientists working in natural language processing, corpus linguistics, and computational linguistics have developed a wide range of tools that have direct application to work in literary studies. Using a technique called “topic modeling,” a group led by David Newman at the University of California–Irvine (UCI) harvested the latent themes, or topics, contained in 330,000 stories published in the New York Times. The topic-modeling procedure they employed required no human preprocessing; it was “unsupervised” in its sifting through a corpus of documents and then identifying patterns of words that were frequently collocated.‡ The software categorizes the words in each document into mathematically correlated clusters, which are described as “topics.” Not surprisingly, the UCI team first presented their research at the Intelligence and Security Informatics conference in San Diego (Newman, Smyth, and Steyvers 2006). More interesting (for scholars of literature) than the

* This process of human intervention is known in data and text mining as “supervised learning.”

† From 2011 to 2012, I served as the project lead on “Phase Two” of the SEASR project. The work was generously funded by the Mellon Foundation.

‡ Andrew McCallum and his team at the University of Massachusetts have done exciting work developing a “Machine Learning for Language Toolkit,” or “MALLET,” which provides functionality for a variety of text-mining applications. The MALLET software includes David Mimno’s topic-modeling code, which is used and described at length in chapter 8.
intelligence applications of topic modeling are the applications to humanities research. Historian Sharon Block, for example, teamed up with Newman and employed topic-modeling routines to explore the entire eighteenth-century run of the *Pennsylvania Gazette*. In her essay “Doing More with Digitization: An Introduction to Topic Modeling of Early American Sources” (2006), Block walks readers through a series of examples of how the technique can assist historians and reveal new avenues for research in the form of unanticipated patterns and trends.* Though not designed with literary scholarship in mind, the topic-modeling tools can be applied to literary texts at the level of the corpus or even at the level of the individual book or poem.†

Still another project working to apply the tools and techniques of text mining and corpus linguistics to literature is the WordHoard project at Northwestern University. Ironically, the WordHoard site describes its software as “an application for the close reading and scholarly analysis of deeply tagged texts” but then goes on to say that it “applies to highly canonical literary texts the insights and techniques of corpus linguistics, that is to say, the empirical and computer-assisted study of large bodies of written texts or transcribed speech” (WordHoard 2006). The descriptive prose that follows adds that the software allows for a deeply “microscopic” and philological inquiry of the text(s). Although it is true that WordHoard provides access to, or tools for, harvesting richly encoded texts, the results being gleaned from the texts are not so much the results of a close reading–like process as they are the results of a macroscopic text-mining process that aggregates a number of relatively small details into a more global perspective. As such, the process seems to have less in common with close-reading practices and more with Moretti’s notion of distant reading. The devil is in the details and in how the details are investigated and aggregated in order to enable a larger perspective. Writing of “detailism” and digital texts, Julia Flanders discusses Randolph Starn’s introduction to a special issue of...

* Cameron Blevins provides another historical example. Blevins uses topic modeling to explore entries in Martha Ballard’s eighteenth-century diary. See http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/.
† David Newman was the first guest speaker in the Beyond Search workshop that I ran at Stanford from 2006 to 2009. Prior to his arrival, I prepared a corpus of texts for Newman to process. Included in those data were the novels of Jane Austen. As part of his presentation, Newman showed how the topic of “sentiment” (composed of words denoting emotion) could be tracked throughout the Austen corpus. Looking at the graphs that he prepared, participants in the workshop could see how Austen employs moments of strong emotion throughout her texts. In some novels, we observed a regular fluctuation, while others showed a steady trend upward: as the novels progressed, the presence of strong emotions increased.
Trad ition

Representations. She notes the effort to connect “detail . . . with a larger historical view.” She goes on to emphasize that detail is used “not as ‘mere facts’ cited as evidence . . . but as the contextually embedded ‘trace, clue, sign, shard’ that carries a specifiable, signifying linkage to some historical genealogy” to some larger system (2005, 43). WordHoard offers a way of aggregating these signs into a coherent argument. The website offers the word love—as it appears in the works of Chaucer, Spenser, and Shakespeare—as an example. Female characters, the data reveal, are “about 50% more likely to speak of love than men.” This conclusion is derived not through a computer-based close reading of the texts, but rather via a quantitative zooming out and away from the texts, a zooming out that allows the user to simultaneously “see” all of the separate occurrences of the word throughout the corpus. The end result is that the WordHoard tool takes us quite far away from the actual occurrences of the words in the texts; our attention is drawn to an examination of the bigger picture, the macroview of love when used as a noun, of love when used as a verb, and in both cases of love as it is used by male or female speakers. This is not close reading; this is macroanalysis, and the strength of the approach is that it allows for both zooming in and zooming out.*

* A relatively recent entry into this realm of micro-macro-oriented text-analysis tools is Aditi Muralidharan’s WordSeer (http://wordseer.berkeley.edu). Australian digital humanist Tim Sherratt offers another variety of similar tools via his “WraggeLabs Emporium” (http://wraggelabs.com/emporium).
4 MACROANALYSIS

Keynes was a great economist. In every discipline, progress comes from people who make hypotheses, most of which turn out to be wrong, but all of which ultimately point to the right answer. Now Keynes, in The General Theory of Employment, Interest and Money, set forth a hypothesis which was a beautiful one, and it really altered the shape of economics. But it turned out that it was a wrong hypothesis.

—Milton Friedman, Opinion Journal, July 22, 2006

The approach to the study of literature that I am calling “macroanalysis” is in some general ways akin to economics or, more specifically, to macroeconomics. Before the 1930s, before Keynes’s General Theory of Government, Interest, and Money in 1936, there was no defined field of “macroeconomics.” There was, however, neoclassical economics, or “microeconomics,” which studies the economic behavior of individual consumers and individual businesses. As such, microeconomics can be seen as analogous to our study of individual texts via “close readings.” Macroeconomics, however, is about the study of the entire economy. It tends toward enumeration and quantification and is in this sense similar to bibliographic studies, biographical studies, literary history, philology, and the enumerative, quantitative analysis of text that is the foundation of computing in the humanities. Thinking about macroanalysis in this context, one can see the obvious crossover with WordHoard. Although there is sustained interest in the micro level, individual occurrences of some feature or word, these individual occurrences (of love, for example) are either temporarily or permanently de-emphasized in favor of a focus on the larger system: the overall frequencies of love as a noun versus love as a verb. Indeed, the very object of analysis shifts from looking at the individual occurrences of a feature in context to looking at the trends and patterns of that feature aggregated over an entire corpus. It is here that one makes the move from a study of words in the context of sentences
or paragraphs to a study of aggregated word “data” or derivative “information” about word behavior at the scale of an entire corpus.

By way of analogy, we might think about interpretive close readings as corresponding to microeconomics, whereas quantitative distant reading corresponds to macroeconomics. Consider, then, the study of literary genres or literary periods: are they macroanalytic? Say, for example, a scholar specializes in early-twentieth-century poetry. Presumably, this scholar could be called upon to provide sound generalizations, or “macroreadings,” of twentieth-century poetry based on a broad familiarity with the individual works of that period. This would be a type of “macro” or “distant” reading.* But this kind of macro-reading falls short of approximating for literature what macroeconomics is to economics, and it is in this context that I prefer the term analysis over reading. The former term, especially when prefixed with macro, places the emphasis on the systematic examination of data, on the quantifiable methodology. It de-emphasizes the more interpretive act of “reading.” This is no longer reading that we are talking about—even if programmers have come to use the term read as a way of naming functions that load a text file into computer memory. Broad attempts to generalize about a period or about a genre by reading and synthesizing a series of texts are just another sort of microanalysis. This is simply close reading, selective sampling, of multiple “cases”; individual texts are digested, and then generalizations are drawn. It remains a largely qualitative approach.† Macroeconomics is a numbers-driven discipline grounded in quantitative analysis, not qualitative assessments. Macroeconomics employs quantitative benchmarks.

* Ian Watt’s impressive study The Rise of the Novel (1957) is an example of what I mean in speaking of macro-oriented studies that do not rise far beyond the level of anecdote. Watt’s study of the novel is indeed impressive and cannot and should not be dismissed. Having said that, it is ultimately a study of the novel based on an analysis of just a few authors. These authors provide Watt with convenient touchstones for his history, but the choice of these authors cannot be considered representative of the ten to twenty thousand novels that make up the period Watt attempts to cover.

† The human aggregation of multiple case studies could certainly be considered a type of macroanalysis, or assimilation of information, and there are any number of “macro-oriented” studies that take such an approach, studies, for example, that read and interpret economic history by examining various case studies. Alan Liu pointed me to Shoshanna Zuboff’s In the Age of the Smart Machine (1988) as one exemplary case. Through discussion of eight specific businesses, Zuboff warns readers of the potential downsides (dehumanization) of computer automation. Nevertheless, although eight is better than one, eight is not eight thousand, and, thus, the study is comparatively anecdotal in nature.
for assessing, scrutinizing, and even forecasting the macroeconomy. Although there is an inherent need for understanding the economy at the micro level, in order to contextualize the macro results, macroeconomics does not directly involve itself in the specific cases, choosing instead to see the cases in the aggregate, looking to those elements of the specific cases that can be generalized, aggregated, and quantified.

Just as microeconomics offers important perspectives on the economy, so too does close reading offer fundamentally important insights about literature; I am not suggesting a wholesale shelving of close reading and highly interpretive “readings” of literature. Quite the opposite, I am suggesting a blended approach. In fact, even modern economics is a synthesis—a “neoclassical synthesis,” to be exact—of neoclassical economics and Keynesian macroeconomics. It is exactly this sort of unification, of the macro and micro scales, that promises a new, enhanced, and better understanding of the literary record. The two scales of analysis work in tandem and inform each other. Human interpretation of the “data,” whether it be mined at the macro or micro scale, remains essential. Although the methods of inquiry, of evidence gathering, are different, they are not antithetical, and they share the same ultimate goal of informing our understanding of the literary record, be it writ large or small. The most fundamental and important difference in the two approaches is that the macroanalytic approach reveals details about texts that are, practically speaking, unavailable to close readers of the texts.

John Burrows was an early innovator in this realm. Burrows’s 1987 book-length computational study of Jane Austen’s novels provided unprecedented detail into Austen’s style by examining the kinds of highly frequent words that most close readers would simply pass over. Writing of Burrows’s study of Austen’s oeuvre, Julia Flanders points out how Burrows’s work brings the most common words, such as the and of, into our field of view. Flanders writes, “[Burrows’s] effort, in other words, is to prove the stylistic and semantic significance of these words, to restore them to our field of view. Their absence from our field of view, their non-existence as facts for us, is precisely because they are so much there, so ubiquitous that they seem to make no difference” (2005, 56–57). More recent is James Pennebaker’s book The Secret Life of Pronouns, wherein he specifically challenges human instinct and close reading as reliable tools for gathering evidence: “Function words are almost impossible to hear and your stereotypes about how they work may well be wrong” (2011, 28). Reviewing Pennebaker’s book for the New York Times, Ben Zimmer notes that “mere mortals, as opposed to infallible computers, are woefully bad at keeping track of the ebb and flow of words, especially the tiny, stealthy ones” (2011, n.p.). At its most basic, the macroanalytic approach is simply another method of gathering details, bits of information that may have escaped our attention because of their sheer multitude. At a more
sophisticated level, it is about accessing details that are otherwise unavailable, forgotten, ignored, or impossible to extract. The information provided at this scale is different from that derived via close reading, but it is not of lesser or greater value to scholars for being such. Flanders goes on: “Burrows’ approach, although it wears its statistics prominently, foreshadows a subtle shift in the way the computer’s role *vis-à-vis* the detail is imagined. It foregrounds the computer not as a factual substantiator whose observations are different in kind from our own—because more trustworthy and objective—but as a device that extends the range of our perceptions to phenomena too minutely disseminated for our ordinary reading” (2005, 57). For Burrows, and for Flanders, the corpus being explored is still relatively small—in this case a handful of novels by Jane Austen—compared to the large corpora available today. This increased scale underscores the importance of extending our range of perception beyond ordinary reading practices. Flanders writes specifically of Burrows’s use of the computer to help him see more in the texts that he was then reading or studying. The further step, beyond Burrows, is to allow the computer to help us see even more, even deeper, to go beyond what we are capable of reading as solitary scholars.*

The result of such macroscopic investigation is contextualization on an unprecedented scale. The underlying assumption is that by exploring the literary record writ large, we will better understand the context in which individual texts exist and thereby better understand those individual texts. This approach offers specific insights into literary historical questions, including insights into:

- the historical place of individual texts, authors, and genres in relation to a larger literary context
- literary production in terms of growth and decline over time or within regions or within demographic groups
- literary patterns and lexicons employed over time, across periods, within regions, or within demographic groups
- the cultural and societal forces that impact literary style and the evolution of style
- the cultural, historical, and societal linkages that bind or do not bind individual authors, texts, and genres into an aggregate literary culture
- the waxing and waning of literary themes
- the tastes and preferences of the literary establishment and whether those preferences correspond to general tastes and preferences

* This approach again resonates with the approaches taken by the Annales historians. Patrick H. Hutton writes that whereas “conventional historians dramatize individual events as landmarks of significant change, the *Annales* historians redirect attention to those vast, anonymous, often unseen structures which shape events by retarding innovation” (1981, 240).
Furthermore, macroanalysis provides a practical method for approaching questions such as:

- whether there are stylistic patterns inherent to particular genres
- whether style is nationally determined
- whether and how trends in one nation's literature affect those of another
- the extent to which subgenres reflect the larger genres of which they are a subset
- whether literary trends correlate with historical events
- whether the literature that a nation or region produces is a function of demographics, time, population, degrees of relative freedom, degrees of relative education, and so on
- whether literature is evolutionary
- whether successful works of literature inspire schools or traditions
- whether there are differences between canonical authors and those who have been traditionally marginalized
- whether factors such as gender, ethnicity, and nationality directly influence style and content in literature

A macroanalytic approach helps us not only to see and understand the operations of a larger “literary economy,” but, by means of scale, to better see and understand the degree to which literature and the individual authors who manufacture that literature respond to or react against literary and cultural trends. Not the least important, as I explore in chapter 9, the method allows us to chart and understand “anxieties of influence” in concrete, quantitative ways.

For historical and stylistic questions in particular, a macroanalytic approach has distinct advantages over the more traditional practice of studying literary periods and genres by means of a close study of “representative” texts. Franco Moretti has noted how “a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn’t a sum of individual cases: it’s a collective system, that should be grasped as a whole” (2005, 4). To generalize about a “period” of literature based on a study of a relatively small number of books is to take a significant leap from the specific to the general. Naturally, it is also problematic to draw conclusions about specific texts based on some general sense of the whole. This, however, is not the aim of macroanalysis. Rather, the macroscale perspective should inform our close readings of the individual texts by providing, if nothing else, a fuller sense of the literary-historical milieu in which a given book exists. It is through the application of both approaches that we reach a new and better-informed understanding of the primary materials.

An early mistake or misconception about what computer-based text analysis could provide scholars of literature was that computers would somehow pro-
vide irrefutable conclusions about what a text might mean. The analysis of big corpora being suggested here is not intended for this purpose. Nor is it a strictly scientific practice that will lead us to irrefutable conclusions. Instead, through the study and processing of large amounts of literary data, the method calls our attention to general trends and missed patterns that we must explore in detail and account for with new theories. If we consider that this macroanalytic approach simply provides an alternative method for accessing texts and simply another way of harvesting facts from and around texts, then it may seem less threatening to those who worry that a quantification of the humanities is tantamount to the destruction of the humanities.

In literary studies, we are drawn to and impressed by grand theories, by deep and extended interpretations, and by complex speculations about what a text—or even a part of a text—might mean: the indeterminacies of deconstruction, the ramifications of postcolonialism, or how, for example, the manifold allusions in Joyce’s *Ulysses* extend the meaning of the core text. These are all compelling. Small findings, on the other hand, are frequently relegated to the pages of journals that specialize in the publication of “notes.” Craig Smith and M. C. Bisch’s small note in the *Explicator* (1990), for example, provides a definitive statement on Joyce’s obscure allusion to the *Iliad* in *Ulysses*, but who reads it and who remembers it?* Larger findings of fact, more objective studies of form, or even literary biography or literary history have, at least for a time, been “out of style.” Perhaps they have been out of style because these less interpretive, less speculative studies seem to close a discussion rather than to invite further speculation. John Burrows’s fine computational analysis of common words in the fiction of Jane Austen is an example of a more objectively determined exploration of facts, in this case lexical and stylistic facts. There is no doubt that the work helps us to better understand Austen’s corpus, but it does so in a way that leaves few doors open for further speculation (at least within the domain of common word usage, or “idiolects,” as Burrows defines them). A typical criticism levied against Burrows’s work is that “most of the conclusions which he reaches are not far from the ordinary reader’s natural assumptions” (Wiltshire 1988, 380).

Despite its complexity, the result of the work is an extended statement of the facts regarding Austen’s use of pronouns and function words. This final statement, regardless of how interesting it is to this reader, has about it a simplicity that inspires only a lukewarm reaction among contemporary literary scholars who are evidently more passionate about and accustomed to deeper theoreti-

Founda
cal maneuverings. To Burrows's credit, Wiltshire acknowledges that the value of Burrows's study is “not that it produces novel or startling conclusions—still less ‘readings’—as that it allows us to say that such ‘impressions’ are soundly based on verifiable facts” (ibid.).

Arguments like those made by Burrows have been, and perhaps remain, underappreciated in contemporary literary discourse precisely because they are, or appear to be, definitive statements. As “findings,” not “interpretations,” they have about them a deceptive simplicity, a simplicity or finality that appears to render them “uninteresting” to scholars conditioned to reject the idea of a closed argument. Some years ago, my colleague Steven Ramsay warned a group of computing humanists against “present[ing] ourselves as the people who go after the facts.”* He is right, of course, in the sense that we ought to avoid contracting that unpleasant disease of quantitative arrogance. It is not the facts themselves we want to avoid; however, we certainly still want and need “the facts.”

Among the branches of literary study, there are many in which access to and apprehension of “the facts” about literature are exactly what is sought. Most obvious here are biographical studies and literary history, where determining what the facts are has a great deal of relevance not simply in terms of explaining context but also in terms of determining how we understand and interpret the literary works within that context: the works of a given author or the works of a given historical period. Then there is the matter of stylistics and of close reading, which are both concerned with ascertaining, by means of analysis, certain distinguishing features or facts about a text.

Clearly, literary scholars do not have problems with the facts about texts per se. Yet there remains a hesitation—or in some cases a flat-out rejection—when it comes to the usefulness of quantification. This hesitation is more than likely the result of a mistaken impression that the conclusions following from a computational or quantitative analysis are somehow to be preferred to conclusions that are arrived at by other means. A computational approach need not be viewed as an alternative to interpretation—though there are some, such as Gottschall (2008), who suggest as much. Instead, and much less controversially, computational analysis may be seen as an alternative methodology for the discovery and the gathering of facts. Whether derived by machine or through hours in the archive, the data through which our literary arguments are built will always require the careful and imaginative scrutiny of the scholar. There will always be a movement from facts to interpretation of facts. The computer is a tool that assists in the identification and compilation of evidence. We must, in turn, interpret and explain that derivative data. Importantly, though, the

* Ramsay made these comments at the “Face of Text” conference hosted by McMaster University in November 2004. See http://tapor1.mcmaster.ca/~faceoftext/index.htm.
computer is not a mere tool, nor is it simply a tool of expedience. Later chapters will demonstrate how certain types of research exist only because of the tools that make them possible.

Few would object to a comparative study of Joyce and Hemingway that concludes that Hemingway's style is more minimalist or more “journalistic” than Joyce's. One approach to making this argument would be to pull representative sentences, phrases, and paragraphs from the works of both authors and from some sampling of journalistic prose in order to compare them and highlight the differences and similarities. An alternative approach would involve “processing” the entire corpus of both authors, as well as the journalistic samples, and then to compute the differences and similarities using features that computers can recognize or calculate, features such as average sentence length, frequent syntactical patterns, lexical richness, and so on. If the patterns common to Hemingway match more closely the patterns of the journalistic sample, then new evidence and new knowledge would have been generated. And the latter, computational, approach would be all the more convincing for being both comprehensive and definitive, whereas the former approach was anecdotal and speculative. The conclusions reached by the first approach are not necessarily wrong, only less certain and less convincing. Likewise, the second approach may be wrong, but that possibility is less likely, given the method.* Far more controversial and objectionable would be an argument along the lines of “Moby Dick is God, and I have the numbers to prove it.” The issue, as this intentionally silly example makes clear, is not so much about the gathering of facts but rather what it is that we are doing with the facts once we have them.

It is this business of new knowledge, distant reading, and the potentials of a computer-based macroanalysis of large literary corpora that I take up in this book. The chapters that follow explore methods of large-scale corpus analysis

* There are those who object to this sort of research on the grounds that these methods succeed only in telling us what we already know. In a New York Times article, for example, Kathryn Schulz (2011) responded to some similar research with a resounding “Duh.” I think Schulz misses the point here and misreads the work she is discussing (my blog post explaining why can be found at http://www.matthewjockers.net/2011/07/01/on-distant-reading-and-macroanalysis/). To me, at least, her response indicates a lack of seriousness about literature as a field of study. Why should further confirmation of a point of speculation engender a negative response? If the matter at hand were not literary, if it were global warming, for example, and new evidence confirmed a particular “interpretation” or thesis, surely this would not cause a thousand scientists to collectively sigh and say, “Duh.” A resounding “I told you so,” perhaps, but not “Duh.” But then Schulz bears down on the straw man and thus avoids the real revelations of the research being reviewed.
and are unified by a recurring theme of probing the quirks of literary influence that push and pull against the creative freedom of writers. Unlike Harold Bloom’s anecdotal, and for me too frequently impenetrable, study of influence, the work presented here is primarily quantitative, primarily empirical, and almost entirely dependent upon computation—something that Bloom himself anticipated in writing *Anxiety of Influence* back in 1973. Bloom, with some degree of derision, wrote of “an industry of source-hunting, of allusion-counting, an industry that will soon touch apocalypse anyway when it passes from scholars to computers” (31). Though my book ends up being largely about literary influence—or, if you prefer, influences upon literary creativity—and to a lesser extent about the place of Irish and Irish American writers in the macro system of British and American literature, it is meant fundamentally to be a book about method and how a new method of studying large collections of digital material can help us to better understand and contextualize the individual works within those collections. The larger argument I wish to make is that the study of literature should be approached not simply as an examination of seminal works but as an examination of an aggregated ecosystem or “economy” of texts. Some may wish to classify my research as “exploratory” or as “experimental” because the work I present here does more to open doors than it does to close them. I hope that this is true, that I open some doors. I hope that this work is also provocative in the sense of provoking more work, more exploration, and more experimentation.

I am also conscious that work classified under the umbrella of “digital humanities” is frequently criticized for failing to bring new knowledge to our study of literature. Be assured, then, that this work of mine is not simply provocative. There are conclusions, some small and a few grand. This work shows, sometimes in dramatic ways, how individual creativity—the individual agency of authors and the ability of authors to invent fiction that is stylistically and thematically original—is highly constrained, even determined, by factors outside of what we consider to be a writer’s conscious control. Alongside minor revelations about, for example, Irish American writing in the early 1900s and the nature of the novelistic genre in the nineteenth century, I continually engage this matter of “influence” and the grander notions of literary history and creativity that so concerned Elliot, Bloom, and the more or less forgotten Russian formalists whose bold work in literary evolution was so far ahead of its time.

The chapters that follow share a common theme: they are not about individual texts or even individual authors. The methods described and the results reported represent a generational shift away from traditional literary scholarship, and away even from traditional text analysis and computational authorship attribution. The macroanalysis I describe represents a new approach to the study of the literary record, an approach designed for probing the digital-textual world as it exists today, in digital form and in large quantities.
Quantitative Analysis and Literary Studies
David L. Hoover

History, Goals, and Theoretical Foundation

Modern quantitative studies of literature begin about 1850, with periods of intense activity in the 1930s and the 1980s. Fortunately, several excellent overviews discuss earlier work in the context of computers and literary studies (Burrows 1992a), stylometry (Holmes 1998), and authorship attribution (Holmes 1994; Love 2002). We can thus concentrate here on recent advances, driven primarily by the huge growth in the availability of electronic texts, increasingly sophisticated statistical techniques, and the advent of much more powerful computers that have produced much more accurate and persuasive analyses.

Quantitative approaches to literature represent elements or characteristics of literary texts numerically, applying the powerful, accurate, and widely accepted methods of mathematics to measurement, classification, and analysis. They work best in the service of more traditional literary research, but recent and current work often necessarily concentrates much of its effort on the development of new and improved methodologies. The availability of large numbers of electronic literary texts and huge natural language corpora has increased the attractiveness of quantitative approaches as innovative ways of "reading" amounts of text that would overwhelm traditional modes of reading. They also provide access to kinds of information that are not available even in principle without them. Quantitative approaches are most naturally associated with questions of authorship and style, but they can also be used to investigate larger interpretive issues like plot, theme, genre, period, tone, and modality.

A concrete example will suggest some of the benefits of quantitative analysis. In To the Lighthouse, Virginia Woolf describes a vacation house that has been closed up for the winter:

Nothing it seemed could break that image, corrupt that innocence, or disturb the swaying mantle of silence Once only a board sprang on the landing; once in the middle of the night with a roar, with a rupture, as after centuries of quiescence, a rock rends itself from the mountain and hurtles crashing into the valley, one fold of the shawl loosened and swung to and fro.

A critic struck by the comparison of a rock hurtling into a valley with a shawl loosening and swinging might also be interested in the apparent self-agency of the rock, the board, and the shawl, and might want to investigate Woolf's use of inanimate objects where animates are expected, a type of personification. The critic would normally deploy a series of supporting examples selected from the novel through careful reading, perhaps including some striking examples like these:

all round the table, beginning with Andrew in the middle, like a fire leaping from tuft to tuft of furze, her children laughed

It was as if the water floated off and set sailing thoughts which had grown stagnant on dry land, and gave to their bodies even some sort of physical relief.

And now in the heat of summer the wind sent its spies about the house again.

The list might be expanded with similar examples involving body parts:
would have liked to reply kindly to these blandishments

Indeed she had been keeping guard over the dish of fruit … hoping that nobody would
touch it until, oh, what a pity that they should do it — a hand reached out, took a pear,
and spoil the whole thing.

For how could one express in words these emotions of the body? … It was one's body
feeling, not one's mind.

Most readers will agree that Woolf's personifications are striking, but their literary functions seem
quite varied. In the first, the comparison of laughter and fire seems an apt and vivid way of
characterizing the spontaneous, variable, and contagious outbreak of humor, while the
personification of the hand in the fifth example, by removing the agency, focuses our attention on
the fruit basket still life.

A careful enough reading can examine all the uses of inanimate objects and body parts in the novel.
If the goal is merely to point out the personifications or to categorize them, there may be little gain
in quantifying the analysis, though categorizing the personifications would seem peculiar without
any indication of the frequencies of the various categories. Examples are rarely significant,
however, unless they are either unusual or characteristic of the novel or the author — otherwise why
analyze them? And the unusual and the characteristic must be validated by counting and
comparison: the bare claim that Woolf uses a great deal of personification is without value and
nearly meaningless unless it is quantified. In rare cases the quantification can be implicit: no
mathematical demonstration is necessary to show that a novel without the word "the" is unusual, but
Woolf's use of inanimate subjects is another matter. A single remarkable use of personification can
certainly be significant and noteworthy, but most stylistic and interpretive observations rest upon
patterns, and, therefore, upon repetition. Basing an argument about To the Lighthouse on the
prevalence of personification, then, requires counting those personifications and at least a rough
comparison of their frequency with some kind of norm or reference point. Finding dozens of odd
inanimate subjects in To the Lighthouse and only a few in other modernist novels of roughly the
same length might be sufficient.

Readers who know Woolf's novel well may doubt the centrality of personification to its
interpretation: in To the Lighthouse, the personification seems to be an aesthetic literary device
rather than an important and integral stylistic characteristic. The same cannot be said of The
Inheritors (Golding 1955). In that strange novel, the extreme prevalence of body parts and
inanimate objects as agents and subjects of verbs of motion (and even verbs of perception) is central
to Golding's creation of the imagined Neanderthal world-view of the text (see Hoover 1999 for
discussion). Many stylistic and interpretive patterns, however, are far more pervasive or far more
subtle, and they require more sophisticated, more powerful, and more explicit quantification.

Methods

Almost any item, feature, or characteristic of a text that can be reliably identified can be counted,
and most of them have been counted. Decisions about what to count can be obvious, problematic, or
extremely difficult, and poor initial choices can lead to wasted effort and worthless results. Even
careful planning leaves room for surprises, fortunately often of the happy sort that call for further or
different quantification. The frequencies of various letters of the alphabet and punctuation marks,
though not of obvious literary interest, have been used successfully in authorship attribution, as
have letter n-grams (short sequences of letters). Words themselves, as the smallest clearly
meaningful units, are the most frequently counted items, and syntactic categories (noun, verb,
infinite, superlative) are also often of interest, as are word n-grams (sequences) and collocations
(words that occur near each other). Thematic or semantic categories (angry words, words related to
time), while more difficult to count, have the advantage of being clearly relevant to interpretation,
and automated semantic analysis may reduce the effort involved. Phrases, clauses, syntactic
patterns, and sentences have often been counted, as have sequences or subcategories of them (prepositional phrases, subordinate clauses, passive sentences). Many of the items listed above are also used as measures of the lengths of other items: word length in characters, sentence or clause length in letters or words, text length in words, sentences, paragraphs, and so forth. Nonlinguistic textual divisions ranging from small units like lines and couplets to larger structural units like paragraphs, stanzas, scenes, acts, and chapters can also sometimes be usefully counted, as can literary categories like narrators and characters (including subcategories like first-person and third-person narrators, and characters divided by age, ethnicity, nationality, class, and gender), and plot elements (marriages, deaths, journeys, subplots).

The most obvious place to count whatever is counted is a single literary text that is of interest, as with the example from Woolf above. The need for some kind of comparative norm suggests that counting more than one text will often be required and the nature of the research will dictate the appropriate comparison text. In some cases, other texts by the same author will be selected, or contemporary authors, or a natural language corpus. In other cases, genres, periods, or parts of texts may be the appropriate focus. Counting may be limited to the dialogue or narration of a text, to one or more speakers or narrators, or to specific passages.

In the simplest quantifications, the numbers are merely presented and interpreted or offered as evidence that further investigation is likely to be productive. A critic interested in how writers differ in their vocabularies may find the raw counts of the numbers of different words (word types) in the first 50,000-word sections of a group of novels worth studying. In the first section of Sinclair Lewis's *Main Street*, for example, about 8,300 different words appear, but only 4,400 in Faulkner's *Light in August*, where the localization of the story may make the huge difference seem comprehensible. The 5,200 different words in the first section of James's *The Ambassadors* and the 6,600 in London's *The Sea Wolf* will require different explanations, and few readers would predict that *Main Street* has an exceptionally large vocabulary or *Light in August* an exceptionally small one.

Quantification does not end with counting or measurement and presentation, of course, and many different kinds of mathematical operations have been applied to the numbers. Among the simplest of these is comparing frequencies or averages among a group of texts, often using an appropriate statistical test of significance, such as Student's T-test or Chi-square, to gauge the likelihood that the observed difference could have arisen by chance. Authorship or style cannot reasonably be analyzed if the differences observed are likely to occur without the author's intervention. Fortunately, the patterns found in literary texts are often so obviously significant that no statistical testing is required, but it is easy to overestimate the oddity of a pattern, and statistical tests help to avoid untenable claims.

The standard deviation (roughly, the average difference, in either direction, of all frequencies from the mean), which measures how widely scattered the values are, and the z-score, which measures the distance of any given value from the mean in standard deviations, are often valuable for questions of textual difference. For example, in a corpus of 46 Victorian novels by six authors, the average rate of occurrence per 10,000 words is about 11 for "upon" and 63 for "on." In Silas *Marner* the frequencies are 4 "upon," 72 "on," and in *Vanity Fair* 17 "upon" and 50 "on," so that the difference between these two novels seems more extreme for "upon" than for "on." The standard deviations for the two words tell a different story: "upon" is quite variable in these six authors, with a standard deviation of about 9 words per 10,000 (not far below its average frequency of 11), but "on" is distributed much more evenly, with a standard deviation of about 15 (less than one-fourth its average frequency). Thus the frequencies of these words in the two novels are well within a single standard deviation from the mean, with z-scores between -0.84 and 0.71. Because they differ less than the average difference from the mean, frequencies in this range are quite likely to occur by chance, though the combination of the differences between the two words is suggestive.

Another simple operation is dividing the frequency of an item in one text by its frequency in another, yielding the distinctiveness ratio (DR), a measure of the difference between the texts.
Ratios below 0.67 or above 1.5 are normally considered worth investigating. Returning to inanimate subjects in Woolf, even my example of 24 such subjects in one novel and 8 in another gives a distinctiveness ratio of 3. But note that in the example of "upon" above the DR between the novels is greater than 4, so that some care should be taken not to over-interpret a DR when the frequencies of the words vary a great deal. Some measures of vocabulary richness or concentration, such as Yule's characteristic constant K, take into account the frequencies of all the word types in a text and require more complex calculations, as do other measures of vocabulary richness based on probabilistic models.

Recent years have seen a trend toward multivariate methods that are especially designed to deal with large amounts of data — methods such as principal components analysis, cluster analysis, discriminant analysis, correspondence analysis, and factor analysis. Statistical programs have made these methods much more practical by performing long sequences of required computations rapidly and without error. Principal components analysis, the most popular of these methods, allows the frequencies of many different items with similar distributions in a group of texts to be combined into a single component. The result is a small number of unrelated measures of textual difference that account for most of the variation in the texts. The first two of these components are typically used to create a scatter plot in which the distance between any two texts is a simple visual measure of their similarity. This technique provides a graphical method of "reading" a large number of frequencies at once, and is much easier to interpret than the list of frequencies themselves.

Delta, a promising new measure of textual difference based on word frequency, has stirred a great deal of interest (Burrows 2002a). Delta is designed to pick the likeliest author of a questioned text from among a relatively large number of possible authors. Burrows begins by recording the frequencies of the most frequent words of a primary set of texts by the possible authors and calculating the mean frequency and standard deviation for each word in this set of texts. He then uses z-scores to compare the difference between the mean and each of the primary authors with the difference between the mean and the questioned text for each of the words. He completes the calculation by averaging the absolute values of the z-scores of all the words to produce Delta, a measure of the difference between the test text and each primary-set author. The primary set author with the smallest Delta is suggested as the author of the test text. A further innovation in Delta is that Burrows expands the set of words analyzed to the 150 most frequent rather than the 30–100 used in earlier work with PCA, and I have shown that further expansion of the list to the 800 or even the 4,000 most frequent words often produces even stronger results on long texts (Hoover 2004a).

Recently Burrows has introduced two further measures, Zeta and Iota, which concentrate on words of moderate and low frequencies, respectively (2006). For both measures, a word frequency list is created for a sample of text by a primary author, and then the sample is divided into several sections of equal size. The heart of the procedure is to record the number of these sections that contain each of the words and then record which of the words occur in samples by other authors and in texts to be tested for authorship. By sorting the word list on the basis of how many of the primary author's text sections contain each word, Burrows eliminates the very frequent words that occur in most texts and concentrates on different parts of the word frequency spectrum. For Zeta, he retains only moderately frequent words, ones that occur in a subset of the primary author's sections. Where only two poets are being compared, he then further reduces the list of words by removing those that exceed a specific frequency in the works of the second poet. Where many authors are being compared, he removes words that appear in the text samples of most of the other authors. Whether there are two or many authors, the result is a list of words that are moderately frequent in the primary author and very infrequent in the other author(s). For Iota, words are removed that appear in most of the sections by the primary author. For two authors, words are also removed that do not appear in the second author's sample, and, where many authors are being tested, the second step removes words that appear in about half or more of the other authors. Both of these methods are remarkably effective in attributing poems as short as 1,000 words, and the discussion of methodology and substance is rich enough to provoke another wave of interest in authorship attribution and stylometry.
Techniques related to artificial intelligence are also increasingly being applied, including neural networks, machine learning, and data mining (see, for example, Waugh, Adams, and Tweedie 2000). These methods, which require more computing power and expertise than many other methods, are sometimes used in authorship attribution, but more often in forensic than in literary contexts. One reason for this is that they treat authorship attribution as a classification problem, and their results are more difficult to extend to traditional literary questions. Many of these techniques, as well as key word analysis, are well suited to the analysis of content, especially in the context of the huge amounts of text being produced on the World Wide Web.

Applications

Many kinds of studies of literary texts use quantitative methods. Quantitative thematic analysis can trace the growth, decay, or development of vocabulary within a thematic domain, or study how authors differ in their expressions of a theme (see Fortier 2002). Many empirical studies of literature translate readers' judgments into numerical scales to study literary response using techniques borrowed from the social sciences. Metrical analysis, because of the inherent reliance of meter on pattern, is a natural area for quantitative study, though there has been less research in this area than one might have expected.

A growing area of research is in the study of manuscript relationships, where techniques designed for the study of genetic relationship among organisms have been ingeniously and fruitfully applied to the study of the manuscripts of Chaucer's Canterbury Tales (see, for example, Spencer et al. 2003). These studies take literally the metaphor of genetic relationships among manuscripts, treating differences among them as if they were differences in DNA sequences. The huge amount of data involved in some manuscript traditions invites and practically requires sophisticated statistical techniques, whether those of evolutionary biology or other multivariate techniques. Genre and period definition and classification also benefit from quantitative approaches, especially factor analysis and other multivariate techniques.

Authorship attribution and statistical stylistics (or stylometry), currently two of the most important areas of quantitative analysis of literature, deserve a fuller treatment. They share many basic assumptions and methods, though some techniques that are effective in distinguishing authors may have no clear interpretive value. A discussion of authorship attribution in the present context necessarily forces a distinction between forensic and literary attribution that is sometimes without a difference. Determining who wrote a text generally requires much the same methodology whether the text is ransom note, a threatening letter, a legal opinion, the federalist papers, a contemporary political novel like Joe Klein's Primary Colors, an anonymous eighteenth-century verse satire, or play by Shakespeare.

Yet two differences between the forensic and literary attribution must be kept in mind. First, in many forensic contexts, the identity of the person who produced the language of the text may be irrelevant, while the identity of the person responsible for sending it may be crucial. A kidnapper may force a victim to write a ransom note, and a manifesto may be cribbed from a series of websites; determining these facts may or may not help to solve the crime. Second, the text in a forensic problem typically has little intrinsic value and becomes irrelevant once the attribution is made and the crime solved. In the case of literary attribution, however, and preeminently for Shakespeare's plays, the aesthetic and contested cultural value of the texts lies at the heart of the problem. One consequence of these differences is that literary attribution is often only a first step, so that methods easily turned to stylistic or interpretive purposes tend to be favored.

Only when external evidence fails is it reasonable to apply quantitative methods, and the presence or absence of a closed set of possible authors and differences in the size and number of documents available for analysis are usually more significant than the kind of text involved. Here I will concentrate on reasonably tractable kinds of literary authorship problems in which the questioned text is of a reasonable size and similar texts by the claimant authors are available.
Authorship attribution has often been based on a single variable like word length, sentence length, or vocabulary richness, and some researchers continue to achieve good results using a small number of such variables. Most current research, however, has turned to more robust multivariate methods such as principal components analysis and cluster analysis, often combining the results of more than one method to solve a problem. In his excellent overview of computers and the study of literature, Burrows (1992a) uses principal components analysis (PCA) of the fifty most frequent words to argue against the possibility that Lady Vane had a hand in the "Memoirs of a Lady of Quality" which Smollet includes in his *Peregrine Pickle*. In other work, he shows that the seventy-five most frequent words can successfully distinguish 4,000-word sections of novels by the Brontë sisters and that the twenty most frequent words can distinguish 500-word sections of letters by Scott and Byron (Burrows 1992b). Even more remarkable, when statistical tests are used to select the words that most effectively discriminate between Scott and Byron, the ten most effective of these do an excellent job of separating the works of Scott and Byron even across several genres. A good recent example of this methodology, using a careful approach that also takes into account traditional methods, persuasively adds additional shore journalism to Stephen Crane's small oeuvre (Holmes et al. 2001). Principal components analysis (PCA) of the fifty most frequent words of the texts shows that Crane's fiction can be distinguished from Conrad's, that his fiction can be distinguished from his shore journalism and New York City journalism, and that these two kinds of journalism are different from his war journalism. The same method shows that Crane's shore and New York journalism are different from that of his brother Townley and two other contemporary journalists. Both PCA and cluster analysis strongly suggest that seventeen pieces of previously unattributed shore journalism (known to be by one of the brothers) is Stephen's rather than Townley's.

Authorship attribution based on n-grams, sequences of various numbers of letters or words, has become increasingly popular, sometimes performing better than word frequency alone, especially on small texts. In a wide-ranging and provocative article, Clement and Sharp (2003) show that both letter and word n-grams perform marginally better than methods based on words. For these experiments, the frequencies of the various items in the known documents are transformed into probabilities of randomly extracting them from the text, and the test document is assigned to the author whose training set maximizes the probability of generating the text. Besides presenting many different methods and varieties of results, Clement and Sharp raise important questions about the relationship between content and style, the effects of text size, and apparently random differences that alter the accuracy of analyses. Although letter n-grams lack any transparent relationship to the meaning or style of a text, and are unlikely to be attractive to researchers who are interested in broader literary questions, word n-grams are likely to become increasingly popular because they may both improve accuracy and allow the critic to focus on meaningful word groups.

**Four Exemplary Studies**

Statistical stylistics or stylometry is the broadest of the areas in which quantitative analysis intersects with literary study, and it might be said to subsume parts or all of the applications just discussed. Its central concerns are closest to those of literary studies in general, with a special emphasis on the patterns that comprise style and how those patterns are related to issues of interpretation, meaning, and aesthetics. Rather than surveying or describing various kinds of stylometric studies, I will focus in a more detail on four recent articles that exemplify some of the most central concerns and methods while treating important literary problems and questions.

"Cicero, Sigonio, and Burrows" (Forsyth, Holmes, and Tse 1999) is about authorship, but it also treats issues of chronology and genre. It applies methods first proven on English to classical Latin and neo-Latin (inflected languages) and examines not only words, but also word length (in syllables) and some information about transitions between words of different lengths. These variables are analyzed using PCA, cluster analysis, and discriminant analysis, and the authors combine careful analysis with useful methodological observations. The central question asked is whether it is likely that the *Consolatio Ciceronis*, which was edited and published by Sigonio in 1583, is really the lost work known to have been written by Cicero about 45 bc and existing only as
fragments quoted in other works, or, as was suggested shortly after its publication, a forgery by Sigonio himself. Can authorship attribution methods distinguish "between Cicero and Ciceronianism" (Forsyth, Holmes, and Tse 1999: 378)?

After collecting more than 300,000 words of classical and neo-Latin by eleven authors and dividing them into 70 sample texts, the authors use PCA based on the 46 most frequent function words to show that Cicero's oratory is distinct from his prose —that, as has often been noted, genre effects sometimes overwhelm authorship effects. The same method distinguishes Cicero well from six other classical authors, as does cluster analysis, and both produce slightly weaker but still broadly accurate results when Sigonio is tested against the other sixteenth-century authors.

Turning to Sigonio, Cicero, and the Consolatio, the authors use stepwise discriminant analysis of known texts by Sigonio and Cicero to determine the words that are most effective in distinguishing the two authors. This technique is especially appropriate in cases like this one where some samples belonging to distinct groups are available. It identifies a small group of discriminators that are quite effective in distinguishing Sigonio and Cicero: they classify only two Ciceronian texts as by Sigonio and attribute both sections of the Consolatio to him. Discriminant analysis is also used to discover variables that distinguish effectively between classical and neo-Latin, discriminators that classify the Consolatio as neo-Latin. Adding information about word length and syllable transition improves the accuracy of the analyses and more firmly identifies the Consolatio as neo-Latin. Finally, discriminant analysis also shows that the Consolatio is enough like Sigonio's other work to suggest that he is its author.

The authorship of this disputed work is inherently significant, and this article does an exceptionally clear job of describing the literary and cultural situation in which the authorship question is to be asked. The careful division of the problem into subproblems provides clarity, and the variety and methodological sophistication of the analyses both strengthen the case against Cicero as the author and serve as a guide to future work.

In "Jonsonian Chronology and A Tale of a Tub," one of several careful and important studies, Hugh Craig also uses discriminant analysis, but he applies it to a very different literary problem (1999). His central question is the position of Ben Jonson's A Tale of a Tub in the chronology of his work, and thus the context in which the play is read. Whether, "it is a late work of pastiche or is in origins an early, naively conventional one" (230–31) has important implications for its significance and its interpretation.

The existence of several datable early, middle, and late comedies allows Craig to set up a discriminant analysis based on the 58 most frequent function words, more heavily weighting those words that discriminate best among the three periods. When the plays are divided into 2,000-word segments, those segments separate clearly into clusters, with very little overlap. Craig's methodology, like the PCA analysis popularized by Burrows, allows a scatter plot of the play segments to be compared with a scatter plot of the variables that produced it, and this in turn allows him to discuss the words and their stylistic and chronological implications.

Analyzing segments of A Tale of a Tub in the same way shows that its segments are very widely dispersed: some appear among the earliest segments, some in the middle, some among the late segments, and one outside all three of these clusters. Though this is consistent with an early play later revised, Craig wisely tests other late plays, showing that the scatter is not an artifact of the analysis and that A Tale of a Tub is much more widely scattered than the others. The sectioning of the plays also allows a discussion of plot and content in relationship to the scattered segments.

Next, the play is repeatedly re-segmented at 100-word intervals and subjected to the same analysis to pinpoint abrupt changes in style, which are discussed with reference to the boundaries between acts and scenes. Finally, rolling segments of A Tale of a Tub are compared with those of other plays, showing that the fluctuations in A Tale of a Tub are much more extreme. Although no firm conclusions can be reached, this careful, innovative, and thorough analysis strongly suggests an
early play reworked by Jonson near the end of his career. Craig's frequent and insightful return to the text and to questions of serious literary significance marks this as model stylometric analysis.

In taking up "Charles Brockden Brown: Quantitative Analysis and Literary Interpretation" (Stewart 2003) we shift genres, continents, centuries, and focus, but retain a strong relationship between quantitative analysis and more traditional literary questions. Rather than authorship or chronology, Stewart focuses on the styles of narration in two novels by Charles Brockden Brown (1771–1810), Wieland or The Transformation and the unfinished Memoirs of Carwin, the Biloquist. He investigates whether Brown successfully creates distinct narrative voices for the four narrators of Wieland and a consistent voice for Carwin, who is both one of those narrators and also the narrator of the unfinished Memoirs of Carwin.

Burrows used PCA successfully in distinguishing the dialogue of Jane Austen's various characters in his classic Computation into Criticism (1987), treating the characters as if they were literally the "authors" of their own speech. One interesting question is whether or not a writer of more modest gifts can successfully create distinct and consistent narrative styles. However important Brown may be to the origins of American literature, he is no Jane Austen. I have investigated a similar question regarding Hannah Webster Foster's 1797 American epistolary novel, The Coquette (Hoover et al. forthcoming), and, in a discussion of Nineteen Eighty-Four, The Inheritors, and The Picture of Dorian Grey, have suggested revisions to the standard methodology that may improve analyses of parts of texts by a single author (2003).

Stewart uses both PCA and cluster analysis and bases them not only on the frequencies of the 30 most frequent words, but also on the frequencies of various punctuation marks, and on the frequencies of words, sentences, and paragraphs of different lengths. He shows that the chapters of Wieland narrated by Clara, Pleyel, and Theodore are generally quite distinct from the chapter narrated by the villainous Carwin, which clusters with his chapters from Memoirs of Carwin. The analysis reveals an anomaly that provides the impetus for a discussion of more traditional literary concerns: both Pleyel's chapter of Wieland and the final chapter of that novel, which is narrated by Clara, cluster with Memoirs of Carwin, and Carwin's chapter of Wieland. (In an article in a similar spirit, McKenna and Antonia [2001] probe the differences among interior monologue, dialogue, and narrative in Joyce's Ulysses, arguing that multivariate analysis of Gerty McDowell's language can contribute to the interpretation of form, meaning, and ideology in that complex and difficult novel.)

There is no space here to do justice to the subtlety of Stewart's integration of these quantitatively anomalous results into the larger critical debate surrounding the interpretation of the early American novel, but he produces some very suggestive reasons for the similarity of the voices of Pleyel and Carwin, including their long years spent in Europe — quite significant in an early American novel — and the fact that both want to dominate and possess Clara. Stewart also suggests connections between Clara's narration of the final chapter of Wieland, a kind of "happy ending" postscript written from Europe after her marriage to Pleyel, and critical views of Brown as a skeptic about the American experiment. He also makes intriguing suggestions of a connection between Carwin's ventriloquism and the similarity between Clara's voice and his own in the final chapter of Wieland. This study not only uses statistics effectively to provide insight into important questions of interpretation, but also "suggests that traditional critical interpretation has a real bearing on how we understand the meaning of those statistics" (138). There is, of course, always the danger of arguing a specious excuse for a real anomaly after the fact, but that is the nature of interpretation, and such speciousness often leads to its own correction.

I conclude this selection of exemplary articles with "The Englishing of Juvenal" (Burrows 2002b), which contrasts in topic and methodology with the three articles just discussed, but continues their serious engagement with traditional literary concerns. Its focus is on translation and style, with a twist of authorship and chronology, and the method is Delta analysis, described above. (For another very interesting look at translation and authorship attribution, see Rybicki 2006.) To a database of more than half a million words of English Restoration poetry, Burrows adds fifteen translations of Juvenal's tenth satire dating from 1646 to 1967. When Delta is used to attribute the translations to
their authors, it is not very successful. D'Urfey is ranked first and Dryden second as the author of Dryden's translation; Johnson is strongly identified as the author of his translation, but Vaughan and Shadwell both rank well down the list of possible authors of theirs. Tests involving other translations give similar spotty results, suggesting that some authors effectively suppress their own styles and others do not; for his other translations, for example, Dryden often appears far down the list of likely authors.

Characteristically, Burrows goes on to a second analysis. This time, rather than asking who is the likeliest author of each translation, this analysis focuses on the authors, asking which of the fifteen Juvenal translations is most like the original work of each translator. In three of the four tests, the results are correct; in the fourth Dryden's comes in a very close second to Higden's as the most similar to the work of Dryden. These impressive results on a very difficult problem show that Delta is capable of capturing subtle authorial markers that persist even when submerged beneath the style of a translation. Another interesting fact is that D'Urfey, who ranks first as author of Dryden's Juvenal X, appears as the most likely author of five of the fifteen translations and as second or third most likely of eight others. Burrows shows that this is not the phenomenon often seen in authorship studies, where the lowest common denominator is, by default, the likeliest author when the true author is not present. The phenomenon is limited to the translations of Juvenal, suggesting that there are real similarities in style between D'Urfey's English and Juvenal's Latin.

Burrows then alters Delta slightly by using the averages of the word frequencies in all of the translations as the test text, treating this average text as a model of Juvenalism, but also retaining the average frequencies as the means against which Delta is calculated. (For other, more extensive alterations to Delta that I have suggested, see Hoover 2004b.) This naturally results in a complete set of zero z-scores for the test text and the mean, but it allows all fifteen translations to be measured against the "model." Shadwell's translation is the most similar to the mean and Johnson's the most different, even more different than the twentieth-century translations and the prose translations. Burrows concludes by using the differences between Johnson and the model to illuminate some of the important characteristics of Johnson's style, noting that Dryden and Johnson lie at opposite ends of a spectrum from versatility to consistency, a spectrum that all students of style would do well to remember. The emphasis on comparison in this article and the telling applications of statistical methods are particularly valuable. The concluding comments about the contrast between close reading and computer analysis emphasize the use of the computer to enhance and extend our ability to "read" literary texts in new ways:

The close reader sees things in a text — single moments and large amorphous movements —to which computer programs give no easy access. The computer, on the other hand, reveals hidden patterns and enables us to marshal hosts of instances too numerous for our unassisted powers. Even in the common case where we do not have fifteen versions of one original to bring into comparison, these principles hold good. 

(Burrows 2002b: 696)

A Small Demonstration: Zeta and Iota and Twentieth-Century Poetry

Given the rapid developments in this field, a small demonstration of the potential of Burrows's newest measures of textual difference, Zeta and Iota, seems appropriate. I began with some Delta tests on very different data, samples of poetry by forty twentieth-century poets, using large samples by twenty-six poets as the primary set and thirty-nine long poems as the secondary set. Twenty-five of these were by poets in the primary set and fourteen by other poets (these poems by primary authors were removed from their main samples). Delta is very accurate on these texts, correctly identifying the authors of all but three of the long poems by members of the primary set. I then took the few errors that occurred and looked for circumstances where a single author erroneously ranks first as the author of one poem and also ranks among the likeliest authors of another poem by the
same poet. Among the most similar of the poets in my study using these criteria are Wallace Stevens, Archibald MacLeish, and T. S. Eliot. Burrows's tests of Waller and Marvell using Zeta and Iota were based on main sets of about 13,000 and 20,000 words, and my set for Eliot is about the same size; the sets for Stevens and MacLeish are much larger, more than 70,000 words. The individual poems to be tested are roughly the same size as those Burrows tested, about 2,000 to 6,000 words.

The results of tests of MacLeish against Stevens using both Zeta and Iota were impressive. The new measures had no difficulty distinguishing the two poets, whether MacLeish or Stevens formed the primary sample. There is space here to discuss only the results of Zeta, which is based on the middle of the word frequency spectrum — words that have largely been ignored in earlier studies. Zeta is even more effective in distinguishing MacLeish and Stevens than it was in distinguishing Waller and Marvell, with the lowest Zeta for the primary author typically twice as large as that for the second author. As Burrows found, the poems by another author, here Eliot, sometimes narrowly outscore some of those of the primary author. Although this may seem disconcerting, it actually suggests that Zeta is narrowly and appropriately tuned to the difference between the authors being tested.

On this set of texts, I found that much more stringent stipulations than those used by Burrows produced some fascinating results: the 26 words that are found in all five sections of MacLeish's sample but do not occur in the Stevens sample seem to be good potential MacLeish authorship markers. Their total frequency is 321 in the five MacLeish sections and 40 in the two individual long MacLeish poems, but only 2 in the Stevens sample and his two long poems combined. Relaxing the restriction to retain words if they appear in more than three, more than two, or more than one section gradually reduces the amount of difference between the poems by MacLeish and Stevens, though all of these analyses are completely accurate. The 40 words remaining based on the same stipulations in the Stevens sample, with a total frequency of 545 in Stevens and 60 in Stevens's two long poems, but only 3 in the MacLeish sample and his two long poems combined, in turn seem to be good potential Stevens authorship markers.

Selecting words that occur in all of the sections of the primary sample seems to violate Burrows's intention of avoiding the 30–150 most frequent words that are so often used in other methods, but the stipulation of a maximum frequency of 3 in Stevens accomplishes this in any case. For example, "answered" is the most frequent of the 26 potential MacLeish markers, but it ranks only 323rd among the most frequent words in the MacLeish samples, and "reality" is the most frequent of the 40 potential Stevens markers, but it ranks only 184th among the most frequent words in the Stevens samples. Both are thus beyond the range normally used in tests of frequent words. The 26 MacLeish words range in rank from 323 to 1,422 and the 40 Stevens words from 184 to 1,378, placing all of them well within the range of words that I normally now include in Delta analyses and placing most of them within the range I normally include in cluster analyses. The presence of powerful discriminators like these may help to explain why expanding the size of the word list that is analyzed so often increases the accuracy of the results.

Finally, a glance at the 26 MacLeish words and the 40 Stevens words suggests that Zeta and Iota may provide useful ways of focusing our attention on interesting words:

**Ubiquitous Stevens — Rare MacLeish**

reality, except, centre, element, colors, solitude, possible, ideas, hymns, essential, imagined, nothingness, crown, inhuman, motions, regard, sovereign, chaos, genius, glittering, lesser, singular, alike, archaic, luminous, phrases, casual, voluble, universal, autumnal, café, inner, reads, vivid, clearest, deeply, minor, perfection, relation, immaculate

**Ubiquitous MacLeish — Rare Stevens**
Besides the obviously greater length and abstractness of the Stevens words, especially the nouns, the Stevens list is saturated with adjectives, while the MacLeish list has very few adjectives and proportionally more verbs and concrete nouns. A search for some of the most frequent of these marker words in each poet's work yields an interesting pair of short poems: MacLeish's "'Dover Beach' — A Note to that Poem" and Stevens's "From the Packet of Anacharsis." Forms of no less than 7 of MacLeish's 26 marker words appear in his short poem (215 tokens, 123 types), including the 3 italicized in the following brief passage:

… It's a fine and a
Wild smother to _vanish_ in: pulling down —
Tripping with outward ebb the urgent inward.
Speaking alone for myself it's the _steep_ hill and the
Toppling _lift_ of the young men I am toward now…

Forms of 6 of Stevens's 40 marker words appear in his even shorter poem (144 types, 91 tokens), including the 3 italicized in the brief passage below (internal ellipsis present in the original):

> And Bloom would see what Puvis did, protest
> And speak of the floridest _reality_…
> In the punctual _centre_ of all circles white
> Stands truly. The circles nearest to it share
> Its _color_…

One of the most difficult challenges for quantitative analyses of literature is preventing the huge numbers of items being analyzed from overwhelming our ability to see the results in insightful ways. By reducing the numbers of words to be examined and selecting sets of words that are particularly characteristic of the authors, Zeta and Iota seem likely to prove very useful for literary analysis as well as authorship attribution, whatever the further developments of them may be once they have been tested and refined. (Zeta and Iota sometimes produce anomalous results in tests including many authors; Burrows suggests [personal communication] that they are better reserved for head-to-head comparisons.)

The Impact, Significance, and Future Prospects for Quantitative Analysis in Literary Studies

As has often been noted, quantitative analysis has not had much impact on traditional literary studies. Its practitioners bear some of the responsibility for this lack of impact because all too often quantitative studies fail to address problems of real literary significance, ignore the subject-specific background, or concentrate too heavily on technology or software. The theoretical climate in literary studies over the past few decades is also partly responsible for the lack of impact, as literary theory has led critics to turn their attention away from the text and toward its social, cultural, economic, and political contexts, and to distrust any approach that suggests a scientific or "objective" methodology. There are, however, signs of progress on both these fronts. The recent increased interest in archives within literary criticism will almost necessarily lead to the introduction of quantitative methods to help critics cope with the huge amount of electronic text now becoming available. Some quantitative studies have also begun to appear in mainstream literary journals, a sure sign of their growing acceptance. The increasing frequency of collaborations between literary scholars and practitioners of quantitative methods of many kinds also promises to produce more research that strikes an appropriate balance between good methodology and significant results. Prospects for the emergence of quantitative approaches as a respected, if not central, branch of literary studies seem bright.
References


Does size matter? Authorship attribution, short samples, big problem

Maciej Eder

Abstract

The aim of this study is to find such a minimal size of text samples for authorship attribution that would provide stable results independent of random noise. A few controlled tests for different sample lengths, languages and genres are discussed and compared. Depending on the corpus used, the minimal sample length varied from 2,500 words (Latin prose) to 5,000 or so words (in most cases, including English, German, Polish and Hungarian novels). Another observation is connected with the method of sampling: contrary to common sense, randomly excerpted ‘bags of words’ turned to be much more effective than the classical solution, i.e. using original sequences of words (‘passages’) of desired size. Although the tests have been performed using the Delta method (Burrows, 2002) applied to the most frequent words (MFWs), some additional experiments have been conducted for SVM and k-NN applied to MFWs, character 3-grams, character 4-grams, and POS-tag 3-grams. Despite significant differences in overall attributive success between particular methods and/or style-markers, the minimal amount of textual data needed for reliable authorship attribution turned out to be method-independent.

1 Introduction

In the field of computational stylistics, and especially in authorship attribution, the reliability of the obtained results becomes even more essential than the results themselves: failed attribution is much better than false attribution (cf. Love, 2002). However, while dozens of outstanding papers deal with increasing the effectiveness of current stylo-metric methods, the problem of their reliability remains somehow underestimated. Especially, the simple yet fundamental question of the shortest acceptable sample length for reliable attribution has not been discussed convincingly.

It is true that the problem is not new. Its importance is stressed, although not directly, by Rudman in his seminal papers concerning reliability in authorship attribution inference (Rudman 1998a, 1998b, 2003). In his investigation of style variation in Golding’s The Inheritors, Hoover noticed that truncating all the samples to the size of the shortest chapter spoils the results, probably due to the short sample effect (Hoover, 2003: 439). In another instance, Rybicki discovered that his own results of remarkable similarities in the patterns of distance between idiolects in two different translations of the same trilogy of novels were due to the gap between talkative and non-talkative characters, the latter simply not saying enough to produce a reliable sample (Rybicki, 2006; 2008).

A few scholars have proposed an intuitive solution of this problem, e.g. that an analyzed text should be ‘long’ (Craig, 2004: 287), that ‘for stylo-metric reliability the minimum sample size allowed is 1,000 words’ (Holmes et al., 2001: 406), that ‘with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship’ (Burrows, 2002: 276), etc. Those statements, however, have not been followed by thorough empirical investigation. Additionally, many otherwise successful attribution studies do not obey even the assumed limit of 1,000 words per sample (Juola
and Baayen, 2005; Burrows, 2002; Jockers et al., 2008, etc.). In those attribution studies based on short samples, despite their well-established hypotheses, good choice of style-markers, advanced statistics applied and convincing results presented, one cannot avoid the simple yet nontrivial question whether those impressive results have not been obtained by chance, or at least have not been positively affected by randomness?

Recently, a few studies concerning different issues in scalability in authorship attribution have been published (Zhao and Zobel, 2005; Hirst and Feigina, 2007; Stamatatos, 2008; Koppel et al., 2009; Mikros, 2009; Luyckx, 2010; Luyckx and Daelemans, 2011). However, the problem addressed in the present study, that of an experimental estimation of the sample length that would provide a reliable attribution, has not been solved exhaustively. Also, there have been no cross-language studies, while this would seem an ideal way to validate the obtained results and to generalize the observed behavior of particular case studies.

2 Hypothesis

Word frequencies in a corpus are not random variables in a strict sense – especially, an occurrence of a given word highly depends on its nearest context: the probability of finding, say, ‘and’ immediately followed by ‘and’ is extremely low. And yet words do display some characteristics of random variables: since the author is not obliged at any rate to distribute words regularly, particular word frequencies might vary substantially across different works, or even different passages (chapters, stanzas) written by the same person. Thus, similarly to other probabilistic phenomena, word frequencies strongly depend on the size of the population (i.e., the size of the text used in the study). Now, if the observed frequency of a single word exhibits too much variation for establishing an index of vocabulary richness resistant to sample length (Tweedie and Baayen, 1998), a multidimensional approach – based on numerous probabilistic word frequencies computed at once – should be even more questionable. On theoretical grounds, we can intuitively assume that the smallest acceptable sample length would be hundreds rather than dozens of words. Next, we can expect that, in a series of controlled authorship experiments with longer and longer samples tested, the probability of attribution success would at first increase very quickly, indicating a strong correlation with the current text size; but then, above a certain value, further increase of input sample size would not affect the effectiveness of the attribution significantly. However, in any attempt to find this critical point on a ‘learning curve’, one should be aware that this point might depend – to some extent – on the language, genre, or even the particular texts analyzed.

3 Data and method of testing

In any approach to the problem of scalability in authorship attribution, an appropriate choice of test data is of a great importance. One possible solution is to perform a contrastive analysis of naturally long vs. naturally short texts (e.g. novels vs. short stories, essays vs. blog posts, etc.), to estimate the possible correlation between sample length and attribution accuracy. The most obvious weakness of this kind of approach, however, is that the results might be biased by inherent cross-genre differences between the two groups of texts. To eschew this limitation, in the present study the same dataset was used for all the comparisons: the goal was to extract shorter and longer virtual samples from the original corpus, using intensive re-sampling in a large number of iterations. The advantage of such a gradual increase of excerpted virtual samples is that it covers a wide range between ‘very short’ and ‘very long’ texts, and makes it possible to capture a break point of the minimal sample size for a reliable attribution.

Several corpora of known authorship were prepared for different languages and genres (used separately): for English, Polish, German, and Hungarian novels, for Latin and Ancient Greek prose (non-fiction), and for English, Latin, and Ancient Greek epic poetry. Within a particular genre (novels, non-fiction, poetry), the collected corpora were roughly similar in size. Keeping rigidly the same
number of texts in each corpus seemed to be unrealistic; additionally, it should be stressed that any cross-corpus (or cross-language) comparison will never be fully objective, even if the collected datasets are identical in size. The corpora used in the present study were as follows:

- 63 English novels by 17 authors,
- 66 German novels by 21 authors,
- 69 Polish novels by 13 authors,
- 64 Hungarian novels by 9 authors,
- 94 Latin prose texts by 20 authors,
- 72 Ancient Greek prose texts by 8 authors,
- 32 English epic poems by 6 authors,
- 32 Latin epic poems by 6 authors,
- 30 Ancient Greek epic poems by 8 authors.

The texts have been gathered from a variety of public domain sources, including Perseus Project, The Latin Library, Bibliotheca Augustana, Project Gutenberg, Literarure.org, Ebooks@Adelaide, and the author’s private collection. The acquired texts have been edited in order to normalize spelling (if applicable), to exclude footnotes, disclaimers, non-authorial prefaces, and so forth.

For each corpus, three discrete controlled attribution experiments aimed to examine three different methods of sampling (discussed below in detail) were performed. To assess the textual data, a few attribution techniques have been applied. As the main methodological basis for all the experiments, however, the widely accepted Delta method (Burrows, 2002) was chosen, with the assumption that the results should be valid, by extension, for other distance-based methods as well. In all the tests, 200 most frequent words (MFWs) were analyzed. For the computation tasks, including text preprocessing, sampling, and classification, a tailored script for the open-source statistical environment R was used.

The reason of choosing Delta based on MFWs was that it combines high accuracy of supervised methods of classification with simplicity of multidimensional techniques using distance measures of similarity. It seemed to be a good compromise between two main approaches to stylometry: literary-oriented studies on stylistic similarities between texts (authors, genres, styles and so forth) on the one hand, and information technology studies on authorship attribution ‘in the wild’ on the other. Whilst the former approach usually involves explanatory distance-based techniques such as multidimensional scaling or cluster analysis, and is aimed to capture stylometric relationships between literary texts, the latter considers attribution as a particular case of a classification problem (where precision is the most important issue), and usually relies on sophisticated machine-learning methods of supervised classification. In attempts to find a balance between those two discrete stylometric worlds, Burrow’s Delta seemed to be the best choice.

It is true that Delta exhibits some methodological pitfalls, including the lack of validation of the obtained results, and the tacit assumption of variables’ independence (Argamon, 2009). Also, it is sometimes claimed to be suboptimal in comparison with other classification algorithms: among computer scientists, there seems to be a consensus that support vector machines (SVM) using character n-grams is presently the single best, language-independent approach in the field of authorship attribution (Koppel et al., 2009; Stamatatos, 2009). On the other hand, however, Delta proved to perform almost equally well when compared with other classification techniques, including SVM (Jockers and Witten, 2010), and the claims about the robustness of character n-grams turned out to be unfounded for some languages (Eder, 2011).

The above arguments summarized, one has to admit that relying on one attribution method alone might lead to biased and/or unreliable results. Thus, apart from Delta based on MFWs, a number of additional tests have been conducted using other classifiers (SVM, k-NN) and other style markers (character 3-grams, character 4-grams, POS-tags 3-grams).

The benchmarks were based on the standard procedure used in machine-learning methods of
classification; namely, all the available texts from a given corpus were divided into two groups. The ‘training’ set consisted of one representative text per author, while all the remaining samples were included into the ‘test’ set. Next, samples of a given length were extracted from the original texts, and each sample from the ‘test’ set was tested against the ‘training’ set to verify the quality of classification. The percentage of correctly ‘guessed’ samples, i.e. those linked to their actual authors, was regarded as a measure of attribution accuracy.

The procedure as described above, strictly following the original implementation of Delta (Burrows, 2002), provides an approximation to many real-case attribution studies that have been published in recent decades. Although this approach should reveal precisely the correlation between sample size and attribution performance, it is not reliable enough to test the attribution performance itself. This is because the ‘training’ texts might not have been ‘representative’ (whatever that means) for their authors, or there might have appeared other irregularities in stylometric profiles. One of the ways to eschew the possible bias is to swap randomly some texts between the ‘training’ and the ‘test’ set, and to replicate the classification stage. This solution is referred to as cross-validation and it will be addressed in the final sections of this paper. However, since the point of gravity in this study was set towards literary studies (digital philology) rather than ‘raw’ authorship attribution, advanced topics of validation, as well as rigorous statistical terminology have been simplified as much as possible.

4 Experiment I: bags of words

The tests discussed in this paper are aimed to excerpt, in many approaches, shorter and longer subsets from the original literary texts. It is arguable, however, that in literary works, and particularly in novels, the vocabulary is distributed slightly differently in narrative and dialogue parts (Burrows, 1987: 163–75; Hoover, 2001: 426). Thus, to avoid the possible impact of these natural inconsistencies in word distributions, for the first experiment a very intuitive way of sampling was chosen, sometimes referred to as ‘bags of words’: the goal was to pick the words randomly, one by one, from the subsequent texts. This type of sampling provides a good approximation to the original vocabulary distribution in the text, but one has to remember that, at the same time, it destroys the original syntactic and lexical contexts of particular words.

The research procedure was as follows. For each text in a given corpus, 500 randomly chosen single words were concatenated into a new sample. These new samples were analyzed using the classical Delta method; the percentage of attributive success was regarded as a measure of effectiveness of the current sample length. The same steps of excerpting new samples from the original texts, followed by the stage of ‘guessing’ the correct authors, were repeated for the length of 500, 600, 700, 800, ..., 20,000 words per sample.

The results for the corpus of 63 English novels are shown on Fig. 1. The observed scores (black circles on the graph; grey circles will be discussed below) clearly indicate the existence of a trend.
Figure 2: Dependence of attribution accuracy and length of text samples: 66 German novels.

Figure 3: Dependence of attribution accuracy and length of text samples: 69 Polish novels.

Figure 4: Dependence of attribution accuracy and length of text samples: 64 Hungarian novels.

(solid line): the curve, climbing up very quickly, tends to stabilize at a certain point, which indicates the minimal sample size for the best attributing rate. Although it is difficult to find the precise position of that point, it becomes quite obvious that samples shorter than 5,000 words provide a poor ‘guessing’, because they can be immensely affected by random noise. Below the size of 3,000 words, the obtained results are simply disastrous (more than 60% of false attributions for 1,000-word samples may serve as a convincing caveat). Other analyzed corpora showed quite similar shapes of the ‘learning curves’, although some interesting differences also could be noticed.

In particular, the overall achieved attribution effectiveness was varying: among the modern novel corpora, Hungarian (Fig. 4) gained lower scores than English, then came German (Fig. 2), and Polish (Fig. 3). This phenomenon has already been observed in previous cross-language studies (Rybicki and Eder, 2011; Eder, 2011; Eder and Rybicki, 2012). The accuracy rates of both Ancient prose corpora (Fig. 5, 6) were fairly satisfying, Latin being slightly less attributable than Greek. Similar divergences could be observed in poetic corpora: ‘guess-
ing’ scores for English epic poems (Fig. 7) were substantially higher than for Greek poetry (Fig. 8), and very similar to Latin poetry (not shown).

The critical point of attributive saturation could be found at about 5,000 words per sample in most of the corpora analyzed (and no significant difference between inflected and non-inflected languages could be observed). However, there were also some exceptions. First of all, the corpus of Latin prose exhibited a significant improvement in resistance to short samples (its minimal effective sample size was of some 2,500 words; cf. Fig. 5, black circles). At the same time, the Latin corpus showed a very clear and distinctive trend of increasing accuracy followed by fairly horizontal scores of statistical saturation. In the other corpora, especially in Polish novels, the ‘learning curves’ gained their saturation somewhat slowly and less distinctively.

The behavior of the poetic corpora of English and Latin should also be commented upon. At first glance, English epic poetry required a promising amount of 3,000 words per sample for reliable attribution (Fig. 7). However, the number of unpredictable and very poor scores scattered randomly on the plot (cf. the outliers much below the trend line) suggest that attributing English poems shorter than 5,000 words might bring about a risk of severe misclassification. The picture of Latin epic poetry (not shown) was surprisingly similar.

Another peculiarity has been observed in both Greek corpora – prose and poetic (Fig. 6, 8). In both cases, after the usual quick advance of performance due to increasing length of text samples, the attributive accuracy happened to decrease for some time, and then to stabilize. It is difficult to explain this phenomenon; however, a similar behavior of Greek has been observed in a recent study (Eder 2011: 103-5), where a systematic error in corpus slightly improved the accuracy of attribution.

Speaking of a performance stabilization above a certain sample length, the ‘guessing’ scores for each corpus analyzed also seem to show that effectiveness would not increase in samples exceeding 15,000 words. This is also a valuable observation, suggesting that there are limits to Hoover’s statement that ‘for statistical analysis, the longer the text the better’ (Hoover, 2001).
5 Experiment II: passages

The results of the first experiment were quite disappointing, to say the least. They might easily lead to the suspicion that the ‘bags of words’ type of sampling was a decisive factor here, since this way of combining samples breaks the original sequences of words with all their possible syntactic relations. A variant of the above experiment was prepared, then, to test the possible impact of sampling on attribution performance.

The way of preparing samples by extracting a mass of single words from the original texts seems to be an obvious solution for the problem of statistical representativeness. In most attribution studies, however, shorter or longer passages, or blocks, of disputed works are usually analyzed, either randomly chosen from the entire text, or simply truncated to the desired size (Hoover, 2003: 349; Reynolds et al., 2012; etc.). The purpose of the second experiment was to test the performance of this typical sampling as opposed to extracted ‘bags of words’. The whole procedure was repeated step by step as in the previous test, but now, instead of collecting individual words, sequences of 500 words (then 600, 700, 800, ..., 20,000) were chosen randomly from the original texts.

The excerpted virtual samples were analyzed using the same classification method, the same number of iterations, and the same number of frequent words as in the first experiment, but the results turned to be significantly different. The differences become evident when the final scores of the two experiments are represented on shared graphs (Fig. 1–8). The grey circles on the graphs and the dashed trend lines show the effectiveness of the ‘passage’ type of sampling, as opposed to the black circles followed by the solid trend lines of the ‘bags of words’. Despite minor discrepancies, three main observations could be made here that seem to be applicable to all the corpora examined:

1. For each corpus analyzed, the attribution accuracy obtained in the second experiment (excerpted passages) was always worse than the scores described in the first experiment, relying on the ‘bags of words’ type of sampling. This is counter-intuitive, and it means that the procedure of ex-
 excerpting a mass of single words as described in the first experiment was not responsible for the considerably poor results. Quite the opposite, this way of sampling turned to be a better solution. The results cannot be simply explained away by the fact that artificially excerpted samples (‘bags of words’) are no longer tied to a specific local context in a text. It is a well-known fact that topic strongly intervenes with authorial style (Stamatatos, 2009; Luyckx and Daelemans, 2011), and the same can be said of narrative and dialogic parts of novels (Burrows, 1987; Hoover, 2001). The observed phenomenon, however, was noticeable irrespective of the assessed topics or genres.

(2) For ‘passages’, the dispersion of the observed scores was always wider than for ‘bags of words’, indicating a bigger impact of random noise for this kind of sampling. Certainly, as above, the effect might be due to the obvious differences in word distribution between narrative and dialogue parts in novels (Fig. 1–4); however, the same effect was similarly strong for non-literary prose in Latin and Greek (Fig. 5–6) and in English poetry (Fig. 7), or even substantially stronger in Greek poetry (Fig. 8).

(3) The distance between the both trend lines – for ‘words’ and for ‘passages’ – was varying noticeably across different corpora. At the same time, however, there was also some regularity in this variance, quite clear in the corpora of novels (Fig. 1–4) and probably related to the degree of inflection of the languages analyzed. Namely, the more inflected the language, the smaller the difference in correct attribution between both types of sampling: the greatest in the English novels (Fig. 1 grey circles vs. black), the smallest in the Hungarian corpus (Fig. 4). Both prose corpora of Latin and Greek – two highly inflected languages – also fitted the model (Fig. 5–6). The only exception were the corpora of English and Greek poetry, where the results were ambiguous (Fig. 7–8).

6 Experiment III: chunks

At times we encounter an attribution problem where extant works by a disputed author are doubtless too short to be analyzed in separate samples. The question is, then, if a concatenated collection of short poems, epigrams, sonnets, book reviews, notes etc. in one sample (cf. Eder and Rybicki, 2009; Dixon and Mannion, 2012) would reach the effectiveness comparable to that presented above? And, if concatenated samples are suitable for attribution tests, do we need to worry about the size of the original texts constituting the joint sample?

To examine the usefulness of concatenated samples, an experiment slightly different from the previous two was prepared. Provided that 8,000 words or so turned to be quite enough to perform a reliable attribution (see above), in the present approach the size of 8,192 words was chosen to combine samples from shorter chunks. In 12 iterations, a number of word-chunks, or n-grams, were randomly selected from each text and concatenated: 4,096 chunks of 2 words in length (bi-grams), 2,048 chunks of 4 words (tetra-grams), 1,024 chunks of 8 words, 512 chunks of 16 words, and so on, up to 2 chunks of 4,096 words. Thus, all the samples in question were 8,192 words long.

The obtained results were roughly similar for all the languages and genres tested, and somehow correlated with the results of the two previous experiments. As shown in Fig. 9 (for the corpus of English novels), the effectiveness of ‘guessing’ depended to some extent on the word-chunk size used. The attribution scores were worse for long chunks within a sample (4,096 words or so) than for bi-grams or 4-word chunks. This decrease of performance was linear: the shorter a chunk, the better the ‘guessing’ scores. Interestingly, the difference between the effectiveness of the shortest and the longest chunks followed the difference between ‘bags of words’ and ‘passages’ in the first two experiments (Fig. 1). Certainly, this is easy to explain, since single words are the extreme case of short chunks, and ‘passages’ are in fact very long chunks. The results of this experiment seem to fill the gap between the two trend lines for ‘words’ and ‘passages’ presented above. This remark applies to all the corpora tested.

The results seem to be fairly optimistic, because there is no substantial difference in attribution between, say, a few chunks of 500 words combined in one sample, and a dozen concatenated chunks.
of 100 words. It suggests that in real attribution studies, concatenated samples would display a very good performance.

One has to remember, however, that the above simulation of concatenated short texts was artificial. The chunks were excerpted randomly from long texts, regardless of sentence delimitation, etc. What is even more important, short literary forms, like epigrams or aphorisms, have their own stylistic flavor, substantially different from typical literary works (Mauntner, 1976). Short literary forms are often masterpieces of concise language, with a domination of verbs over adjectives, particles and so on, with a proverbial witty style, and with a strong tendency to compression of content. Thus, a collection of aphorisms will certainly have different word frequencies than a long essay written by the same author; similarly, a collection of short mails will differ from an extensive epistle, even if they have been addressed to the same addressee. For that reason, further investigation is needed here.

### 7 Evaluation

This section can be safely skipped by most readers. It does not contribute to the general picture as discussed in this study; it is aimed to provide some insight into the evaluation procedures behind the experiments. Also, it introduces a number of cross-check tests for other machine-learning techniques and alternative style markers. It should be stressed, however, that discussing all the tests that have been conducted in the evaluation stage (dozens of tests involving some 2 million iterations and almost 100 million single ‘guesses’) is simply unrealistic. Thus, the corpus of English novels and the experiment with ‘bags of words’ will be used as a case study in this section.

Among the aforementioned drawbacks of Delta, the particularly painful one is that the choice of texts to be included in the ‘training’ set is arbitrary, or depending on subjective decisions which works by a given author are ‘representative’ for his/her stylometric profile. Even if this choice is fully automatic (e.g. when the samples for the ‘training’ set are chosen randomly by the machine), it is still
very sensitive to local authorial idiosyncrasies. In consequence, the estimated classification accuracy might overfit the actual behavior of input data.

Advanced machine-learning methods of classification routinely try to eschew the problem of model overfitting due to possible inconsistencies of training samples. The general idea of such ‘cross-validation’ tests is to replicate the original experiment multiple times with random changes to the composition of both the ‘training’ and ‘test’ sets: in a number of random swaps between the samples, followed by the stage of classification, one obtains an average behavior of the corpus. A commonly accepted solution, introduced to stylometry from exact sciences, is 10-folded cross-validation (Zhao and Zobel, 2005; Juola and Baayen, 2005; Stamatatos, 2008; Koppel et al., 2009; Jockers and Witten, 2010; Luyckx and Daelemans, 2011, and so forth). It has been shown, however, that mere 10 swaps might be far too little to betray potential model overfitting: a set of cross-language attribution tests with a large number of random re-compositions of the ‘training’ and ‘test’ sets have shown substantial inconsistencies for some of the analyzed corpora (Eder and Rybicki, 2013).

To avoid homeopathic solutions, then, and to perform a robust cross-validation, the evaluation procedure applied in this study was as follows. For each type of sampling assessed, for each corpus, and for each sample size, the texts included into the ‘training’ set were chosen randomly (one text per author) in 100 independent iterations. In consequence, the classification test was applied 100 times, and the average attributive success rate was recorded. The whole procedure was repeated for every single sample size tested: 500, 600, 700, ..., 20,000 words (using ‘bags of words’ in the first experiment, ‘passages’ in the second one). This approach could be compared to an extreme version of a 100-fold cross-validation (extreme, because in each iteration the whole ‘training’ set was thoroughly re-composed). Certainly, the whole task was computationally very intensive.

The cross-validated attribution accuracy scores for the English corpus are shown on Fig. 10 (averaged performance rates represented by grey circles, a trend – by black solid line). The overall attributive success is indeed worse than for the non-validated variant (Fig. 1), but model overfitting is not substantial. Much more important for the scope of this study, however, is the shape of the ‘learning curve’ and the point where the curve becomes saturated – and they are almost identical (Fig. 1 vs. 10). Thus, the conclusions concerning the minimal sample length seem to be validated, at least for Delta.

To test the possible impact of different classification algorithms on the sample size effect, another series of check tests, followed by 100-fold cross-validation, have been performed. The methods tested were SVM (claimed to be the most accurate attribution method so far), and k-NN. The comparison of their performance is shown again on Fig. 10. It turns out that SVM indeed outperforms other methods, Delta being an undisputed runner-up (Fig. 10, dashed line vs. solid line). Unexpected bad performance of k-NN (Fig. 10, dotted line) can be explained by the fact that the ‘training’ set contained one sample per author only – and these settings are a priori suboptimal for k-NN. Despite the differences between particular classifiers, the shapes of the ‘learning curves’ remain stable. This shows that the short sample problem cannot be easily by-passed by switching to sophisticated machine-learning algorithms.

Yet another series of check tests have been conducted to examine the performance of alternative style-markers as confronted with short samples. Character n-grams seem to be a particularly promising proposition here. They are claimed to be robust and language-independent (Koppel et al., 2009; Stamatatos, 2009), and resistant to untidily prepared corpora, e.g. containing a mass of misspelled characters (Eder, 2013). Besides character n-grams, potentially powerful markers also include syntax-based features, such as automatically recognized

---

2 Since the classical Delta procedure counts z-scores based on the ‘training’ set alone, and then applies the variables’ means and standard deviations to the ‘test’ set, the permutation of both sets somewhat impacts the z-scores and thus, possibly, the final results as well. In view of this, a parallel series of tests with z-scores calculated for both sets has been performed. The differences between those two approaches were noticeable yet not significant.
The results of 100-fold cross-validation for twelve possible combinations of four different style-markers (MFWs, character 3-grams, character 4-grams, POS-tag 3-grams) and three classifiers (Delta, SVM, k-NN) are presented on Fig. 11. No matter which classifier was used, MFWs proved to be the most accurate solution; then came POS-tag 3-grams, then character-based markers. The attribution success scores for SVM based on MFWs clearly suggest that this particular combination provides the best performance in the corpus of English novels; Delta combined with MFWs is almost as good. The shape of the ‘learning curves’ does not betray any increase of resistance to short samples when alternative style-markers are used.

8 Discussion

8.1 methodological remarks

The experiments described above were designed to be as reliable as possible; however, some arbitrary choices were inescapable. These are as follows:

(1) Choice of style-markers. In non-traditional authorship attribution, many different features of style have been tested: indexes of vocabulary richness, measures of rhythmicity, frequencies of function words, frequencies of parts-of-speech tags, and so on (an extensive list of possible style-markers is provided by Koppel et al., 2009: 11–13). In the present study, a few possible types of markers were tested, with preference given to the classical solution, i.e. vectors of frequencies of the most frequent words (MFWs). It is possible that in some languages, alternative style-markers might exhibit better performance.

(2) Number of style-markers to be analyzed. It is true that the effectiveness of nearest neighbor

3For a number of reasons, ranging from availability of well-trained taggers for particular languages (or particular historic periods), to substantially different grammar between the languages addressed in the present study, the check tests with POS-tag n-grams were limited to the corpus of English novels and the corpus of Latin poetry only. The software used for tagging were the Stanford NLP Tools (for English) and TreeTagger (for Latin).
classifications, including Delta, are very sensitive to the number of features analyzed (Jockers and Witten, 2010). Unfortunately, as has been shown (Rybicki and Eder, 2011), there is no universal vector of MFWs suitable for different languages or genres. In the present study, 200 MFWs were used for each test; this arbitrary choice was a compromise between the small number of function words most effective for Latin, and the very long vectors of 1,000 or more words optimal for the corpus of English novels. To test the possible impact of this chosen number of 200 MFWs, an additional experiment was prepared using different MFW vectors. As shown in Fig. 12 (English novels), the overall attribution effectiveness depends indeed on the vector of MFWs analyzed, but – importantly – the shape of the ‘learning curves’ and the point of statistical saturation are stable regardless different MFWs settings.

(3) Number of texts tested and choice of ‘training’ and ‘test’ set members. It has been proven that the effectiveness of attribution depends on corpus size and particularly on the number of authors tested (Luyckx, 2010; Luyckx and Dealemans, 2011). In short: a 2-class authorship attribution case needs less textual data than a 100-class case. Despite the importance of this problem, it was not addressed in the present study. Since in particular corpora the number of authorial classes was strictly constant for all the tests performed, the obtained results are not affected by this issue (it is true, however, that any cross-language conclusions are limited, since the number of authorial classes was not fixed across the corpora).

(4) Choice of particular texts included in each corpus. One of the common beliefs demonstrated in technically-oriented authorship studies is the slightly naive assumption that keeping the number of authorial classes constant and/or using strictly the same amount of training data guarantees the reliability of the experiment. Unfortunately, collecting textual data is probably the most unreliable stage of any corpus study; the uncontrollable factors are countless here. Unequal availability (representativeness) of texts in electronic version, stylistic differentiation in particular national literatures, possible existence of writers-outliers displaying either exceptional stylistic imagery or extreme dull-
ness in style, inherent linguistic differences between national corpora, and so on – they all make any definite conclusions simply unfeasible. This corpus-related problem might have had a significant impact on the results presented in the present study, with little hope to eschew what is an inherent feature of textual data. This was also the reason why the corpora were not strictly of the same size: the cost of acquiring such a collection of allegedly ‘comparable’ corpora would be really high, and the results would be still questionable to some extent.

(5) Choice of a technique of attribution. The scores presented here, as obtained with classical Delta procedure, turned out to be slightly different when solved with other nearest neighbor classification techniques (Fig. 11). However, similarly to different vectors of MFWs (Fig. 12), the shapes of all the curves and the points where the attribution success rates become stable are identical for each of these methods. The same refers to different combinations of style-marker settings – although different settings provide different ‘guessing’, they never affect the shape of the curves. Thus, since the obtained results are method-independent, this leads us to a conclusion about the smallest acceptable sample size for future attribution experiments and other investigations in the field of stylometry.

A few words should be added about explanatory multidimensional methods that are traditionally used in authorship attribution: Principal Components Analysis, Multidimensional Scaling, and Cluster Analysis. In these methods, there is no direct way of examining the short sample effect and its impact on attribution. However, a very simple test might be performed to show the importance of this problem. Using a method of excerpting ‘bags-of-words’ as introduced above, one can perform a series of, say, Principal Components Analyses and compare the obtained plots. The results of two PCA tests of 1,000 randomly chosen words from the same corpus of 28 English novels are shown in Fig. 13 and 14. In both cases, 100 MFWs were used. Without deciding which of the two pictures is more likely to be ‘true’, the substantial differences between them are quite obvious. In the first picture, one can distinguish discrete clusters for Eliot and Austen, a group of texts in the bottom part of the plot (including Fielding and Sterne), and a common cloud of the remaining samples. In the second picture, besides the central cloud, three distinguishable clusters are noticeable: for Charlotte Bronté, for Thackeray, and for Richardson/Fielding. The results of this considerably simple experiment show how misleading an explanatory interpretation of points scattered on a plot might be. What is worse, there are a number of real-life attribution cases based on samples of about 1,000 words (the Federalists Papers, and Scriptores historiae Augustae, to name but a few); approaching them with PCA or MDS might bring about a risk of being substantially mistaken.

8.2 sample size

A characteristic paradox of non-traditional authorship attribution is that the most accurate state-of-the-art methods need very long samples to show their power, while in real life, there are not so many anonymous novels, as opposed to countless anonymous ballads, limericks, or critical notes. Thus, paradoxically, the more helpful stylometry could be to supplement traditional literary criticism, the more unhelpful it seems to be in many cases. For that reason, the results obtained in the present study – a few thousand words per sample, at least, to perform an attribution test – will not satisfy most literary scholars.

Certainly, this leads to the question how to interpret the obvious discrepancy between these unsatisfactory results and several classic attribution studies where short samples have been used with success. An extreme example is provided by Burrows, who observed that a poem of only 209 words by Aphra Behn was correctly assigned (Burrows, 2002: 278). The study by Koppel et al. (2009) goes even further, showing that a corpus of very short blog posts (of 217–745 words in length) displays an accuracy of 38.2–63.2%, depending on the classification method used.4 These results are very impressive; they show how much authorial information can be retrieved

4The scores in question were obtained using 512 function words as style-markers. More sophisticated features, such as parts-of-speech tags or 1,000 character trigrams with ‘highest information gain in the training corpus’, displayed
from just a couple of paragraphs of running text. On the other hand, one should also remember that the remaining 61.8–36.8% of samples used in this study were wrongly classified, a crucial pitfall in real attribution cases. The commonly known thing is that natural language processing tools and techniques, such as parts-of-speech taggers or syntax parsers, easily achieve a few dozen percent of accuracy, but the actual problem is to gain every next percent (and the difficulty increases exponentially rather than proportionally). The same can be said about the accuracy of stylometric methods.

An interesting insight to this problem is provided by a detailed inspection of final rankings of candidates obtained in the above two experiments. Namely, for all the corpora, no matter which method of sampling is used, the rankings are considerably stable: if a given text is correctly recognized using an excerpt of 2,000 words, it will be also ‘guessed’ in most of the remaining iterations; if a text is successfully assigned using 4,000 words, it will be usually attributable above this word limit, and so on. At the point of statistical saturation, where increasing the length of sample does not improve the attribution effectiveness, only a few remaining texts are finally linked to their actual authors. E.g., in the case of English novels, a fingerprint of both Charlotte and Anne Brontës was considerably well recognizable even for short samples, then came Richardson, Dickens (David Copperfield, Pickwick Papers), Eliot, Galsworthy, Austen, again Dickens (Hard Times, Oliver Twist, Great Expectations). It was very hard to attribute Hardy, but particularly long samples were needed to distinguish novels written by James.

In other words: in some texts, the authorial fingerprint is rather easily traceable, even if very short samples are used. This was the case of the Latin prose corpus (Fig. 5), where some particularly strong style-markers made it possible to distinguish authors using only 2,500 excerpted words per sample. However, there are also some other texts where the authorial signal is hidden, or covered by noise, and it needs to be carefully extracted using very long samples (as was the case of many texts in the Polish corpus, cf. Fig. 3). The only problem is, however, that in real authorship attribution we have no a priori knowledge which category an anonymous text belongs to. Thus, it seems that a minimal sample size for a reliable attribution does not begin at the point where the first correct ‘guesses’ appear, but where the most problematic samples finally recognize their own authors.

9 Conclusions

The main research question approached in this study was how much data is sufficient to recognize authorial uniqueness. There was no clear answer, though. It seems that for corpora of modern novels, irrespective of the language tested, the minimal sample size is some 5,000 words (tokens). Latin prose required only 2,500 words, and Ancient Greek prose just a little more to display their optimal performance. The results for the three poetic corpora (Greek, Latin, English) proved ambiguous, suggesting that some 3,000 words or so would be usually enough, but significant misclassification would also occur occasionally. Thus, the results depended on genre rather than on language examined.

Another conclusion is connected with the method of sampling. Contrary to common sense, randomly excerpted ‘bags of words’ turned to be much more effective than the classical solution, i.e. using original sequences of words (‘passages’) of a desired size. This means that dealing with a text of 20,000 words in length, it is better to select 10,000 words randomly than to rely on the original string of 20,000 words. Again, it is better to excerpt 10 samples of 10,000 randomly chosen words from a whole novel than to rely on 10 subsequent chapters as samples. Certainly, the obtained results are partially dependent on the language tested (the level of inflection being one of the suspected factors), genre, literary epoch, number of assessed texts, number of authors, and probably also on particular selection of texts included in a corpus. Nonetheless, the results provide a convincing argument in favor of using randomly excerpted ‘bags of words’ rather than relying on arbitrarily chosen ‘passages’ (blocks of words).
This also means that some of the recent attribution studies should be at least re-considered. Until we develop style-markers more precise than word frequencies, we should be aware of some limits in our current approaches, the most troublesome of these being the limits of sample length. As I tried to show, using 2,000-word samples will hardly provide a reliable result, to say nothing of shorter texts.

The present study does not contradict the soundness of undertaking difficult attribution cases. Quite the contrary, it tries to show that stylometry has not said its last word, and there is an urgent need to develop more reliable techniques of attribution. Promising propositions include using part-of-speech tags instead of word frequencies (Hirst and Feiguina, 2007), sophisticated techniques of estimating level of uncertainty of word counts (Hinneburg et al., 2007), and last but not least, the method of using recall/precision rates as described in the above-cited study by Koppel et al. (2009). Attributing short samples is difficult, but arguably possible, if it is approached with the awareness of the risk of misclassification.

References


Jockers, M. L., Witten, D. M. and Criddle,


Rudman, J. (2003). Cherry picking in nontradi-
‘Stylo’: a package for stylometric analyses

Maciej Eder  
Pedagogical Univ. of Kraków

Jan Rybicki  
Jagiellonian University

Mike Kestemont  
University of Antwerp

November 12, 2015

Abstract

The ‘stylo’ package (Eder et al. 2013) provides easy-to-use implementations of various established analyses in the field of computational stylistics, including non-traditional authorship attribution, genre recognition, style development (“stylochronometry”), etc. The package includes a number of explanatory methods provided by the function \texttt{stylo()} (multidimensional scaling, principal component analysis, cluster analysis, bootstrap consensus trees). Additionally, a number of supervised machine-learning methods are available via the function \texttt{classify()} (Delta, support vector machines, naive Bayes, k-nearest neighbors, nearest shrunken centroids). The \texttt{rolling.delta()} function analyses collaborative works and tries to determine the authorship of fragments extracted from them. The function \texttt{rolling.classify()} offers a more flexible interface to sequential classification of collaborative works. The \texttt{oppose()} function performs a contrastive analysis between two given sets of texts: among other things, it generates lists of words significantly \textit{preferred} and \textit{avoided} by one or more authors in comparison to the texts by another author (or a set of them).

Keywords

stylometry, computational stylistics, authorship attribution, cluster analysis, dendrogram, bootstrap consensus trees, PCA, MDS, k-NN, SVM, NSC, naive Bayes, Delta, Zeta, rolling stylometry

Contents

1 Introduction 3
2 Installation 4
3 Functions provided 5
4 stylo() 7
4.1 Corpus preparation 7
4.2 Starting the function 9
4.3 Options available on GUI 9
4.3.1 Input 10
4.3.2 Language 11
4.3.3 Features 11
4.3.4 MFW settings 12
4.3.5 Culling 13
4.3.6 Statistics 14
4.3.7 Distances 15
4.3.8 Sampling 17
4.3.9 Graphs 18
4.3.10 Plot area 19
4.3.11 PCA/MDS 19
4.3.12 PCA flavour 20
4.3.13 Various 20

5 classify() 21
5.1 Corpus preparation 21
5.2 Calling the function 22
5.3 Options 22
5.3.1 Options inherited from stylo() 22
5.3.2 Statistics 22
5.3.3 General 23
5.3.4 SVM options 23
5.3.5 k-NN options 23
5.3.6 Output: general 23

6 rolling.delta(), rolling.classify() 24
6.1 Corpus preparation 25
6.2 Calling the function 25
6.3 Options 25
6.3.1 Features 25
6.3.2 Sampling 25
6.3.3 Colors 25

7 oppose() 26
7.1 Corpus preparation 26
7.2 Calling the function 26
7.3 Options 26

8 Options unavailable on GUI 27
8.1 Cluster analysis linkage 27
8.2 Network analysis support 27
8.3 Undocumented arguments 28
Stylometric studies, in all their variety of material and method, have two features in common: the electronic texts they study have to be coaxed to yield numbers, and the numbers themselves have to be processed via statistics. Sometimes, the two actions are two independent parts of a given study. To give the simplest example, one piece of software is used solely to compile word frequency lists; then, one of the many commercial statistics packages takes over to extract meaning from this mass of words, draw graphs etc.

Yet, as stylometrists have begun to produce statistical methods of their own – to name but a few, Burrows’s Delta, Zeta and Iota (Burrows, 2002, 2007) and their modifications by other scholars (Argamon, 2008, Craig and Kinney, 2009, Hoover, 2004a, 2004b) – commercial software, despite its wide array of accessible methods, becomes something of a straightjacket. This is why a number of dedicated stylometric solutions have appeared, targeting the specific analyses frequently used in this community. Hoover’s Delta, Zeta and Iota Excel spreadsheets are pioneering examples of this approach (Hoover, 2004b). Constantly developed since 2004, they have at least two major assets: they do exactly what the stylometrist wants (with several optional procedures) and they only require fairly standard – although proprietary – spreadsheet software. This has been especially helpful for uses in specialist workshops and classrooms: the student only needs additional (and, often, free) software to produce word frequency lists and (s)he is ready to go. Yet Excel imposes one important limitation: it is very demanding from the point of view of memory usage. Moreover, the two-stage nature of the process (a separate piece of software prepares word lists that can be later automatically imported into the spreadsheet) might be problematic, because it takes an experienced Visual Basic programmer to make Excel itself extract the various frequency dictionaries needed.

In this respect, Juola’s JGAAP can directly import texts in a variety of formats and perform a variety of authorship attribution tasks using an impressive variety of statistical methods (Juola et al., 2008). These can be further expanded by experienced programmers in Java. Java is also the language of another software solution which takes an even broader approach: Craig’s Intelligent Archive is able to perform a number of standard stylometric procedures, but it can also be used as a corpus organizer. Once the initial work of registering texts is done, it enables a versatile combination of individual texts and groups of texts (Craig and Kinney, 2009).

Since contemporary stylometry uses either stand-alone dedicated programs custom-made by stylometrists, or applies existing software, the stylo package can be situated somewhere in-between: the powerful open-source statistical programming environment
R provides, on the one hand, the opportunity of building statistical applications from scratch, and, on the other, allows less advanced researchers to use ready-made scripts and libraries (cf. http://www.R-project.org). In our own stylometric adventure with R, one of the aims was to build a tool (or a set of tools) that would combine sophisticated state-of-the-art algorithms for classification and/or clustering with a user-friendly, “point-and-click” interface. In particular, we wanted to implement a number of popular multidimensional methods to be used by scholars without advanced programming skills. It soon became evident that once our R scripts were provided with a graphical user interface and modest documentation, they lent themselves well to classroom use. In our experience, this suite of tools offers an excellent way to work around R’s typically steep learning curve, without losing anything of the power of the environment – namely R’s considerable computing power and speed.

A crucial point in building the interface was to make sure that all stages of a typical stylometric analysis – from loading texts to visualizing the results – could be performed from within a single function. The \texttt{stylo()} function, for instance, does all the work: it processes electronic texts to create a list of all the words used in all texts studied, with their frequencies in the individual texts; normalizes the frequencies with $z$-scores (if applicable); selects words from the desired frequency ranges; performs additional procedures that might improve attribution, such as Hoover’s (2004a, 2004b) automatic deletion of personal pronouns and “culling” (automatic removal of words too characteristic for individual texts); compares the results for individual texts; performs a variety of multivariate analyses; presents the similarities/distances obtained in tree diagrams; and finally, produces a bootstrap consensus tree (a new graph that combines many tree diagrams for a variety of parameter values). It was our aim to develop a general platform for multi-iteration stylometric tests; for instance, an alternative script derived from the function \texttt{classify()} produced heatmaps to show the degree of Delta’s success in attribution at various intervals of the word frequency ranking list (Rybicki and Eder, 2011).

The last stage of the interface design was, firstly, to add a GUI (since some humanists might be allergic to the command-line mode provided by R) and, secondly, a host of various small improvements (like saving and loading the parameters for the most recent analysis, a wide choice of graphic output formats, etc.). Nevertheless, advanced users could still easily switch off the GUI and embed the functions provided by the “stylo” library in their own scripts.

2 Installation

Make sure you are connected to the internet. Launch R. Type \texttt{install.packages("stylo")} in the console. Whenever you start a new R session, type \texttt{library(stylo)}. This will automatically load all the functionality provided in the package (see below). If you are very lazy and only use R for stylometric purposes, you can find your \texttt{Rprofile.site} configuration file (in \texttt{R/R-<your R version here>/etc}), open it with administrator privileges and insert the line \texttt{library(stylo)} there. In this case, the “stylo” library will be loaded at the start of each R session and you can start invoking the particular functions right away.
3 Functions provided

The most important tools included in this package are distributed over the following functions:

- `stylo()`
- `classify()`
- `oppose()`
- `rolling.delta()`
- `rolling.classify()`

The next sections of this manual describe these four functions together with all the different input options they can take. If you want to get a general overview of these four functions, type `help(stylo)`, `help(classify)`, etc., and a help window will appear. More advanced users might be interested in some other functions provided by the library. Generally speaking, they are a great deal of lower-level functions which are called automatically from inside the upper-tier functions, such as `classify()`, `oppose()`, etc. This lower-level functionality can of course be used for developing your own scripts and functions. These include:

- `assign.plot.colors`
- `define.plot.area`
- `delete.markup`
- `delete.stop.words`
- `dist.argamon`
- `dist.cosine`
- `dist.delta`
- `dist.eder`
- `dist.simple`
- `draw.polygons`
- `gui.classify`
- `gui.oppose`
- `gui.stylo`
- `load.corpus.and.parse`
- `load.corpus`
- `make.frequency.list`
- `make.ngrams`
• make.samples
• make.table.of.frequencies
• parse.corpus
• parse.pos.tags
• perform.culling
• perform.delta
• perform.knn
• perform.naivebayes
• perform.nsc
• perform.svm
• stylo.default.settings
• stylo.pronouns
• txt.to.features
• txt.to.words.ext
• txt.to.words
• zeta.chisquare
• zeta.craig
• zeta.eder

In most cases, these lower-level functions provide very basic processing functionality and they are therefore not intended to be invoked by everyday users. Hence, they will not be discussed in this manual. However, if you are interested how they work and how to use them, you can invoke the help pages for these functions: help(load.corpus), help(make.ngrams), etc. Help pages routinely contain some insightful examples as to how to use the code: refer to them if you want to understand what a particular function does. The examples can be copy-pasted into an active R console. (Don’t be afraid of the lines ‘## Not run’ – they prevent R to run some automatic checks on interactive functions; you can use these examples safely).

Apart from functions, the package ‘stylo’ (ver. 0.6.1) contains three datasets that can be used to start playing with stylometric methods without any actual texts. The datasets are as follows:

• novels
• galbraith
• lee

The first dataset contains 9 full-size novels by Jane Austen and the Brontë sisters. The second and the third set contains computed tables of word frequencies for 26 and 28, resp., contemporary novels that for copyright-related reasons could not be made available in their original format. A detailed description of the datasets can be retrieved via help(novels), help(galbraith) and help(lee).
4 stylo()

This is currently the main tool in the package. The function `stylo()` is meant to enable users to automatically load and process a corpus of electronic text files from a specified folder, and to perform a variety of stylometric analyses from multivariate statistics to assess and visualize stylistic similarities between input texts. This function provides explanatory analyses; any users interested in machine-learning supervised methods might want to skip this section and go to `classify()`, below.

`stylo()` will typically be used to produce a most-frequent-word (MFW) list for the entire corpus. Next, it will acquire the frequencies of the MFWs in the individual texts to create an initial matrix of words (rows) by individual texts (columns): each cell will contain a single word’s frequency in a single text. Subsequently, it will normalize the frequencies: it selects words from the the desired frequency ranges for an analysis (this is also saved to disk as `table_with_frequencies.txt`) and it will perform additional processing procedures (automatic deletion of personal pronouns and culling, see 4.3.5 below) to produce a final wordlist for the actual analysis (this information is saved to disk in the current working directory as `wordlist.txt`). It then compares the results for individual texts, performing e.g. distance calculations and using various statistical procedures (cluster analysis, multidimensional scaling, or principal components analysis). Finally, the function will produce graphical representations of distances between texts and it will write the resulting authorship (or similarity) candidates to a logfile (`results.txt`) in the current working directory. When the consensus tree option is selected, the script produces virtual cluster analyses for a variety of parameters, which then produce a final diagram that reflects a compromise between the underlying cluster analyses.

4.1 Corpus preparation

The procedure of loading corpora as described immediately below is probably the best way to start doing your first analyses. However, experienced users of R sooner or later will discover that input data structures (corpora, vectors of features, tables of frequencies) can be passed as R objects directly from, say, other functions, without any interaction with texts files. Refer to section 9.1 for details.

Each project requires a separate and dedicated working folder. You will want to give it a meaningful name (like `SanskritPoetry11` rather than `Blah-blah`), since the
name of the folder will appear as the title in your graphs generated by the function. By default, the results of your analyses and other useful files will be written automatically to this folder. The actual text files for your analyses must be placed in a subfolder in the working directory, named corpus (Note: all file names are case sensitive!). All functions in this tool suite expect to find at least two input texts for their analyses.

The text files need to follow the following naming syntax: category_title.txt. For people working in authorship attribution, the category will capture a text’s authorial signature; other users, perhaps interested to compare a translators’ styles, should name their files translatorname_title.txt. Likewise, if you are looking for stylistic similarities between writers of the same gender, use gender_title.txt, etc. It is really important to use an underscore “_” (underscore) as a delimiter: e.g. colors on the final graphs will also be assigned according to strings of characters up to the first underscore in the input files’ names. (For further details and examples, type help(assign.plot.colors)). Consider the following examples, in which the classes are the authors’ names and authors’ gender, respectively:

ABronte_Agnes.xml
ABronte_Tenant.xml
Austen_Emma.xml
Austen_Pride.xml
Austen_Northanger.xml
Conrad_Nostromo.xml
Conrad_Lord.xml
Dickens_Pickwick.xml
...
M_Conrad_Lord_Jim.txt
M_Joyce_Dubliners.txt
F_Woolf_Night_and_day.txt
F_Woolf_Waves.txt
...

Everything that comes after the underscore (say, the short titles of novels) can be followed by any other information. Be careful with long names, however, since these might not fit in the graphs that will be generated. The texts must either be all in plain text format, or all in HTML, or all in TEI-XML (the latter two options have not been extensively tested so far, and should be used carefully).

A concise remark about possible encoding issues should also be added. If the operating system you use is Linux or Mac, you just need to make sure the texts are all in UTF-8 (aka Unicode). If your operating system is Windows, you have two options. Firstly, you might want to save all the texts in ANSI codepage, but you have to tread carefully if your machine runs one charset, say, Central European (1250) and your texts are in the Western European codepage (1252); in this respect, for instance, French is notoriously difficult (nous sommes vraiment désolés). Alternatively, you can convert your texts into Unicode (a variety of freeware converters are available on the internet), and to use an appropriate encoding option when launching the function, say, stylO() either by clicking the “UTF-8” button on GUI (beginners), or passing the argument encoding = “UTF-8” directly to the function (advanced users).
4.2 Starting the function

Start up R. At the prompt (where you see the cursor blinking), move to your folder (the main folder you will be working in, not the corpus subfolder) using the command `setwd()`. E.g.:

```
setwd("/Users/virgil/Documents/disputed-works-of-mine")
```

You can use either absolute paths (as in the above example), or relative ones, i.e. you can navigate directly from the current working directory. If you want to go, say, two levels up and then descend to a folder `first_experiment`, type:

```
setwd("../../first_experiment")
```

You can always check your current working directory typing `getwd()`. (If you use R app for Windows, you can set your directory by clicking the File menu: see Fig. 1; Mac OS users – click the Misc menu on your R console). Call the function by typing `stylo()` at the prompt and hitting enter. After a while, you should see a GUI box appear on the screen. Change as many options as you need. Since there are multiple tabs in the GUI, make sure you only click the OK button after you’ve set the parameters in all the tabs. Shortly afterwards, you will see the names of the files processed appear in the R console, followed by other (technical) information. Depending on the size of your corpus, this step might take a few minutes. When the process is completed without major errors, you will typically see a diagram on your screen; otherwise, a graphic file (you can choose one or more format if you like) will be saved in your working directory (at better resolution than the onscreen version, so use this for your publication), and you can start exploring the other `stylo()` output files there.

4.3 Options available on GUI

As a first step, beginners should learn how to use the graphical user interface (GUI), which allows you to control the script’s main parameters without having to tamper with the actual code. However, if you do prefer to tamper with the code, you can call the function in batch mode: `stylo(gui=FALSE)`. In that case, before you start, you might want to visit the help pages via typing the command `help(stylo)`. Also, you should
be familiar with additional options that can, or rather should, be passed as arguments; they are listed on the margins of this document.

Whenever you use the GUI, each successful execution or “run” of the script will generate a `stylo_config.txt` file (saved in your working folder) which you can review (for instance, should you have forgotten the parameters you used in your last experiment). The parameter settings specified in this file will be retrieved at each subsequent run of the script, so that the user won’t have to re-specify their favorite settings every time. Please note that when you hover your cursor over the labels of each of the entries in the GUI, tool tips will appear that will help you understand the GUI. In the following sections we will discuss each of the different tabs in the `stylo()` GUI.

No matter if you decide using GUI or not, you can pass additional arguments from command-line. If the graphic mode is on, these “new” values will appear in the GUI and thus they will be still modifiable. Some examples include:

```r
stylo(mfw.min=300, mfw.max=300, analyzed.features="c", ngram.size=3)
stylo(gui=FALSE, analysis.type="MDS", write.png.file=TRUE)
stylo(mfw.min=100, mfw.max=1000, mfw.incr=100, analysis.type="BCT")
```

### 4.3.1 Input

This is where you specify the format of your corpus (see 4.1 above for more details about corpus preparation, and mind possible encoding issues). The available choices are:

- **plain text**: plain text files.
- **xml**: XML files; this option will remove all tags and TEI headers.
- **xml (plays)**: XML files of plays; with this option, all tags, TEI headers, and speakers’ names between `<speaker>...</speaker>` tags are removed.
- **xml (no titles)**: XML contents only: all tags, TEI headers, and chapter/section (sub)titles between `<head>...</head>` tags are removed.
- **html**: the option will attempt to remove HTML headers, menus, links and other tags.
• UTF-8: if you use Linux or Mac, this option is immaterial; however, if your
operating system is Windows, then you need to set it depending whether your
dataset is encoded in Unicode (then check the option), or in ANSI (then leave it
unchecked).

4.3.2 Language

This setting makes sure that pronoun deletion (see below) works correctly. If you decide
not to remove pronouns from your corpus (which is known to improve authorship
attribution in some languages), this setting is immaterial (unless you are using English; see immediately below).

• English: this setting makes sure that contractions (such as “don’t”) are not
treated as single words (thus “don’t” is understood as two separate items, “don” and “t”),
and that compound words (such as “topsy-turvy”) are not treated as one word
(thus “topsy-turvy” becomes “topsy” and “turvy”).

• English (contr.): this setting makes sure that contractions (such as “don’t”) are
treated as single words (thus “don’t” is understood as “don’t” and counted
separately), but compound words (such as “topsy-turvy”) are still not treated as
one word (thus “topsy-turvy” becomes “topsy” and “turvy”).

• English (ALL): this setting makes sure that contractions (such as “don’t”) are
treated as single words (thus “don’t” is understood as “don’t” and counted
separately), and that compound words (such as “topsy-turvy”) are treated as one
word (thus “topsy-turvy” becomes “topsy” and “turvy”).

• Latin: this setting makes sure that “v” and “u” are treated as discrete character
signs in Latin texts.

• Latin.corr: since some editions do not distinguish between “v” and “u”, this option
provides a consistent conversion of both characters to “u” in each text.

• CJK: Chinese, Japanese and Korean scripts, provided that the input data is
encoded in Unicode.

• Other: non-Latin scripts: Hebrew, Arabic, Cyryllic, Coptic, Greek, Georgian, Latin
phonetic, so far. Make sure your input data is in Unicode!

Please do note that for all other languages, apostrophes do not join words and
compound (hyphenated) words are split. This is not the ideal solution and will be
addressed as soon as we get to it.

4.3.3 Features

In many established approaches to stylometry, the (relative) frequencies of the most
frequent words (MFW) in a corpus are used as the basis for multidimensional analyses.
It has been argued, however, that other features are also worth considering, especially
word and/or character n-grams. The general idea behind such n-grams is to combine a
string of individual items into a partially overlapping, consecutive sequences of n of these
individual items. Given a sample sentence “This is a simple example”, the character
2-grams (“bigrams”) are as follows: “th”, “hi”, “is”, “s”, “i”, “is”, “s”, “a”, “a”, “a”, “s”, “i”, “im”, “mp”, etc. The same sentence split into bigrams of words reads “this
is”, “is a”, “a simple”, “simple example”. It has been heavily debated in the secondary literature whether the use of \( n \)-grams really increases the accuracy of stylometric tests (Hoover, 2002, 2003, 2012; Koppel et al., 2009; Stamatatos, 2009; Eder, 2011; Alexis et al., 2014). However, it has been shown (Eder, 2013) that character \( n \)-grams are impressively robust when one deals with a “dirty” corpus (one with a high number of misspelled characters, or one with bad \( \text{OCR} \)). The ideal combination of parameters in this section is another bone of contention between scholars; in fact, Eder and Rybicki (2013) maintain that this differs not only from language to language but also from one collection of text to another.

- **words**: words are used as the unit. Naturally, the higher the \( n \) you specify, the less repetitive your \( n \)-grams there will be, and this means poor statistics (data sparseness).

- **characters**: characters are used as the unit.

- **\( n \)-gram size**: this is where you can specify the value of \( n \) for your \( n \)-grams. Certainly, setting this option to 1 makes sure that individual words/chars will be used instead of higher-order \( n \)-grams.; of course, single-letter counts do not seem like a good idea.

- **preserve case**: normally, all the words from the input texts are turned into lowercase, no matter if they are proper nouns or not – e.g. the sentence *The family of Dashwood had long been settled in sussex* will be turned into *the family of dashwood had long been settled in sussex*. In some situations, however, you might be interested in preserving the case. That’s the option to do it.

- **select files manually**: normally, the script performs the analysis on *all* files in your *corpus* subfolder. If this option is checked, a dialogue window will appear enabling the user the make a selection of input files from the subfolder. Obviously, you can achieve the same results by simply removing the unwanted texts from the *corpus* subfolder. Again, note that this function will expect you to select at least two different input files.

### 4.3.4 MFW settings

This is where you specify the size of the most-frequent-word list that will be used for your analysis. Actually, the name is slightly misleading, since you are not at the mercy of...
“most frequent words” only. You can use most frequent word pairs (bigrams), character sequences, etc. We keep the name “MFW” because... Well, we don’t really remember why we keep it; probably, there was no-one around to propose a better solution.

- Minimum: this setting determines how many words (or features) from the top of the frequency list for the entire corpus will be used in your analysis in the first (and possibly, only) run of the function. With a value of 100 for this parameter, your analysis will be conducted on the 100 most frequent words (features) in the entire corpus.

- Maximum: this setting determines how many words from the top of the word frequency list for the entire corpus will be used in your analysis in the last (and possibly, only) run of the function. Thus, a setting of 1000 results in your (final) experiment being conducted on 1000 most frequent words in the entire corpus. (This parameter setting is especially important when working with the bootstrap consensus trees in \texttt{stylo()}, a procedure which involves running several analyses in a row. See immediately below under “Increment”).

- Increment: this setting defines the value by which the value of Minimum will be increased at each subsequent run of your analysis until it reaches the Maximum value. Thus, a setting of 200 (at a Minimum of 100 and a Maximum of 1000) provides for an analysis based on 100, 300, 500, 700 and 900 most frequent words. (As above, this parameter setting is especially important when working with the bootstrap consensus trees in \texttt{stylo()}, a procedure which involving running several analyses in a row).

- Start at freq. rank: sometimes you might want to skip the very top of the frequency list. With this parameter, you can specify how many words from the top of the overall frequency rank list should be skipped. Normally, however, users will want to set this at 1.

N.B. For all statistical procedures (see 4.3.6 below) except the Consensus Tree, it is advisable to set Minimum and Maximum to the same value (this makes the Increment setting immaterial), unless you want to produce a large series of cluster analysis, multidimensional scaling or principal components analysis graphs in a row, for instance to observe how/if the results change for various lengths of the MFW list.

4.3.5 Culling

“Culling” refers to the automatic manipulation of the wordlist (proposed by Hoover 2004a, 2004b). The culling values specify the degree to which words that do not appear in all the texts of your corpus will be removed. Thus, a culling value of 20 indicates that words that appear in at least 20% of the texts in the corpus will be considered in the analysis. A culling setting of 0 means that no words will be removed; a culling setting of 100 means that only those words will be used in the analysis that appear in \textit{all} texts of your corpus at least once.

- Minimum: this setting specifies the first (and possibly, only) culling setting in your analysis (cf. the minimum MFW setting).

- Maximum: this setting specifies the last (and possibly, only) culling setting in your analysis (cf. the maximum MFW setting). (This parameter setting is esp-
Figure 5: \texttt{stylo()} GUI, the third tab: Statistics, Distances

... especially important when working with the bootstrap consensus trees in \texttt{stylo()}, a procedure which involves running several analyses in a row).

- **Increment**: this defines the increment by which the value of Minimum will be increased at each subsequent run of your analysis until it reaches the Maximum value. Thus a setting of 20 (at a Minimum of 0 and a Maximum of 100) provides for an analysis using culling settings of 0, 20, 30, 60, 80 and 100. (This parameter setting is especially important when working with the bootstrap consensus trees in \texttt{stylo()}, a procedure which involves running several analyses in a row).

- **List cutoff**: Usually, it is recommended to cut off the tail of the overall wordlist; if you do not want to cut the list and analyze vectors of thousands of words at once, then the variable may be set to an absurdly big number (although this can be computationally demanding for your machine). This setting is independent from the culling procedure.

- **Delete pronouns**: (this setting too is independent of the culling procedure). If this option is checked, make sure you have selected the correct language for your corpus (see 4.3.2 above). This will select a list of pronouns for that language inside the script. Advanced users can use this part of the tool to remove any words they want. So far, we have pronoun lists for English, Dutch, Polish, Latin, French, German, Spanish, Italian, and Hungarian.

N.B. As had been mentioned above, for all statistical procedures (see 4.3.6 below) except consensus trees, it is advisable to set Minimum and Maximum to the same value (this makes the Increment setting immaterial), unless you want to produce a large series of cluster analysis, multidimensional scaling or principal components analysis graphs etc. in a row.

### 4.3.6 Statistics

This is the very last moment to emphasize one important thing: the function \texttt{stylo()} provides a bunch of \textit{unsupervised} methods used in stylometry, such as principal components analysis, multidimensional scaling or cluster analysis. The results are represented either on a scatterplot, or a tree-like diagram (dendrogram); the last stage of the analysis involves a human interpretation of the generated plots. The results obtained using these techniques “speak for themselves”, which gives a practitioner an opportunity to notice with the naked eye any peculiarities or unexpected behavior in the analyzed corpus. Also, given a tree-like graphical representation of similarities between particular samples, one
can easily interpret the results in terms of finding out which group of texts a disputable sample belongs to. On the other hand, however, these methods cannot be validated in terms of an automatic verification of a given method’s reliability. Thus, if you feel you’d better use one of machine-learning techniques, refer to the function classify(), below.

- Cluster Analysis: Performs cluster analysis and produces a dendrogram, or a graph showing hierarchical clustering of analyzed texts. This option makes sense if there is only a single iteration (or just a few). This is achieved by setting the MFW Minimum and Maximum to equal values, and doing the same for Culling Minimum and Maximum.

- MDS: Multidimensional Scaling. This option makes sense if there is only a single iteration (or just a few). This is achieved by setting the MFW Minimum and Maximum to equal values, and doing the same for Culling Minimum and Maximum.

- PCA (cov.): Principal Component Analysis using a covariance matrix. This option makes sense if there is only a single iteration (or just a few). This is achieved by setting the MFW Minimum and Maximum to equal values, and doing the same for Culling Minimum and Maximum.

- PCA (corr.): Principal Component Analysis using a correlation matrix (and this is possibly the more reliable option of the two, at least for English). This option makes sense if there is only a single iteration (or just a few). This is achieved by setting the MFW Minimum and Maximum to equal values, and doing the same for Culling Minimum and Maximum.

- Consensus Tree: this option will output a statistically justified “compromise” between a number of virtual cluster analyses results for a variety of MFW and Culling parameter values.

- Consensus strength: For Consensus Tree graphs, direct linkages between two texts are made if the same link is made in a proportion of the underlying virtual cluster analyses. The default setting of 0.5 means that such a linkage is made if it appears in at least 50% of the cluster analyses. Legal values are 0.4 – 1. This setting is immaterial for any other Statistics settings.

4.3.7 Distances

This is where the user can choose the statistical procedure used to analyze the distances (i.e. the similarities and differences) between the frequency patterns of individual texts in your corpus. Although this choice is far from trivial, some of the following measures seem to be more suitable for linguistic purposes than others. On theoretical grounds, Euclidean Distance and Manhattan Distance should be avoided in stylometry based on word frequencies (unless the frequencies are normalized; see: Delta). Canberra Distance is quite troublesome but effective e.g. for Latin; it is very sensitive to rare vocabulary, and thus might be a good choice for inflected languages, with sparse frequencies (it should be combined with careful culling settings and a limited number of MFWs taken into analysis). For English, usually Classic Delta is a good choice: mathematically speaking (Argamon, 2008), it is simply Manhattan distance applied to normalized (z-scored) word frequencies. A theoretical explanation of the measures implemented in this function is pending. The available distance measures are as follows:
• Euclidean Distance: basic and the most “natural”. It is an obvious choice when
your variables are similarly distributed. However, since word distributions are
not similar at any rate (e.g. compare the huge difference between the frequencies
of “the” and “dactyloscopy”), this distance measure is not appropriate to testing
vectors of dozens of most frequent words. Or, to be precise, it could be used to
assess less frequent (content) words. According to Zipf’s law, these words are
distributed more or less similarly in a corpus since, by being less common than
function words, they appear in the flattened sections of a Zipf curve.

\[
\delta_{(AB)} = \sqrt{\sum_{i=1}^{n} |(A_i)^2 - (B_i)^2|}
\]

where:
- \(n\) = the number of MFWs (most frequent words),
- \(A, B\) = texts being compared,
- \(A_i\) = the frequency of a given word \(i\) in the text \(A\),
- \(B_i\) = the frequency of a given word \(i\) in the text \(B\).

• Manhattan Distance: obvious and well documented. It shares the pros and cons of
Euclidean Distance.

\[
\delta_{(AB)} = \sum_{i=1}^{n} |A_i - B_i|
\]

• Classic Delta as introduced by Burrows (2002). Since this measure relies on z-
scores – i.e. normalized word frequencies – it is dependent on the number of texts
analyzed and on a balance between these texts: if a corpus contains, say, a large
number of plays by Lope de Vega and only one play by Calderón de la Barca, the
final results might be biased.

\[
\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|
\]

where:
- \(n\) = the number of MFWs (most frequent words or other features),
- \(A, B\) = texts being compared,
- \(A_i\) = the frequency of a given feature \(i\) in the text \(A\),
- \(B_i\) = the frequency of a given feature \(i\) in the text \(B\),
- \(\mu_i\) = mean frequency of a given feature in the corpus,
- \(\sigma_i\) = standard deviation of frequencies of a given feature.

Argamon (2008) showed that the above formula can be simplified algebraically:

\[
\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - B_i}{\sigma_i} \right|
\]

• Argamon’s Linear Delta, or Euclidean distance applied to normalized (z-scored)
word frequencies (Argamon, 2008). The distance is sensitive to the number of
texts in a corpus.

\[
\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \sqrt{\left| \frac{(A_i)^2 - (B_i)^2}{\sigma_i} \right|}
\]
• Eder’s Delta: it is a modification of standard Burrows’s distance; it slightly increases the weights of frequent words and rescales less frequent ones in order to suppress discriminative strength of some random unfrequent words. The distance was meant to be used with highly inflected languages. It is sensitive to the number of texts in a corpus.

\[
\Delta_{(AB)} = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{A_i - B_i}{\sigma_i} \times \frac{n - n_i + 1}{n} \right)
\]

where:

\(n_i = \) the position of a given feature on a frequency list (i.e. its rank).

• Eder’s Simple: a type of normalization as simple as can be (independent on the size of the corpus), intended to convert the implications of Zipf’s law. The normalization used in this distance is so obvious and so widely-spread in exact sciences that naming it “Eder’s Simple Distance” is an abuse, so to speak.

\[
\delta_{(AB)} = \sum_{i=1}^{n} \left| \sqrt{A_i} - \sqrt{B_i} \right|
\]

• Canberra Distance: sometimes amazingly good. It is very sensitive to differences in rare vocabulary usage among authors. On the other hand, this can be a disadvantage, since sensitiveness to minute differences in word occurrences also means significant sensitiveness to noise. Last but not least, Canberra Distance is very sensitive to the number of words (features) analyzed.

\[
\delta_{(AB)} = \sum_{i=1}^{n} \left| \frac{A_i - B_i}{|A_i| + |B_i|} \right|
\]

• Cosine Distance: a classical measure, introduced to this package quite recently.

\[
\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}
\]

• It is also possible to use any custom distance measure. This option is discussed below, in the section 9.4.

4.3.8 Sampling

When the default setting of “No sampling” is checked, each of the texts in its entirety is treated as a single sample. The second option, that of “Normal sampling”, performs the analysis on equal-sized consecutive sections of each text, and the size is determined by the setting immediately below. Eder (2015) suggests that even better attribution results can be achieved with “Random sampling”, where samples are made up of words each randomly selected from anywhere in the text (“bag of words”); here, too, the sample size must be set below.
4.3.9 Graphs

Do you want to display the graph on the screen? Do you want to write the graph directly to a graphics file? Which format? If you’ve been thinking about any plots, these are the options to fiddle with. You can display the graph on the screen and write to a file (the latter will be done with much better quality). The produced files are saved in your working directory. As has already been mentioned, the name of your working directory is used both as the title on top of the graph (see 4.3.10 below) and as the name of the graph files; some important parameter settings (for MFW list size, culling, pronoun deletion, statistical method) are also placed on the graph and in the name of the file. However, remember that if you perform another analysis with the same parameters on a slightly modified corpus, this will overwrite the earlier graph.

- **Onscreen**: check this if you want to display the graph on the screen.
  
  \[\text{display.on.screen=TRUE|FALSE}\]

- **PDF**: check this to obtain a PDF file with your graph.
  
  \[\text{write.pdf.file=TRUE|FALSE}\]

- **JPG**: check this to obtain your graph in JPEG format.
  
  \[\text{write.jpg.file=TRUE|FALSE}\]

- **SVG**: check this to produce a SVG vector file (XML-based scalable image format). Useful if you want to embed your plot in HTML code, or to edit it further with, say, Inkscape.
  
  \[\text{write.svg.file=TRUE|FALSE}\]

- **PNG**: check this to obtain your graph in PNG format (probably the safest option if you want publication-ready resolution).
  
  \[\text{write.png.file=TRUE|FALSE}\]
4.3.10 Plot area

Further graphic options. Here you can specify the dimensions of the plot area (expressed in inches, yes, even though the makers of the package are all honest Continental Europeans and usually think in centimeters), font size, thickness of the lines used to plot the graphs, etc. Since it is usually hard to remember all the values, an additional option is provided to reset the picture options – if this is checked, the remaining options will be overwritten and the defaults restored.

- Set default: this is the above-mentioned “amnesia button”. If you fiddle too much with different graphic settings and forget, say, the default size of scatterplots, this option is exactly what you need. (Remember, however, that some computers hate the “amnesia buttons”, HAL from 2001: A Space Oddyssey being the most convincing caveat).

- Plot height: self-evident. The valid units are inches.

- Plot width: self-evident. The valid units are inches.

- Font size: self-evident. The unit is “points” (a detailed explanation what type of “points” it is and how they are related to “typographic points” , “Didot points”, “Postscript points” etc. can be skipped here; the only thing worth noting is that they are the same “points” as in popular office programs, like MS Office).

- Line width: if you plan to re-scale your plot, you might also want to increase the thickness of the lines in your graph. This value is expressed in R generic units (1 is default).

- Colors: when this option is checked, the script will automatically assign the same colors to texts with the same first segment of their file names (the first string ending in “_”).

- Grayscale: select this option to have automatic color coding (in greyscale). The file names convention: see above. While the graphs become less pretty than when colored, this might be your method of choice if the journal you’re planning to publish in makes you pay for color illustrations their weight (or, rather, resolution) in gold.

- Black: select to have a black & white graph. No file naming convention is required with this option.

- Titles: if this is checked, the graphs will contain a main (top) title (based on the name of your folder and your choice of the statistics option) and a subtitle (bottom) showing the distance metric which you selected, as well as the MFW, Culling, Feature and Consensus settings. If you don’t want to have the title decided automatically, you need to pass an argument custom.graph.title.

4.3.11 PCA/MDS

More graphic options. The following settings, however, apply to scatterplots only, i.e. to the plots produced by principal component analysis and multidimensional scaling.

- Labels: to identify samples on MDS/PCA plots using their labels, or names, as specified in filenames (see 4.1 above).
• Points: instead of labels, use points to show the positions of the particular samples. This option might be helpful if you want to represent a large number of samples on one plot.

• Both: in some cases, you might be interested to get precise positions of samples on a graph and to keep the samples’ labels on. This option pinpoints the exact locations of the samples using points, and slightly offsets the labels (see the option ‘Label offset’ immediately below).

• Margins: if you use particularly long samples’ names (labels) and/or simply want to add some blank space on your plot, set custom margin size (in percentage of plot area). You can find out your best setting by trial and error.

• Label offset: set custom offset between label and point (in percentage of plot area).

4.3.12 PCA flavour

Even more graphic options. Should you want to use PCA, you can choose one of the four visualization flavors.

• Classic: original PCA visualization using (colored, if applicable) sample names or points.

• Loadings: display PCA feature (word etc.) loadings. This type of visualization is aimed to show the discriminative strength of particular features (e.g. words) across two first principal components. Some visual similarity to popular “word clouds” makes this approach attractive and comprehensive.

• Technical: technical greyscale PCA visualization, showing feature loadings as well as a PC barplot. Potentially useful for greyscale printing in traditional publications.

• Symbols: select to display the samples in your PCA with a group symbol (instead of their entire name). Potentially useful when dealing with lots of samples.

4.3.13 Various

• Horizontal CA tree: select to have your cluster analysis graph oriented horizontally. Probably the better option for clarity, especially if there are a lot of samples to fit on one dendrogram.

• Save distance table: save final distance table(s) in separate text file(s). In most cases, you will not need to use this option.

• Save features: save final feature (word, n-gram) list(s), e.g. the words actually used in the analysis. Use this option to have more control over the experiment: if you feel that your results are suspicious or too good to be true, you can open the generated file and check the features used.

• Save frequencies: this option gives you even more control: you can inspect frequencies of each word across the whole corpus. Switching this option saves frequency table(s) in separate text file(s).
• Dump samples: if you still feel that your experiment in not supervised enough, you might be interested in a “post mortem” inspection of all the samples used: this option dumps the original samples (either whole texts, or chunks specified using the “Sample size” option above) to an external file. Be aware, however, that a corpus containing dozens of full-sized novels might produce a huge dump file.

5 classify()

This function performs a number of machine-learning algorithms of classification: Delta (Burrows, 2002), $k$-nearest neighbors, support vectors machines, naïve Bayes, and nearest shrunken centroids (Jockers and Witten, 2010). Most of the options are derived from the above-mentioned stylo() function.

Unlike explanatory methods as supported by stylo(), this approach involves two stages of the analysis. In the first step, the traceable differences between samples produce a set of rules, or a classifier, for discriminating authorial “uniqueness” in style. The second step is of predictive nature – using the trained classifier, the machine assigns other text samples to the authorial classes established by the classifier; any disputed or anonymous samples will be assigned to one of the classes as well, provided that such a classification is usually based on probabilistic grounds.

The procedure described above relies on an organized corpus of texts. Namely, the clue is to divide all the available texts into two groups: primary (training) set and secondary (test) set. The first set, being a collection of texts written by known authors (“candidates”), serves as a sub-corpus for finding the best classifier, or discrimination rules, while the second set is a pool of texts of known authors, anonymous texts, disputed ones, and so on. The better the classifier, the more samples from the test set are attributed (“guessed”) correctly and the more reliable the attribution of the disputed samples.

The function writes the resulting authorship (or similarity) candidates to a logfile (results.txt) in the current working directory.

5.1 Corpus preparation

Since machine-learning methods involve two sets of texts instead of one, you need to create two subdirectories within your working directory. You don’t really need to name this directory in any special way – no graphs will be generated and thus no titles on graph will be used. However, the names of both subfolders are very important: the one containing training samples should be named primary_set, and the test set should be titled secondary_set (all file names are case sensitive!). In a usual authorship attribution study, the training set will contain at least one text by each candidate author, preferably one “representative” for his/her work, whatever that is (thus, for Goethe, something else than Farbenlehre; for Dickens, probably not The Pickwick Papers). In the test set, you usually put the anonymous or disputed samples you want to analyze, but in most cases you also include a number of known texts to test the robustness of a particular method. To keep the things simple: if you collect a number of texts written by women and men in the training set, you should also put some other text written by both groups (to provide “unseen” or “fresh” data) to the test set to see if they are correctly recognized. A sample corpus of English writers might be split into two subsets as follows:

ABronte_Agnes.xml
Austen_Emma.xml
The number of samples to be kept in these subfolders depends on the method you are going to use. For Delta, support vector machines, and k-NN, the minimal number of texts per class is 1 (as in the above example). Both naive Bayes and nearest shrunken centroids require at least two samples per class to be put into the training set. (Certainly, if you are short of texts, you can cheat: by putting a given sample twice into the training set under two different names, e.g. Swift_Tub-1.txt, Swift_Tub-2.txt). However, it is a commonly accepted fact that the more training data the better – whenever you have enough texts available, put a good portion of them into the training set.

Another aspect of the above question is the nature of your stylometric test. If you want to assess authorship, then a couple of texts per “candidate” should be fine, but if you want to find a rule of gender differentiation, then you probably should collect quite a lot of textual data written by men and women. And if you believe it is possible to separate left-handed and right-handed writers, you need to take tons of training texts, and even then your experiment will lack some methodological rigour (it is sometimes called “the risk of modeling a noise”).

5.2 Calling the function

The function is evoked by the command classify().

5.3 Options

Most of the options are derived from the stylo() function. Refer to the section 4.3 for further details.

5.3.1 Options inherited from stylo()

Input (4.3.1), Language (4.3.2), Features (4.3.3), MFW settings (4.3.4), Culling (4.3.5), Delta Distances (4.3.7), Sampling (4.3.8), Output: various (4.3.13).

5.3.2 Statistics

- Delta: Burrows’s Delta. classification.method="delta"
- k-NN: k-nearest neighbor classification. "knn"
- SVM: support vector machines. "svm"
- NaiveBayes: naive Bayes classification. To use this method, you should have at least two texts of each class (author, genre, etc.) in the primary (training) set. "naivebayes"
- NSC: nearest shrunken centroid classification. To use this method, you should have at least two texts of each class (author, genre, etc.) in the primary (training) set. "nsc"
5.3.3 General

- **ALL culling**: the culling procedure (cf. 4.3.5) is based on the percentage of samples containing a given word. To compute this ratio, one might want to use the texts from the first set only, or from both sets.

- **ALL wordlists**: the both tables of frequencies are build using the pre-prepared word list of the whole primary set (default). Alternatively, one might want to prepare this list of both sets by activating this option.

- **ALL z-scores**: how the z-scores should be calculated. If the variable is set to FALSE, the z-scores are relying on the primary set only (this should be better in most cases; after all, this is the classical solution used by Burrows and Hoover). Otherwise, the scaling is based on all the values in the primary and the secondary sets. (This option is applicable to Delta only).

5.3.4 SVM options

Support vector machines classification settings: refer to `help(svm)` from `library(e1071)` for details.

- **Linear**: linear kernel of SVM; probably the best choice in stylometry, since the number of variables (e.g. MFWs) is many times bigger than the number of classes.

- **Polynomial**: polynomial kernel of SVM.

- **Radial**: radial kernel of SVM.

- **Degree**: parameter needed for kernel of type “polynomial” (default: 3).

- **Coef0**: parameter needed for kernel of type “polynomial” (default: 0).

- **Cost**: cost of constraints violation (default: 1); it is the C-constant of the regularization term in the Lagrange formulation.

5.3.5 k-NN options

- **k value**: the k value in k-Nearest Neighbour algorithm, or number of neighbours to be considered. Certainly, the bigger the number the better the performance, but, on the other hand, to set this value to 3, you need to have at least three samples per class in the training set. If you keep just one text per class in the training set, the performance is ex definitione suboptimal.

- **l value**: minimum vote for definite decision, otherwise “doubt”. (More precisely, less than k − l dissenting votes are allowed, even if k is increased by ties).

5.3.6 Output: general

- **Misclassifications**: here you can specify whether you want to list the misclassified samples into the log file. Certainly, in most cases you will want to have them listed. However, if you plan to perform a multi-iterated large-scale experiment to test performance of a given method, you will probably prefer to switch off all that verbosity.
• Count good guesses: report the number of correct guesses for each iteration (written to the log file). To say the truth, this option is a bit obsolete, since using `classify()` you are almost always interested in the number of correctly classified samples.

• No. of candidates: final ranking of candidates is directed to a file. You may specify the number of final ranking candidates to be displayed (at least 1). This option works for Delta only.

6 rolling.delta(), rolling.classify()

The procedure “Rolling Delta”, supported by the function `rolling.delta()` is reminiscent of a number of earlier applications of the metric (e.g. van Dalen-Oskam and van Zundert, 2007; Kestemont, 2010; Burrows, 2010). The general goal is to use the Delta metric to reliably visualize stylistic shifts in texts, for instance in order to study the stylistic evolution in texts, to detect plagiarism or to pinpoint authorial takeovers in the case of collaborative authorship.

The first step involves a “windowing” procedure in which each reference text is segmented into consecutive, equal-sized samples or “windows” (window.size parameter). The samples are allowed to partially overlap (step.size parameter). If we specify a window.size of 5,000 and a step.size of 100, for example, the first sample of a text contains words 1–5,000, the second 101–5,100, the third sample 201–5,200, and so forth (see 6.3.2 below). Like Delta, our procedure uses the relative frequencies of a (preferably small) set of $n$ words which were most frequent in the entire collection of reference texts. Subsequently, we compute a representative centroid for each reference text that consists of the mean relative frequency for each of the $n$ words in the windows extracted from the text, as well as the standard deviation.

We then proceed to the analysis of the test text. We divide it into windows too and iteratively compute the difference in style (the Delta) between each test window and each reference centroid. In order to calculate the Delta with the $n$ most frequent words we employ the following formula. Let $C$ be an author’s centroid and let $W$ be the window we wish to compare it to. For each of the $n$ words, we calculate the absolute difference between its average frequency in $C$ and its frequency in $W$. Next, we weigh this difference using each word’s standard deviation in the centroid. (Words whose frequencies display significant fluctuation in a reference text’s windows are thus assigned a lower weight.) The final Delta is the summation of these $n$ weighted differences.

$$\Delta(C, W) = \sum_{i=1}^{n} \frac{1}{\sigma_i C} |\mu_i(C) - f_i(W)|$$

After “rolling” through the test text we can plot the resulting series of Deltas for each reference text in a graph. The relatively lower the Deltas for a given reference text, the relatively more similar the style in the test windows – and vice versa (cf. Hoover, 2004b: 471). If the curve for a text would show a sudden drop at a given position, this could be indicative of a stylistic change in the text (which might, for instance, be caused by one author taking over from another. One can use vertical lines in the plot to mark the position of certain events in the test text as an aid in interpreting the graph (e.g. chapter beginnings).
6.1 Corpus preparation

You will need two subfolders in your working directory: primary_set should contain the test texts: the individual writings by each of the authors who collaborated on the test text – the latter goes into the secondary_set subfolder (once again, the names are case-sensitive). In the pilot study for this method (Rybicki et al., 2014), the primary set was composed of individual writings by Joseph Conrad and individual writings by Ford Madox Ford, such as, respectively, *The Heart of Darkness* and *The Good Soldier*; the secondary set only contained a single text at a time: one of the texts written in collaboration by the two writers, such as *The Inheritors*. To study another Conrad/Madox collaboration, *The Inheritors* had to be removed from the secondary set and replaced by, say, *Romance*.

6.2 Calling the function

The function is evoked by the command `rolling.delta()`.

6.3 Options

While many of the options are derived from the main `stylo()` function – especially in the “Input & Language”, “Statistics”, and “Output” tabs, some differences must be emphasized here.

6.3.1 Features

Contrarily to this section in `stylo()`, MFW settings only use a single value (Maximum) since only one analysis is performed, and the same is true of the Culling value.

6.3.2 Sampling

The “Slice length” parameter sets the size of the text “window” or of consecutive samples cut out one by one from the test text. “Stepsize” controls the size of the overlap between two consecutive windows. For the default settings, a slice length of 5,000 and a stepsize of 1,000 takes the first 5,000 words from the beginning of the test text as the first sample, the section from the 1001th word in the text to its 6,000th word, and so forth. Beware of very small stepsize values: we have not yet seen a computer that would not hang R at a setting of 1!

6.3.3 Colors

An additional tab has been added to control the colors of the curves for each training set text. Two courses of action are advisable here. If you only want to differentiate the texts by author, you might want to set a single color from the pull-down fields for that author’s texts, and another for the texts by the second writer. “Color 1” sets the color for the text that comes first alphabetically in the ordinary listing of filenames, “Color 2” colors the second text, etc. Thus, to use the above examples, all texts by Conrad would precede those by Ford, and *conrad_heart.txt* (for *The Heart of Darkness*) would precede *conrad_jim.txt* (for *Lord Jim*), etc. The other alternative is to use different shades of the same basic color for each writer so that the similarity between the particular sets can be more visible. The colours in the pull-down fields can be replaced with other text names of colors in R’s pallette; you can get a listing of these by invoking R’s `colors()` command.
7 oppose()

It performs a contrastive analysis between two given sets of texts, using Burrows’s Zeta (2007) in its different flavors, including Craig’s extensions (Craig and Kinney, 2009). Also, the Whitney-Wilcoxon procedure as introduced by Kilgariff (2001) is available. The function generates a vector of words significantly preferred by a tested author, and another vector containing the words significantly avoided.

7.1 Corpus preparation

Suppose you want to find out the characteristic words of men and women, and then to see which of the anonymous books in a corpus might have been written by men, and which by women. In order to do this, you put all the women into the primary_set subfolder of your working directory, and all the men into the secondary_set folder, and then you place the anonymous texts in the test_set subfolder (the test_set folder is optional; the script will run if it is not there or if you are not interested in the anonymous texts).

7.2 Calling the function

The function is evoked by the command oppose().

7.3 Options

- Text slice length (in words), the default is 10,000 words. This parameter refers to the size of the samples into which each text is “sliced” in order to perform zeta.

- Text slice overlap (in words). The default, 0, means that the first sample will contain words 1 to 10,000 in the text, the second sample 10,001 to 20,000, etc. If you set it to 2000, the first sample will contain words 1 to 10,000 in the text, the second sample 8001 to 18,000, etc. Beware of low values: setting it to 1 will result in a huge number of samples and R might eventually crash.

- Rare occurrences threshold: the default, 2 prevents hapax legomena and dislegomena from appearing in the resulting zeta wordlist file; 1 gets rid of just hapax legomena, 10 makes sure only words that appear at least 11 times in the corpus are included in the list.

- Filter threshold (default 0.1) gets rid of word of weak discrimination strength (it’s like $p$, the degree of statistical significance, in various standard statistical texts). The higher the number, the less words appear in the final wordlist. It does not normally exceed 0.5. To quote Maciej Eder: “if the ‘craig.zeta’ method is selected, you might probably want to filter out some words of weak discrimination strength. Provided that 2 means the strongest positive difference and 0 the strongest negative difference (Hoover, 2009), the values either just above or just below 1 are not significant and thus can be (or rather should be) omitted. If chisquare method was chosen, all the differences of $p$ value below 0.05 were filtered out, in pure Zeta, there is no a priori solution. Threshold 0.5 would filter out a vast majority of words, threshold set to 1 would filter all the words in a corpus.”

- Method: we have 3 zeta flavors so far: Craig’s (as described by him and Hoover); Eder’s (not documented yet, but basically derived from Canberra distance measure); chi-square (not documented yet).
• Output: self-evident, except that “Identify points” only works (if it does work) with output set to “Onscreen”.

The script outputs two files: a list of words_preferred.txt, which are words significantly preferred by the primary author(s); and a list of words_avoided.txt, which are words significantly avoided by the primary author(s).

The graph plots the frequencies of both word categories, preferred in avoided, for each sample into which the texts have been sliced; normally, primary set samples (marked as circles; colors correspond to the authors of the texts from which they were taken) should appear in the top left corner of the plot (high words preferred frequencies, low words avoided frequencies), while the secondary set samples (marked as triangles) should gather in the bottom right (low words preferred frequencies, high words avoided frequencies). Samples from texts by authors in the test set are marked as crosses; if they overlap with either the primary or the secondary set samples, this shows the stylistic similarity. Also, a polygon marking the outside limit of the primary set samples, and another one for the secondary set, are drawn on the graph. Of course, you can now combine the words preferred and avoided files into a single wordlist.txt file and use the stylo() function for better discrimination between two groups of texts.

8 Options unavailable on GUI

8.1 Cluster analysis linkage

• linkage: algorithm for establishing clusters in a dendrogram; choose one of the following linkage methods: "nj", "ward" (default), "single", "complete", "average", "mcquitty", "median", "centroid".

8.2 Network analysis support

The package “stylo” does not produce any networks per se, however, it does generate tables of edges/nodes (or, edges alone), using two Eder’s algorithms of establishing connections between the nodes (Eder, forthcoming). The table can be loaded into Gephi (https://gephi.org). To get such a table, invoke the function stylo() with an argument network=TRUE, and optionally with some other arguments, as listed below. E.g.:

```r
stylo(network=TRUE, network.type="undirected")
```

• network: an output file (or files) will be generated when this option is set TRUE, if this is set FALSE, the options immediately below are immaterial. (Default: FALSE).

• network.tables: one of two flavors of output: either one table (edges), or two (edges and nodes). Choose “edges” (default), or “both”, respectively. Using both tables instead of one allows you to edit the table of nodes in, say, Excel, in order to set some node attributes. When you use two tables, however, make sure you import edges to Gephi first; also, make sure you uncheck (in Gephi) the option for creating missing nodes.

• network.type: when “undirected” type of network is chosen (default), then the connections from and to are counted together (summed into one stronger connection). When "directed" network is chosen, then the incoming connections and the outgoing ones are counted separately.
Figure 8: 124 Greek texts represented as connected nodes of a network

- linked.neighbors: if this value is set to 1, then a link between a given sample and its nearest neighbor is established; when it is set to 2, two neighbors are connected (the nearest neighbor and the first runner-up), etc. Default value is 3, which means that the nearest neighbor and two runners-up are taken into consideration.

- edge.weights: the connections' weights are always differentiated: the nearest neighbor has the strongest link, then comes the first runner-up, and so forth. The assigned weights might be "linear" = 1, 2, 3, ..., n; "quadratic" = 1, 4, 9, ..., n^2; or "log" (logarithmic) = log(1 + (1, 2, 3, ..., n)).

Network analysis plots might be useful for visualizing textual relations in large datasets. Particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes (Fig. 8). The procedure of linking is twofold (details: Eder, forthcoming). One of the involved algorithms computes the distances between analyzed texts, and establishes, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and weaker connections to the 1st and the 2nd runner-up (i.e. two texts that get ranked immediately after the nearest neighbor). The second algorithm performs a large number of tests for similarity with different number of features to be analyzed (e.g. 100, 200, 300, ..., 1,000 MFWs). Finally, all the connections produced in particular “snapshots” are added, resulting in a consensus network.

8.3 Undocumented arguments

- features
- frequencies
- training.frequencies
- test.frequencies
9 Advanced topics

9.1 Batch mode

As mentioned somewhere in this document, it is possible to pass input data into `stylo()` and `classify()` from inside R, without relying on any external files. Also, the final results – apart from being plotted on screen and saved to the hard-drive – are accessible as R objects. One can imagine a very complex stylometric experiment computed remotely on a high-performance server without any file reading/writing involved. The next sections provide an outline of such a pipeline design.

The datasets that can be piped into `stylo()` and `classify()` from other R functions include: (1) pre-processed corpus, in terms of a list containing vectors of words (or other countable units); see `help(load.corpus.and.parse)` for further details, see also an example discussed in `help(stylo)`, (2) table of word frequencies: an R matrix or data frame with variables (words) formatted vertically as columns, and samples (texts) ordered horizontally as rows, (3) words (MFWs) or other features to be analyzed: an R vector containing the features as elements. In the case of `classify()`, two corpora and/or two frequency tables are piped instead. The following executable toy example, quoted after `help(classify)`, shows how the training and the test corpus should be passed in a pipeline:

```r
# preparing a training corpus
txt1 = c("now", "i", "am", "alone", "o", "what", "a", "slave", "am", "i")
txt2 = c("what", "do", "you", "read", "my", "lord")
corpusTRAIN = list(txt1, txt2)
names(corpusTRAIN) = c("hamlet_sample1", "polonius_sample1")

# preparing a test corpus
txt4 = c("to", "be", "or", "not", "to", "be")
txt5 = c("though", "this", "be", "madness", "yet", "there", "is", "method")
txt6 = c("the", "rest", "is", "silence")
corpusTEST = list(txt4, txt5, txt6)
names(corpusTEST) = c("hamlet_sample2", "polonius_sample2", "uncertain_1")

# launching the classification
classify(training.corpus = setTRAIN, test.corpus = setTEST)
```
One can pass the features to be analyzed (e.g., MFWs) into `stylo()` or `classify` in a similar way. Consider the following example:

```r
my.selection.of.function.words = c("the", "and", "of", "in", "if", "into", "within", "on", "upon", "since")
stylo(features = my.selection.of.function.words)
```

In any attempts to use existing tables of frequencies, one needs to remember that R in general, and the package `stylo` in particular, accepts tabular datasets with variables stored as *columns*. Certainly, this might be slightly confusing, since in other statistical programs the variables are usually stored in *rows*. If your dataset is formatted as follows:

<table>
<thead>
<tr>
<th></th>
<th>ABronte</th>
<th>Austen</th>
<th>CBronte</th>
<th>Conrad</th>
<th>Dickens</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;the&quot;</td>
<td>4.57</td>
<td>4.24</td>
<td>4.25</td>
<td>4.19</td>
<td>4.47</td>
</tr>
<tr>
<td>&quot;to&quot;</td>
<td>3.11</td>
<td>3.29</td>
<td>3.43</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>&quot;and&quot;</td>
<td>3.19</td>
<td>3</td>
<td>3.08</td>
<td>2.85</td>
<td>2.81</td>
</tr>
<tr>
<td>&quot;of&quot;</td>
<td>2.6</td>
<td>3</td>
<td>2.63</td>
<td>2.43</td>
<td>2.86</td>
</tr>
<tr>
<td>&quot;I&quot;</td>
<td>2.17</td>
<td>2.2</td>
<td>2.13</td>
<td>2.42</td>
<td>2.22</td>
</tr>
<tr>
<td>&quot;a&quot;</td>
<td>2.24</td>
<td>1.92</td>
<td>1.92</td>
<td>2.21</td>
<td>1.92</td>
</tr>
</tbody>
</table>

then you should do a tiny tweak before piping it further:

```r
my.rotated.dataset = t(my.dataset)
stylo(frequencies = my.rotated.dataset)
```

For the sake of compatibility, the tables produced by the package `stylo` are always saved in the transposed format – do not be confused, then.

When it comes to the final results stored as R objects, you should remember that each function returns its value after evaluation. The main functions of the package `stylo` do not clutter your screen with thousands of values, but the results are there anyway. They are made invisible. To have an access to these values, pipe the function, say, `classify` to a variable:

```r
my.results = classify()
```

The variable `my.results` is a list containing a number of elements, including tables of frequencies, classification results, and so forth. E.g. type `my.results$frequencies.training.set` to get one of the tables.

Now, since the functions accept R objects as input data, and at the same time their output is stored in R’s memory as a list of R objects, it is possible to connect two functions “on the fly”. Consider an experiment involving the function `oppose()`, aimed at extracting the most characteristic words for subcorpora A and B, which will be later used to perform a cluster analysis on some other samples. The solution is straightforward:

```r
# launching 'oppose' to extract significant words
zeta.test.results = oppose()

# combining two lists of words into one vector
```
set1 = zeta.test.results$words.preferred
set2 = zeta.test.results$words.avoided
words.from.zeta = c(set1, set2)

# using the above vector as features for 'stylo'
stylo(features = words.from.zeta)

9.2 Custom splitting rule

When calling stylo, users can define their tokenization rules. The argument `splitting.rule` takes the form of a regular expression which will be used to split input character strings into discrete units, usually words. In obsolete versions of the package ‘stylo’, the default splitting sequence of chars was "^[[:alpha:]]+" on Mac/Linux, and "\W+" on Windows. Two different splitting rules were used, because regular expressions are not entirely platform-independent. In the version 0.5.6, then, assumed letter characters have been indicated explicitly. Type `help(txt.to.words)` to see the actual character codes that were used as default rule.

If you are sure that your corpus contains clean spaces as word delimiters (i.e. there is no punctuation), you can use a very simple splitting rule, such as the one listed below. Even better, the rule might allow spaces, tab characters and newlines as word delimiters (the second variant):

```
classify(splitting.rule = "[:space:]+")
classify(splitting.rule = "[ \t\n]+")
```

Some other regular expressions may include:

```
"^[\u0041-\u005A\u0061-\u007A]+"  # (standard Latin-1)
"^[\u00C0-\u00D6\u00D8-\u00F6\u00F8-\u00FF]+"  # (Western European)
"^[\u0100-\u017F]+"  # (Central European)
```

The custom splitting rule option can be also used if your input “texts” contain sequences of POS-tags rather than words, e.g.:

```
N-voc ADJ N-gen N-gen N-nom N-gen ADJ-NUM ABBR N-nom ADJ ...
```

Then your splitting rule might look as follows:

```
"^[A-Za-z-]+"
```

9.3 Splitting rule and batch mode combined

In many cases, the default functionalities provided by the library `stylo` can turn out to be insufficient for your needs. Then, you can prepare your own corpus using some of the functions provided by the library, then pipe it into other R functions and/or packages, and take it back to `stylo` after relevant modifications. Let’s consider the following example in which the aforementioned custom regular expression solution was used. Suppose there is a collection of texts stored in a subdirectory “corpus”. However, these texts are tagged:

```
All_NNP true_JJ histories_NNS contain_VBP instruction_NN ;_: though_RB ,., in_IN some_DT ,., the_DT treasure_NN may_MD be_VB hard_JJ to_TO find_VB ,., and_CC when_WRB found_VBN ,_, ...```

Suppose one wants to drop the lemmas and get tags only: NNP JJ NNS VBP NN RB IN DT DT NN MD VB JJ TO VB CC WRB VBN RB... It is possible to extract the grammatical annotation via the function `parse.pos.tags()`, using the following code:

```r
# loading all input texts from the directory 'corpus':
my.raw.data = load.corpus(files = dir(), corpus.dir = "corpus")

# we invoke the function 'parse.pos.tags'
my.cool.data = parse.pos.tags(my.raw.data, tagger = "stanford", feature = "pos")

# now, we launch stylo() with an argument:
stylo(parsed.corpus = my.cool.data)
```

The above function is not supported by the older versions of `stylo` (ver. 0.6.2), but it can be easily worked around:

```r
# loading all input texts from the directory 'corpus':
my.raw.data = load.corpus(files = dir(), corpus.dir = "corpus")

# now, it's time for some substitutions and regular expressions:
my.slightly.better.data = lapply(my.raw.data, gsub,
pattern = "\[[[:alpha:]],;’-]+_\]", replacement="")

# to get rid of punctuation marks, another regexp might be helpful:
my.acceptable.data = lapply(my.slightly.better.data, gsub,
pattern = "\[[[:punct:]]\]", replacement = "")

# next, using the function 'txt.to.words' (from the "stylo" package)
# one can tokenize the whole corpus:
my.cool.data = lapply(my.acceptable.data, txt.to.words)

# the last stage is to launch the stylo() function with an argument:
stylo(parsed.corpus = my.cool.data)
```

Similarly, custom tables of frequencies can be build separately, and used in a form of external R objects piped into `stylo()`:

```r
# external frequencies:
my.table = make.table.of.frequencies(my.cool.data, words = c("nn",
"jj", "dt", "prp")
stylo(frequencies = my.table)
```

### 9.4 Custom similarity measures

The package `stylo` in ver. 0.6.0 provides a socket for defining and plugging in custom distance measures. Suppose you want to test the Cosine Delta (or, Würzburg Delta) distance discussed by Jannidis, Schöch and Pielstrom (2015). Their measures is basically a regular Cosine Distance applied to z-scored data. To use it with `stylo`, one has to prepare a custom function that will compute the distance out of table of frequencies. The following function does the job, even if it could be slightly optimized:
wurzburg.cosine = function(x)
# z-scoring the input matrix of frequencies
x = scale(x)
# computing cosine dissimilarity
y = as.dist( x %*% t(x) / (sqrt(rowSums(x^2) %*% t(rowSums(x^2)))) )
# then, turning it into cosine similarity
z = 1 - y
# getting the results
return(z)

Now, the code has to be typed (or copy-pasted) to the R console so that it is visible for other functions. We are ready to use the usual functions, supplemented with an additional argument:

stylo(distance.measure = "wurzburg.cosine")

classify(distance.measure = "wurzburg.cosine")

rolling.classify(distance.measure = "wurzburg.cosine")

Other possible applications include e.g. testing if the Entropy Distance outperforms other similarity measures. The code for the similarity function is straightforward:

dist.entropy = function(x)
A = t(t(x + 1) / colSums(x + 1))
B = t(t(log(x + 2)) / -(colSums(A * log(A))))
y = dist(B, method="manhattan")
return(y)

stylo(distance.measure = "dist.entropy")

Etc. etc. The are plethora of possible distance measures. The users are encouraged to examine them all!

9.5 Large-scale stylometric tests

Suppose that one wants to conduct a large experiment: the goal is to perform multiple runs with different lengths of the wordlist, increasing gradually the ‘start.at’ variable. This option is not implemented in the package stylo. However, using the batch mode it is just a few steps to success. Consider the following tailored script:

### the script begins ###
library(stylo)

# assume we want to perform a series of tests using 50 words
# and gradually moving the starting point on the wordlist
# e.g., from 100 to 1000 by 50 (i.e. for 100, 150, 200, 250, ... 1000)
# this is a vector of the start points we want to test:
where.to.start = seq(100,1000,50)
for(current.start.point in where.to.start) {

    #############
    # CORE CODE:
    # in each iteration, 'stylo' will be launched in batch mode
    # the option "start.at" will be incremented
    stylo(gui = FALSE,
          display.on.screen = FALSE,
          use.existing.freq.tables = TRUE,
          corpus.lang = "English.all",
          mfw.min = 50, mfw.max = 50,
          start.at = current.start.point)
    #############

    # now, we want to get the table of distances
    current.results = results.stylo$distance.table

    # what about saving this table?
    # first, we have to create a unique file name to prevent overwriting
    # the same file in each iteration:
    current.filename = paste("distances_starting_at_",
                              current.start.point, ".txt", sep="")

    # now, it's time to save the results in their files
    write.table(file = current.filename, current.results)
}

# a short message on screen, followed by a newline char:
cat("what about another stylometric test?\n")
### the script is done ###

10 Error messages and troubleshooting

[TBD]

References


Hoover, D. L. (2012). The rarer they are, the more they are, the less they matter. In Digital Humanities 2012: Conference Abstracts. Hamburg University, Hamburg, pp. 218–21.


Stylometry with R: A Package for Computational Text Analysis
by Maciej Eder, Jan Rybicki and Mike Kestemont

Abstract This software paper describes ‘Stylometry with R’ (stylo), a flexible R package for the high-level analysis of writing style in stylometry. Stylometry (computational stylistics) is concerned with the quantitative study of writing style, e.g. authorship verification, an application which has considerable potential in forensic contexts, as well as historical research. In this paper we introduce the possibilities of stylo for computational text analysis, via a number of dummy case studies from English and French literature. We demonstrate how the package is particularly useful in the exploratory statistical analysis of texts, e.g. with respect to authorial writing style. Because stylo provides an attractive graphical user interface for high-level exploratory analyses, it is especially suited for an audience of novices, without programming skills (e.g. from the Digital Humanities). More experienced users can benefit from our implementation of a series of standard pipelines for text processing, as well as a number of similarity metrics.

Introduction

Authorship is a topic which continues to attract considerable attention with the larger public. This claim is well illustrated by a number of high-profile case studies that have recently made headlines across the popular media, such as the attribution of a pseudonymously published work to acclaimed Harry Potter novelist, J. K. Rowling (Juola, 2013), or the debate surrounding the publication of Harper Lee’s original version of To Kill a Mocking Bird and the dominant role which her editor might have played therein (Gamerman, 2015). The authorship of texts clearly matters to readers across the globe (Love, 2002) and therefore it does not come as a surprise that computational authorship attribution increasingly attracts attention in science, because of its valuable real-world applications, for instance, related to forensics topics such as plagiarism detection, unmasking the author of harassment messages or even determining the provenance of bomb letters in counter-terrorism research. Interestingly, the methods of stylometry are also actively applied in the Humanities, where multiple historic authorship problems in literary studies still seek a definitive solution – the notorious Shakespeare-Marlowe controversy is perhaps the best example in this respect.

Authorship attribution plays a prominent role in the nascent field of stylometry, or the computational analysis of writing style (Juola, 2006; Stamatatos et al., 2000; Stamatatos, 2009; Koppel et al., 2009; Van Halteren et al., 2005). While this field has important historical precursors (Holmes, 1994, 1998), recent decades have witnessed a clear increase in the scientific attention for this problem. Because of its emergent nature, replicability and benchmarking still pose significant challenges in the field (Stamatatos, 2009). Publicly available benchmark data sets are hard to come across, mainly because of copyright and privacy issues, and there are only a few stable, cross-platform software packages out there which are widely used in the community. Fortunately, a number of recent initiatives lead the way in this respect, such as the recent authorship tracks in the PAN competition (http://pan.webis.de), where e.g. relevant data sets are efficiently interchanged.

In this paper we introduce ‘Stylometry with R’ (stylo), a flexible R package for the high-level stylistic analysis of text collections. This package explicitly seeks to further contribute to the recent development in the field towards a more advanced level of replicability and benchmarking in the field. Stylometry is a multidisciplinary research endeavor, attracting contributions from divergent scientific domains, which include researchers from Computer Science – with a fairly technical background – as well as experts from the Humanities – who might lack the computational skills which would allow them easy access to the state-of-the-art methods in the field (Schreibman et al., 2004). Importantly, this package has the potential to help bridge the methodological gap luring between these two communities of practice: on the one hand, stylo’s API allows to set up a complete processing pipeline using traditional R scripting; on the other hand, stylo also offers a rich graphical user interface which allows non-technical, even novice practitioners to interface with state-of-the-art methods without the need for any programming experience.

Overview of stylometry

Stylometry deals with the relationship between the writing style in texts and meta-data about those texts (such as date, genre, gender, authorship). Researchers in ‘stylochronometry’, for instance, are interested in inferring the date of composition of texts on the basis of stylistic aspects (Stamou, 2008;
Juola, 2007). Authorship studies are currently the most popular application of stylometry. From the point of view of literary studies, stylometry is typically concerned with a number of recent techniques from computational text analysis that are sometimes termed ‘distant reading’, ‘not reading’ or ‘macroanalysis’ (Jockers, 2013). Instead of the traditional practice of ‘close reading’ in literary analysis, stylometry does not set out from a single direct reading; instead, it attempts to explore large text collections using computational techniques (and often visualization). Thus, stylometry tries to expand the scope of inquiry in the humanities by scaling up research resources to large text collections in order to find relationships and patterns of similarity and difference invisible to the eye of the human reader.

Usually, stylometric analyses involve a complex, multi-stage pipeline of (i) preprocessing, (ii) feature extraction, (iii) statistical analysis, and finally, (iv) presentation of results, e.g. via visualization. To this end, researchers presently have to resort to an ad hoc combination of proprietary, language-dependent tools that cannot easily be ported across different platforms. Such solutions are difficult to maintain and exchange across (groups of) individual researchers, preventing straightforward replication of research results and reuse of existing code. stylo, the package presented, offers a rich, user-friendly suite of functionality that is ideally suited for fast exploratory analysis of textual corpora as well as classification tasks such as are needed in authorship attribution. The package offers an implementation of the main methods currently dominant in the field. Its main advantage therefore lies in the integration of typical (e.g. preprocessing) procedures from stylometry and statistical functionality by other, external libraries. Written in the R language, the source code and binaries for the package are freely available from the Comprehensive R Archive Network, guaranteeing a straightforward installation process across different platforms (both Unix- and Windows-based operating systems). The code is easily adaptable and extensible: the developers therefore continue to welcome user contributions, feedback and feature requests. Our code is open source and GPL-licensed: it is being actively developed on GitHub.

In the rest of this paper, we will first illustrate the functionality of the package for unsupervised multivariate analysis through the high-level function stylo(). Secondly, we will discuss a number of graphical user interfaces which we provide for quick exploration of corpora, in particular by novice users or students in an educational setting, as well as for scholars in the Humanities without programming experience. Next, we move on to the function classify(), implementing a number of supervised classification procedures from the field of Machine Learning. Finally, we concisely discuss the oppose(), rolling.deltal() and rolling.classify() functionality which allow, respectively, to inspect differences in word usage between two subsets of a corpus, and to study the evolution of the writing style in a text.

Overview of the package

Downloading, installing and loading stylo is straightforward. The package is available at CRAN and at GitHub repository. The main advantages and innovative features of stylo include:

Feature extraction

Crucial in stylometry is the extraction of quantifiable features related to the writing style of texts (Sebastiani, 2002). A wide range of features have been proposed in the literature, considerably varying in complexity (Stamatatos, 2009). ‘Stylometry with R’ focuses on features that can be automatically extracted from texts, i.e. without having to resort to language-dependent preprocessing tools. The features that the package allows to extract are n-grams on token- and character level (Houvardas and Stamatatos, 2006; Kjell, 1994). Apart from the fact that this makes the package considerably language-independent, such shallow features have been shown to work well for a variety of tasks in stylometry (Daelemans, 2013; Kestemont, 2014). Moreover, users need not annotate their text materials using domain-specific tools before analyzing them with ‘Stylometry with R’. Apart from the standard usage, however, the package does allow the users to load their own annotated corpora, provided that this is preceded by some text pre-processing tasks. An example of such a non-standard procedure will be shown below. Thus, stylo does not aim to supplant existing, more targeted tools and packages from Natural Language Processing (Feinerer et al., 2008) but it can easily accommodate the output of such tools as a part of its processing pipeline.

https://github.com/computationalstylistics/stylo

1
Metrics

A unique feature of *stylo* is that it offers reference implementations for a number of established distance metrics from multivariate statistical analysis, which are popular in stylometry, but uncommon outside the field. Burrows's Delta is the best example here (Burrows, 2002); it is an intuitive distance metric which has attracted a good share of attention in the community, also from a theoretical point of view (Hoover, 2004a,b; Argamon, 2011).

Graphical user interface

The high-level functions of the package provide a number of Graphical User Interfaces (GUIs) which can be used to intuitively set up a number of established experimental workflows with a few clicks (e.g. unsupervised visualization of texts based on word frequencies). These interfaces can be easily invoked from the command line in R and provide an attractive overview of the various experimental parameters available, allowing users to quickly explore the main stylistic structure of corpora. This feature is especially useful in an educational setting, allowing (e.g. undergraduate) students from different fields, typically without any programming experience, to engage in stylometric experimentation. The said high-level functions keep the analytic procedure from corpus pre-processing to final results presentation manageable from within a single GUI. More flexibility, however, can be achieved when the workflow is split into particular steps, each controlled by a dedicated lower-level function from the package, as will be showcased below.

Example workflow

An experiment in stylometry usually involves a workflow whereby, subsequently, (i) textual data is acquired, (ii) the texts are preprocessed, (iii) stylistic features are extracted, (iv) a statistical analysis is performed, and finally, (v) the results are outputted (e.g. visualized). We will now illustrate how such a workflow can be performed using the package.

Corpus preparation

One of the most important features of *stylo* is that it allows loading textual data either from R objects, or directly from corpus files stored in a dedicated folder. Metadata of the input texts are expected to be included in the file names. The file name convention assumes that any string of characters followed by an underscore becomes a class identifier (case sensitive). In final scatterplots and dendrograms, colors of the samples are assigned according to this convention; common file extensions are dropped. E.g. to make the samples colored according to authorial classes, files might be named as follows:

- ABronte_Agnes.txt
- ABronte_Tenant.txt
- Austen_Pride.txt
- Austen_Sense.txt
- Austen_Emma.txt
- CBronte_Professor.txt
- CBronte_Villette.txt
- EBronte_Wuthering.txt

All examples below can be reproduced by the user on data sets which can be downloaded from the authors’ project website. For the sake of convenience, however, we will use the datasets that come with the package itself:

```r
data(novels)
data(galbraith)data(lee)
```

Our first example uses nine prose novels by Jane Austen and the Brontë sisters, provided by the dataset *novels*.

Preprocessing

*stylo* offers a rich set of options to load texts in various formats from a file system (preferably encoded in UTF-8 Unicode, but it also supports other encodings, e.g. under Windows). Apart from raw text, *stylo* allows to load texts encoded according to the guidelines of the Text Encoding Initiative, which is relatively prominent in the community of text analysis researchers. To load all the files saved in a directory (e.g. ‘corpus_files’), users can use the following command:

```r
# Example command to load all txt files from corpus_files directory
load_corpus_files(directory = 'corpus_files')
```
raw.corpus <- load.corpus(files = "all", corpus.dir = "corpus_files",
encoding = "UTF-8")

If the texts are annotated in e.g. XML, an additional pre-processing procedure might be needed:

corpus.no.markup <- delete.markup(raw.corpus, markup.type = "xml")

Since the dataset that we will use has no annotation, the markup deletion can be omitted. We start
the procedure with making the data visible for the user:

data(novels)
summary(novels)

To preprocess the data, stylo offers a number of tokenizers that support a representative set of
European languages, including English, Latin, German, French, Spanish, Dutch, Polish, Hungarian,
as well as basic support for non-Latin alphabets such as Korean, Chinese, Japanese, Hebrew, Arabic,
Coptic and Greek. Tokenization refers to the process of dividing a string of input texts into countable
units, such as word tokens. To tokenize the English texts, e.g. splitting items as ‘don’t’ into ‘do’ and
‘n’t’ and lowercasing all words, the next command is available:

tokenized.corpus <- txt.to.words.ext(novels, language = "English.all",
preserve.case = FALSE)

The famous first sentence of Jane Austen's *Pride and Prejudice*, for instance, looks like this in its
tokenized version (the 8th to the 30th element of the corresponding vector):

tokenized.corpus$Austen_Pride[8:30]

[1] "it" "is" "a" "truth" "universally"
[6] "acknowledged" "that" "a" "single" "man"
[11] "in" "possession" "of" "a" "good"
[16] "fortune" "must" "be" "in" "want"
[21] "of" "a" "wife"

To see basic statistics of the tokenized corpus (number of texts/samples, number of tokens in
particular texts, etc.), one might type:

summary(tokenized.corpus)

For complex scripts, such as Hebrew, custom splitting rules could easily be applied:

tokenized.corpus.custom.split <- txt.to.words(tokenized.corpus,
splitting.rule = "[A-Za-z\U0500-\U055F\U0560-\U05FF\U05F0-\U05F2]+",
preserve.case = TRUE)

A next step might involve 'pronoun deletion'. Personal pronouns are often removed in stylometric
studies because they tend to be too strongly correlated with the specific topic or genre of a text
(Pennebaker, 2011), which is an unwanted artefact in e.g. authorship studies (Hoover, 2004a,b). Lists
of pronouns are available in stylo for a series of languages supported. They can be accessed via for
example:

stylo.pronouns(language = "English")

[1] "he" "her" "hers" "herself" "him"
[6] "himself" "his" "i" "me" "mine"
[11] "my" "myself" "our" "ours" "ourselves"
[16] "she" "thee" "their" "them" "themselves"
[21] "they" "thou" "thy" "thyself" "us"
[26] "we" "ye" "you" "your" "yours"
[31] "yourself"

Removing pronouns from the analyses (much like stopwords are removed in Information Retrieval
analyses) is easy in stylo, using the delete.stop.words() function:

corpus.no.pronouns <- delete.stop.words(tokenized.corpus,
stop.words = stylo.pronouns(language = "English"))

The above procedure can also be used to exclude any set of words from the input corpus.
Features

After these preprocessing steps, users will want to extract gaugeable features from the corpus. In a vast majority of approaches, stylometrists rely on high-frequency items. Such features are typically extracted in the level of (groups of) words or characters, called \textit{n}-grams (Kjell, 1994). Both word-token and character \textit{n}-grams are common textual features in present-day authorship studies. \textit{Stylo} allows users to specify the size of the \textit{n}-grams which they want to use. For third order character trigrams (\(n = 3\)), for instance, an appropriate function of \textit{stylo} will select partially overlapping series of character groups of length 3 from a string of words (e.g., ‘tri’, ‘rig’, ‘igr’, ‘gra’, ‘ram’, ‘ams’). Whereas token level features have a longer tradition in the field, character \textit{n}-grams have been fairly recently borrowed from the field of language identification in Computer Science (Stamatatos, 2009; Eder, 2011). Both \textit{n}-grams at the level of characters and words have been listed among the most effective stylistic features in survey studies in the field. For \(n = 1\), such text representations model texts under the so-called ‘bag-of-words’ assumption that the order and position of items in a text is negligible stylistic information. To convert single words into third order character chains, or trigrams:

\begin{verbatim}
corpus.char.3.grams <- txt.to.features(corpus.no.pronouns, ngram.size = 3, features = "c")
\end{verbatim}

Sampling

Users can study texts in their entirety, but also draw consecutive samples from texts in order to effectively assess the internal stylistic coherence of works. The sampling settings will affect how the relative frequencies are calculated and allow users to normalize text length in the data set. Users can specify a sampling size (expressed in current units, e.g. words) to divide texts into consecutive slices. The samples can partially overlap and they can be also be extracted randomly. As with all functions, the available options are well-documented:

\begin{verbatim}
help(make.samples)
\end{verbatim}

To split the current corpus into non-overlapping samples of 20,000 words each, one might type:

\begin{verbatim}
sliced.corpus <- make.samples(tokenized.corpus, sampling = "normal.sampling", sample.size = 20000)
\end{verbatim}

Counting frequent features

A crucial point of the dataset preparation is building a frequency table. In stylometry, analyses are typically restricted to a feature space containing the \(n\) most frequent items. It is relatively easy to extract e.g. the 3,000 most frequent features from the corpus using the following function:

\begin{verbatim}
frequent.features <- make.frequency.list(sliced.corpus, head = 3000)
\end{verbatim}

After the relevant features have been harvested, users have to extract a vector for each text or sample, containing the relative frequencies of these features, and combine them into a frequency table for the corpus. Using an appropriate function from \textit{stylo}, these vectors are combined in a feature frequency table which can be fed into a statistical analysis (external tables of frequencies can be loaded as well):

\begin{verbatim}
freqs <- make.table.of.frequencies(sliced.corpus, features = frequent.features)
\end{verbatim}

Feature selection and sampling settings might interact: an attractive unique feature of \textit{stylo} is that it allows users to specify different ‘culling’ settings. Via culling, users can specify the percentage of samples in which a feature should be present in the corpus in order to be included in the analysis. Words that do not occur in at least the specified proportion of the samples in the corpus will be ignored. For an 80\% culling rate, for instance:

\begin{verbatim}
culled.freqs <- perform.culling(freqs, culling.level = 80)
\end{verbatim}

Analysis

\textit{Stylo} offers a seamless wrapper for a variety of established statistical routines available from R's core library or contributed by third-party developers; these include t-Distributed Stochastic Neighbor Embedding (van der Maaten and Hinton, 2008), Principal Components Analysis, Hierarchical Clustering and Bootstrap Consensus Trees (a method which will be discussed below). An experiment can be initiated with a pre-existing frequency table with the following command:
When the input documents are loaded directly from text files, the default features are most frequent words (MFWs), i.e. 1-grams of frequent word forms turned into lowercase. Also, by default, a standard cluster analysis of the 100 most frequent features will be performed. To perform e.g. a Principal Components Analysis (with correlation matrix) of the 200 most frequent words, and visualize the samples position in the space defined by the first two principal components, users can issue the following commands:

```r
stylo(frequencies = culled.freqs, gui = FALSE)
```

In Fig. 1, we give an example of how Principal Components Analysis (the first two dimensions) can be used to visualize texts in different ways, e.g. with and without feature loadings. Because researchers are often interested in inspecting the loadings of features in the first two components resulting from such an analysis, **stylo** provides a rich variety of flavours in PCA visualizations. For an experiment in the domain of authorship studies, for instance, researchers will typically find it useful to plot all texts/samples from the same author in the same color. The coloring of the items in plots can be easily controlled via the titles of the texts analyzed across the different R methods that are used for visualization – a commodity which is normally rather painful to implement across different packages in R. Apart from exploratory, unsupervised analyses, **stylo** offers a number of classification routines that will be discussed below.

The examples shown in Fig. 1 were produced using the following functions:

```r
stylo(frequencies = culled.freqs, analysis.type = "PCR",
     custom.graph.title = "Austen vs. the Bronte sisters",
     pca.visual.flavour = "technical",
     write.png.file = TRUE, gui = FALSE)
```

```r
stylo(frequencies = culled.freqs, analysis.type = "PCR",
     custom.graph.title = "Austen vs. the Bronte sisters",
     write.png.file = TRUE, gui = FALSE)
```

```r
stylo(frequencies = culled.freqs, analysis.type = "PCR",
     custom.graph.title = "Austen vs. the Bronte sisters",
     pca.visual.flavour = "symbols", colors.on.graphs = "black",
     write.png.file = TRUE, gui = FALSE)
```

```r
stylo(frequencies = culled.freqs, analysis.type = "PCR",
     custom.graph.title = "Austen vs. the Bronte sisters",
     pca.visual.flavour = "loadings",
     write.png.file = TRUE, gui = FALSE)
```

**Return value**

**Stylo** makes it easy to further process the objects returned by an analysis. To cater for the needs of less technical users, the results returned by an analysis are saved by default to a number of standard files and outputted on screen. Advanced users can easily use the returned objects in subsequent processing:

```r
stylo.results = stylo() # optional arguments might be passed
```

```r
print(stylo.results)
summary(stylo.results)
```

The list of features created, for instance, can be easily accessed (and manipulated) subsequently, and the same applies to tables of frequencies or other results:

```r
stylo.results$features
stylo.results$table.with.all.freqs
stylo.results$distance.table
stylo.results$pca.coordinates
```
CONTRIBUTED RESEARCH ARTICLES

Figure 1: Illustration of different visualization options for the first two dimensions outputted by a Principal Components Analysis (applied to 9 novels by 4 authors from our dummy corpus). Four different visualization flavours are presented: ‘Technical’ (Fig. 1a), ‘Classic’ (Fig. 1b), ‘Symbols’ (Fig. 1c) and ‘Loadings’ (Fig. 1d). Users whose file names follow stylo’s naming conventions can easily exploit different coloring options.

GUI mode

Apart from the various functions to perform actual stylometric tasks, the package comes with a series of GUIs that can be used to set up typical experimental workflows in a quick and intuitive fashion. This unique feature renders stylo especially useful in educational settings involving students and scholars without programming experience. The cross-platform graphical user interface (automatically installed along with the rest of the package) has been written for Tcl/Tk and can be easily invoked from the command line. Four GUIs are currently available, which all come with extensive tooltips to help users navigate the different options. In this section, we will illustrate the use of these GUIs via an unsupervised stylometric experiment involving Bootstrap Consensus Trees.

The currently most widely used GUI component of ‘Stylometry with R’ is the eponymous GUI for stylo(), which is useful for the unsupervised stylistic exploration of textual corpora. It can be easily invoked using a single intuitive command (without the need to specify additional arguments):

stylo()

The various tabs of the stylo GUI (see Figure 2) present in a clear fashion the various parameters which can be specified before running the analysis by clicking the OK button. Users can freely switch between tabs and revisit them before running an experiment. Moreover, stylo will remember the experimental settings last used, and automatically default to these when users re-launch the GUI (which is useful for authors running a series of consecutive experiments with only small changes in parameters).

To illustrate the GUI mode, we will now concisely discuss a sample experiment involving Bootstrap Consensus Trees (BCT, selectable under the STATISTICS tab in the GUI). In stylometry, BCT exploits the...
idea that the results become stable when one divides the list of MFW in non-identical, yet potentially overlapping frequency bands and analyzes these independently from each other (Eder, 2012). BCT were originally borrowed by Eder from the field of Language Evolution and Genetics; since a number of successful applications of the technique have been reported in the literature (Rybicki and Heydel, 2013; van Dalen-Oskam, 2014; Stover et al., 2016). If the user specifies that different frequency bands should be used on the FEATURES tab, the bootstrap procedure will run different (virtual) cluster analyses and aggregate the results into a single (unrooted) consensus tree. This visualization will only consider nodes for which there exists a sufficiently large consensus among the individual cluster analyses. The user in the corresponding text field (e.g. 0.5, which comes down to a majority vote for the cluster nodes). As such, users can assess the similarities between texts across different frequency bands.

Under the FEATURES tab, users can define the minutes of the MFW division and sampling procedure, using the increment, the minimum and maximum parameters. For minimum = 100, maximum = 3000, and increment = 50, stylo will run subsequent analyses for the following frequency bands: 100 MFW, 50–150 MFW, 100–200 MFW, ..., 2900–2950 MFW, 2950–3000 MFW. This is an attractive feature because it enables the assessment of similarities between texts across different bands in the frequency spectrum. A parallel logic underpins the CULLING text fields, where experiments will be carried out iteratively for different culling rates.

We illustrate the working of the BCT procedure in stylo using the recently covered case study on Go Set a Watchman, the second novel by Harper Lee, written before To Kill a Mockingbird. The novel itself attracted a reasonable attention worldwide, also because of its alleged authorship issues. Suspicion resurfaced about the strange fact that one of the greatest bestsellers in American history was its author’s only completed work; Lee’s childhood friendship with Truman Capote (portrayed as Dill in To Kill A Mockingbird) and their later association on the occasion of In Cold Blood fueled more speculations on the two Southern writers’ possible, or even just plausible, collaboration; finally, the role of Tay Hohoff, Lee’s editor on her bestseller, was discussed.

The stylometric study on this novel, featured in Wall Street Journal (Gamerman, 2015), revealed that the truth proved to be at once much less sensational than most of the rumors. Very strong stylometric evidence shows clearly that Harper Lee is the author of both To Kill A Mockingbird and Go Set A Watchman. In our replication of the experiment, the following code was used to produce the plots:

data(lee)

stylo(frequencies = lee, analysis.type = "CA",
    write.png.file = TRUE, custom.graph.title = "Harper Lee",
gui = FALSE)

stylo(frequencies = lee, analysis.type = "CA",
    mfw.min = 1500, mfw.max = 1500, custom.graph.title = "Harper Lee",
    write.png.file = TRUE, gui = FALSE)

stylo(frequencies = lee, analysis.type = "BCT",
    mfw.min = 100, mfw.max = 3000, custom.graph.title = "Harper Lee",
    write.png.file = TRUE, gui = FALSE)
Figure 3: Analysis of the corpus of 28 novels by Harper Lee, Truman Capote as well as a number of comparable control authors writing in the American South. A frequency table of this corpus is provided by the package *stylo*, so that all our experiments can be replicated. In all plots, Lee’s writing style is clearly very consistent, even if for some input parameters Lee’s novels are close to Capote’s. Figure panel 3a-3b: Traditional dendrograms outputted by cluster analyses with Burrows’s Classic Delta Metric for 100 MFW and 1,500 MFW respectively (default settings; entire novels). Figure panel 3c: Bootstrap consensus tree for 100 MFW to 3,000 MFW (with an incremental step size of 50 words). Unrooted tree which combines clade information from analyses such as the ones presented in Fig. 1a-1b. The tree collapses nodes which were observed in at least 50% of the underlying trees (majority vote).

Classify

Apart from the already-discussed explanatory multivariate tests and the associated visualizations, stylometry has borrowed a number of advanced classification methods from the domain of Machine Learning. Some of them have simply been transferred to stylometry (e.g. Support Vector Machines or Naïve Bayes Classifier); others have been tailored to the needs of humanities researchers. The best example in this respect is Delta, a so-called ‘lazy’ learner developed by Burrows (Burrows, 2002). The *stylo* package offers an interface to a selection of established classifiers: including Burrows’s original Delta and other distance-based classifiers, Nearest Shrunken Centroids, Support Vector Machines and Naïve Bayes Classifier. These are available through a single function:

```r
classify()  # optional arguments might be passed
```

If any non-standard text preprocessing procedures are involved, the above function can be fed with the result of a multi-stage custom pipeline. Combining the function classify() with spreadsheet tables of frequencies is also possible.

In a typical classification experiment, the analysis is divided in two stages. In the first stage, representative text samples for each target category (e.g. authorial group) are collected in a training corpus. The remaining samples form the test corpus. The first set, being a collection of texts, e.g. written by known authors (‘candidates’), serves as a sub-corpus for fine-tuning the hyperparameters of a classifier and model architecture selection. The second set is a pool that consists of test texts of known...
authorship and anonymous texts of disputed authorial provenance. The classifier’s performance can be measured by applying a standard evaluation metric to the classifier’s output on the test set (e.g. the number of correct attributions to authors in the training set). In \textit{stylo}, users can divide their data over two subdirectories (or input custom-created R objects using the low-level functions discussed above); one directory should contain the training samples, the other the test samples. Other options can be specified via the parameters that run parallel to those of the \texttt{classify()} function, such as the desired feature type or culling rate. Function-specific parameters for \texttt{classify()} include the number of cross-validation folds or the type of classifier (e.g. Support Vector Machine).

We illustrate the performance of classification methods in \textit{stylo} using the well-known case study of the pseudonymous author Galbraith/Rowling, which recently attracted a good deal of press attention. In July 2013, the \textit{Sunday Times} (UK) revealed that J. K. Rowling, the successful author behind the bestselling series of \textit{Harry Potter} novels, had published a new detective novel (\textit{The Cuckoo’s Calling}) under the pseudonym of ‘Robert Galbraith’. (The paper had received an anonymous tip with respect to this pen name over Twitter). For covering this case study, the \textit{Sunday Times} has collaborated with Patrick Juola, an authority in the field of authorship attribution, and Peter Millican (Juola, 2013). They reported in a blog post on the Language Log that their stylometric analysis showed the writing style (e.g. on the level of function words) found in \textit{The Cuckoo’s Calling} to be broadly consistent with Rowling’s writing in other works. Below, we report on a dummy attribution experiment which illustrates a supervised procedure.

In this experiment we will confront Galbraith’s \textit{The Cuckoo’s Calling} with 25 other fantasy novels and thrillers by 4 famous novelists: H. Coben (e.g. \textit{Tell No One}), C. S. Lewis (e.g. \textit{The Chronicles of Narnia}), J. R. R. Tolkien (e.g. the \textit{Lord of the Rings} trilogy) and J. K. Rowling (e.g. the \textit{Harry Potter} series). Our replication experiments indeed confirm that Galbraith’s writing style is more consistent with that of Rowling than that of any other author included. Instead of loading particular text files, we will use a computed table of frequencies provided by the package; the table has to be split into two tables (training set and test set). As an illustration, we specify the training set manually (with two training texts per class):

```r
# specify a table with frequencies:
data(galbraith)
freqs <- galbraith

# specify class labels:
training.texts <- c("coben_breaker", "coben_dropshot", "lewis_battle", "lewis_caspian", "rowling_casual", "rowling_chamber", "tolkien_lord1", "tolkien_lord2")

# select the training samples:
training.set <- freqs[rownames(freqs) %in% training.texts,]

# select remaining rows as test samples:
test.set <- freqs[!(rownames(freqs) %in% training.texts),]
```

To perform Delta on the Rowling corpus (50 MFWs, no sampling), we type:

```r
classify(training.frequencies = training.set, test.frequencies = test.set, mfw.min = 50, mfw.max = 50, classification.method = "delta", gui = FALSE)
```

The results are automatically outputted to a log file ‘final_results.txt’:

```
galbraith_cuckoos --> rowling rowling coben

50 MFW, culled @ 0%, 17 of 17 (100%)

General attributive success: 17 of 17 (100%)

MFWs from 50 to 50 @ increment 100
Culling from @ to @ @ increment 20
Pronouns deleted: FALSE; standard classification
```

The overall performance of the classifier for our dummy corpus is optimal, since 100% of the test samples were correctly attributed to the correct authors. The experiment adds support to the identification of the author of \textit{The Cuckoo’s Calling} as Rowling. To combat model overfitting, cross-validation on the training data can be applied. It has been shown that for linguistic datasets a standard
10-fold cross validation might overestimate the performance of models, especially if languages other than English are assessed (Eder and Rybicki, 2013). To neutralize class imbalance, `stylo` therefore provides stratified cross-validation protocols for stylometric experiments. To perform a classification with a 'plain vanilla' 20-fold CV, using Nearest Shrunken Centroids classification and a series of tests for 50, 100, 150, 200, ..., 500 MFWs, one might type:

```r
results <- classify(training.frequencies = training.set,
                    test.frequencies = test.set,
                    mfw.min = 50, mfw.max = 500, mfw.incr = 50,
                    classification.method = "nsc", cv.folds = 20, gui = FALSE)
```

To inspect the classification accuracy for particular cross-validation folds, the user can type:

```r
results$cross.validation.summary
```

Average scores of the cross-validation outcome (note that the overall performance is now slightly worse, ca. 95%) can be accessed via:

```r
colMeans(results$cross.validation.summary)
```

## Miscellaneous other functions

Apart from the above discussed functions, the package offers miscellaneous other, less established functions to stylometrically analyze documents. With the `oppose()` function, users can contrast two sets of documents and extract the most characteristic features in both sets of texts. The most discriminative features can be visualized and fed into other components of the package as part of a pipeline. Several metrics are implemented that can select features which display a statistically significant difference in distributions between both sets. Craig’s Zeta, for instance, is an extension of the Zeta metric originally proposed by Burrows (Burrows, 2007), which remains a popular choice in the stylometric community to select discriminative stylometric features in binary classification settings (Craig and Kinney, 2009). An example of another more widely used metric for feature selection in corpus linguistics is the Mann-Whitney ranks test (Kilgariff, 2001). As a dummy example, we can confront the above mentioned texts; be it the novels by Jane Austen and Anne Brontë:

```r
data(novels)
corpus.all <- txt.to.words.ext(novels, language = "English.all",
                              preserve.case = TRUE)
corpus.austen <- corpus.all[grep("Austen", names(corpus.all))]
corpus.abronte <- corpus.all[grep("ABronte", names(corpus.all))]
zeta.results <- oppose(primary.corpus = corpus.austen,
                        secondary.corpus = corpus.abronte, gui = FALSE)
```

As can be seen in the results (first 20 most discriminating words), Jane Austen is an enthusiast user of terms related to socio-cultural phenomena (e.g. *situation*, *opinion*, *party*, *engaged*, ...), whereas Anne Brontë’s vocabulary can be characterized by a variety of auxiliary verbs with contractions, as well as religious and light-related vocabulary (e.g. *bright*, *dark*).

```r
zeta.results$words.preferred[1:20]
```

```
[1] "Her"    "farther"  "behaviour" "opinion"  "party"
[6] "point"  "perfectly" "afterwards" "Colonel" "directly"
[11] "spirits" "situation" "settled"  "hardly"   "Jane"
[16] "Emma"   "equal"    "family"   "engaged" "They"
```

```r
zeta.results$words.avoided[1:20]
```

```
[1] "don’t"  "I’ve"    "I’ll"     "beside"  "Arthur"
[6] "can’t"  "it’s"    "won’t"    "Huntingdon"
[11] "presence" "Helen"  "face"    "bright" "God"
[16] "mamma"  "further" "heaven"  "dark"    "feet"
```

Of course, the above results of this simple feature selection tool can be fed into one of the package’s classification routines:
Figure 4: The Rolling Stylometry visualization. The medieval French allegoric story *Roman de la Rose* assessed using Rolling SVM and 100 MFWs; window size: 5,000 words, sample overlap: 4,500 words. Sections attributed to Guillaume de Lorris are marked red, those attributed to Jean de Meun are green. The level of certainty of the classification is indicated by the thickness of the bottom stripe. The commonly-accepted division into two authorial parts is marked with a vertical dashed line ‘b’.

```
combined.features <- c(zeta.results$words.preferred[1:20],
                       zeta.results$words.avoided[1:20])
stylo(parsed.corpus = corpus.all, features = combined.features, gui = FALSE)
```

Other functionality worth mentioning are rolling.delta() and rolling.classify(). These functions implement a procedure meant to progressively analyze the development of a style in a text, using e.g. one of the stylometric distance metrics discussed (Rybicki et al., 2014; Eder, 2016). In many works, specific parts of the text are conjectured to have been plagiarized or contributed by other authors: rolling.delta() and rolling.classify() offer an easy way to visualize local stylistic idiosyncrasies in texts. In Fig. 4 we have plotted a rolling.classify() analysis of the well-known French allegorical romance *Roman de la Rose* from the Middle Ages. It has been written by two authors: Guillaume de Lorris is the author of the opening 4,058 lines (ca. 50,000 words), and the second part by Jean de Meun consists of 17,724 lines (ca. 218,000 words). This knowledge is supported by the text itself, since Jean de Meun explicitly points out the takeover point (it is marked with a dashed vertical line ‘b’ in Fig. 4). In this example, the aim is to verify whether two authorial styles can indeed be discerned in the text, that is, before and after the authorial takeover. First a Support Vector Machine classifier is trained on four 5,000-word samples: two extracted from the beginning of the text and two near the middle of the text (yet well beyond the hypothesized takeover: they are marked with the dashed line ‘a’ and ‘c–d’, respectively). Next, we apply a windowing procedure and we extract consecutive and partially overlapping samples from the entire text. Finally, the trained classifier is applied to each of these ‘windows.’ In Fig. 4 we plot the respective classification scores for both authors in each sample: in this case, these scores represent the probability, estimated by a Support Vector Machine, that a particular sample should be attributed to one of the two authors involved. Although the result is not flawless, a clear shift in authorial style can be discerned around the position of the takeover, as indicated verbatimly in the text by one the authors.

The dataset to replicate the test can be downloaded from this page: https://sites.google.com/site/computationalstylistics/corpora/Roman_de_la_Rose.zip. The following code should be typed to perform the classification:

```
# unzipping the dataset
unzip("Roman_de_la_Rose.zip")

# changing working directory
setwd("Roman_de_la_Rose")

rolling.classify(write.png.file = TRUE, classification.method = "svm", mfw = 100,
                 training.set.sampling = "normal.sampling", slice.size = 5000,
                 slice.overlap = 4500)
```
Conclusion

‘Stylometry with R’ targets two distinct groups of users: experienced coders and beginners. Novice users have found it useful to work with the intuitive Graphical User Interface (GUI), which makes it easy to set and explore different parameters without programming experience. We wish to emphasize, however, that stylo is useful beyond these high-level functions and GUIs: it also offers experienced users a general framework that can be used to design custom processing pipelines in R, e.g. in other text-oriented research efforts. The current version of stylo (version number 0.6.3) is available from GitHub under a GPL 3.0 open-source licence; binary installation files are available from CRAN. stylo has been used in a number of innovative studies in the field of computational stylistics (Kestemont et al., 2013; van Dalen-Oskam, 2014; Lauer and Jannidis, 2014; Anand et al., 2014; Oakes and Pichler, 2013; Boot, 2013), and we encourage the future application of stylo to challenging new problems and languages in stylometry.

Acknowledgments

We would like to thank the users of stylo for the valuable feedback and feature requests which we have received over the past years. MK was partially founded for this research as a postdoctoral fellow by The Research Foundation of Flanders (FWO). ME was partially supported by Poland’s National Science Centre (grant number 2014/12/W/ST5/00592).

Bibliography


Maciej Eder  
**Institute of Polish Language**  
**Polish Academy of Sciences**  
al. Mickiewicz 31, 31-120 Kraków  
**Poland**  
maciejeder@gmail.com

Jan Rybicki  
**Institute of English Studies**  
**Jagiellonian University**  
al. Mickiewicza 9A, 31-120 Kraków  
**Poland**  
jkrybicki@gmail.com

Mike Kestemont  
**Department of Literature**  
**University of Antwerp**  
**Prinsstraat 13, B-2000 Antwerp**  
**Belgium**  
mike.kestemont@uantwerp.be
Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?

Jan RYBICKI, Maciej EDER
Pedagogical University, Kraków, Poland
Contact: jkrybicki@gmail.com

Abstract
This paper examines the success of authorship attribution of Burrows’s Delta in several corpora representing a variety of languages and genres. Contrary to the approaches of our predecessors, who only investigated the attributive effectiveness of the very top of the list of the most frequent words, hundreds of possible combinations of word vectors were tested in this study, not solely starting with the most frequent word in each corpus. The results show that Delta works best for prose in English and German and less well for agglutinative languages such as Polish or Latin.

Introduction
In 2007, John Burrows identified three regions in word frequency lists of corpora in authorship attribution and stylometry. The first of these regions consists of the most frequent words, for which his Delta has become the best-known method of study. This is evidenced by a varied body of research with interesting modifications of the method (e.g. Argamon, 2008; Hoover, 2004a, 2004b). At the other end of the frequency list, Iota deals with the lowest-frequency words, while ‘the large area between the extremes of ubiquity and rarity’ (Burrows, 2007) is now the target of many studies employing Zeta or its modifications, such as Craig’s Zeta (e.g. Craig and Kinney, 2009; Hoover, 2007).

Due to the popularity of the three methods it was only a matter of time before Delta (and, to a lesser extent, Zeta and Iota) were applied to texts in languages other than Modern English: Middle Dutch (Dalen-Oskam and Zundert, 2007), Old English (García and Martín, 2009),
2007) and Polish (Eder and Rybicki, 2009). Delta has also been used in translation-oriented papers, including Burrows’s own work on Juvenal (Burrows, 2002) and Rybicki’s attempts at translator attribution (2009, 2011).

It has been generally – and mainly empirically – assumed that the use of methods relying on the most frequent words in a corpus should work just as well in other languages as it does in English; this question has not been approached in any detail until very recently (Juola, 2009). We cannot fail to observe that its success rates in Polish, although still high, fall somewhat short of its detection rate in English (Rybicki, 2009; Eder and Rybicki, 2011). Also, to further complicate the issue of multilingualism, the study by Rybicki mentioned above (2009) seems to suggest that, in a corpus of translated literary texts, Delta is much better at recognizing the author of the original than the translator. Or, to be more precise: with only two candidate translators of the same author, Delta fares well; however, at higher numbers of translators and of authors of the original, Delta’s guessing favors the latter rather than the former. Additionally, genre differences between texts have often been blamed for worse (or better) results in authorship attribution by Delta. This was yet another good reason for a more in-depth look into the workings of Burrows’s method not only in its ‘original’ English and in a variety of other languages, but also in a variety of genres.

Methods
The software we used provides several flavours of Delta (as well as other distance measures); however, the one consistently used in the final results of this study was Burrows’s classic Delta, for the reason that it was the classic method and, perhaps more importantly, because tentative results obtained with the other Delta varieties were very similar.

In this study, a single major modification was applied to the usual Delta process. According to the standard Delta procedure, each corpus was divided into ‘training’ samples in a primary set (one representative sample per each author) and the remaining ‘test’ samples in a secondary
set. The goal of such a procedure was to test how many samples of known authorship were ‘guessed’, or correctly classified to the proper ‘training’ sample.

Each analysis was first made with the top 50 most frequent words in the corpus; then the 50 most frequent words would be omitted and the next 50 words (i.e. words ranked 51 to 100 in the descending word frequency list) would be taken for analysis; then the next 50 most frequent words (those ranked 101 to 150), and so on until the required limit (usually the 5000th most frequent word) would be reached. Then the procedure would restart with the first 100 words (1-100), the second 100 words (101-100), and so on. At every subsequent restart, the number of the words omitted from the top of the frequency list would be increased by 50 until the length of this ‘moving window’ descending down the word frequency list reached another limit (usually 5000). This was done with a single 1000-line script, written by Eder, for the statistical programming environment R. The script produced word frequency tables, calculated the myriad Delta iterations and produced ‘heatmap’ graphs of Delta’s success rate for each of the frequency list intervals, showing the best combinations of initial word position in wordlist and size of window, including variations of pronoun deletion and culling parameters. In fact, the heatmaps are probably the only feasible way of presenting such an amount of results in a comprehensive way. In the resulting graphs below, the horizontal axis presents the size of each wordlist used for one set of Delta calculations (the ‘moving window’); the vertical axis shows how many of the most frequent words were omitted (or where the ‘moving window’ began for each iteration). Each of the runs of the script produced an average of ca. 3000 Delta iterations.

Material
The texts that constitute the corpora used in this study were taken from a variety of good-quality Internet sources (mostly, various national electronic libraries), cleaned of paraphernalia (such as extra titles or Project Gutenberg’s legal disclaimers) and saved as
Unicode text files; at this point, human editing ceased and the script took over to split the strings into words and perform the entire analysis.

The project included the following corpora (used separately).

<table>
<thead>
<tr>
<th>Code</th>
<th>Language</th>
<th>Texts</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>E1</td>
<td>English</td>
<td>65 novels from Swift to Conrad</td>
<td>Author</td>
</tr>
<tr>
<td>E2</td>
<td>English</td>
<td>32 epic poems from Milton to Tennyson</td>
<td>Author</td>
</tr>
<tr>
<td>P1</td>
<td>Polish</td>
<td>69 19th- and early 20th-century novels from Kraszewski to Żeromski</td>
<td>Author</td>
</tr>
<tr>
<td>F1</td>
<td>French</td>
<td>71 19th- and 20th-century novels from Voltaire to Gide</td>
<td>Author</td>
</tr>
<tr>
<td>L1</td>
<td>Latin</td>
<td>94 prose texts from Cicero to Gellius</td>
<td>Author</td>
</tr>
<tr>
<td>L2</td>
<td>Latin</td>
<td>28 hexameter poems from Lucretius to Jacopo Sannazaro</td>
<td>Author</td>
</tr>
<tr>
<td>G1</td>
<td>German</td>
<td>66 literary texts from Goethe to Thomas Mann</td>
<td>Author</td>
</tr>
<tr>
<td>H1</td>
<td>Hungarian</td>
<td>64 novels from Kemény to Bródy</td>
<td>Author</td>
</tr>
<tr>
<td>I1</td>
<td>Italian</td>
<td>77 novels from Manzoni to D’Annunzio</td>
<td>Author</td>
</tr>
<tr>
<td>S1</td>
<td>English</td>
<td>42 works by Shakespeare</td>
<td>Genre</td>
</tr>
</tbody>
</table>

**Results**

The English novel corpus (E1, Fig. 1) was the one with the best attributions for all available sample sizes starting at the top of the reference corpus word frequency list; it was equally easy to attribute even if the first 2000 most frequent words were omitted in the analysis – or even the first 3000 for longer samples. This was also the only corpus where a perfect attributive score (100%) was achieved almost constantly, which is reflected, in the graph, by the widespread and smooth dark colour in the heatmap. The English epic poems (E2, Fig. 2), on the other hand, while displaying a 100% accuracy in some ‘pockets,’ attributed in general significantly worse than the English novels. For less frequent words, i.e. below the 2000th on the frequency list, the guessing effectiveness begun decreasing very quickly; the area of best attributive success was removed away from the top of the word frequency list, into the 1000th-2000th most-frequent-word region.
The Polish corpus of 69 19th- and early 20th-century classic Polish novels (P1, Fig. 3) showed marked improvement in Delta attribution rate when the wordlist started at some 450 words down the frequency list; the most successful sample sizes were relatively small: no more than 1200 words long.
The French corpus proved difficult to interpret because there was no clear smooth area of good accuracy (F1, Fig. 4): Delta was very successful mainly for small-sized pockets from the top of the overall frequency wordlist. In contrast, the graph for the German corpus (G1, Fig. 5) presented a success rate akin to that for the English novels, with a consistently high correct attribution in most of the studied spectrum of sample size and for samples beginning anywhere between the 1st and the 1000th word in the corpus frequency list. The best attribution was achieved in a narrow region around 1000 MFWs from the top of the list.
Fig. 5. Heat Attribution accuracy for 66 German prose texts.

Of the two Latin corpora, the prose texts (L1, Fig. 6) could serve as excellent evidence for a minimalist approach in authorship attribution based on most frequent words, as the best (if not perfect) results were obtained by staying close to the axis intersection point: no more than 750 words, taken no further than from the 50th place on the frequency rank list. The top score, 75%, was in fact achieved only once – at 250 MFWs from the top of the list.

Fig. 7. Attribution accuracy for 28 Latin hexameter poems.

Fig. 8. Attribution accuracy for 64 Hungarian novels.
The other Latin corpus, that of hexameter poetry (L2, Fig. 7), paints a much more heterogeneous picture: Delta was only successful for top words from the frequency list at rare small (150), medium (700) and large (1700) window sizes, and for a few isolated places around the 500/500 intersection point in the graph. Again, the best score of 75% is represented by two pockets at 110 and 120 MFWs counting from the top of the list.

The corpus of 19th-century Hungarian novels (H1, Fig. 8) exhibited good success for much of the studied spectrum and an interesting hotspot of short samples at ca. 4000 words from the top of the word frequency list. What was even more interesting, the hotspot was surrounded by an area of a very weak attributive success.

Fig. 9. Attribution accuracy for 77 Italian novels.  

Fig. 10. Accuracy in genre recognition for 42 works by Shakespeare.

With the Italian novels (I1, Fig. 9), Delta was at its best for a broad variety of sample sizes, but only when some 1000 most frequent words were eliminated from the reference corpus. The top Italian score, 76%, appeared only a few times for wordlists of 400, 450 and 500 words starting at the 350th and the 400th most frequent word.
The final corpus used in this series of analyses was that of 42 works by Shakespeare (S1, Fig. 10). It was also the single case where Delta was tested for genre recognition – the works were categorized as poems, tragedies, comedies, romances or histories. And while the overall reliability was poor, there is a smallish yet visible darker region in Fig. 10 for pockets of some 2500 most frequent words starting at the top, or near the top, of the word frequency list.

**Conclusions**

The graphs presented above seem to confirm the suspicions that, while Delta is still the most successful method of authorship attribution based on word frequencies, its success is not independent of the language of the texts studied. This has not been noticed so far for the simple reason that Delta studies have been done, in a great majority, on English-language prose. Yet even the switch from prose to poetry within the language of Dickens and Milton has consequences for the best-attrition region – perhaps for the simple reason that poetic texts (even those brought together in E2, a corpus of epic poetry, i.e. works of some length) provide less adequate statistics than material gathered from full novels.

Thus Delta’s high success for prose texts in general is a positive and optimistic result of this series of experiments; less cause for optimism – and less uniformity – can be seen in Delta’s behaviour in prose texts in other languages. Its high and consistent attributions throughout the frequency regions studied for the 66 German novels allows a hypothesis that Germanic languages might provide the best material for authorial attribution, and that their shared characteristics can be thanked for this. The relatively poorer results for Latin and Polish – both highly inflected in comparison with English and German – suggests the degree of inflection as a possible factor. This would make sense in that the top strata of word frequency lists for languages with low inflection contain more uniform words, especially function words; as a result, the most frequent words in languages such as English are
relatively more frequent than the most frequent words in agglutinative languages such as Latin.

While diagrams for most other languages in this series of experiments seem, at the very least, not to disprove this working hypothesis, a severe blow to its simple elegance has been dealt by the Hungarian corpus, i.e. a collection of texts in a language generally deemed the most inflected one of those under study here. To make matters worse, Delta’s success in this unlikely collection of texts was even more remarkable due to their relative similarity as representatives of the same trend in 19th-century Hungarian fiction. What seemed a difficult corpus in a difficult (i.e. highly agglutinative) language scored visibly better than the ‘easier’ corpora of Polish or Latin prose. At this point, it is worth mentioning that any statements on the relative ease and difficulty of corpora collected from various languages and literatures can be tentative at best and require further study.

The greatest methodological problem that this study shows as far as Delta is concerned is that, while ‘pockets’ of good attribution reliability can be found at a variety of parameters of culling, wordlist length and/or number of the most frequent words omitted (or not) from the top of the frequency list, ‘pockets’ of similar size can be found nearby where attribution is anything but good. This study shows that obtaining near-perfect results for, say, the top 1000 most frequent words does nothing to guarantee similar success for the top 2000 words (with the possible exception of English or German corpora, where Delta’s success has been shown to be more uniform than in the other languages studied). And that, while it might be a good idea to manipulate the above-mentioned parameters, it is not yet known how to manipulate them for a given corpus, language, genre or attribution type. It seems so far that there is no ‘best,’ or ‘most reliable,’ or ‘universal,’ value for either the moving window or its initial position in the most-frequent-word lists. This is frustrating. And this calls for finding a way to even out the pockets of better and worse parameter combinations – to average them out and
thus to eschew cherry-picking – possibly, with bootstrapping, as suggested by initial results of our recent studies (Eder, 2011; Rybicki, 2011; Eder and Rybicki, 2011). But even more frustrating is the fact that we do not know why Delta in Hungarian performs oddly compared to English because, simply, no one knows why.

References


Rybicki, J. (2009). Translation and Delta Revisited: When We Read Translations, Is It the Author or the Translator that We Really Read?, *Digital Humanities 2009: Conference Abstracts*, College Park, MD, pp. 245-47.

Since much of the testing of the script was done by one author’s graduate students, the script included a simple Tcl/Tk GUI by Rybicki (for easier operation). Both authors wish to take this opportunity to thank the happy helpers: Barbara Bajak, Izabela Jakus, Magdalena Jamrych, Monika Jaworska, Agnieszka Jucha, Małgorzata Kozieł, Malwina Kuraś, Izabela Leoniak, Anna Mikulec, Monika Obrzut, Jakub Piasecki, Agnieszka Rybus, Alicja Usień, Katarzyna Szosta, Paulina Zegar, Agnieszka Zgoll.

Colour versions of the heatmaps generated for this study can be found in the online version of this paper.
Visualization in stylometry: cluster analysis using networks

Maciej EDER
Pedagogical University, Kraków, Poland
Polish Academy of Sciences, Institute of Polish Language, Kraków, Poland

Correspondence: Maciej Eder, Institute of Polish Studies, Pedagogical University of Kraków, ul. Podchorążych 2, 30-084 Kraków, Poland
E-mail: maciejeder@gmail.com

Abstract
The aim of this paper is to discuss reliability issues of a few visual techniques used in stylometry, and to introduce a new method that enhances the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods. A promising way of extending cluster analysis dendrograms with a self-validating procedure involves producing numerous particular “snapshots”, or dendrograms produced using different input parameters, and combining them all into the form of a consensus tree. Significantly better results, however, can be obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities onto a form of a network. Additionally, network analysis seems to be a good solution for large datasets.

1 Introduction
Most of the computational methods used in stylometry have been originally introduced to solve authorship attribution problems. This fact had an immense influence on the further development of the whole discipline. The seminal study by Mosteller and Wallace (2007 [1964]) showed in a very convincing way that authorship attribution based on statistical analysis of style is ultimately the problem of classification. In its standard form, attribution is aimed at extracting a unique authorial profile from a disputed text and from texts written by possible “candidates”; the goal is to compare the profiles and to single out the matching “candidate”. Even if one deals with an open-set attribution case – where the list of possible
candidates cannot be reliably established – the general idea does not differ substantially from other classification problems.

Exact science has developed a number of well-performing, sophisticated machine-learning algorithms, suitable for classification tasks, derived mostly from the field of biometrics, nuclear physics, or software engineering, that could be easily adopted to authorship attribution. They include naïve Bayes classification, support vector machines, nearest shrunken centroids, or random forests, to name but a few (Mosteller and Wallace, 2007 [1964]; Koppel et al., 2009, Jockers et al., 2008; Tabata, 2012).

Independently, a ground-breaking monograph on Jane Austen published by Burrows (1987) ushered stylometry into literary criticism. It turned out that from a literary perspective, matching profiles of “candidates” is not as important as obtaining a broader picture of relations between different novels, types of narration, main characters’ voices, and so forth. The methods adopted or introduced by Burrows, Hoover, Craig, and others (Burrows, 1987, 2002, 2007; Hoover, 2003a, 2003b; Craig and Kinney, 2009) were very intuitive and easily-applicable to literary studies. These include principal components analysis, multidimensional scaling, cluster analysis, Delta, Zeta and Iota. Despite their limitations (the lack of validation of the obtained results being the most obvious), they are still widely used.

The reason of their popularity is that they meet the needs of literary scholars, also because they offer convincing visualizations.

Needless to say, visualization has an undeniable explanatory power. Scatterplots, maps, trees and diagrams provide an insight into the whole corpus at one glance. Moreover, they allow to draw conclusions about literature from a distant-reading perspective, through a visual interpretation of groupings and separations of several samples. Certainly, this is particularly desired in stylometry beyond authorship attribution. The attractiveness of visualization in computational literary criticism is confirmed not only by the aforementioned studies by Burrows or Hoover, but also by immense popularity of beautiful yet relatively simple plots presented by Moretti, Jockers, Posavec and others (Morretti, 2005; Jockers, 2013; Posavec, 2007; Sinclair and Rockwell, 2014). The aim of this paper is to discuss reliability issues of a few visual techniques, and to enhance the explanatory power of visualization with a procedure of validation inspired by advanced statistical methods.
2 Reliability in computational stylistics

The question of reliability in non-traditional authorship attribution has been extensively discussed by Rudman (1998a, 1998b, 2003), who formulated a number of caveats concerning corpus preparation, sampling, selection of style-markers, interpreting the results, etc. Rudman’s fundamental remarks, however, have not been preceded by empirical investigation. Experimental approaches to the problem of reliability include an application of recall/precision rates as a way of assessing the level of (un)certainty (Koppel et al., 2009), a study on different scalability issues in stylometry (Luyckx, 2010), a paper discussing the short sample effect and its impact on authorship attribution reliability (Eder, 2015), an experiment using intensive corpus re-composition to test whether the attribution accuracy depends on particular constellation of texts used in the analysis (Eder and Rybicki, 2013), a study aimed to examine the performance of untidily prepared corpora (Eder, 2013a), and so forth.

Sophisticated machine-learning methods of classification routinely try to estimate the amount of potential error that may be due to inconsistencies in the analyzed corpus. A standard solution here is a 10-fold cross-validation, or 10 random swaps between two parts of a corpus: a subset of training texts and a subset of texts used in the testing procedure.

Most unsupervised methods used in stylometry, such as principal components analysis, multidimensional scaling or cluster analysis, lack this important feature. On the other hand, however, the results obtained using these techniques “speak for themselves”, which gives a practitioner an opportunity to notice with the naked eye any peculiarities or unexpected behavior in the analyzed corpus. Also, given a tree-like graphical representation of similarities between particular samples, one can easily interpret the results in terms of finding out the group of texts to which a disputed sample belongs.

Hierarchical cluster analysis – as discussed in the present study – is a technique which tries to find the most similar samples (e.g. a literary text, etc.) and builds a hierarchy of clusters, using a “bottom-up” approach. What makes this method attractive is the very intuitive way of graphical representation of the obtained results: contrarily to the scatterplots as produced by multidimensional scaling or principal components analysis, where the goal is to interpret relative positions of several points settled on a rectangular plot, cluster analysis produces explicit links between neighboring items (see Fig. 1–4). However, despite obvious advantages, some problems still remain unresolved. The final shape of a dendrogram highly depends on many factors, the most important being (1) the particular distance measure applied.

---

1An earlier version of the Section 2 has been published in a paper discussing relations between the Greek New Testament and its Latin translation (Eder, 2013c).
to the data, (2) the algorithm of grouping the samples into clusters, and (3) the number of variables (e.g. the most frequent words) to be analyzed. These factors will be briefly discussed below.

(1) In a study of multivariate text analysis using dendrograms, Burrows concludes, “my many trials suggest that, for such data as we are examining, complete linkages, squared Euclidean distances, and standardized variables yield the most accurate results” (Burrows, 2004: 326). The distance used by Burrows is a widely accepted solution in the field of computational stylistics; there are no studies, however, that would satisfactorily explain the principles of using this particular measure. Presumably, “standardized variables” mean, in this context, relying on z-scores (i.e. scaled values) rather than on relative word frequencies. If this is true, the distance used here is in fact equivalent to the Linear Delta measure introduced by Argamon (2009: 134), a slightly modified version of the classic Delta measure as developed by Burrows (2002). There is no denying that Delta, and ipso facto the distance measure embedded in it, proved to be very effective – a fact confirmed by numerous stylometric studies; and thus it should be also applicable to hierarchical cluster analysis procedure. Even if convincing at first glance, however, the choice of this particular measure needs to be theoretically justified and confirmed by empirical comparisons with other distances.

(2) Another factor affecting the final shape of a dendrogram is the method of linkage used. In the above-cited statement, Burrows favors the complete linkage algorithm as the most effective one. We do not know, however, which were the other algorithms considered by Burrows, and we do not know what method of comparison was used to test their effectiveness. In a similar study, Hoover argues that the best performance is provided by Ward’s linkage (Hoover, 2003b); his claim is confirmed by a concise comparison of Ward’s, complete, and average linkages. Good performance of Ward’s method has been also proven in many other applications within the field of quantitative linguistics, corpus linguistics, and related disciplines. Although it seems to be accurate indeed, there is no awareness, however, that this method has been designed for large-scale tests of more than 100 samples: for the sake of speed, the optimal clustering was not a priority (Ward, 1963: 236).

(3) Even if some issues still remain unresolved, scholars roughly agree that Euclidean (normalized) distance and Ward’s linking algorithm provide acceptable results. However, the same cannot be said about the third factor cluster analysis depends on, which is the number of features (e.g. frequent words) to be analyzed, and the type of countable features (e.g. words, word n-grams, etc.).
The question how many features should be used for stylometric tests has been approached in many studies, but no consensus has been achieved: some scholars suggest using a small number of carefully selected words (often, function words), others prefer long vectors of words, and so on. Although all these solutions are reasonable and theoretically justified, the final choice of the number of features to analyze is \textit{a priori} arbitrary. This problem is sometimes referred to as “cherry-picking” (Rudman, 2003). Awareness of this issue, followed by partial solution, can be observed in the studies by Hoover (2003a, 2003b), who assesses a given corpus with a few discrete cluster analyses for different MFW values. Even if still subject to arbitrary choices, this approach gives a fairly good insight into variability of the input data. This way of dealing with uncertainty will be discussed below in detail, with its possible extension to other visualization techniques.

\section*{3 Multilayer model of written text}
As will shortly be demonstrated, even the slightest change in the experiment setup might cause a severe reshaping of the final dendrogram. Without deciding which of the three factors discussed in the previous section – linkage algorithm, distance measure, and the number of words analyzed – is more likely to affect the final shape of a dendrogram, one must admit that the first two are related to the method of clustering, while the third factor is inherently linked to certain linguistic features of analyzed texts.

Endless discussions of how many frequent words or $n$-grams should be taken into account (e.g., Mosteller and Wallace, 1964; Koppel et al., 2009; Hoover, 2003a; Burrows, 2007; Eder, 2013b; Schöch, 2013, etc.) show rather clearly that there is no universal frequency strata where the authorial fingerprint is hidden. Just the opposite, it seems that the authorial signal is spread throughout the whole frequent and not-so-frequent words spectrum, but at the same time it may become obscured by additional and unpredictable signals, which are considered noise in classical approaches to attribution. In stylometry beyond attribution, however, this “noise” is worth a closer look. Why are some authors misclassified? Which texts are wrongly attributed to a given author, and why are they linked to this very author and not to others? These and similar questions are probably much more interesting than the neverending fine-tuning of the parameters of this or that classification algorithm in order to neutralize the impact of the “noise”.

Obviously, the problem is not new. Cross-genre authorship attribution, for one, has always been a major challenge (Kestemont \textit{et al.}, 2012; Schöch, 2013). Also, there have been a few attempts to extract particular signals hidden in texts: author’s nationality (Jockers,
On theoretical grounds, function words should be responsible for authorial recognition, while content words should be more topic- and genre-related. The above-mentioned empirical studies, however, do not really confirm this assumption. There is no clear rule here, and the same words are sometimes claimed to reveal different signals. For instance, the definite article “the” is considered to discriminate British vs. American flavors of English in one study (Jockers, 2013: 105), and female vs. male language in another (Pennebaker, 2011: 42).

The difficulties with separating one specific signal suggest that a text (written or spoken) is a multi-layer phenomenon, in which particular layers are correlated. These layers include authorship, chronology, personality, gender, topic, education, literary quality, translation (if applicable), intertextuality, literary tradition (e.g. sources of inspiration), and probably many more. Arguably, literary quality somehow depends on education, genre depends on topic, authorial voice is affected by chronology, gender affects personality, and so on. Some layers might be barely noticeable, some other might become surprisingly strong. In authorship attribution, this complex system of uncontrollable layers is a problem of unwanted noise, in literary-oriented computational stylistics – an opportunity to see more.

4 Dendrogram, or one snapshot at a time

Since particular frequency strata are responsible, to some extent, for different signals hidden in a literary text, the dendrograms generated using longer or shorter MFW vectors presumably will also be heterogenous. And they actually are (Fig. 1–4); the only problem is that their variability is much bigger than one could expect and – what is worse – the changes in dendrograms’ shapes are unpredictable. Different combinations of linkage algorithms, number of MFWs and distance measures applied, one obtains a convincing example of how unstable the final results might be.
Worth noticing, however, that the authorial “leaves” on the dendrograms are usually correctly clustered regardless of the parameters used. In Fig. 1 (Ward’s linkage, 100 MFWs), most of the authors are recognized to be stylistically homogenous; the exceptions include Charles Dickens and Henry James. When the number of features increases to 300 MFWs, the “leaves” of the dendrogram are matched with no misattributions (Fig. 2). In any attempts to visualize larger groupings of texts, however, one needs to admit that the “branches” are significantly less predictable than the “leaves”: is Galsworthy stylistically similar to George Eliot or to Joseph Conrad? Is Thackeray linked to Walter Scott or to Charles Dickens? What does the main division into two large clusters mean? Figs. 1–4 might support many contradictory hypotheses.
The problems do not end here: a detailed inspection of multiple dendrograms generated for gradually increasing number of features (MFWs) shows that substantial rearrangements might occur quite suddenly. An example of this behavior is shown in Fig. 3 and 4. Cluster analysis using McQuitty’s linkage and 136 MFWs (not shown) reveals a perfect authorial recognition, but when 137 MFWs are used, the cluster for Joseph Conrad is split into two parts and remains detached (along many other substantial rearrangements of the corpus) until the same corpus is assessed at 969 MFWs (Fig. 3). *Almayer’s Folly* jumps back from Kipling’s branch to Conrad’s cluster exactly between the word 969 and 970 on the frequency list (Fig. 4). The knowledge that this 970th word is “wine” does not help much, however, since multivariate analyses take into consideration a great number of features at a time. The word “wine”, not very discriminative itself, was the factor to tip the scale in favor of Conrad. What is more important here it is the side-effect: apart from the local Kipling/Conrad change, the whole dendrogram has been severely affected and, in consequence, significantly reshaped. Such abrupt changes seem to be a rule rather than the exception, at least for textual datasets.

![Image 3](image3.png)

*Fig. 3* Cluster analysis of sixty-six English novels, 969 MFWs, classic Delta distance, McQuitty’s linkage.
The decision which of the dendrograms presented above reveal the actual separation of the samples and which show fake similarities is not trivial at all. Generating hundreds of dendrograms covering the whole spectrum of MFWs, a variety of linkage algorithms, and a number of distance measures, would make this choice even more difficult. At this point, a stylometrist inescapably faces the above-mentioned cherry-picking problem (Rudman, 2003). When it comes to choosing the plot that is the most likely to be “true”, scholars are often in danger of more or less unconsciously picking the one that looks more reliable than others, or that simply confirms their hypotheses. If common sense is used to evaluate the obtained plots, any counter-intuitive results will be probably dropped simply because they do not fit the scholars’ expectations. An interesting variant of cherry-picking is discussed by Vickers, who writes about the “visual rhetoric” of different lines, arrows, colors and so forth added to a graph; while helpful, at the same time they suggest apparent separations of samples (Vickers, 2011: 127).

4 Consensus tree, or many dendrograms combined

A partial solution of the cherry-picking problem involves combining the information revealed by numerous dendrograms into a single consensus plot. This technique has been developed in phylogenetics (Paradis et al., 2004) and later used to assess differences between Papuan languages (Dunn et al., 2005). It has been also introduced into stylometry (Eder, 2013b) and applied in a number of stylometric studies (Rybicki, 2012; Rybicki and Heydel, 2013; van Dalen-Oskam, 2014). This approach assumes that, in a large number of “snapshots” (e.g. for 100, 200, 300, 400, …, 1,000 MFWs), actual groupings tend to reappear, and apparent

*Fig. 4 Cluster analysis of sixty-six English novels, 970 MFWs, classic Delta distance, McQuitty’s linkage.*
similarities are likely to remain accidental. The goal, then, is to capture the robust patterns across a set of generated snapshots. The procedure is aimed at producing a number of virtual dendrograms, and then at evaluating robustness of groupings across these dendrograms. If a given link – say, between Richardson’s *Pamela* and Fielding’s *Tom Jones* – turns out to appear frequently enough, it is reproduced on a consensus plot. In other words, several regular (yet virtual) dendrograms “vote” for the most robust links – the procedure summarizes the information on clustering from particular plots.

*Fig. 5* Consensus tree of sixty-six English novels, 100–1000 MFWs, classic Delta distance, Ward’s linkage.
In Fig. 5, a consensus tree of the corpus of 66 English novels has been shown (the “snapshots” were computed for 100, 200, 300 etc. up to 1,000 MFWs). Some texts groupings can be easily identified, including, among others, an expected cluster of the three Brontë sisters, and a branch of Kipling/Conrad – clearly subdivided into two distinct authorial voices. Unlike typical dendrograms, however, the established links do not represent stylometric distances between samples. Instead, they indicate the strength of the consensus, or the repetitiveness across a number of virtual “snapshot” dendrograms.

Upgrading the procedure from a cherry-picked cluster analysis into a consensus tree is a significant step towards reliable stylometry. Such a tree captures the average behavior of a corpus for a given frequency strata (in this case, 100–1000 MFW). More importantly, it filters out local disturbances (artifacts) that could otherwise be considered as valid results. Some arbitrary decisions cannot be avoided, though. They include the number of features to be assessed, the number of iterations (“snapshots”) to produce a consensus tree, and – last but not least – the linkage algorithm embedded in the whole procedure. A considerably simple way to neutralize these issues is to reproduce a given experiment using different settings. Sooner or later, however, other limits of consensus tree approaches become painful, especially when the number of analyzed texts increases. The technique introduced below is aimed at overcoming these limits.

5 Consensus network, or importance of runners-up

Although the problem of unstable results can be partially by-passed using consensus techniques, two other issues remain unresolved. Firstly, when the number of analyzed samples exceeds a few dozen, the plot becomes cluttered and thus illegible. Secondly, the procedure of hammering out the consensus is aimed at identifying nearest neighbors only, which means extracting the strongest patterns (usually, the authorial signal) and filtering out weaker textual similarities. Consequently, samples on a consensus tree are very likely to be grouped into many discrete authorial clusters rather than into a few larger branches. When the number of analyzed texts is considerably small, the granulation of clusters is barely noticeable (Fig. 5), in large corpora; however, numerous little branches are linked directly to the root of the dendrogram. Usefule in explanatory authorship attribution, such a plot will not support stylometric interpretations of similarities between texts, authors, genres, styles or literary epochs. Arguably, large-scale stylometry will be interested in deeper textual relations rather than in mere nearest neighborhood.
To overcome the two aforementioned issues, it seems reasonable to leverage the idea of consensus, in terms of embedding it into a flexible way of visualization. Techniques of network analysis seem to be particularly promising.

The concept of network has already been used to assess linguistic data: the applications included an analysis of syntactic structures in English (Ferrer i Cancho, 2005), syntactic structures in Czech, German and Romanian (Ferrer i Cancho et al., 2004), commonly occurring English adjectives and nouns (Newman, 2006: 14), word associations (Lancichinetti, 2011: 17; Lai et al., 2004). Network analysis has been also used to compare differences between several texts in a corpus, namely, to investigate the process of word network growth given a number of $n$ sequences (Caldeira et al., 2006), and recently to visualize relations in a corpus of a few hundreds English novels (Jockers, 2013). The method introduced below is somewhat inspired by these studies. It relies on the assumption that particular texts can be represented as nodes of a network, and their explicit relations as links between these nodes. The most significant difference, however, between the approaches applied so far and the present study is the way in which the nodes are linked. This new procedure of linking is twofold: one of the involved algorithms computes the distances between analyzed texts, the other is responsible for establishing a consensus of links.

A typical approach to authorship attribution involves a comparison of a disputed (anonymous) sample against a reference corpus, in order to identify the nearest neighbor of the disputed sample. To do this, stylometric distance between each pair of samples is estimated, then the texts are ordered from the most to the least similar. To give an example: in the case of The Jungle Book by Kipling, the ranking begins with Kim (the nearest neighbor), the next is Captains Courageous, then Lord Jim by Conrad, and so on, and the last place in this procession is given to Gulliver’s Travels by Swift. Each text in the corpus is associated with its own ranking of neighbors, from the nearest to the farthest one.

Now, these rankings can be re-used to produce a stylometric network. In a simple variant, the links would be established between nearest neighbors only: Kipling’s The Jungle Book connected to Kim, Hardy’s Far from the Madding Crowd connected to Jude the Obscure, and so forth. However, since in literature-oriented studies, weaker or hidden textual relations are potentially more interesting than explicit similarities, it makes sense to use the rankings more extensively. In stylometric terms, it means that runners-up (i.e. a few texts that have been ranked immediately after the nearest neighbor) should not be excluded from the analysis, even if, in typical approaches to classification, these runners-up are considered as unwanted noise and routinely filtered out.
Let the algorithm establish, then, for every single node, a strong connection to its nearest neighbor (i.e. the most similar text), and two weaker connections to the 1st and the 2nd runner-up. The outline of the algorithm is represented in Fig. 6 (top). Consequently, the final network will contain a number of weighted links, some of them being thicker (close similarities), some other revealing weaker connections between samples. Arguably, in most literary analyses, the thick connections will betray authorial similarities (usually the strongest stylometric signal), while thin links will reflect hidden layers of subtle intertextual correlations. In this paper, it is assumed that three neighbors – a nearest one and its two runners-up – provide enough information about weaker similarities. However, one can set any number of neighbors to be connected. An empirical comparison of different ways of connecting the nodes will be discussed in a separate study.

The second algorithm (Fig. 6, bottom) is aimed at overcoming the problem of unstable results. It is an implementation of the idea of consensus dendrograms as discussed above into network analysis. The goal is to perform a large number of tests for similarity with different number of features analyzed (e.g. 100, 200, 300, ..., 1,000 MFWs). Finally, all the connections produced in particular “snapshots” are added, resulting in a consensus network. Weights of these final connections tend to differ significantly: the strongest ones mean robust nearest neighbors, while weak links stand for secondary and/or accidental similarities. Validation of the results – or rather self-validation – is provided by the fact that consensus of
many single approaches to the same corpus sanitizes robust textual similarities and filters out apparent clusterings.

The two algorithms combined, one is presented with a robust picture of actual (strong) clusterings, emerging from an ethereal web of weaker stylistic similarities in the background. The above two-fold procedure of linking is implemented in the package ‘stylo’, an open-source stylometric library written in the R programing language (R Core Team, 2013) and available at CRAN repository (http://cran.r-project.org).

The next crucial step in network analysis is to arrange the nodes on a plane in such a way that they reveal as much information about linkage as possible. Apart from very small networks that can be arranged manually, usually an algorithmic layout is applied. In the present study, one of the force-directed layouts was chosen, namely the algorithm ForceAtlas2 embedded in GEPHI, an open-source tool for network manipulation and visualization (Bastian et al., 2009). Force-directed layouts perform gravity-like simulation and pull the most-connected nodes (i.e. the ones that have several links and/or their links are very strong) to the center of the network, while the least connected nodes are pushed outside.

A network produced using the above procedure is fairly informative per se: it usually reveals some clusterings discoverable with the naked eye, some centrally located nodes as well as peripheries, some denser and sparser areas, and so forth. At the same time, however, such a network can be subjected to a variety of standard measures used in networks analysis, which make the interpretation of the results more complete. These include measures of network size, its density, centrality of the nodes (closeness, betweenness, degree), and others. The measure of modularity, used as a community detection tool, might be particularly helpful to interpret clusters of stylistically similar texts.

In Fig. 7, a network of sixty-six English novels produced using the above procedure is shown. Spatial arrangement of the nodes was established by the said force-directed layout, the nodes’ colors were assigned according to the modularity measure. The network is clearly split into a few groups that obviously confirm the predominance of authorial signal in the dataset. What is more interesting, however, is the relations between particular authorial clusters – and this is one notable advantage of networks over consensus trees. The outliers include Austen, Trollope, James and Conrad while the central parts are occupied mostly by the works of Dickens and Sterne. A circle of immediate satellites formed by Hardy,

\[2\] The newest versions of the package ‘stylo’ are posted at the Computational Stylistics Group webpage (https://sites.google.com/site/computationalstylistics/), with a concise manual, installation instructions, and other supplementary materials.
Galsworthy, the Brontës, Richardson, Fielding and Thackeray is also noteworthy. Moreover, modularity-based color assignment sheds new light onto the already-interesting picture: while different works of a given author are usually recognized to form a distinct group, notable exceptions include a common cluster for Richardson, Fielding, Swift and Scott; another common cluster is formed by the Brontë sisters, and the Dickensian oeuvre is split into two discrete groups (quite well connected with each other, though). Last but definitely not least, the network clearly shows a chronological pattern undiscoverable using consensus trees: a diagonal timeline beginning at the left side of the network, i.e. the late 18th-century area occupied by Fielding, Richardson and Swift, through the Victorians (roughly in the middle), all the way to the early modernist Joseph Conrad.

![Consensus network of sixty-six English novels](image)

**Fig. 7** Consensus network of sixty-six English novels: classic Delta distance, 100-1000 MFWs, modularity 0.5.

Modularity is not the only way in which stylistic properties of particular texts/nodes can be assessed. Another useful yet extremely simple measure is the degree or the number of connections that a particular node has. The real potential of this measure, however, comes on stage when the nodes are re-linked to form a directed consensus network.

### 6 Directed network, or seeking stylistic hubs

In the variant of a network discussed so far, all the connections of particular “snapshots” were simply added, regardless of their direction. It means that any two nodes are connected no matter if the node $A$ points to $B$ as its neighbor, or if is pointed to by $B$. It is true that in most
cases the relation between the nodes is mutual. However, since the rankings of candidates are calculated independently for every single text in a corpus, some non-symmetrical relations might occur as well. This is particularly the case when untypical texts are analyzed: such a text will point to its nearest neighbors anyway, but it would hardly be pointed to by other texts. Arguably, a directed network will discover such situations.

The procedure of establishing the connections does not differ from the undirected variant as introduced above, except that the direction of the links is recorded. Also, any mutual relations are not summed into one connection, but kept as two independent links: $A \rightarrow B$ and $A \leftarrow B$. Consequently, every single node will have, by definition, at least three outcoming links pointing to the nearest neighbor and to two runners-up. It is possible, however, that a minority of well-defined nodes might send numerous links in different directions, while others would constantly point to but three neighbors. And the other way around: it is possible that some nodes receive a vast majority of links from the entire network, while other nodes remain unpointed. In other words, measuring the number of connections of particular nodes should lead to identifying “hubs”, or texts that are stylistically followed (high incoming degree), and the stylistic followers (high outcoming degree).

![Consensus network of sixty-six English novels (directed): the degree of outcoming links marked in color.](image)
In Fig. 8, a directed consensus network with node coloring according to outdegree is shown. One can easily identify a few hotspots – they represent the “radiating” hubs, or the texts from which the number of outcoming links is the highest. These are: *Dorian Gray* by Wilde (12 links), *Sentimental Journey* by Sterne (10), *Kim* by Kipling (10), *Tom Jones* by Fielding (9), and *Agnes Grey* by Anne Brontë (9).

It is easy to explain the behavior of *Dorian Gray* and *Tom Jones*, one might say, since these are the only novels by Wilde and Fielding, respectively, included into the corpus. In the absence of natural nearest neighbors – i.e. other texts written by the same author – the analyzed novels blindly seek any similarities around. On the other hand, however, this does not apply to *Wuthering Heights*, the only novel Emily Brontë: she turns out to be surprisingly introvert, with her mere 5 outcoming links, while her elder sister sends links to 9 novels by Austen, Eliot, Trollope, Dickens, and Charlotte Brontë. It is also surprising to see the extroversion of Sterne’s *Sentimental Journey*, especially when compared with a very modest behavior of *Tristram Shandy*.

Since the procedure of linking the nodes is based on classification principles, the existence of radiating hubs betrays the texts likely to be misclassified in a real-case authorship attribution study. A provisional interpretation of this phenomenon is that a given text turns into a radiating hub whenever it lacks in strong authorial signal, or when its authorial voice is overshadowed by other signals: genre, gender, chronology and so forth. Needless to say that the ability of detecting radiating hubs makes this technique a potentially useful addition to authorship attribution toolbox – as a straightforward way to identify unstable samples.
From a literary point of view, however, the incoming links are potentially much more interesting, especially when they happen to form any “absorbing” hubs. Such a hub represents a text pointed out as the nearest neighbor by several other texts from the corpus. Measure of incoming links, or indegree, applied to the corpus of sixty-six English novels is represented in Fig. 9. Two major absorbing hubs can immediately be spotted; they focus on two novels by Dickens, *David Copperfield* and *Little Dorrit*. Two other hotspots are also fairly noticeable, namely *Middlemarch* by Eliot and *Nicolas Nickleby*, again by Dickens. Poorly connected novels found their place on the other pole of the indegree measure: *Dorian Gray* by Wilde (no incoming links at all), Sterne’s *Sentimental Journey* (a single yet very strong incoming link from *Tristram Shandy*), and Swift’s *Gulliver* (a single strong link from *A Tale of a Tub*).

Unlike radiating hubs, the absorbing ones are harder to interpret. In social sciences, physics etc., the hubs are usually considered to betray the most important events/agents/phenomena. In stylistics, however, what they really mean remains largely open to dispute. Jockers’s approach to the question of literary influence seems to assume that the hubs indicate the most influential works (Jockers, 2013: 154–168). Arguably, however, the picture is far more complex here.

The most striking observation is that according to the incoming links, Dickens would have had to live much earlier to have influenced Richardson, Sterne or Swift. Is it the method,
then, that is wrong, or the interpretation? In the aforementioned study on literary imitation, Jockers filters out all textual similarities that could not have happened due to chronological reasons, before undertaking actual analysis (ibid., 163). However, discarding the backward time links cannot deny the fact that they do appear in the corpus.

It seems reasonable to assume that the absorbing hubs should be interpreted as sources of stylistic influence in a very broad sense, for instance as witnesses of stylistic mood of an entire literary epoch. It is true that these hubs might indeed indicate the most influential texts (copied, paraphrased, sequelled, consciously/unconsciously imitated, and so forth). At the same time, however, they might also reflect texts stylistically “average”, typical for their times rather than exceptional. In any case, the absorbing hubs betray texts lacking in a single, distinct stylistic signal.

A slightly oversimplified interpretation of both types of hubs might be as follows. The absorbing hubs stand for receivers of stylistic appreciation (regardless of their actual stylistic quality), radiating hubs represent emitters of stylistic appreciation (not mere followers, though, since they do not follow a single author).

7 Conclusions
In the present study, a few reliability issues of explanatory methods used in stylometry were discussed. They include unstable output – because final results highly depend on the setup of the experiment – as well as lack of validation. A promising way of extending cluster analysis dendrograms with a self-validating procedure involved producing numerous particular “snapshots”, or dendrograms produced using different input parameters, and combining them all into the form of a consensus tree. This approach, however, inherits some drawbacks of cluster analysis – dependence on a chosen linkage algorithm being the most painful – and introduces a few new pitfalls: granulation of clusters, and cluttered visualization when a corpus becomes large.

Significantly better results were obtained using a new visualization technique, which combines the idea of nearest neighborhood derived from cluster analysis, the idea of hammering out a clustering consensus from bootstrap consensus trees, with the idea of mapping textual similarities into a network. Additionally, network analysis seems to be a good solution for large datasets.

The added value of consensus trees over standard dendrograms is the reliability of the results represented in a plot, the added value of stylometric consensus networks is at least three-fold: the reliability inherited from consensus trees, insight into a more complete picture
of textual relations beyond mere nearest neighborhood, and, last but not least, the capability of handling dozens, or even hundreds of text samples in a single plot. The only limitation here seems to be the paper size one wants to use for drawing a literary network. Regardless of the printing issues, however, the aim of this study was to encourage stylometrists to produce a reliable map of literature in its entirety, and to propose a methodological background for such a map.

References


SHAKESPEARE, COMPUTERS, AND THE MYSTERY OF AUTHORSHIP

HUGH CRAIG AND ARTHUR F. KINNEY
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of figures</td>
<td>viii</td>
</tr>
<tr>
<td>List of tables</td>
<td>xii</td>
</tr>
<tr>
<td>Notes on contributors</td>
<td>xiv</td>
</tr>
<tr>
<td>Preface and acknowledgments</td>
<td>xv</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Hugh Craig and Arthur F. Kinney</td>
<td></td>
</tr>
<tr>
<td>2 Methods</td>
<td>15</td>
</tr>
<tr>
<td>Hugh Craig and Arthur F. Kinney</td>
<td></td>
</tr>
<tr>
<td>3 The three parts of Henry VI</td>
<td>40</td>
</tr>
<tr>
<td>Hugh Craig</td>
<td></td>
</tr>
<tr>
<td>4 Authoring Arden of Faversham</td>
<td>78</td>
</tr>
<tr>
<td>Arthur F. Kinney</td>
<td></td>
</tr>
<tr>
<td>5 Edmond Ironside and the question of Shakespearean authorship</td>
<td>100</td>
</tr>
<tr>
<td>Philip Palmer</td>
<td></td>
</tr>
<tr>
<td>6 The authorship of The Raigne of Edward the Third</td>
<td>116</td>
</tr>
<tr>
<td>Timothy Irish Watt</td>
<td></td>
</tr>
<tr>
<td>7 The authorship of the Hand-D Addition to The Book of Sir Thomas More</td>
<td>134</td>
</tr>
<tr>
<td>Timothy Irish Watt</td>
<td></td>
</tr>
<tr>
<td>8 The 1602 Additions to The Spanish Tragedy</td>
<td>162</td>
</tr>
<tr>
<td>Hugh Craig</td>
<td></td>
</tr>
<tr>
<td>9 Transforming King Lear</td>
<td>181</td>
</tr>
<tr>
<td>Arthur F. Kinney</td>
<td></td>
</tr>
</tbody>
</table>
Contents

Conclusion 202
   Arthur F. Kinney

Appendix A  Plays in the corpus 212
Appendix B  A list of 200 function words 221
Glossary 223
Index 228
In almost anything we read there are phrases that seem to resonate with an authorial voice. Over longer stretches of writing all sorts of small signals confirm that we are in touch with a recognizable originating consciousness, even if we realize that this imagined source is our own conjecture, created from the indirect but familiar indications in what we read or hear. Without thinking about it too much, we perform an intuitive calculus on a new work or passage to test it for likeness to an authorial style we have previously internalized. The passages seem to us either authentic or not.

But what, say, of Hal, Hotspur, and Falstaff? They inhabit the same play, but each has a recognizable style that sets him apart from the other two. Can there be a single, identifiable Shakespearean language that unites their three very different kinds of speech? What could it be about the way they speak that would actually unite them, and separate them from characters created by other writers of the same time writing in the same genre, and even drawing on the same conventions of the wily villain, the vain-glorious soldier, the braggart hero, and so on?

There are other impediments to identifying an authorial style in a systematic way. For instance, how can we be sure that some of the distinctive Shakespearean phrasings are not in fact common expressions from his own time? How confident are we that we can detect authorship as a steady patterning, rather than through occasional highlights? And when we explain our judgments on authorship to others, how can we do more than make some unsatisfactory generalizations, or offer some sample passages trusting that they will strike the same chord?

Consider the following extracts from three plays from the late sixteenth and early seventeenth centuries:

I meant indeed to pay you with this, which if like an ill venture it come unluckily home, I break, and you, my gentle creditors, lose. Here I promised you I would be, and here I commit my body to your mercies. Bate me some, and I will pay you some, and (as most debtors do) promise you infinitely.
This is a sleepy tune. O murd’rous slumber!
Layest thou thy leaden mace upon my boy,
That plays thee music? Gentle knave, good night;
I will not do thee so much wrong to wake thee.

But since you have made the days and nights as one,
To wear your gentle limbs in my affairs,
Be bold you do so grow in my requital
As nothing can unroot you.

There are numerous aspects of these passages that might help connect them with Shakespeare, even if we were not familiar with the passages themselves (from 2 Henry IV, Julius Caesar, and All’s Well that Ends Well, respectively¹). They are all rich in metaphors; they are linguistically fanciful and playful; in manner, they are all more or less courtly and engaging. Their syntax is relaxed and varied, but always resolves into definite closures of structure and sense.

They share a further element linking them to regular Shakespeare practice that readers are less likely to notice: they all use the word gentle. This, it turns out, is a favourite Shakespeare word. It was equally available to him and to his contemporaries. It was by no means unusual, or neologistic, or archaic (OED). Yet gentle occurs nearly twice as regularly in Shakespeare plays as in Early Modern English dialogue generally. This is not just a matter of particular scenes, with a cluster of instances used for a special local effect, or by one character. It is a persistent preference, a minor but recurring thread in Shakespeare’s linguistic fabric. To confirm this we can divide Shakespeare plays, and plays by others, into short sections, and then check how many (or how few) of the sections have an instance of the word. (In this way a particular cluster of occurrences is discounted: we are just counting whether there is an instance in a given segment, or not.) If we take 27 Shakespeare plays, and 109 plays from the period by other playwrights, and divide each into 2000-word segments, we find that while more than half of the Shakespeare segments have at least one example of gentle, the word appears only in 3 of every 10 of the others (Figure 2.1). It seems fitting, given this pattern of difference, that the word gentle was used of Shakespeare as a person. Ben Jonson calls him ‘My gentle Shakespeare’ in his commendatory poem in the First Folio. For Shakespeare’s audience the word would have had associations first with the gentry and nobility, and only secondly with attributes of personal

¹ 2 Henry IV, Epilogue, lines 10–16; Julius Caesar, IV.iii.267–70; All’s Well that Ends Well, V.i.3–6. Lineation of Shakespeare plays and quotations from Shakespeare here and below are from G. B. Evans et al., eds., The Riverside Shakespeare, 2nd edn (Boston: Houghton Mifflin, 1997).
behaviour. In 1596 Shakespeare’s father was granted a coat of arms, giving male members of the family the right to write ‘gentleman’ after their names. We might speculate that Shakespeare’s preference for the word gentle is connected with this interest in family status.

Such speculations would never fully explain the motivations lying behind this or any other facet of Shakespeare’s vocabulary. But in seeking an empirically based profile of this vocabulary, gentle is a beginning, pointing the way to a method that might be effective. What we need to do is to find more markers like this one, and to collect other markers that indicate that Shakespeare is not the author of a given passage. With a multitude of such markers we could transform the interestingly lop-sided distribution of Figure 2.1 into something more like a comprehensive picture. We would be hoping for success in separating all Shakespeare segments from all others, something on which we could build a reliable test for any segment of sufficient length.

Some such markers have already been identified in discussions of authorship. MacDonald P. Jackson, for instance, has observed that Shakespeare’s characters use yes (rather than yea or aye) only where a special emphasis is required, or to respond to a negative, where other playwrights like Jonson and Fletcher adopted yes as the standard form.² In the present set of play segments, yes emerges as the strongest of all the markers of others’ work as opposed to Shakespeare’s. Brave is also high on the list. It occurs in one

in five segments of Shakespeare’s plays, compared to two in five in the run of contemporary drama. To these we can add more instances of the distinctiveness of Shakespearean vocabulary. *Answer* and *beseech* come up much more often in Shakespeare dialogue than the dialogue of others. He had relatively little recourse to the word *sure* and to the plural form *hopes*. There is the beginning of an individual Shakespeare lexicon here, the first indications of a profile in word choice that marks him out from all his fellow dramatists.

A disputed passage in which some or all of *gentle, answer, and beseech* occur, and in which there are no instances of *yes, brave, sure, and hopes*, shows an affinity with known Shakespeare. If we add to these hundreds of other Shakespeare ‘marker’ words, we can be more confident in assigning a mystery passage to Shakespeare, or away from him. This is the basis for the method devised by John Burrows and called Zeta. In the current chapter we present one variety of the Zeta test, a principle of selection to allow us to make use of hundreds of words that are unusually common in the author of interest, and of hundreds that are not, to make two axes of authorial differentiation.

We can begin by taking the set of plays by Shakespeare, and the other group of plays by his peers, and seeing if we can separate the two simply by tracking the appearances and absences of sets of words like *gentle* and *brave*. Our usual practice is to establish two sets, each of 500 words, one of which in this case will be markers of Shakespeare, and the other markers of the plays by his peers.

For this purpose we regard *hope* as a different word from *hopes*, and *beseech* as a different word from *beseeching*. We treat words the way a concordance does, rather than organizing them in the manner of a dictionary: the basic unit for our statistical work is thus the word-form rather than the headword. From one perspective this means we are missing a ‘true’ total for *hope* or *laugh*. On the other hand, this method means we are preserving an aspect of the detail of the language of the texts in our counts, and avoiding a series of decisions that might introduce their own arbitrariness and inconsistency.

As mentioned earlier, we work with 27 ‘core’ Shakespeare plays, and 109 well-attributed single-author plays by others (there is a full list in Appendix A). We divide all these into segments each of 2000 words, ignoring speech, act, and scene boundaries. Residual portions we add to the end of the last segment of each play. We are aiming to examine the plays

---

down to quite small units – as it were at the level of the scene, rather than
the level of the act or whole play – but using actual scene divisions would
make for segments of wildly uneven length. Hence the decision to choose
an arbitrary length in words, regardless of any other divisions within the
play. In all, our set of plays yielded 291 Shakespeare segments and 1009
non-Shakespeare ones.

There are two different ways to assess the power of the method. The first
is to see how well the counts of Shakespeare and non-Shakespeare marker
words differentiate between the Shakespeare and non-Shakespeare seg-
ments. Is it true that the lowest-scoring Shakespeare segment still con-
tains more of the Shakespeare words than the highest-scoring segment
by another writer? On the opposite side, do all the segments in our set
by Shakespeare’s peers contain more of what we have declared to be non-
Shakespeare words than any Shakespeare segment? If this complete separ-
ation is not achieved, how many exceptions are there? If the two sets have
large numbers of scores that overlap each other, we will know that the
diversity of vocabulary patterns within each of the two groups, and the
degree of common ground between the two, have defeated our attempts
to make a clear separation.

The second test is to see how well the method can discriminate between
freshly introduced segments that we know are either by Shakespeare or
by some other writer. These are segments that have not played any part
in the selection of marker words, reflecting the ultimate goal of testing
a new segment that is of disputed or unknown authorship. There is the
danger that a method may be too closely tailored to the separation of the
particular segments we use to choose the marker words, rather than (as
we would wish) set up as a general method for distinguishing between
any Shakespeare segment and any other. With this in mind we choose
at random one indisputable Shakespeare play and one play confidently
attributed to another playwright and put them aside for a second test.
In effect, we strip these two plays of authorship and transform them
into plays of unknown provenance, so that they can substitute for the
anonymous or disputed plays and play sections that we will want to
assign to an author.

For our first experiment, the lot fell to *Coriolanus*. We removed its 13
segments from the Shakespeare set, leaving 26 Shakespeare plays and the
full set of 109 others. We then found the 500 words that are most charac-
teristic of the slightly reduced Shakespeare set compared to the others. For
this purpose the criterion for a ‘characteristic’ word was that it appeared in
many of the Shakespeare segments and in few of the segments by others.
The formula we use for this purpose takes account of two quantities, the proportion of Shakespeare segments where the word appears, and the proportion of non-Shakespeare segments where it does not. Thus we count up all the Shakespeare segments containing the word, and divide this count by the total number of Shakespeare segments. Then we find the number of non-Shakespeare segments where the word does not appear, and divide that by the total number of non-Shakespeare segments. We then add these two proportions together to get a score that reflects the degree to which a given word is more common in Shakespeare than in the work of his peers. The highest possible score is 2, for words that appear in every Shakespeare segment (giving a score for that part of the index of 1), and in none of the non-Shakespeare ones (giving a score of 1 for that part also). The lowest possible score is zero, for words that never appear in Shakespeare and occur in every one of the segments by other authors. In practice we find no words at these two extremes; we simply choose the 500 words with the highest scores on this formula.

The word *gentle* appears at the head of the list, with a score of 1.24 (it appears in 69 per cent of the Shakespeare segments, and does not appear in 55 per cent of the segments by others). We have already discussed its lop-sided distribution between these two groups. Our cut-off for the full list is at 500. The word in this position is *heaven*, which has a score on the index of 1.03, coming much closer to the neutral score of 1 but still appearing more regularly in the Shakespeare segments. It is found in 71 per cent of the latter and in 68 per cent of the segments by other writers.

We then do the same thing in reverse to find our non-Shakespeare markers. This time we are seeking words that often appear in segments by other authors, but appear only rarely in segments by Shakespeare. This time *yes* heads the list (with a score of 1.27) and our last word is *discourse* (with a score of 1.03, the same as that for *heaven* in the other list).

We are concentrating on lexical words for this particular authorship test, words with semantic content like *gentle* and *brave*. There is a separate group of words, the function words, those with grammatical force rather than meaning, like *the* and *but*, most of which are too common for our purposes here. We are looking for words that are absent from some segments, whereas there are only a limited number of function words, and they occur very regularly, so that one can be sure of finding most of their number in any passage of reasonable length. (We employ this group in a second method, described later in this chapter.) It is convenient to establish a list of the function words and exclude them from the test from the beginning. There is a much larger range of different lexical words available
Methods

and most of them occur irregularly and in small numbers. In our text of the 1604 *Hamlet*, for instance, there are nearly 160 words that occur 25 times or more, but more than 4500 that appear fewer than 25 times.

Having compiled lists of the Shakespeare and non-Shakespeare marker words in this way, we counted how many of them appeared in each segment. For this purpose, as before, one appearance was enough; the only distinction was between no instances and some. Each segment's count of Shakespeare words and non-Shakespeare words can then be compared to the counts for other segments.

To take one example, 84 of the 500 Shakespeare marker words appear in the first segment of *All's Well that Ends Well*. It does not have *gentle* but it does have three instances of *answer*, and so on down the full list of marker words. On the non-Shakespeare list this segment's score is 63: no instance of *brave*, but one of *hopes*, another prominent non-Shakespeare marker. (The particular numbers are only important in relation to the general pattern of such counts for Shakespeare segments and for segments by others.) We cannot rest simply with the totals of appearances, since some segments — the final segments of each play, which include the remainder after we have divided the total of words by 2000 word-units — are longer than others, and this needs to be taken into account. We therefore divide the number of marker words that appear in each segment by the total number of different words in that segment. For each segment, then, we ask two questions. Firstly, what proportion of the various different words it uses are Shakespeare ‘favourites’? And, secondly, what proportion are words he generally avoids? Taking results for all 1298 segments like this we plot the counts on two axes, one for the Shakespeare and one for the non-Shakespeare markers. Figure 2.2 shows the results. This is an unfamiliar way to think about the plays. It is unusual to think of so many plays in a systematic relation to each other. It is odd to divide them up into even-sized portions. Out of all the myriad possible linguistic, literary, and dramatic aspects of the segments we have chosen just two to represent them in a chart, both counts of their word-use, a positive and a negative measurement of closeness to Shakespeare’s overall vocabulary preferences. Yet these counts consolidate a great deal of information about individual word-use, and give us a common basis on which to compare all the segments. Each of the segments now has a place in a map that combines the two counts. We have established a region to which we can expect genuine Shakespeare segments to gravitate — the bottom right of the graph, defined by high scores on Shakespeare words, and low scores on the words he habitually neglects. The upper-left quadrant, correspondingly, is the
region where we expect segments by others will appear, with low scores on the Shakespeare words and high scores on the others.

If we look first at the two large sets of entries, the Shakespeare diamonds and the non-Shakespeare dots, we can see that the differences are large enough, and consistent enough across the two groups, to separate most of the segments. There are dots that appear in the area dominated by diamonds, but they are few in number and they are in the upper-left part of the cluster formed by the diamonds. Similarly, there are diamonds with scores that mean that they appear within the cluster of the dots, but these are isolated and well outside the heart of the dots’ cluster. We can make a simple numerical estimate of the success rate by drawing an imaginary straight line between the two clusters. Calculation shows that 17 of the non-Shakespeare segments fall to the Shakespeare side of this line, and five of the Shakespeare segments fall on the non-Shakespeare side, overall 2 per cent of the total, giving a success rate of 98 per cent. It is remarkable that one author’s preferences in vocabulary are so consistent that counting

---

Figure 2.2 Lexical-words test: 2000-word Shakespeare segments versus 2000-word segments by others, with 2000-word segments of Coriolanus.

---

4 For this purpose we use the perpendicular bisector of the line joining the centroids of the two clusters, as discussed in Chapter 3, below (p. 54).
words from two lists gives us a way of classifying the vast majority of them correctly.

The graph shows that neither of the two measures on its own would suffice to separate the two clusters. On the horizontal axis, which shows the counts of Shakespeare words, one can see by eye that the average Shakespeare count is much higher than the average non-Shakespeare one; but there is still a large overlap, a span occupied by both dots and diamonds. When we include the counts on non-Shakespeare markers – the vertical axis – though, the two groups separate almost completely. To tell the Shakespeare segments from the others, we need to know how many of the words that Shakespeare usually avoids appear in a given segment, as well as the number of his favourites. Once we have done this, the choices the authors habitually make within a shared vocabulary serve as a simple and powerful means of telling them apart. Admittedly, even with both counts in play, there are some exceptions, creating a band of the graph shared between the two groups. Given that there are 278 Shakespeare segments, and 1009 non-Shakespeare ones, though, this is an impressive performance. Combining the counts of Shakespeare and non-Shakespeare marker words gives us a simple way of distinguishing the two kinds of segment, which we can expect will fail in only a handful of cases.

If we turn to the Coriolanus segments, the circles, which played no part in drawing up the two lists of words, we see that they do indeed fall in Shakespeare territory. We can conclude that with this method we could assign a Shakespeare segment to Shakespeare even if it had played no part in arriving at the word lists, not with absolute certainty, but with considerable confidence.

We can try the same procedure again, this time including a randomly chosen play by a playwright other than Shakespeare, Thomas Middleton’s Hengist, King of Kent. We compiled new lists, drawing on all twenty-seven Shakespeare plays this time, compared with a non-Shakespeare set without Hengist. The results are shown in Figure 2.3. The Hengist segments, like the Coriolanus ones, were free to fall wherever their counts of the marker words took them. Nine of the ten Hengist segments are placed in unambiguously non-Shakespeare territory, while one appears in a region shared between Shakespeare and the others. We are reminded that the segments used to establish the pattern of distinction between Shakespeare and the others do overlap. Grey diamonds appear well to the left and higher up than the main group, and black dots stray down and to the right. There is certainly a strong tendency to separate, but the evident variation within the two groups means that occasionally in a given segment Shakespeare
Hugh Craig and Arthur F. Kinney

uses fewer of his favourite words, and more of those he normally eschews, so that on these measures it could on this basis be mistaken for a passage from one of his contemporaries. (The grey diamond in the upper-left-hand corner of the group is in fact the third segment of All’s Well that Ends Well, extending from a point within I.iii.186 to a point within II.i.194.) In the same way one of the Hengist segments uses sufficiently few of the non-Shakespeare words, and at the same time just enough of the Shakespeare words, to overlap the Shakespeare group.

It is clear that the method can draw out patterns from the lexical words data that serve to distinguish Shakespeare segments from others, with effectiveness even in the more difficult case of test segments from outside the ones used to establish the groups of marker words. In the odd case, we must concede, a segment is not decisively placed in what we know to be the correct cluster. We have a method that is powerful but not infallible.

This is a crucial point. The results of computational stylistics are always matters of probability, not of certainty. Writers are free agents, and language is an endlessly flexible instrument. Writers tend to remain within a defined band of style, but this is a propensity, not an iron law. In the past, quantitative work in literary studies has sometimes suffered from exaggerating the reliability of its findings. In 1989, for example,
Donald W. Foster presented results that showed a newly discovered poem falling again and again within the Shakespearean range. After further elaboration by Foster, and a considerable controversy, the poem was subsequently included in several editions of Shakespeare’s works. In the end, however, some of Foster’s results did not stand up to scrutiny, and it proved that even on his chosen measures Shakespeare’s style did not always match the style of the poem. The poem has since been definitively attributed elsewhere when an obviously superior candidate, John Ford, was found.5

Our method relies on the careful collection and weighing of data, and must be rigorously pursued. Even at its best it can mislead, by deceptively unequivocal results. The possibility of error must always be borne in mind. Yet none of this should obscure the main facts. The graphs confirm our intuitions that authorial individuality is powerful. They also establish, in a way readers’ impressions can never do, that this effect is consistent and marked even in the mixed dialogue of plays produced in a highly collaborative and tightly interconnected theatre practice like the one Shakespeare knew.

The distribution of strategically chosen lexical words evidently offers the foundation for some useful authorial discrimination. It makes sense that writers have preferences for some words, and a tendency to neglect others. The analysis of large amounts of text shows how consistent and marked these patterns are throughout a career, even one as long as Shakespeare’s. There are good reasons for looking for a second test as well. There will always be an element of error in any single method, and the graphs above show this, in those occasional data points of known origin that declare themselves as ambiguous, or frankly as members of the wrong side of the classification. After all, we are dealing with writers who are at liberty to imitate each other, to try new styles, and to write differently for a particular occasion or in a new genre, and we are trying to corral this vast conscious and unconscious literary activity with what are, at heart, very simple rules: either the segment in question uses a given word, or it does not. It is no wonder that success in discrimination is not complete. If we could find a second method, working independently of this one, so much the better. This also makes life more complicated, however. When

the methods fail to support each other, we are abruptly reminded of the residual uncertainty in each one (a French proverb says that ‘Someone with a watch knows what time it is; someone with two watches is never sure’). If the two methods agree, on the other hand, we start to build a really strong case.

The obvious place to look for a second computational-stylistics method is in the function words, those like and or you, which have vital grammatical functions but little or no semantic content. As we have already noted, these words, unlike the lexical words, are limited in number, and many of them are very abundant. They form a closed set, most of which will appear in almost any 2000-word segment of dialogue. The commonest of all, the, makes up one word in thirty in the plays we are analysing. In fact these words have been the commonest base data for computational authorship work. It is well established that writers use them at different rates, so that, if we combine a number of them together, they can offer effective authorial discriminations. In this case calculating frequencies makes more sense than simply recording appearance and non-appearance.

To demonstrate a method based on frequencies of the function words we can examine a convenient series of problems where the solution is known with a degree of certainty. Brian Vickers, in Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays, discusses five cases of putative Shakespearean collaboration in detail. In each one a single collaborator is identified and the division of the play is clear-cut. Vickers details a strong scholarly consensus for each, and adds his own synthesizing and supplementing tests and discussion. It would seem safe to work on the basis that collaboration in these plays, and the details of the boundaries between parts, are so well established that they can serve as a test bed for the methods of the present book. We treat here Titus Andronicus, Timon of Athens, Henry VIII, and The Two Noble Kinsmen. We omit Pericles, where the collaborator, George Wilkins, has only one play to his name as independent work, and thus the method of establishing general stylistic differences over a wide sample cannot be applied. With George Peele (for Titus Andronicus), Thomas Middleton (for Timon of Athens) and John Fletcher (for Henry VIII and The Two Noble Kinsmen) there is a substantial canon to provide a basis for comparison with Shakespeare.

In the remainder of this chapter we will concentrate on these four cases, where we are almost sure Shakespeare collaborated with another author. This will demonstrate something of the technicalities of the methods and also give an indication of their effectiveness in an area that, on the face of it, presents great challenges for authorship study. In a collaboration the playwrights (we assume) are aiming at a single style for their play. They are working in the same genre, and on the same material. We can expect individual characteristics to be restrained. Writers may well be working outside their normal range and the demands of a particular assignment may well lead to departures from the patterns established over work in plays where they have a free hand.

In this demonstration we will use the same method in each case, forgoing the improvements that might come from adapting the methods to the particular instance for the sake of a benchmark test applied across the board. Having determined some rules for testing, we use them without variation to provide a strict estimate of their power to separate segments of known authorship. In later chapters we will adapt the methods to particular cases, taking account of genre and date in the authorial canons used to establish authorial signatures, bringing in additional statistical perspectives to wring more discriminating power out of the methods, and essaying some new methods adapted to the problem at hand.

The aim is to test methods, rather than to investigate, so we want sample material that is as uncontroversial as possible. With this in mind, we use whole scenes as defined in early editions, and only those that are generally regarded as the unaided work of one or other collaborator (in this Vickers is once again our guide). We also use only those that have 1500 words or more of dialogue, to minimize the effects of local variation. It is easy to see that very short passages can be dominated by a particular situation or subject, so that they depart from the persistent balance of elements that constitute a writer’s style. As passages get longer, the local pressures tend to balance each other out and underlying habits and preferences assert themselves more strongly. Alvar Ellegård suggested that this sort of variation stabilizes at 4000 words in the kind of expository prose he was concerned with.⁹ We have to balance the competing requirements of ideal sample length and applicability to our particular cases. A minimum of 1500 words yields us a number of scenes in each of our chosen plays, where a minimum of 4000 would yield none.

Table 2.1 shows the division of the 1594 Quarto of *Titus Andronicus* into scenes, with the length in words of each scene, as counted by our programs, and the attribution of each to Shakespeare or to George Peele, according to Brian Vickers’ chapter in *Shakespeare, Co-Author* (pp. 148–243). An extra scene appears in Act III in the Folio version of the play, but this is excluded here to keep the focus on a single early printed version. Vickers’ division builds on a long scholarly tradition of authorship work on the play. He draws on multiple markers of authorial difference, from the number of alliterations and the stress patterns of the verse to the number of polysyllabic words. We can proceed with the five scenes with more than 1500 words: I.i, attributed to Peele, and II.iii, III.i, V.ii, and V.iii, attributed to Shakespeare. This time instead of a comparison between Shakespeare and a mixed group of other writers we are interested in a Shakespeare–Peele contrast.

With the function words we cannot rely on counting appearances and non-appearances in segments as we did with the lexical words. We already know that many of the function words will appear in every segment. We need a procedure that works in frequencies and combines them so as to bring out more subtle patterns of use. We first select words that do show differences between the two authors (Shakespeare and Peele), and then use a mathematical technique to concentrate the discriminating power of the frequencies of the chosen words into a few composite factors.

A simple way to find the function words that are used most differently by two authors is Student’s *t*-test. This procedure is based on two of the fundamental statistical quantities, the mean or average and the standard deviation. The mean is simply the total of the counts for any variable divided by the number of counts. It is a summary of the underlying tendency of the group. In terms of prediction, it is the best guide to the count of any newly introduced sample from the same parent group. It is just a single number, however, and may conceal quite a wide variation: the same mean may have come about from a group of very similar values, or from a group that includes some very high and very low values. Hence the usefulness of the standard deviation, which measures the dispersion of values around the mean. The standard deviation is in the same units as the original variable (in the present work, the frequency, or proportional frequency, of a given word). It is calculated as the square root of the total variance, which is the average of the squared differences between the

Methods

samples (in our case, the segments) and the mean. The standard deviation has the property that in a statistically normal distribution approximately two-thirds of the values will lie within one standard deviation above or below the mean.

The $t$ value of a variable is the difference between the means of the two groups, divided by the standard deviation of all the counts. The test compares the means of the two groups (in this case, counts for a given word in the Shakespeare and Peele groups of segments), takes standard deviations into account, and produces a composite measure of the strength and consistency of the difference. The variables with the highest $t$ values, those in which we are most interested, will have large differences in means between the two groups and small overall standard deviations, implying a genuine difference in rates of use between the two groups.

This $t$ value can be compared to a published probability table to estimate how likely the difference is to have come about by chance alone, taking into account the number of the samples. A low probability means that the two groups are markedly, and consistently, different in their frequencies. We tested 200 function words in our collection of 27 Shakespeare plays and 4 Peele ones. Of these, 55 were regarded by the $t$-test as consistently higher or lower in one author or the other, a difference expressed as a probability of less than one in 10,000 that the groups derive from the same parent population. (The absolute level of probability is in fact not of great

<table>
<thead>
<tr>
<th>Scene</th>
<th>Length in words</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.i</td>
<td>3763</td>
<td>Peele</td>
</tr>
<tr>
<td>II.i</td>
<td>1038</td>
<td>Peele</td>
</tr>
<tr>
<td>II.ii</td>
<td>213</td>
<td>Peele</td>
</tr>
<tr>
<td>II.iii</td>
<td>2414</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>II.iv</td>
<td>484</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>III.i</td>
<td>2454</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>IV.i</td>
<td>1030</td>
<td>Peele</td>
</tr>
<tr>
<td>IV.ii</td>
<td>1433</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>IV.iii</td>
<td>947</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>IV.iv</td>
<td>860</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.i</td>
<td>1338</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.ii</td>
<td>1670</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.iii</td>
<td>1556</td>
<td>Shakespeare</td>
</tr>
</tbody>
</table>
interest here, since we are using the test primarily as a way to identify variables to be used in a further test.11)

At the top of this list of the words that Peele uses more than Shakespeare are *and* and *thy*. At the top of the list of words that Shakespeare uses more than Peele are *it* and *very*. We might suspect some broader stylistic patterns behind this: Peele’s dialogue prosier and more archaic in its grammatical forms; Shakespeare’s including more casual and domestic exchanges. A method called principal components analysis, or PCA, can do a purely mathematical equivalent, comparing patterns of individual word-use to extract some composite ‘factors’, combinations of the original word-count variables.12

The PCA method seeks to find the strongest such factor, then the second strongest independent one, and so on. A strong factor is one that accounts for a large proportion of the total variance in a table of counts. This is a ‘data reduction’ method, since it aims to create a new composite variable, a so-called principal component, which represents a good deal of the underlying contrasts and similarities in the counts. If, for instance, we have a table of the heights and weights for the adult inhabitants of a given town, then we can make a new variable that is simply the sum of the two counts, and we will find that this combined variable represents with a good deal of accuracy the patterns of variation within the two original variables. Shorter people will tend to be lighter, and taller people heavier. The new variable we could call ‘size’.13 It will not account for all the variations in the height and weight table, since there are some short individuals who are heavy, and some tall individuals who are relatively light, but it will capture a basic fact about the table, one that in a sense is the most important fact in it.

Principal components analysis is most useful where there are numerous variables, rather than just two. We may have counts of rainfall, daily

---

11 Thus the probability level does not translate into any very useful inference about these populations, and should be regarded as no more than a convenient threshold, especially as there is no presumption that the word-frequencies follow a normal distribution. The *t*-test was done on 2000-word segments, 301 from Shakespeare plays and 25 from Peele plays. We used the ‘equal variance not assumed’ results from the statistical package SPSS (version 16.0, copyright SPSS Inc., 1989–2007).


maxima and minima of temperature, barometric pressure, wind speed, and so on for hundreds of locations across a continent over a year. We ask PCA to create the combination of these variables that represents the strongest ‘latent’ factor. The result might be one that corresponds with seasonal difference. We can then say that a lot of what is happening in the fluctuations in rainfall, temperature, and so on can be summed up in the difference between summer and winter. Then PCA can look for a second such composite variable. This might turn out to be one that is correlated with distance from the coast, or height above sea level. Either way we have found a way of summing up what is happening within the local variations in the table in some broad axes of difference.

Each principal component has a weighting for each variable. Variables that are irrelevant to the broad difference being isolated have low weightings; the most important ones have high weightings; and the ones that behave quite differently from each other are given opposite signs. Each sample (or segment) can then be given a score made up of the sum of its counts on each of the constituent variables, each one multiplied by its own weighting.

The lexical words method used for earlier experiments in this chapter simply adds counts of appearances of words together. PCA gives each word-frequency variable a weighting so as to highlight cumulative similarities and dissimilarities. The idea is familiar from stock market indexes or consumer price indexes, in which some elements are weighted more heavily than others to take account of their importance in the basket of stocks or consumer items.

In our case the variables are the frequencies of individual words. We chose the words that we know are used most differently by the two authors, and then asked the method to combine these word-variables so as to show the strongest underlying affinities and contrasts. Figure 2.4 presents the scores on the first two principal components for three groups of samples: 2000-word segments from the 27 confidently attributed Shakespeare plays in our archive, 2000-word segments from the 4 Peele plays we include, and the 5 eligible Titus scenes. The first principal component arrays the Peele segments (the black circles) to the left, and the Shakespeare segments (the grey diamonds) to the right. The Peele segments are also in the lower half of the graph, while the Shakespeare segments spread from top to bottom. The procedure has combined the information from the chosen word-variables and created two indexes. Each of the segments has a score on either index. When plotted together as in Figure 2.4 they provide a map that defines a Shakespeare region (to the right, characterized by high
scores on the first component) and a Peele one (to the lower left, characterized by low scores on both components). As with the lexical tests shown in Figures 2.2 and 2.3 in this chapter, this is a basis for classifying new segments as Shakespeare or Peele, depending on their use of the chosen function words. We are still using 2000-word segments of known authorship to provide the basis for classification, but this time we are contrasting one author with another, rather than one author with a larger mixed set of other authors.

The *Titus* scene usually ascribed to Peele – Act I, Scene i – marked with a grey triangle, is placed to the lower left of the graph. The other four scenes, ascribed to Shakespeare, and marked with solid black shapes, are all above it and to the right. A basis for a broad discrimination between the two authors is established by the two principal components, and on the basis of this one of the *Titus* scenes is placed with Peele and the other four with Shakespeare, a separation that is consistent with the division of the play shown in Table 2.1. One might wonder whether counts of these function words would vary so much within authors, from early works to late, or from one genre to another, or from any other cause, that overall authorial differences would be overwhelmed. Figure 2.4 shows that there is some overlap, some areas where circles are found mixed with diamonds, but generally the distinctions hold up. Moreover, in the case of the

Figure 2.4 Function-words test: 2000-word Shakespeare segments, 2000-word Peele segments, and 5 scenes from *Titus Andronicus*. 
Methods

Table 2.2. *Division of Timon of Athens* between Shakespeare and Middleton according to Vickers, Shakespeare, Co-Author, Table 4.6.

<table>
<thead>
<tr>
<th>Scene</th>
<th>Length in words</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.i</td>
<td>2087</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>I.ii</td>
<td>2003</td>
<td>Middleton</td>
</tr>
<tr>
<td>II.i</td>
<td>295</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>II.ii</td>
<td>1754</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>III.i</td>
<td>497</td>
<td>Middleton</td>
</tr>
<tr>
<td>III.ii</td>
<td>711</td>
<td>Middleton</td>
</tr>
<tr>
<td>III.iii</td>
<td>363</td>
<td>Middleton</td>
</tr>
<tr>
<td>III.iv</td>
<td>835</td>
<td>Middleton</td>
</tr>
<tr>
<td>III.v</td>
<td>914</td>
<td>Middleton</td>
</tr>
<tr>
<td>III.vi</td>
<td>848</td>
<td>Middleton</td>
</tr>
<tr>
<td>IV.i</td>
<td>305</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>IV.ii</td>
<td>422</td>
<td>Shakespeare and Middleton</td>
</tr>
<tr>
<td>IV.iii</td>
<td>4248</td>
<td>Shakespeare and Middleton</td>
</tr>
<tr>
<td>V.i</td>
<td>1860</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.ii</td>
<td>130</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.iii</td>
<td>88</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.iv</td>
<td>678</td>
<td>Shakespeare</td>
</tr>
</tbody>
</table>

*Titus* segments, the method has discriminated between two authors’ work within the same play and displays a pattern entirely consistent with the division between them presented in Vickers’ book. As the two playwrights worked on their joint assignment, writing dialogue for the same characters in the same settings in a shared plot, their contrasting use of *and, thy, it, very* and the rest was nevertheless inscribing two deeply entrenched individual styles in their scenes.

We can deal more quickly with the three other collaborative plays. Vickers’ book details a long tradition of scholarship on *Timon of Athens* that detects a second author besides Shakespeare in the writing of some scenes, and identifies this writer as Thomas Middleton. As with *Titus*, we have a neat hypothesis involving just two authors and many discrete sections. Table 2.2 presents the divisions with their lengths in our texts according to our counting procedures, and their authorship as Vickers records it. Once we have discarded the two scenes with mixed authorship, and the smaller scenes, we are left with I.i, I.ii, II.ii, and V.i. Of these I.i, II.ii, and V.i are attributed to Shakespeare, and I.ii to Middleton.

The *t*-test identifies eighty-seven function words as highly distinctive in use in the Shakespeare and Middleton sets. The highest values of all are for
doth and hath (much more common in Shakespeare than in Middleton), and for that as a demonstrative and now (more frequent in Middleton than in Shakespeare).\textsuperscript{14} Figure 2.5 shows the segments and scenes on the first two principal components that combine the frequencies of all eighty-seven words, using two different sets of weightings. The Shakespeare and Middleton clusters have some overlap. One Shakespeare segment is placed well into Middleton territory, for instance, with a score of 0.59 on the first component, and −0.96 on the second. This is the second segment of Romeo and Juliet, spanning I.ii.6 to I.iv.46. The scene from Timon attributed to Middleton (marked by a grey triangle) is placed in the Middleton cluster, while the other three scenes (the solid black shapes) are well to the Shakespeare side, consistent with the accepted ascription.

The last Shakespeare collaboration we will consider is with John Fletcher. We begin with the history play Henry VIII, first performed in 1613 and first published in the Folio of 1623. In 1850 James Spedding divided the play scene by scene between Shakespeare and Fletcher, and this division has stood up remarkably well to a great variety of tests, as Vickers shows (pp. 336–96). Table 2.3 shows this ascription with the length of each scene in words in the text we use for counting. In Act V we

have adopted the scene division of the *Riverside Shakespeare*, which follows the Folio in this case; some modern editions divide V.ii after line 34. Spedding ascribed the first part of III.ii, up to the exit of the King at III. ii.203, to Shakespeare, and the rest of the scene to Fletcher. In line with our policy of examining only scenes wholly by one collaborator for this test of our methods we have omitted III.ii from our analysis. As before, we identify the function words that are used most differently by the two authors. There are fifty-nine that are rated by the t-test in the top category. Fletcher uses *ye* and *all* much more than Shakespeare. Shakespeare uses the preposition *in* and the verb form *hath* much more than Fletcher. (*Ye* and *hath* have long been known as markers of the difference between Shakespeare’s and Fletcher’s styles.) Figure 2.6 shows the results of a

---

PCA using the full list of fifty-nine words. The known Shakespeare and Fletcher segments are well separated. Just one – which we can identify as the fifth segment of *The Faithful Shepherdess* – crosses the no-man’s-land between the two clusters.\(^\text{18}\) The method assigns the *Henry VIII* scenes tested according to the usual ascription, with III.i and V.ii both with the Fletcher cluster.

As a last exploration of the efficacy of the methods in separating scenes from a collaborative play we examine *The Two Noble Kinsmen*. The play was first printed in 1634 and described then as the joint work of Fletcher and Shakespeare. Table 2.4 lists the scenes of the play, with their length in words and their authorship as summarized by Vickers as before. Most of the scenes are too short for our purpose here. A 1500-word minimum leaves us three: I.i, which Vickers gives to Shakespeare, and II.ii and III.vi, which he gives to Fletcher.

We use the fifty-nine words already identified which satisfy the \(t\)-test standard for difference in use between Shakespeare and Fletcher. We can then include the three *Two Noble Kinsmen* scenes in a PCA with the Shakespeare and Fletcher segments. Figure 2.7 shows the results for the first two principal components, in the manner of Figures 2.4, 2.5 and 2.6.

\(^{18}\) This segment begins within III.i.51 and ends within III.i.296, following the lineation in the version in F. Bowers, ed., *The Dramatic Works in the Beaumont and Fletcher Canon*, 10 vols. (Cambridge: Cambridge University Press, 1976), Vol. 3.
Methods

The scores for Act I, Scene i of *The Two Noble Kinsmen* place this scene well away from Act II, Scene ii and Act III, Scene vi of the same play. The method shows that I.i has a Shakespearean pattern of function-word use, while the other two scenes are placed with Fletcher segments. This corresponds with the attribution shown in Table 2.4. As with the other three plays discussed here, the results are entirely consistent with one well-supported theory about their creation. The method and the theory offer clear-cut mutual support.

The graphs of the present chapter provide a visual summary of the power of the two computational-stylistic methods. Both prove capable of abstracting patterns of word-use that are consistent enough to separate the segments of known provenance, often into almost completely distinct authorial clusters, sometimes into clouds of points that merge at their

---

**Table 2.4. Division of *The Two Noble Kinsmen* between Shakespeare and Fletcher according to Vickers, Shakespeare, Co-Author, Table 6.28.**

<table>
<thead>
<tr>
<th>Scene</th>
<th>Length in words</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.i</td>
<td>1810</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>I.ii</td>
<td>949</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>I.iii</td>
<td>803</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>I.iv</td>
<td>412</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>I.v</td>
<td>108</td>
<td>Fletcher</td>
</tr>
<tr>
<td>II.i</td>
<td>496</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>II.ii</td>
<td>2400</td>
<td>Fletcher</td>
</tr>
<tr>
<td>II.iii</td>
<td>744</td>
<td>Fletcher</td>
</tr>
<tr>
<td>II.iv</td>
<td>287</td>
<td>Fletcher</td>
</tr>
<tr>
<td>II.v</td>
<td>569</td>
<td>Fletcher</td>
</tr>
<tr>
<td>II.vi</td>
<td>343</td>
<td>Fletcher</td>
</tr>
<tr>
<td>III.i</td>
<td>1051</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>III.ii</td>
<td>343</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>III.iii</td>
<td>503</td>
<td>Fletcher</td>
</tr>
<tr>
<td>III.iv</td>
<td>249</td>
<td>Fletcher</td>
</tr>
<tr>
<td>III.v</td>
<td>1212</td>
<td>Fletcher</td>
</tr>
<tr>
<td>III.vi</td>
<td>2706</td>
<td>Fletcher</td>
</tr>
<tr>
<td>IV.i</td>
<td>1351</td>
<td>Fletcher</td>
</tr>
<tr>
<td>IV.ii</td>
<td>1346</td>
<td>Fletcher</td>
</tr>
<tr>
<td>IV.iii</td>
<td>878</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.i</td>
<td>1392</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.ii</td>
<td>1033</td>
<td>Fletcher</td>
</tr>
<tr>
<td>V.iii</td>
<td>1201</td>
<td>Shakespeare</td>
</tr>
<tr>
<td>V.iv</td>
<td>1159</td>
<td>Shakespeare</td>
</tr>
</tbody>
</table>
Hugh Craig and Arthur F. Kinney

fringes. Neither method is absolutely reliable. Part of the reason for this is the degree of overlap in these texts. The passages we are dealing with share conventions, a limited set of genres, and a common heritage. They were all written in an uninterrupted span for audiences in one city over just two generations. Within this common ground they are all striving for some innovation and development. This context of variation within broad limits means that if we take hundreds of samples of dramatic dialogue the length of a medium-sized scene, as we do here, we are almost bound to find authors occasionally writing sufficiently unlike themselves, and sufficiently like each other, to defy the authorial boundaries that do divide the vast bulk of one writer’s work from the other’s. But against this is a steady pressure of differentiation, which gives us a very useful pair of tools for assessing the origins of mysterious or disputed texts. It is worth remembering, too, that the fallibility of the methods individually can be reduced by using both the methods on a single case, as we do in the studies presented in succeeding chapters.

The results shown here are evidence of forces operating within language on a large scale: forces we could certainly have guessed at, but could never have measured without the computer. For example, Shakespeare’s characters, with a remarkable degree of consistency, use the words gentle and beseech much more often than the characters of his contemporaries do. Shakespeare’s characters use yes less often. They are users of hath rather

Figure 2.7 Function-words test: 2000-word Shakespeare segments, 2000-word Fletcher segments, and 3 scenes from The Two Noble Kinsmen.
than *has*. They are fond of the conjunction *that*, but not of *all* or *now*. Putting together these patterns of use, with those of other marker words, allows us to differentiate the plays as a group from plays by others (as the graphs show). It is remarkable that these regularities obtain even in the work of a famously diverse writer like Shakespeare, writing in a mode like drama, made up of multiple contrasting voices.

When we turn from the general patterns to particular cases we need to keep a proper scepticism, however. The right balance of confidence and caution in interpreting the results of methods like these can only come with repeated experiments with samples of known origin. These will help establish how much faith one should have in the findings, especially when they contradict the scholarly consensus, or one’s own judgment as a reader. We aim to keep this in mind in the chapters that follow, and to continue to provide indexes of reliability from tests on samples whose origins we know. We can see from the experiments in this chapter that occasionally the methods get things wrong, or at least fail to give an unambiguously right answer. Most often, though, they get it right, arriving through their quite different means at the same conclusions as the consensus of the scholarly tradition. With this experience behind us we can move from problems where the answer is known to areas of doubt and dispute.
All the Way Through: Testing for Authorship in Different Frequency Strata

John Burrows
University of Newcastle, Australia

Abstract
This article describes the operation of two new tests of authorship and offers some results. Both tests rely on controlled contrasts of word-frequency and both exclude the very common words, which have been put to such good use in recent years. One test treats of words used with some consistency by a target-author but more sporadically by others. The second treats of words used sporadically by the target-author but not by most others. (The inclusion of words that some other authors use avoids the strict constraint that has impoverished this form of evidence.) In suitable cases, both tests prove very accurate. The fact that evidence of authorship can be detected in these three distinct frequency-strata helps to explain why such tests should work at all and so encourages the development of even better ones.

1 Rationale
Computational stylistics deals in patterns derived from the relative frequencies of many words across a chosen range of texts. In recent years, much attention has been given, in this area of scholarly inquiry, to the words that occur most often. These, of course, are words that every writer of English will call upon in almost any form of written expression. We now know that there are strong concomitant variations of frequency among many of these words. The consistency of the variations is usually sufficient to bear the weight of statistical analysis. This makes it possible, among other things, to distinguish the writings of different authors from each other and, accordingly, to test the authorship of doubtful texts.

Such idiosyncratic patterns of occurrence in the top stratum of frequency are scarcely apparent to a reader. A given writer’s habitual choices among comparatively rare words, on the other hand, are easily seen, often attract comment, and lend themselves to parody. In the large area between the extremes of ubiquity and rarity, it is also possible to identify many words (not all of them unusual) that a given writer employs with some consistency while most others do not. Now if, in comparable bodies of work, some writers use certain words that others do not—and the more so if those words recur—it seems fair to suppose that the users are the most likely to employ them in new specimens of their work. It would follow that each writer’s use of such words might yield a distinctive profile.

If the words that make up the second and third of these frequency strata are to yield evidence of authorship, appropriate statistical procedures must be brought to bear. In the third stratum, the inherent fragility of low-frequency statistics is made worse, in our field of inquiry, by a tendency to focus on words not used by anyone except the target-author. The presence of such words in a given
text can undoubtedly be of evidential value. Their absence may easily be adventitious. And they are usually too few to bear much weight statistically. It is nevertheless possible, as I hope to show, to support (not displace) this form of evidence by relaxing one’s stipulations and adding words that few (rather than no) others use. The evidential power of each word is thus reduced but there is more room for a cumulative effect.

In studies of putative authorship, the middle frequency stratum has scarcely been addressed. A body of words, most of them lexical, that we take up and discard according to the needs of a topic or an occasion does not seem likely to offer useful evidence of authorship. Among these words, however, there are many that respond to simple rules of consistency and contrast. Once the words that almost everybody uses are excluded, it is possible to identify words that many writers seldom use but that recur, with some consistency, in a wide-ranging set of work by a given author. Not all of them will reappear in everything that he or she writes; but it seems that more of them can be expected here than in the work of other writers.

Statistical tests directed at each of the three strata are proving extremely accurate in identifying the true author of authentic texts 2,000 words or more in length. The level of accuracy gradually falls away with shorter texts. The errors that do arise, even with longer texts, differ from stratum to stratum. The evidence of the most frequent words can be distorted when an author makes an uncharacteristic choice of genre or literary form, as when Henry Fielding turns his hand to verse. The evidence of the middle range of words can be distorted by a radically different choice of subject, as when Aphra Behn turns her mind from love to death. The evidence of the unusual words is susceptible to aberration as when circumstance carries a particular word into a little flurry of occurrences. Some judicious culling of the word-list alleviates these difficulties. But, while the tests are still being developed, it is better not to interfere with the evidence.

Since there need be no overlapping among the words tested in each stratum, it might appear that the three sorts of test are independent of each other. While that is an important part-truth, it is offset, to an extent we do not much understand, by innumerable links, overt and covert, among the occurrence-rates of different words. An emphasis on the feminine, for example, is to be observed at every level of the vocabulary of most female writers. It is present again, in an altered guise, in the vocabulary of Rochester’s literary circle. Despite this restriction of their apparent independence, a concurrence in the test-results from the different strata clearly strengthens a given outcome.

Why should such tests work as accurately and reliably as they do? Provided they are properly used, the proper answer, I believe, is ‘How could it be otherwise?’ Any writer’s vocabulary is a selection from the full resources of a given language as used at that time. His or her preferences will reflect such differentiae as level of education, gender, chosen audience, topics of customary choice, and so on. If it is an international language like English, there will be signs of a given national variety. The writer’s preferences will also reflect idiosyncrasies too subtle for such broad categories. Such a set of preferences will amount, in short, to one major facet of a Saussurian parole, drawn from the larger resources of the langue itself. As such, like every other meaningful aspect of our behaviour, they will display not only an underlying likeness, greater or lesser, to our various fellows but also our differences from them. Whether that line of thought is an incipient theory or merely an idea of a certain generality, I am too simple an empiricist to judge. I note, however, that de Saussure’s distinction is acceptable to Chomsky, whose distinction between ‘competence’ and ‘performance’ is germane. That, I take it, should put it in good standing among adherents of the high-priori sort of language study. Yet, even if de Saussure were denied that privilege, one might still do worse than follow him and bring a little further evidence to show that (in written as well as spoken language) his distinction is as useful as it is plausible.

Let me put it boldly in the form of a postulate. Evidence of authorship pervades whatever anybody writes. Provided appropriate procedures are employed in the analysis of an appropriate set of texts, it can almost always be elicited. It is inherent,
however, not merely in statistical principle but in human behaviour at large, that such evidence cannot be absolute. The consistencies we observe are trends, not universals. Our many stabilities are offset by our capacity for change.

2 Texts to be Assessed

Some of the procedures to be described here are new and need extensive testing. For the more limited purposes of this introductory sketch, the putative authorship of eight poems of the English Restoration era will be assessed. Three are of unquestioned authorship. These are Edmund Waller’s ‘On the Danger his Majesty (being Prince) Escaped’ and ‘Instructions to a Painter’ and Andrew Marvell’s ‘The First Anniversary of the Government under O. C.’ The other five poems are rejoinders to Waller’s ‘Instructions,’ political satires treating harshly of James, Duke of York (Waller’s hero of the moment) and the conduct of naval affairs. Presumably because an author of such work might well be tried for treason, all five were published anonymously and their authorship is still not definitely known. But the ‘Last Instructions to a Painter’ is accepted as Marvell’s. Current scholarly opinion, as reflected in the most recent edition of Marvell’s poetry (Smith, 2003) favours him as author of the ‘Second’ and the ‘Third Advice to a Painter.’ (See also Patterson, 2000.) The authorship of the ‘Fourth’ and ‘Fifth’ remains open but the idea that they may be Marvell’s has no support.1 The case is set out more fully in a recent article (Burrows, 2005), where the evidence of the very common words is brought to bear. The series of tests undertaken below begins with a specimen of the results offered by the very common words.

These eight poems have particular advantages for assessing the effectiveness of tests directed at the less common words. As to Waller, the experiment is fairly straightforward. Two of the eight poems are his, one is not, and the other five cannot be. All eight treat subjects that he favours and (though some are hostile) in a manner broadly related to his. As to Marvell, the matter is more difficult because these political poems stand well apart from most of his acknowledged work. Yet the question is worth asking. Can the tests we shall be examining offer results that accord with scholarly opinion? That would entail an outcome in which the poem known to be Marvell’s and the three considered his were distinguished from the two by Waller and the two by unidentified poets. When the proposed contest between Waller and Marvell is resolved, we shall consider whether the authorship of these poems can also be established in a contest where many other poets of the period are introduced.

For some of the necessary comparisons, I have had recourse to a large and diverse database of half a million words of Restoration poetry.2 A further forty-one independent texts have also been introduced. Twenty-one are poems (or long excerpts from poems) by authors included in the main database. The remaining twenty are by Restoration poets who are not members of the main set.3 All the texts have been modernized to overcome the vagaries of seventeenth-century spelling. Contractions like ‘I’ll’ and ‘don’t’ have been expanded. A few words are tagged so as to distinguish among their main grammatical functions.

3 The Evidence of very Common Words: The Delta Test

In the multivariate statistical procedures used to elicit intelligible patterns in the frequency-distribution of the very common words, the texts are specimens and the words are the variables under scrutiny. To allow for differences in length between text and text, the raw frequencies for each word-type are standardized, usually as percentages of all the word-tokens in each text.4 The standardized frequencies are arranged in descending order as a frequency-profile for each text in turn.

The Delta procedure compares the upper range of the frequency-profile of a given text with those of many authors and shows which of them is least unlike it. The operation of the procedure is described elsewhere and a large body of results is shown. (Burrows, 2002, 2003; Hoover, 2004)5. Especially with texts of more than 2,000 words, the results point towards the more likely candidates and allow the elimination of the more unlikely. The errors that occur from time to time stem from
unusually strong changes in an author’s style, as between the early and the late work of Cowley; as between the satires and the lyrics of Robert Gould; or as when a gifted translator like Dryden submerges many of his usual stylistic propensities as he tries to catch the spirit of his foreign model.

The outcome of each Delta test stands free of all the others in the sense that each specimen text, in turn, is matched against the same collection of authorial sets. If those sets are changed, the various specimens are all affected; but, in contrast to other procedures like cluster analysis and principal component analysis, the specimens tested do not affect each other.

A ‘delta-score’, like those shown under that heading in Table 1, can be defined as ‘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’. When they are calculated, as here, for a sufficient number of authors, they can be ranked and ‘delta z-scores’ can be derived to allow more meaningful comparisons among the results for different texts.

Table 1 shows that Waller and Marvell, respectively, rank first of the twenty-five authors for each of the three unquestioned poems. Marvell also ranks first for the ‘Last Instructions’ and for the ‘Second’ and the ‘Third Advice’. Almost all the scores are strong enough to bear real weight and at no point is there a close contest between Waller and Marvell. Marvell, so it proves, ranks well down the list of twenty-five for the two ‘Advices’ thought not to be his work. Robert Gould, much too young to be a contender, ranks first for the ‘Fourth Advice’, followed by Denham, whose claim can be taken seriously. As for the ‘Fifth Advice’, a close contest among several unlikely candidates is a typical result in cases where the true author lies outside our group of twenty-five. The weak delta z-scores at the head of the list also support that possibility. Table 1, in short, is a fair specimen of the evidence in the article mentioned (Burrows, 2005) and the results are entirely in keeping with current literary scholarship. Taking all this as our basis, we can go on to consider two new approaches to such matters.

4 The Evidence of Less Common Words: The Zeta and Iota Tests

4.1 One on one: Waller versus Marvell; Marvell versus Waller

The Delta procedure focuses on a text and seeks to identify the correct one of many possible authors. The new ‘Zeta’ and ‘Iota’ tests, on which I have been working in recent months, focus on a single author and seek to identify which of many texts are most likely to be his or hers. The point of departure for this series of tests is the complete word-list for a large sample of a particular author’s work. The 13,838 word-tokens of the Waller set used here embrace 2,876 distinct word-types. (The 20,151 of Marvell embrace 4,323.) Corresponding lists are then established and tabulated, showing the incidence of these same word-types in each of the other members of a large group of authorial sets and also in such independent texts as are to be assessed. In my experiments to date, around 10,000 word-tokens seem to suffice as a reliable minimum for an authorial set, 500 (but preferably many more) for an independent text.

The tests rest upon stipulated contrasts between a base-set (the main sample of the current target-author) and a counter-set (comprising one or more of the remaining authorial samples). Table 2, a tiny Microsoft Excel worksheet, offers a concise but limited model of the procedure. The hierarchical array of word-types shown in Columns F and G runs down as usual from ‘the’ and ‘and’. Their incidence in Waller, the base-set for this first series, is shown in Column I. In Columns A to E, the Waller set is broken into five segments of almost equal size, with the remainder added to segment five, increasing it from 2,767 to 2,770. The figures in Column H indicate how many of the five segments contain each word and run down, accordingly, from 5 to 1. This makes a simple measure of Waller’s consistency in the use of each word-type in turn. (Even in the present list of extremely common words, ‘your’ appears in only three of the five segments. In the full list of 2,876 word-types, of course, 1 is much the most common count.) Column J and K treat of the chosen counter-set, the main set of Marvell. Column J shows which of the
<table>
<thead>
<tr>
<th>Author</th>
<th>Delta</th>
<th>Delta z</th>
<th>Delta</th>
<th>Delta z</th>
<th>Delta</th>
<th>Delta z</th>
<th>Delta</th>
<th>Delta z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waller, Of the Danger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1322 Words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waller 1st: Marvell 11th</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waller</td>
<td>1.549</td>
<td>-1.684</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>1.579</td>
<td>-1.492</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Milton</td>
<td>1.584</td>
<td>-1.459</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dryden</td>
<td>1.637</td>
<td>-1.114</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Durfey</td>
<td>1.667</td>
<td>-0.916</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congreve</td>
<td>1.704</td>
<td>-0.676</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cowley</td>
<td>1.708</td>
<td>-0.654</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tate</td>
<td>1.712</td>
<td>-0.628</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Settle</td>
<td>1.727</td>
<td>-0.528</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Denham</td>
<td>1.730</td>
<td>-0.512</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marvell</td>
<td>1.762</td>
<td>-0.298</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swift</td>
<td>1.766</td>
<td>-0.272</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Butler</td>
<td>1.793</td>
<td>-0.101</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shadwell</td>
<td>1.804</td>
<td>-0.027</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oldham</td>
<td>1.858</td>
<td>0.323</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gould</td>
<td>1.873</td>
<td>0.421</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dorset</td>
<td>1.880</td>
<td>0.468</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cotton</td>
<td>1.904</td>
<td>0.624</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rochester</td>
<td>1.906</td>
<td>0.637</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sedley</td>
<td>1.945</td>
<td>0.888</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behn</td>
<td>1.959</td>
<td>0.982</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radcliffe</td>
<td>1.967</td>
<td>1.035</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phillips</td>
<td>2.025</td>
<td>1.408</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radcliffe</td>
<td>2.030</td>
<td>1.445</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brome</td>
<td>2.136</td>
<td>2.132</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Second Advice</th>
<th>Third Advice</th>
<th>Fourth Advice</th>
<th>Fifth Advice</th>
</tr>
</thead>
<tbody>
<tr>
<td>2867 Words</td>
<td>3638 Words</td>
<td>1103 Words</td>
<td>1209 Words</td>
</tr>
<tr>
<td>Marvell 1st: Waller 20th</td>
<td>Marvell 1st: Waller 21st</td>
<td>Waller 3rd: Marvell 14th</td>
<td>Marvell 9th: Waller 23rd</td>
</tr>
<tr>
<td>Delta</td>
<td>Delta z</td>
<td>Delta</td>
<td>Delta z</td>
</tr>
<tr>
<td>1</td>
<td>Marvell</td>
<td>1.253</td>
<td>1.924</td>
</tr>
<tr>
<td>2</td>
<td>Durfey</td>
<td>1.304</td>
<td>1.488</td>
</tr>
<tr>
<td>3</td>
<td>Prior</td>
<td>1.332</td>
<td>1.248</td>
</tr>
<tr>
<td>4</td>
<td>Cowley</td>
<td>1.360</td>
<td>1.013</td>
</tr>
<tr>
<td>5</td>
<td>Congreve</td>
<td>1.370</td>
<td>0.922</td>
</tr>
<tr>
<td>6</td>
<td>Butler</td>
<td>1.405</td>
<td>0.625</td>
</tr>
<tr>
<td>7</td>
<td>Oldham</td>
<td>1.411</td>
<td>0.575</td>
</tr>
<tr>
<td>8</td>
<td>Settle</td>
<td>1.425</td>
<td>0.457</td>
</tr>
<tr>
<td>9</td>
<td>Dryden</td>
<td>1.425</td>
<td>0.456</td>
</tr>
<tr>
<td>10</td>
<td>Denham</td>
<td>1.426</td>
<td>0.452</td>
</tr>
<tr>
<td>11</td>
<td>Tate</td>
<td>1.439</td>
<td>0.338</td>
</tr>
<tr>
<td>12</td>
<td>Swift</td>
<td>1.453</td>
<td>0.221</td>
</tr>
<tr>
<td>13</td>
<td>Gould</td>
<td>1.458</td>
<td>0.173</td>
</tr>
<tr>
<td>14</td>
<td>Milton</td>
<td>1.465</td>
<td>0.121</td>
</tr>
<tr>
<td>15</td>
<td>Shadwell</td>
<td>1.466</td>
<td>0.111</td>
</tr>
<tr>
<td>16</td>
<td>Dorset</td>
<td>1.490</td>
<td>0.092</td>
</tr>
<tr>
<td>17</td>
<td>Sedley</td>
<td>1.496</td>
<td>0.144</td>
</tr>
<tr>
<td>18</td>
<td>Cotton</td>
<td>1.567</td>
<td>0.755</td>
</tr>
<tr>
<td>19</td>
<td>Brome</td>
<td>1.576</td>
<td>0.85</td>
</tr>
<tr>
<td>20</td>
<td>Waller</td>
<td>1.605</td>
<td>1.073</td>
</tr>
<tr>
<td>21</td>
<td>Behn</td>
<td>1.627</td>
<td>1.261</td>
</tr>
<tr>
<td>22</td>
<td>Phillips</td>
<td>1.627</td>
<td>1.263</td>
</tr>
<tr>
<td>23</td>
<td>Rochester</td>
<td>1.637</td>
<td>1.346</td>
</tr>
<tr>
<td>24</td>
<td>Radcliffe</td>
<td>1.671</td>
<td>1.641</td>
</tr>
<tr>
<td>25</td>
<td>Wharton</td>
<td>1.681</td>
<td>1.719</td>
</tr>
</tbody>
</table>
Table 2 Waller and Marvell. Outline of the new procedure

<table>
<thead>
<tr>
<th>Segments of Base-set</th>
<th>Base-set</th>
<th>Counterset</th>
<th>Trial-pieces</th>
<th>Test-pieces</th>
</tr>
</thead>
<tbody>
<tr>
<td>A B C D E F G</td>
<td>H I</td>
<td>J K</td>
<td>L M</td>
<td>N O</td>
</tr>
<tr>
<td>Waller1 Waller2 Waller3 Waller4 Waller5 WORDS</td>
<td>Waller Waller Waller Waller Waller Waller Waller</td>
<td>Marvell Marvell Marvell Danger FirstAnn Painter LastInst</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2767 2767 2767 2767 2770</td>
<td>13838</td>
<td>20151</td>
<td>1322</td>
<td>2065</td>
</tr>
<tr>
<td>COUNT</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>7.23 7.23 7.23 6.87 6.86</td>
<td>1.45</td>
<td>0.99</td>
<td>14.37</td>
<td>6.39</td>
</tr>
<tr>
<td>801 743 740 792 740</td>
<td>3816</td>
<td>4799</td>
<td>361</td>
<td>826</td>
</tr>
<tr>
<td>2767 2767 2767 2767 2770</td>
<td>13838</td>
<td>20151</td>
<td>1322</td>
<td>3131</td>
</tr>
<tr>
<td>2767 2767 2767 2767 2770</td>
<td>13838</td>
<td>20151</td>
<td>1322</td>
<td>3131</td>
</tr>
<tr>
<td>289.48 268.52 267.44 286.23 267.15</td>
<td>275.76</td>
<td>238.15</td>
<td>273.07</td>
<td>263.81</td>
</tr>
<tr>
<td>153 169 147 132 123</td>
<td>724</td>
<td>1024</td>
<td>96</td>
<td>199</td>
</tr>
<tr>
<td>100 98 91 96 74</td>
<td>459</td>
<td>704</td>
<td>43</td>
<td>125</td>
</tr>
<tr>
<td>56 57 44 52 52</td>
<td>261</td>
<td>295</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>56 57 44 52 52</td>
<td>261</td>
<td>295</td>
<td>47</td>
<td>55</td>
</tr>
<tr>
<td>50 25 42 49 38</td>
<td>204</td>
<td>263</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>23 44 31 69 35</td>
<td>202</td>
<td>194</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>35 36 33 54 37</td>
<td>195</td>
<td>269</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>15 36 33 54 37</td>
<td>195</td>
<td>269</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>34 29 45 29 40</td>
<td>177</td>
<td>293</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>48 17 32 32 37</td>
<td>166</td>
<td>70</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>48 17 32 32 37</td>
<td>166</td>
<td>70</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>50 25 42 49 38</td>
<td>204</td>
<td>263</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>23 44 31 69 35</td>
<td>202</td>
<td>194</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>35 36 33 54 37</td>
<td>195</td>
<td>269</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>34 29 45 29 40</td>
<td>177</td>
<td>293</td>
<td>15</td>
<td>32</td>
</tr>
<tr>
<td>48 17 32 32 37</td>
<td>166</td>
<td>70</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>50 25 42 49 38</td>
<td>204</td>
<td>263</td>
<td>12</td>
<td>23</td>
</tr>
<tr>
<td>23 44 31 69 35</td>
<td>202</td>
<td>194</td>
<td>31</td>
<td>30</td>
</tr>
<tr>
<td>35 36 33 54 37</td>
<td>195</td>
<td>269</td>
<td>13</td>
<td>58</td>
</tr>
<tr>
<td>34 29 45 29 40</td>
<td>177</td>
<td>293</td>
<td>15</td>
<td>32</td>
</tr>
</tbody>
</table>

NB: (inf) = infinitive particle; (p) = preposition; (rp) = relative pronoun.
The only operative criterion reflected in the descending order of Waller’s scores in Column I. The fact that this is Waller’s word-list, not Marvell’s, is evident in the many perturbations of rank-order in Column K. But the comparatively weak distinctions shown in Row 10 need to be strengthened and controlled.

If all the data in the full work-sheet for Waller, from Row 12 down to Row 2,887, are sorted on the basis of the counts in Column H, it is possible to discard all those word-types that do not meet a stipulated level of consistency across the five segments of the main Waller-set as represented in Columns A to E. If the remaining data are then sorted on the basis of Column K, it is possible to discard all those word-types that exceed a specified level of frequency in Marvell. If the stipulations are too weak, no consistent authorial difference will emerge. But if they are too strict, the surviving word-list can be too impoverished to yield a reliable outcome.

Table 3 shows the outcome of tests in which Waller and Marvell are opposed. In the top half of the table, Waller provides the base-set, Marvell the counter-set. In the bottom half, their roles are reversed. In each half of the table, we open the series with a simple overview. Beyond the fact that these are the words of one or other of the two opposed authorial sets, no further stipulation is imposed. In both overviews, the occurrence-rate for the chosen base-set is, by definition, 1,000 per 1,000. The occurrence-rate for each of the counter-sets and for such other specimens as may be introduced will never reach that level. In practice, an occurrence-rate between 750 and 850 per 1,000 is usual.

Consider the present case, beginning with Overview A. Among the eight specimen-texts, the two by Waller show higher occurrence-rates than the four associated with Marvell. But the ‘Fourth Advice’ breaks the pattern, with an occurrence-rate higher than that for Waller’s ‘Of the Danger’. In Overview B, Marvell’s ‘First Anniversary’ shows the highest score of the eight but the pattern of the other seven scores is not authorial. These overviews, then, do not support the idea that a given writer’s overall word-list might serve as it stands as an accurate authorial discriminator.

The main difficulty here is that any truly idiosyncratic features of the two authorial frequency-lists are buried among the high frequencies of words...
### Table 3: Waller and Marvell. Two tests on eight poems

#### Waller versus Marvell

<table>
<thead>
<tr>
<th>Length</th>
<th>Overview A</th>
<th>Test 2A: Zeta test</th>
<th>Test 3A: Iota test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Counter-set: Ditto</td>
<td>Counter-set: Frequency &lt; 3</td>
<td>Counter-set: Frequency Zero</td>
</tr>
<tr>
<td>Types</td>
<td>Tokens</td>
<td>per 1000</td>
<td>Types</td>
</tr>
<tr>
<td>Waller</td>
<td>Base-set</td>
<td>13838</td>
<td>2876</td>
</tr>
<tr>
<td>Marvell</td>
<td>Counter-set</td>
<td>20151</td>
<td>1644</td>
</tr>
</tbody>
</table>

#### Marvell versus Waller

<table>
<thead>
<tr>
<th>Length</th>
<th>Overview B</th>
<th>Test 2B: Zeta test</th>
<th>Test 3B: Iota test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Counter-set: Ditto</td>
<td>Counter-set: Frequency &lt; 3</td>
<td>Counter-set: Frequency Zero</td>
</tr>
<tr>
<td>Types</td>
<td>Tokens</td>
<td>per 1000</td>
<td>Types</td>
</tr>
<tr>
<td>Marvell</td>
<td>Base-set</td>
<td>20151</td>
<td>4323</td>
</tr>
<tr>
<td>Waller</td>
<td>Counter-set</td>
<td>13838</td>
<td>1644</td>
</tr>
</tbody>
</table>

**Authorship in Different Frequency Strata**

---

that everybody uses. As the little array of raw
data in Table 2 reveals, the twenty most common
word-types of Waller’s set amount to about a
quarter of all the word-tokens in each of the
specimens. The steepness and the uniformity of the
descending hierarchy were observed by George K.
Zipf (1949) and enshrined in ‘Zipf’s Law’. But as a
recent article (Hoover, 2004) has shown, the Delta
procedure continues to be effective for at least the
top 600 word-types of a well-founded word-list.
A merely additive procedure, however, like that
used for Overviews A and B, is too crude for the task
in hand.

A second difficulty, I believe, affects the use of
any measure of divergence for the analysis of
complete frequency hierarchies. It stems not from
the tyranny of the large numbers at the head of the
list but from subtler arithmetical differences among
the data at different levels of a given frequency-table.
In the topmost stratum, robust statistical compar-
isons can be derived from differently sized diver-
gences from a mean or median score. For each
common word-type in a range of specimens, a bell-
shaped curve provides a sound footing for a range of
useful operations. These curves are always skewed to
the right or positive side because such data yield
zeroes and very low scores for particular word-types
more often than very high scores. Sentences of some
length where ‘the’ does not occur are more common
than those where it occurs in abundance. But the
upward range is the more extensive. It is not hard,
for instance, to devise meaningful sentences where
‘the’ makes up a quarter or more of all the word
tokens: ‘The bald Anglo the blonde had eyed the
night before lay on the dirty mat. The haft of the
Greek’s dagger stood in the hairy chest. The blind
flapped and the scattered papers stirred in the breeze
from the broken pane. The blonde officer sighed at
the enormity of the task ahead. The upside? Yeah,
right—the Anglo sure was off the hot-list.’ Although
such levels cannot sensibly be sustained, many a
dreary, over-circumstantial novel suffers in the
attempt.

In the next stratum from the top, zeroes and
singletons are common but are not infrequently
matched against scores that run up from a handful
into double figures. In the lowest and most
extensive stratum, zeroes by far predominate but
are usually matched against scattered singletons
in simple binary contrasts. Means and standard
deviations can certainly be extracted from the
frequency-patterns of these lower strata. To put
them to the same uses as those of the top stratum
seems inappropriate in principle and can yield
disconcerting results. Let us return to Table 3 and
consider some examples of tests where the two
lower strata are separated from the top one and also
from each other. I have been labelling them as ‘Zeta’
and ‘Iota’ tests as a matter of convenience and as a
way of emphasizing that Omega, the last word on
these matters, is not yet within reach.

In the Zeta tests 2A and 2B (‘Delta’ being Test 1),
two stipulations are imposed. After breaking each
base-set, in turn, into five equal segments, we retain
only those word-types that occur in at least three of
the five. Of these, we discard those that occur fewer
than three times in the counter-set. In Test 2A,
Waller’s list of word-types is reduced from 2,876 to
185. In Test 2B, Marvell’s list of 4,323 is reduced
to 354. As stipulated, the contrast between each
base-set and the corresponding counter-set is very
marked. The real point of interest, however, lies in
the behaviour of the eight independent specimens.
In Test 2A, the occurrence-rates for the two Waller
entries, at 46.90 and 46.83 per 1,000, comfortably
exceed all their rivals. In Test 2B, the occurrence-
rates for the four poems associated with Marvell
exceed the other four.

The Iota tests 3A and 3B focus on the lowest
stratum of the word-lists. The first stipulation is that
a word-type should not appear in more than two
of the five segments of the appropriate base-set.
The second is that it should not occur at all in
the corresponding counter-set. The Waller-list is
thus reduced to 1,169 uncommon word-types, the
Marvell to 2,521. In Test 3A, the scores for the two Waller
entries exceed the rest. In Test 3B, three of the texts associated with Marvell exceed
the rest. But the authorial pattern is broken by the
‘Fifth Advice’, which scores a trifle higher than the
‘Second Advice’. The range of the eight scores
indicates that this breach is due to a higher than
expected score for the ‘Fifth Advice’. The ‘Second
Advice’ does not fall far short of its true partners.
The score for the ‘Fifth Advice’ makes a useful point. It reflects the fact that the tests in Table 3 are specifically designed to distinguish Waller and Marvell from each other and not from anybody else. A text by any other poet is not governed by such differentiae and remains a wild-card, free to include words from the base-set and not necessarily much affected by the exclusion of words from the counter-set. The stipulated contrasts may operate successfully on such texts but cannot be expected to do so. It is as if, after establishing a set of differentiae to distinguish chalk from cheese, one tried them on a specimen of bone. The likeness to chalk in hardness, density, and colour is adventitious. This is not chalk at all. A more broadly based set of differentiae is required for the more general task.

4.2 One against many: Waller and twenty-four others

Simplicity of exposition has its place in a demonstration-piece. In practice, however, head to head contests between two putative authors are best reserved for situations where no other candidates need be considered. That situation can arise from outside knowledge, as when only John Dryden and the Earl of Mulgrave need be considered as possible authors of the celebrated ‘Essay on Satire’. It can arise, by experiment, from the use of exploratory tests like principal component analysis and the Delta procedure. But, as we have just seen, a premature head to head contest can easily go awry.

Once a putative author has been identified, by whatever means, it is possible to verify his or her claim by using variant forms of the Zeta and Iota tests. This can be done by matching a chosen authorial base-set, as before, against a multi-author counter-set. A table corresponding to Table 2 is established. Columns A to I are retained unaltered. But Column K is now occupied by the sum of all the scores, for each successive word-type, of as many other authorial sets as may be desired. Each of those sets occupies its own column and contributes to the sum. And Column J is now occupied by counts of the number of those authors who use each of the word-types. Any stipulated contrast between base-set and counter-set can now be applied to whatever specimens may seem appropriate.

In the present case, Waller supplies the initial base-set. The counter-set is composed of the twenty-four other authorial sets, Marvell among them, that make up our main database. The specimens for examination are the same eight poems as before. They are tested against the twenty-four separate authorial sets; twenty-one independent texts by those same poets; and twenty texts by other Restoration poets.

In Table 4, the Zeta test is based, as before, on two stipulations. The first excludes all those word-types that occur in fewer than three of Waller’s five segments. The second excludes those word-types that occur in more than twenty-two of the twenty-four other authorial sets. The second stipulation serves not only to help establish a firm contrast between Waller and the rest but also to exclude those very common words that are used in other statistical procedures. The effect of the two stipulations is to reduce Waller’s 2,876 word-types to 259. These yield him 1,319 word-tokens at a rate of 95.32 per 1,000. The half-million words of the counter-set yield 16,517 word-tokens at a rate of only 31.72 per 1,000.

In the top central panel of Table 4, it can be seen that, at 72.94 and 59.00, the scores for the two Waller poems easily exceed those for the other six. The scores for Group A, in the panel below it, show that the twenty-four authorial components of the counter-set all fall well short of the two Waller poems but that some of them exceed the least Waller-like of the other six specimens. The mean of these twenty-four sets is 31.24 per 1,000 and the highest of them is only 41.84. The low standard deviation of 6.06 suggests that the stipulations employed have fallen with rather an even hand across this group and also helps to emphasize the sharp divergence of the Waller pieces from the rest.

But the true force of the test is felt in Groups B and C, where the scores for the independent specimens are arrayed. These two lowest sections of the central panel show that the scores for the two Waller poems are still not matched by any of these forty-one texts. Neither those by members of the set nor those by outsiders produce a serious rival. Two of the pieces by outsiders score over 50 per 1,000. They are Blackmore’s King Arthur and Flatman’s...
Table 4 Waller and 24 others. Two tests on eight poems: Waller’s world-list

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Zeta test</th>
<th>Iota test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Types</td>
<td>Tokens</td>
<td>per 1000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waller</td>
<td>Base-set</td>
<td>13838</td>
<td>259</td>
</tr>
<tr>
<td>24 poets</td>
<td>Counter-set</td>
<td>520762</td>
<td>4350</td>
</tr>
<tr>
<td>Test-pieces</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Waller</td>
<td>Of the Danger</td>
<td>1322</td>
<td>59</td>
</tr>
<tr>
<td>Marvell</td>
<td>First Anniv</td>
<td>3131</td>
<td>81</td>
</tr>
<tr>
<td>Marvell?</td>
<td>Last Inst.</td>
<td>7693</td>
<td>126</td>
</tr>
<tr>
<td>Marvell??</td>
<td>2nd Advice</td>
<td>2867</td>
<td>60</td>
</tr>
<tr>
<td>Marvell??</td>
<td>3rd Advice</td>
<td>3638</td>
<td>79</td>
</tr>
<tr>
<td>??</td>
<td>4th Advice</td>
<td>1103</td>
<td>43</td>
</tr>
<tr>
<td>??</td>
<td>5th Advice</td>
<td>1209</td>
<td>30</td>
</tr>
</tbody>
</table>

Group A. Components of counter-set

<table>
<thead>
<tr>
<th>Poet</th>
<th>MAX</th>
<th>MEAN</th>
<th>ST DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behn</td>
<td>41.84</td>
<td>31.24</td>
<td>6.06</td>
</tr>
<tr>
<td>Brome</td>
<td>16.66</td>
<td>15.14</td>
<td>2.85</td>
</tr>
<tr>
<td>Butler</td>
<td>30.96</td>
<td>23.49</td>
<td>41.06</td>
</tr>
<tr>
<td>Cotton</td>
<td>25.43</td>
<td>19.67</td>
<td>41.06</td>
</tr>
<tr>
<td>Cowley</td>
<td>19.72</td>
<td>20.67</td>
<td>2.85</td>
</tr>
<tr>
<td>Denham</td>
<td>21.70</td>
<td>21.70</td>
<td>2.85</td>
</tr>
<tr>
<td>Dorset</td>
<td>25.43</td>
<td>21.70</td>
<td>41.06</td>
</tr>
<tr>
<td>Dryden</td>
<td>30.96</td>
<td>30.96</td>
<td>41.06</td>
</tr>
<tr>
<td>Durfee</td>
<td>34.06</td>
<td>34.06</td>
<td>41.06</td>
</tr>
<tr>
<td>Gould</td>
<td>35.58</td>
<td>35.58</td>
<td>41.06</td>
</tr>
<tr>
<td>MarvellB</td>
<td>36.99</td>
<td>36.99</td>
<td>41.06</td>
</tr>
<tr>
<td>Milton</td>
<td>31.24</td>
<td>31.24</td>
<td>6.06</td>
</tr>
<tr>
<td>Oldham</td>
<td>32.46</td>
<td>32.46</td>
<td>6.06</td>
</tr>
<tr>
<td>Phillips</td>
<td>36.99</td>
<td>36.99</td>
<td>6.06</td>
</tr>
<tr>
<td>Prior</td>
<td>32.00</td>
<td>32.00</td>
<td>6.06</td>
</tr>
<tr>
<td>Radcliffe</td>
<td>35.58</td>
<td>35.58</td>
<td>6.06</td>
</tr>
<tr>
<td>Rochester</td>
<td>34.06</td>
<td>34.06</td>
<td>6.06</td>
</tr>
<tr>
<td>Sedley</td>
<td>30.96</td>
<td>30.96</td>
<td>6.06</td>
</tr>
<tr>
<td>Settle</td>
<td>23.49</td>
<td>23.49</td>
<td>6.06</td>
</tr>
<tr>
<td>Shadwell</td>
<td>35.58</td>
<td>35.58</td>
<td>6.06</td>
</tr>
<tr>
<td>Swift</td>
<td>36.99</td>
<td>36.99</td>
<td>6.06</td>
</tr>
<tr>
<td>Tate</td>
<td>31.24</td>
<td>31.24</td>
<td>6.06</td>
</tr>
<tr>
<td>Wharton</td>
<td>32.46</td>
<td>32.46</td>
<td>6.06</td>
</tr>
</tbody>
</table>

Group B. Independent texts by these poets

<table>
<thead>
<tr>
<th>Poet</th>
<th>MAX</th>
<th>MEAN</th>
<th>ST DEV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behn16</td>
<td>36.99</td>
<td>27.92</td>
<td>6.67</td>
</tr>
<tr>
<td>Brome16</td>
<td>28.25</td>
<td>18.31</td>
<td>5.41</td>
</tr>
<tr>
<td>Behn16</td>
<td>36.99</td>
<td>27.92</td>
<td>6.67</td>
</tr>
<tr>
<td>Brome16</td>
<td>28.25</td>
<td>18.31</td>
<td>5.41</td>
</tr>
</tbody>
</table>

(Continued)
Table 4 Continued

<table>
<thead>
<tr>
<th>Length</th>
<th>Zeta test</th>
<th>Iota test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base-set: &gt;2ex5 segs.</td>
<td>Base-set: &lt;3ex5 segs.</td>
</tr>
<tr>
<td></td>
<td>Counter-set: &lt;23 poets</td>
<td>Counter-set: &lt;11 poets</td>
</tr>
<tr>
<td></td>
<td>Types</td>
<td>Tokens</td>
</tr>
<tr>
<td>Butler01</td>
<td>Plagiaries</td>
<td>1217</td>
</tr>
<tr>
<td>Butler02</td>
<td>Weakness</td>
<td>1401</td>
</tr>
<tr>
<td>Congreve14</td>
<td>Imposs. Thing</td>
<td>1376</td>
</tr>
<tr>
<td>Cotton09</td>
<td>Ireland2</td>
<td>1830</td>
</tr>
<tr>
<td>Cotton11</td>
<td>Epistle</td>
<td>1537</td>
</tr>
<tr>
<td>Cowley17</td>
<td>Royal Society</td>
<td>1315</td>
</tr>
<tr>
<td>Cowley18</td>
<td>Davidis2</td>
<td>6812</td>
</tr>
<tr>
<td>Dryden03</td>
<td>Absalom</td>
<td>7824</td>
</tr>
<tr>
<td>Dryden02</td>
<td>Epistle15</td>
<td>1650</td>
</tr>
<tr>
<td>Durfey</td>
<td>Malecontent</td>
<td>7817</td>
</tr>
<tr>
<td>Gould13</td>
<td>BeauxEsprits</td>
<td>6020</td>
</tr>
<tr>
<td>Gould16</td>
<td>Fanaticism</td>
<td>1995</td>
</tr>
<tr>
<td>Milton02</td>
<td>PR</td>
<td>15694</td>
</tr>
<tr>
<td>Milton03</td>
<td>SA</td>
<td>12885</td>
</tr>
<tr>
<td>Oldham07</td>
<td>Ona Woman</td>
<td>1298</td>
</tr>
<tr>
<td>Oldham13</td>
<td>Ben Johnson</td>
<td>2237</td>
</tr>
<tr>
<td>Phillips</td>
<td>Lady E.C.</td>
<td>1210</td>
</tr>
<tr>
<td>Prior18</td>
<td>Henry&amp;Emma</td>
<td>6033</td>
</tr>
<tr>
<td>Swift18</td>
<td>Verses</td>
<td>3206</td>
</tr>
</tbody>
</table>

Group C. Independent texts by other poets

<table>
<thead>
<tr>
<th></th>
<th>Types</th>
<th>Tokens</th>
<th>per1000</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker</td>
<td>1077</td>
<td>24</td>
<td>26</td>
<td>24.14</td>
</tr>
<tr>
<td>Blackmore</td>
<td>K.Arthur</td>
<td>6986</td>
<td>127</td>
<td>352</td>
</tr>
<tr>
<td>Caryll</td>
<td>Naboth</td>
<td>3892</td>
<td>69</td>
<td>119</td>
</tr>
<tr>
<td>Chamb’l’ne</td>
<td>Pharonn</td>
<td>3617</td>
<td>74</td>
<td>114</td>
</tr>
<tr>
<td>Chudleigh</td>
<td>Gloucester</td>
<td>2743</td>
<td>74</td>
<td>109</td>
</tr>
<tr>
<td>Davenant</td>
<td>Gondibert</td>
<td>5167</td>
<td>97</td>
<td>198</td>
</tr>
<tr>
<td>Duke</td>
<td>Paris</td>
<td>3486</td>
<td>73</td>
<td>121</td>
</tr>
<tr>
<td>Fane</td>
<td>Pindar</td>
<td>1909</td>
<td>62</td>
<td>93</td>
</tr>
<tr>
<td>Flatman</td>
<td>Rupert</td>
<td>1065</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>Garth</td>
<td>Dispens5</td>
<td>2704</td>
<td>59</td>
<td>88</td>
</tr>
<tr>
<td>Heyrick</td>
<td>Atlantis</td>
<td>8797</td>
<td>150</td>
<td>364</td>
</tr>
<tr>
<td>Hutchinson</td>
<td>Order</td>
<td>6792</td>
<td>108</td>
<td>227</td>
</tr>
<tr>
<td>Mulgrave</td>
<td>Poetry</td>
<td>2765</td>
<td>71</td>
<td>103</td>
</tr>
<tr>
<td>Norris</td>
<td>Passion</td>
<td>1319</td>
<td>36</td>
<td>46</td>
</tr>
<tr>
<td>Pordage</td>
<td>Med. Rev.</td>
<td>3103</td>
<td>52</td>
<td>103</td>
</tr>
<tr>
<td>Sprat</td>
<td>Oliver</td>
<td>2522</td>
<td>73</td>
<td>114</td>
</tr>
<tr>
<td>Thompson</td>
<td>Mid. Moon</td>
<td>3250</td>
<td>66</td>
<td>87</td>
</tr>
<tr>
<td>Tutchin</td>
<td>Honesty</td>
<td>2287</td>
<td>46</td>
<td>64</td>
</tr>
<tr>
<td>Wase</td>
<td>Divination</td>
<td>2156</td>
<td>70</td>
<td>92</td>
</tr>
<tr>
<td>Wild</td>
<td>Iter Boreale</td>
<td>3321</td>
<td>62</td>
<td>87</td>
</tr>
</tbody>
</table>

Authorship in Different Frequency Strata

pindaric ode ‘On the Death of the Illustrious Prince Rupert’.

Both poems share in the nationalistic rodomon-tade in which Waller is among the more extravagant of his generation. He is a strong exponent of attitudes that (springing from a narrower tribalism) were gaining ground in Europe, were later to cross the Atlantic, and (chiefly to the detriment of humanity at large) were to influence European and world history in the ensuing 300 years.

Sing we the Glory of triumphant Arms.
So shall all Tyrants yield.
May the like Fortune meet all those
Who vainly dare oppose
Our Monarch’s sacred Law,
Our Nation’s noble Rage.

This fragment is from a pindaric ode ‘On our late Victories by Land and Sea’. In admitting authorship (while disavowing the bombastic sentiments), I note that it is studded with words which Waller uses to like purposes and which help to distinguish him from most others. Of Waller’s 2,876 word-types, 259 meet the stipulations stated above. The ten most common of them are: sacred, rage, glory, sing, law, fortune, equal, reign, yield, triumph. The most common such function-word is ‘like’, used as an adjective.

The effectiveness of the Zeta test on this occasion is undeniable. Its more general reliability remains in doubt. A set of stipulations that identifies texts characteristic of Waller’s maturity yields a word-list that might well be less accurate in identifying the love-songs of his youth or the immense religious poems of his dotage. But such words as the adjective ‘like’ might persist throughout his long career.

The right-hand panels of Table 4 show corresponding results for the Iota test. The stipulations employed on this occasion exclude all those words that occur in more than two of Waller’s five segments and all those that occur in more than ten of the twenty-four authorial components of the counter-set. Waller’s 2,876 word-types are reduced to 1,127. The 1,380 word-tokens occur at a rate of 99.73 per 1,000 in the base-set. They are matched by 8,406 at a rate of 16.14 in the counter-set. At 34.80 and 39.54 per 1,000, the two Waller poems easily outscore the other six. They leave the members of Group A far behind though the highest of these lie above some of the six non-Waller pieces. Some of the independent texts in Groups B and C score higher than any in Group A. But none of them approach Waller’s pair. The standard deviations for all three groups are low.

Of the individual texts, the least remote are Absalom and Achitophel and Christopher Wase’s Divination. The former is a celebrated political satire, the latter a pro-Wallerian contribution to the debate focussed on his ‘Instructions to a Painter’. In this less common range of words, we are obviously tapping a different vein from that of Waller’s patriotic effusions. The word-list now opens with the following ten word-types: portion, armies, exceed, fishes, indite, maker, Christians, tragedy, Chloris, Turks. These and their successors have little in common save for the salient point: they are words used by Waller but not by most of his contemporaries. It is worth noticing, moreover, that they are not truly rare words. There are few rarities even among the 127 word-types used by Waller alone. None of them occur more than twice in Waller but the first six appear in two of his segments. The list begins: repressed, Isaiah, Pandora, daw, piracy, displaying, vizier, enlargement, Antarctic, forebodes.

4.3 One against many: Marvell and twenty-four others

The two tests whose results are set out in Table 5 have Marvell as base-set and the other twenty-four poets as counter-set. The size of the counter-set is altered by the replacement of Marvell by Waller among the twenty-four. The stipulations are exact Marvellian equivalents of those used for Table 4.

As the right-hand panel of Table 5 makes clear, Marvell responds almost as well as Waller to the Iota test. Yet the 2,344 word-types that satisfy our stipulations are miscellaneous indeed. As with the corresponding set in Waller, few of them are rare. The upper range of the list, the words that Marvell shares with ten of the other poets, is marked by the language of pastoral. But, as it must, the list grows more idiosyncratic as more of the other poets are excluded. The set of words used by Marvell alone
Table 5 Marvell and twenty four others. Two tests on eight poems: Marvell’s word-list

<table>
<thead>
<tr>
<th></th>
<th>Length</th>
<th>Zeta test</th>
<th>Iota test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Types</td>
<td>Tokens per 1000</td>
</tr>
<tr>
<td>Marvell</td>
<td>Base-set</td>
<td>20151</td>
<td>389</td>
</tr>
<tr>
<td>24 poets</td>
<td>Counter-set</td>
<td>514359</td>
<td>6473</td>
</tr>
<tr>
<td>Test-pieces</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marvell</td>
<td>Base-set</td>
<td>20151</td>
<td>389</td>
</tr>
<tr>
<td>24 poets</td>
<td>Counter-set</td>
<td>514359</td>
<td>6473</td>
</tr>
</tbody>
</table>

Group A. Components of counter-set

<table>
<thead>
<tr>
<th>Author</th>
<th>Type</th>
<th>Tokens</th>
<th>Per 1000</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behn</td>
<td>21705</td>
<td>272</td>
<td>1248</td>
<td>57.50</td>
</tr>
<tr>
<td>Brome</td>
<td>29539</td>
<td>261</td>
<td>996</td>
<td>33.72</td>
</tr>
<tr>
<td>Butler</td>
<td>30932</td>
<td>270</td>
<td>1206</td>
<td>38.99</td>
</tr>
<tr>
<td>Congreve</td>
<td>30917</td>
<td>343</td>
<td>1877</td>
<td>60.71</td>
</tr>
<tr>
<td>Cotton</td>
<td>12625</td>
<td>238</td>
<td>951</td>
<td>46.81</td>
</tr>
<tr>
<td>Cowley</td>
<td>19272</td>
<td>293</td>
<td>1074</td>
<td>55.73</td>
</tr>
<tr>
<td>Denham</td>
<td>30094</td>
<td>311</td>
<td>1484</td>
<td>49.31</td>
</tr>
<tr>
<td>Dorset</td>
<td>9586</td>
<td>175</td>
<td>344</td>
<td>35.89</td>
</tr>
<tr>
<td>Dryden</td>
<td>18238</td>
<td>285</td>
<td>930</td>
<td>50.99</td>
</tr>
<tr>
<td>Durfey</td>
<td>18757</td>
<td>269</td>
<td>782</td>
<td>41.69</td>
</tr>
<tr>
<td>Gould</td>
<td>29110</td>
<td>286</td>
<td>1139</td>
<td>39.13</td>
</tr>
<tr>
<td>Milton</td>
<td>18924</td>
<td>303</td>
<td>1214</td>
<td>64.15</td>
</tr>
<tr>
<td>Oldham</td>
<td>32462</td>
<td>305</td>
<td>1437</td>
<td>44.27</td>
</tr>
<tr>
<td>Phillips</td>
<td>29004</td>
<td>312</td>
<td>1419</td>
<td>48.92</td>
</tr>
<tr>
<td>Prior</td>
<td>32000</td>
<td>308</td>
<td>1567</td>
<td>48.97</td>
</tr>
<tr>
<td>Radcliffe</td>
<td>11889</td>
<td>203</td>
<td>479</td>
<td>40.29</td>
</tr>
<tr>
<td>Rochester</td>
<td>12725</td>
<td>230</td>
<td>503</td>
<td>39.53</td>
</tr>
<tr>
<td>Sedley</td>
<td>10304</td>
<td>213</td>
<td>428</td>
<td>41.54</td>
</tr>
<tr>
<td>Settle</td>
<td>20333</td>
<td>285</td>
<td>1071</td>
<td>52.67</td>
</tr>
<tr>
<td>Shadwell</td>
<td>14540</td>
<td>237</td>
<td>589</td>
<td>40.51</td>
</tr>
<tr>
<td>Swift</td>
<td>30974</td>
<td>315</td>
<td>1321</td>
<td>42.65</td>
</tr>
<tr>
<td>Tate</td>
<td>13838</td>
<td>240</td>
<td>684</td>
<td>49.43</td>
</tr>
<tr>
<td>Waller</td>
<td>12511</td>
<td>226</td>
<td>684</td>
<td>54.67</td>
</tr>
</tbody>
</table>

Group B. Independent texts by these poets

<table>
<thead>
<tr>
<th>Author</th>
<th>Type</th>
<th>Tokens</th>
<th>Per 1000</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Behn16</td>
<td>Isle of Love</td>
<td>16419</td>
<td>231</td>
<td>979</td>
</tr>
<tr>
<td>Brome16</td>
<td>Answer</td>
<td>1385</td>
<td>38</td>
<td>54</td>
</tr>
<tr>
<td>Butler01</td>
<td>Plagiaries</td>
<td>1217</td>
<td>29</td>
<td>33</td>
</tr>
</tbody>
</table>

(Continued)
<table>
<thead>
<tr>
<th>Length</th>
<th>Types</th>
<th>Tokens</th>
<th>per1000</th>
<th>Rank</th>
<th>Types</th>
<th>Tokens</th>
<th>per1000</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butler02</td>
<td>Weakness</td>
<td>1401</td>
<td>36</td>
<td>44</td>
<td>31.41</td>
<td>19</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Congreve14</td>
<td>Imposs. Thing</td>
<td>1376</td>
<td>44</td>
<td>54</td>
<td>39.24</td>
<td>14</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>Cotton09</td>
<td>Ireland2</td>
<td>1830</td>
<td>51</td>
<td>78</td>
<td>42.62</td>
<td>9</td>
<td>39</td>
<td>49</td>
</tr>
<tr>
<td>Cotton11</td>
<td>Epistle</td>
<td>1537</td>
<td>35</td>
<td>44</td>
<td>28.63</td>
<td>20</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Cowley17</td>
<td>Royal Society</td>
<td>1315</td>
<td>48</td>
<td>61</td>
<td>46.39</td>
<td>6</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Cowley18</td>
<td>Davideis2</td>
<td>6812</td>
<td>199</td>
<td>380</td>
<td>55.78</td>
<td>3</td>
<td>165</td>
<td>209</td>
</tr>
<tr>
<td>Dryden03</td>
<td>Absalom</td>
<td>7824</td>
<td>160</td>
<td>308</td>
<td>39.37</td>
<td>13</td>
<td>176</td>
<td>232</td>
</tr>
<tr>
<td>Dryden02</td>
<td>Epistle15</td>
<td>1650</td>
<td>42</td>
<td>57</td>
<td>34.55</td>
<td>17</td>
<td>57</td>
<td>62</td>
</tr>
<tr>
<td>Durfey</td>
<td>Malecontent</td>
<td>7817</td>
<td>182</td>
<td>341</td>
<td>43.62</td>
<td>8</td>
<td>162</td>
<td>191</td>
</tr>
<tr>
<td>Gould13</td>
<td>BeauEsprits</td>
<td>6020</td>
<td>128</td>
<td>241</td>
<td>40.03</td>
<td>12</td>
<td>118</td>
<td>123</td>
</tr>
<tr>
<td>Gould16</td>
<td>Fanaticism</td>
<td>1995</td>
<td>62</td>
<td>80</td>
<td>40.10</td>
<td>11</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>Milton02</td>
<td>PR</td>
<td>15694</td>
<td>231</td>
<td>732</td>
<td>46.64</td>
<td>5</td>
<td>341</td>
<td>300</td>
</tr>
<tr>
<td>Milton03</td>
<td>SA</td>
<td>12885</td>
<td>193</td>
<td>605</td>
<td>46.95</td>
<td>4</td>
<td>281</td>
<td>412</td>
</tr>
<tr>
<td>Oldham07</td>
<td>Ona Woman</td>
<td>1298</td>
<td>46</td>
<td>59</td>
<td>45.50</td>
<td>16</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Oldham13</td>
<td>Ben Johnson</td>
<td>2237</td>
<td>72</td>
<td>94</td>
<td>43.20</td>
<td>10</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>Phillips</td>
<td>Lady E.C.</td>
<td>1210</td>
<td>36</td>
<td>44</td>
<td>36.36</td>
<td>16</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Prior18</td>
<td>Henry&amp;Emma</td>
<td>6033</td>
<td>164</td>
<td>368</td>
<td>61.00</td>
<td>1</td>
<td>129</td>
<td>164</td>
</tr>
<tr>
<td>Swift18</td>
<td>Verses</td>
<td>3206</td>
<td>74</td>
<td>104</td>
<td>32.44</td>
<td>18</td>
<td>62</td>
<td>88</td>
</tr>
</tbody>
</table>

Group C. Independent texts by other poets

MAX 65.62 44.60
MEAN 48.35 30.17
ST DEV 13.72 7.65

 Opens with: ungirt, practising, departure, ambergris, Thwaites, tulip, tinkling, melons, perpetration, architects. While this little sample is not an encouraging beginning, the full list of 2,344 allows a cumulative effect strong enough to identify Marvell’s work with almost perfect accuracy.

Of our first eight specimens, those associated with Marvell range down from 47.45 to 39.31 per

<table>
<thead>
<tr>
<th>Length</th>
<th>Types</th>
<th>Tokens</th>
<th>per1000</th>
<th>Rank</th>
<th>Types</th>
<th>Tokens</th>
<th>per1000</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Butler02</td>
<td>Weakness</td>
<td>1401</td>
<td>36</td>
<td>44</td>
<td>31.41</td>
<td>19</td>
<td>36</td>
<td>38</td>
</tr>
<tr>
<td>Congreve14</td>
<td>Imposs. Thing</td>
<td>1376</td>
<td>44</td>
<td>54</td>
<td>39.24</td>
<td>14</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>Cotton09</td>
<td>Ireland2</td>
<td>1830</td>
<td>51</td>
<td>78</td>
<td>42.62</td>
<td>9</td>
<td>39</td>
<td>49</td>
</tr>
<tr>
<td>Cotton11</td>
<td>Epistle</td>
<td>1537</td>
<td>35</td>
<td>44</td>
<td>28.63</td>
<td>20</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Cowley17</td>
<td>Royal Society</td>
<td>1315</td>
<td>48</td>
<td>61</td>
<td>46.39</td>
<td>6</td>
<td>37</td>
<td>39</td>
</tr>
<tr>
<td>Cowley18</td>
<td>Davideis2</td>
<td>6812</td>
<td>199</td>
<td>380</td>
<td>55.78</td>
<td>3</td>
<td>165</td>
<td>209</td>
</tr>
<tr>
<td>Dryden03</td>
<td>Absalom</td>
<td>7824</td>
<td>160</td>
<td>308</td>
<td>39.37</td>
<td>13</td>
<td>176</td>
<td>232</td>
</tr>
<tr>
<td>Dryden02</td>
<td>Epistle15</td>
<td>1650</td>
<td>42</td>
<td>57</td>
<td>34.55</td>
<td>17</td>
<td>57</td>
<td>62</td>
</tr>
<tr>
<td>Durfey</td>
<td>Malecontent</td>
<td>7817</td>
<td>182</td>
<td>341</td>
<td>43.62</td>
<td>8</td>
<td>162</td>
<td>191</td>
</tr>
<tr>
<td>Gould13</td>
<td>BeauEsprits</td>
<td>6020</td>
<td>128</td>
<td>241</td>
<td>40.03</td>
<td>12</td>
<td>118</td>
<td>123</td>
</tr>
<tr>
<td>Gould16</td>
<td>Fanaticism</td>
<td>1995</td>
<td>62</td>
<td>80</td>
<td>40.10</td>
<td>11</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>Milton02</td>
<td>PR</td>
<td>15694</td>
<td>231</td>
<td>732</td>
<td>46.64</td>
<td>5</td>
<td>341</td>
<td>300</td>
</tr>
<tr>
<td>Milton03</td>
<td>SA</td>
<td>12885</td>
<td>193</td>
<td>605</td>
<td>46.95</td>
<td>4</td>
<td>281</td>
<td>412</td>
</tr>
<tr>
<td>Oldham07</td>
<td>Ona Woman</td>
<td>1298</td>
<td>46</td>
<td>59</td>
<td>45.50</td>
<td>16</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Oldham13</td>
<td>Ben Johnson</td>
<td>2237</td>
<td>72</td>
<td>94</td>
<td>42.02</td>
<td>10</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>Phillips</td>
<td>Lady E.C.</td>
<td>1210</td>
<td>36</td>
<td>44</td>
<td>36.36</td>
<td>16</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>Prior18</td>
<td>Henry&amp;Emma</td>
<td>6033</td>
<td>164</td>
<td>368</td>
<td>61.00</td>
<td>1</td>
<td>129</td>
<td>164</td>
</tr>
<tr>
<td>Swift18</td>
<td>Verses</td>
<td>3206</td>
<td>74</td>
<td>104</td>
<td>32.44</td>
<td>18</td>
<td>62</td>
<td>88</td>
</tr>
</tbody>
</table>
1,000 and so outscore the other four. They also outscore the twenty-four authorial sets of Group A. Of the twenty-one specimens in Group B, Congreve’s comic tale, ‘An Impossible Thing’, at 40.70 per 1,000, outscores the ‘Third Advice’. Of the twenty poems in Group C, John Tutchin’s satire ‘A Search after Honesty’, at 44.60 per 1,000, also outscores the ‘Third Advice’. The other three Marvell texts outscore all others.

Tutchin’s brand of rough satire often yields unexpected resemblances to satirical poems by other authors. Congreve, like Milton, has a richer vocabulary than most even when his large authorial set is cut down. One effect is that he often uses more of the uncommon words than most and, accordingly, tends to encroach on the word-lists of his fellows. Since Tutchin was a child of about six and Congreve was yet unborn when the ‘Third Advice’ was first published, neither is a candidate for its authorship.

To resolve the question without benefit of such external evidence, the current set of stipulations can be modified. In the present instance, perfect accuracy is obtainable by discarding all the words that are used by more than four of the other twenty-four poets. The other obvious course is to set up one-on-one contests (as illustrated earlier) between the claimants. When the words they share are excluded by the usual sort of stipulations, the remainder afford a basis for weighing up their respective claims. I have yet to encounter a case where the Zeta and Iota tests fail when they are used in a genuine one-on-one end game.

The central panel of Table 5 shows the results of a Zeta test in which Marvell’s base-set is matched against the counter-set of twenty-four authors. As with Waller, the exclusion of words that occur in fewer than three segments of the base-set and in more than twenty-two members of the counter-set yields a solid foundation. The rates per thousand are 100.89 and 47.08 for base-set and counter-set respectively.

The power of the Zeta test is seen once more in the fact that, at 66.43 per 1,000, the rate for Marvell’s ‘First Anniversary’ surpasses that of every other specimen examined. Here, as with the Waller poems in Table 4, the test ranks the authentic text ahead of more than forty rivals. Of the main eight poems, moreover, the ‘Fourth’ and ‘Fifth Advice’, the two not considered to be Marvell’s, rank below all of his. These are all much stronger results than might reasonably have been expected of a test treating of the middle frequency stratum. In that hitherto neglected stratum, as I remarked at the beginning of this article, the demands of subject and occasion might be expected to prevail over the effects of authorial habit.

For the Marvell satires, however, the central panel of Table 5 shows the effect of those very demands and is a stern reminder that the orientation of the Zeta test is not necessarily—and therefore not always—authorial. Apart from ‘The First Anniversary’, the poems associated with Marvell are all outscored by many texts by various other authors.

This sudden outcrop of failures needs to be understood. Although Marvell’s corpus is diverse in literary form, little of the unquestioned work that comprises his base-set is satirical. Several of his longer poems bear on affairs of state but his most persistent note is pastoral. Given the stipulation of consistent recurrence employed here, his word-list for the Zeta test includes many words that occur more often in pastoral than in other literary forms. Poems of that cast are marked not only by some loosely related lexical words but also by a tendency to use function words that were already becoming archaic. The twenty most common words of Marvell’s Zeta-list are: flowers, doth, green, grass, lest, unto, who (interrogative), straight, O, Heaven’s, pure, grief, Oh, thine, hence, roses, equal, stay, under, trees. Many of these twenty word-types (like others from further down the list) are more at home in Marvell’s pastoral lyrics than in his satires where they occur, as a group, at less than half the rate they attain in his main set. Taken individually, several of them do show frequencies normal for Marvell while others do not occur in any of the political satires associated with him. The leading examples of the first sort are ‘lest’, ‘straight’, ‘who’, and ‘under’. These can be taken as representing many other words, from further down the list, words that Marvell is always inclined to use. The leading examples of the second sort, which do not
occur in any of these satires, are ‘grass’, ‘pure’, and ‘roses’. Other members of my little set of twenty, like ‘green’ and ‘trees’, put in an appearance or two. These, too, can be taken as representing a large number of non-political words from further down the list.

It must be emphasized that we are dealing not in absolutes but in lexical probabilities. Most English words can put in an appearance in quite unexpected contexts. But such instances are heavily outnumbered, in any given context, by words more usual there. Our language is so often metaphorical in cast that even a satire on naval affairs is not proof against the language of pastoral. Here is a fragment from my ‘Dutch Comfort, or Our late Reverses at Sea’. Even here, the unexpected words soon begin to be outnumbered by those that might be expected. And, thinking of what is likely rather than what is conceivable—thinking, that is, statistically—one would expect more sails than trees, more guns than meadows in any long poem of this kind:

Far o’er the Atlantick Meadows, pure and green,
A threat’ning Clowd of Trees was to be seen.
Close-haul’d, but favour’d by the mounting Gale,
Van Tromp’s main Squadron of some forty Sail!
Our Duke, who strode the Poop with haughty Mien
Had been outwitted by a Foeman keen.
Proud Phoebus but a Phae¨thon, we found,
Our Fates with his inextricably bound.

Consider the four word-types that I purposely implanted here: pure, green, trees, and meadows, together with the singular forms of the two nouns. All told, they occur six times in the 14,198 words of the three ‘Advice’ poems associated with Marvell. In the other ‘Advices’, which comprise 2,312 words, they do not occur at all. They occur only sparsely in three of Marvell’s long poems, ‘A Poem upon the Death of O. C.’, ‘An Horatian Ode’, and ‘Fleckno, an English Priest at Rome’. Of their 4,503 words, these word-types make up only six. Three of them, all from ‘On the Death of O. C.’, are instances of ‘tree(s)’. The rest are scattered single instances. But Marvell’s ‘Upon Appleton House’ sets all these figures in high relief. In that poem, they occur twenty-seven times in 4,845 words.

In situations of this kind, little is gained by altering the stipulations so as to modify the word-list. To relax them is to weaken their power to differentiate. To tighten them is likely to intensify a given effect. Of the forty or so word-types that occur in all five segments of Marvell’s main set, many are redolent of pastoral. It is fair to suppose that any pastoral poem, from Lycidas to Pope, would be likely to achieve a high score in this particular exercise and that no satire, from Dryden to Pope, would be likely to match it.

Ultimately, of course, the question is not whether I am justified in supposing that the dearth of ‘pastoral’ words in the ‘Advices’ is the main reason why the Zeta test fails to identify three of those poems as Marvell’s. The question is what to make of a test where perfect accuracy is suddenly offset by utter failure. Our three failures make it clear that Zeta scores are not always reliable authorial markers. Our three successes, in each of which the scores for a specified poem surpass the scores for more than forty others, suggest that the Zeta test should not be discarded. That position is supported by the fact that other trials I have made have shown a high level of accuracy.

The difficulty, it seems, is that there is an intractable weakness in our ‘model’, a select word-list that does not reflect the full range of Marvell’s repertoire. In such cases, a particular test may need to be set aside. It is as if a doctor were to say, ‘I’m pretty sure this is a new mutation. If I’m right, there is no point in using our standard test this time. We’ll just have to try another approach’. Far better for him to do so than to go on against the grain and discover the truth in an autopsy. Most poets usually write within their customary repertoires. Whenever they do so, the Zeta test shows high levels of accuracy. When they do not, a reader should know it. The language of the three main ‘Advices’ is Marvellian enough to enable the Delta and Iota tests to operate successfully. The content-words of the middle frequency stratum set these poems apart from the main body of his work. Now whereas even a good reader cannot easily see trends in the frequency of very common words, any reader of Marvell will recognize this large shift of subject.
But is it appropriate to appeal to the reader’s judgment in this way? What of the argument that computational methods are meant to settle questions on which informed readers disagree? To make that a \textit{sine qua non} is an unreasonable demand. Computational methods can often shed new light on vexed questions. But, like all other forms of inductive reasoning, they yield inferences, and not absolute proofs. The best inferences will be drawn by the best students of the full range of evidence: with new methods, as with old, we are always obliged to read and think. When disagreement persists, our best recourse is a return to the beginning. (One of Wittgenstein’s dicta comes to mind: ‘Back to the rough ground. Look and see.’) It may be possible to find a weakness in reasoning, to form better inferences, or to show that the burden of proof is altered by the new evidence. It may be necessary to frame the question better and test it afresh. In twenty-five years, however, I have scarcely seen a case where well-founded computational work remained seriously at odds with the scholarly consensus.

5 Conclusions and Suggestions

One way of measuring the accuracy of these tests is to declare an error whenever the score for a text by the target-author is surpassed by that of any other test-piece. In Tables 4 and 5, the members of Group A, being components of the counter-set, should not be used for this purpose. Where they are involved, an ‘error’ is certainly a danger-signal: but it would be specious to register ‘successes’ of this kind. For Waller, therefore, in the present case, we are measuring two poems of his authorship against forty-seven others. The results of the Zeta test yield no errors out of ninety-four comparisons. So, too, for the Iota test. At no point does either of these rather characteristic pieces encounter a near-rival. Matched against forty-five poems by other authors, Marvell’s ‘First Anniversary’ also yields a perfect record of success for both Zeta and Iota. Its margin of advantage, however, is often much less than Waller’s was. The Iota test registers forty-five successes out of forty-five for the ‘Last Instructions’ and the ‘Second Advice’, forty-three out of forty-five for the ‘Third Advice’. (Even this last score becomes forty-five out of forty-five when the second stipulation is tightened.)

All told, the Iota test yields 272 successful comparisons out of 274, a success-rate of over 99%. The Zeta test is 100% successful for the two Waller poems and ‘The First Anniversary’. With the political satires, however, the Zeta test is unable to distinguish Marvell from many of his fellows. The errors reach double figures for all three poems and the test is clearly inappropriate in such cases.

The one-on-one trials illustrated in Table 3 have yet to yield a genuine error. But, as the only aberrant score in that set indicated, such trials are better not undertaken until wider-ranging tests (or firm external evidence) have identified the main claimants. Stipulations designed to distinguish one author from one other cannot be expected to work properly on a third.

Further trials will show whether our only cluster of errors can be set aside as the effect of a predictable mismatch. If it can, we are looking overall at very high success-rates. That offers good reason for undertaking further work. And the Iota results, in particular, add to previous evidence that three, but only three, of the ‘painter satires’ are the work of Andrew Marvell.

Whoever undertakes them (myself I hope included), such further trials should obviously incorporate a range of variations on the stipulations employed here. Experience so far suggests some limits. It is always desirable to frame the second stipulation, which governs the counter-set, so as to exclude all the words that are used in common-words analyses. It is sometimes necessary to tighten this stipulation further in order to differentiate between authors. If the first stipulation, which governs the base-set, imposes too strict a rule of consistency few words will survive. When (as was noted earlier) the field is impoverished, the test is open to adventitious effects. Too strict a rule of consistency can also yield errors with genuine but uncharacteristic texts. Another worthwhile variation, for use with texts too short to allow the effective use of both Zeta and Iota, is what one might call ‘Iota plus’. In this situation, the select list might exclude only those words that occur in, say, four or more segments of the base-set and also
exceed a fairly strict quantum for the counter-set. Like Iota itself, this approach yields a completely accurate differentiation between our four Marvell poems and all the rest. For short texts, then, it might well offer a useful check on the common-words procedures.

Apart from short texts and over-strict stipulations, aberrant frequencies most often arise in words that might reasonably be culled. Henry Fielding is more given than his sister to the adjective ‘simple’. But her novel, *David Simple*, has a spectacular effect upon the word-count. The compelling objection to *ad hoc* forms of culling is that they make it too easy to tip the balance in one direction or another. In recent articles (e.g. 2004), David Hoover successfully culled those word-types that exceed a given frequency in any of the specimens. That approach seems an effective way of controlling aberrant behaviour in any of the specimen-texts without favouring a preferred outcome.

Further trials should also establish whether it is necessary, for the best effects, to work with text-sets of uniform length. The obvious advantage is offset by a sort of extravagance. This is less damaging with drama and prose fiction than with poetry and personal letters. Poetic texts and poetic corpora both vary too much in length to lie easily upon a procrustean bed. To obtain sets of uniform size, we must either reduce the larger to match the smaller or turn away from the smaller. Too little Milton and Congreve or no Rochester and Sedley? Put like that, the choice is too invidious to contemplate. A better way to put it is that we need to determine what is most appropriate, accepting only such limitations as we must, and resisting them when we can.

With so many variations still in play, it is too soon to offer a binding definition of a Zeta or an Iota score. It is better, at least for the present, to think of Zeta and Iota as working-labels for two little families of tests based upon select word-lists. The lists are formed on the basis of stipulated contrasts between a base-set and a counter-set. In Zeta, the stipulations admit only those word-types that attain a specified level of consistency in the base-set while failing to reach a specified level, whether of consistency or of frequency, in the counter-set. Iota rests upon the residue, embracing word-types that do not meet the first stipulation while occurring even more rarely in the counter-set.9

The results offered in this article, along with those obtained in other trials, suggest that these tests themselves may be of real use in cases of doubtful authorship. Although no two word-frequency tests can ever be entirely independent of each other, each of these tests employs frequencies from a different stratum and, again, from a stratum other than that used in the more familiar common-words procedures. For good and ill, however, they are very simple additive measures. Their ultimate value may be less direct, lying rather in demonstrating that evidence of authorship is indeed present in every frequency stratum. Our task now is to find the best ways of deploying it.

**Acknowledgements**

The author is indebted to those who reviewed this article for many helpful comments and suggestions. In their different ways, as on many similar occasions, Hugh Craig and Harold Love have given invaluable assistance and encouragement.

**References**


Notes

1 The provenance of the main ‘Advice’ poems is set out in Margoliouth (3rd ed., 1971), where only the ‘Last Instructions’ is accepted as Marvell’s. The most accessible set of texts is in Lord (1963).

2 The corpus of about 540,000 words has been reduced by some six thousand by setting aside three poems by Waller and Marvell for use as independent specimens. The corpus ranges widely across the work of the following twenty-five poets: Aphra Behn (1640–89) 21,705 words; Alexander Brome (1620–66) 29,539; Samuel Butler (1612–80) 30,932; William Congreve (1670–1729) 30,917; Charles Cotton (1630–87) 12,625; Abraham Cowley (1618–67) 19,272; Sir John Denham (1615–69) 30,092; Charles Sackville, Earl of Dorset (1638–1706) 9,586; John Dryden (1631–1700) 18,238; Thomas D’Urfey (1653–1723), 18,757; Robert Gould (1660?–1709?) 29,110; Andrew Marvell (1621–78) 23,282; John Milton (1608–74) 18,924; John Oldham (1653–83) 32,462; Katherine Phillips (1631–64) 29,004; Matthew Prior (1664–1721) 32,000; Alexander Radcliffe (*floruit* 1669–96) 11,889; John Wilmot, Earl of Rochester (1648–80) 12,725; Sir Charles Sedley (1639?–1701) 10,304; Elkanah Settle (1648–1724) 24,080; Thomas Shadwell (1642?–92) 14,540; Jonathan Swift (1667–1745) 30,974; Nahum Tate (1652–1715) 20,333; Edmund Waller (1606–87) 16,443; Anne Wharton (1659–85) 12,511. Most of the corpus was prepared by John Burrows and Harold Love, assisted by Alexis Antonia and Meredith Sherlock. The Marvell subset was added by Christopher Wortham assisted by Joanna Thompson.

3 Like the greater part of the main database, some of the independent texts were entered, by keyboard, from standard texts. The rest were downloaded from the Chadwyck-Healey archive, to which my university subscribes. I am much in debt to all those whose work has made mine possible.

4 It is often useful to distinguish between word-types and word-tokens. The many occurrences of ‘the’ in any given text are called word-tokens, instances of the word-type ‘the’.

5 See Burrows (2002, 2003) and Hoover (2004). A 1999 article (first shown to me in 1998) has a brief passage in which absolute z-scores are used as a measure of distance. Although I do not remember noticing it (in an article whose weight lies elsewhere), it may have given me a clue that came to mind when it was needed. I am content, in any case, to yield precedence on this point. See Forsyth, Holmes, and Tse (1999, 393).

6 An outline of the calculation and use of z-scores can be found in introductory manuals of statistics. But readers in need of such help may be best served by the lucid plain-language account in Kenny (1982, 57–8).

7 My colleague, Hugh Craig, has been experimenting successfully on these lines.

8 Willard McCarty has been a persistent and eloquent advocate of the idea that we learn most from our failures. One must hope that it applies to the following remarks. See, for example, McCarty (2005), 286, s. v. ‘failure’.

9 The notion of constructing an authorial word-list and then excluding those used by a stipulated number of other writers has not, I think, been tried. But there is an extensive literature on rare words, especially those that occur only once or twice in a given text. David Holmes (1994) offers a helpful summary of these procedures (among others). More recently, access to electronic archives has assisted in identifying words peculiar to a given author. The work of Donald W. Foster offers notable examples, embracing both success and failure. The protracted controversy surrounding a failed Shakespeare attribution (Foster, 1989) need not occupy us here.
Teasing Out Authorship and Style with T-tests and Zeta

Hoover, David L.
david.hoover@nyu.edu
New York University, USA

Most computational stylistics methods were developed for authorship attribution, but many have also been applied to the study of style. Investigating Wilkie Collins's *Blind Love* (1890), left unfinished at his death and completed by Walter Besant from a long synopsis and notes provided by Collins, requires both authorship attribution and stylistics. External evidence indicates that Besant took over after chapter 48 (Collins 2003), which provides an opportunity to test whether Besant was successful in matching Collins's style and to investigate the styles of Collins and Besant. This divided novel also facilitates the comparison of two computational methods: the T-test and Burrows's Zeta.

The t-test is a well-studied method for determining the probability of a difference between two groups arising by chance (a classic use in authorship and stylistics is Burrows 1992.) Here I use t-tests to identify words used very differently by Collins and Besant. After showing that those word frequencies accurately identify the change of authorship, I examine the words themselves for stylistically interesting characteristics.

I created a combined word frequency list for four novels by Besant and three by Collins, then deleted words occurring only once or twice, personal pronouns (too closely related to the number and gender of characters), all words with more than 90% of their occurrences in one text (almost exclusively proper names), and words limited to one author (required for t-testing). I divided the novels into 167 4,000-word sections, and performed t-tests for the remaining 6,600 words (using a Minitab macro). I cleaned up the results and sorted them on the p value in Excel (with another macro), and retained only the 1719 words with p < .05, about 1,000 for Collins and 700 for Besant (see https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html for detailed instructions and the macros).

I tested these words on six new texts for each author, a novel and five stories for Besant and six novels for Collins. Beginning with the 500 most distinctive words for each author, I deleted a few words that were absent from these texts and used the remaining 993 words to perform a cluster analysis (Fig. 1). (To keep the graph readable, I divided the novels into 10,000-word sections, retaining only half the sections.) Obviously, these marker words are quite characteristic of the authors.

When sections of *Blind Love* are tested along with the texts above, the authorship change after chapter forty-eight is starkly apparent (Fig. 2). This graph is based on the sums of the frequencies of the 500 most distinctive words for each author in each section. (The texts are divided into 1,000-word sections; only a few sections of the novels are shown; the frequencies...
of Collins's marker words are multiplied by -1 for clarity.) Although Besant was working from extensive notes, his style is distinctly different. Had we not known which was Besant's first chapter, these t-tested marker words would have easily located it.

Because the styles of Collins and Besant are so distinct, these marker words should also characterize them. Consider the twenty most distinctive words for each author:

Besant: upon, all, but, then, and, not, or, very, so, because, great, thing, things, much, every, there, man, everything, is, well

Collins: answered, to, had, Mrs, on, asked, in, Miss, mind, suggested, person, resumed, excuse, left, at, reminded, creature, inquired, reply, when

Obviously, more of Besant's words are high frequency function words, and many Collins words are related to speech presentation (answered, asked, inquired, resumed, suggested, reply, and reminded). The presence of added, begged, declared, exclaimed, explained, expressed, muttered, rejoined, and said as likely speech markers among the other Collins marker words, but only gasped, groaned, murmured, replied, and stammered for Besant, suggests they have different ways of presenting speech.

Sorting all of each author's marker words alphabetically immediately reveals word families that each author favors, as thing, things, and everything among the twenty most distinctive Besant words already suggests (anything and nothing are also Besant markers). His every and everything are joined by everybody and everywhere; anything by any and anywhere; nothing and not by never, no, nobody, none, and nor; and much by more, moreover, most, and mostly among his markers. Collins's answered is joined by answer, answering, and unanswerable; and five of his twenty words are joined by two others: ask, asked, asks; inquired, inquiries, inquiry; leave, leaving, left; person, personally, persons; suggest, suggested, suggestion.

About 600 of the 1,700 distinctive words form groups favored by one author, but only about 175 form split groups, many of which fall into intriguing patterns. Collins uses more contractions, so didn't, doesn't, and don't are Collins words, but did and does are Besant words, and similarly for must, need, should, and would and their negative contractions. The singular and possessive forms of brother, friend, sister, and son are Collins's words and the plural forms are Besant's; the singular vs. plural pattern continues almost without exception in split noun groups. Verbs in -ing are Collins words and 3rd singular present forms Besant's. Finally, all nineteen cardinal number marker words are Besant's, including the numbers one to ten (note that Besant's preferred plural nouns often follow numbers). This extraordinary patterning may not seem particularly surprising, but, so far as I know, it has never been noticed before, and cries out for investigation.

Two problems with t-testing are its privileging of relatively uninteresting high-frequency words and its inability to cope with words absent from one author. John Burrows’s Zeta addresses both of these problems (Burrows 2006). (The specific form used here was developed by Hugh Craig (Craig and Kinney, 2009); for an automated spreadsheet and instructions for performing
Zeta analysis see https://files.nyu.edu/dh3/public/TheZeta&IotaSpreadsheet.html.

Zeta's simple calculation begins with the same novels and the same word frequency list used for the t-test, except that personal pronouns and words present in only one author are now included. Zeta is simply the sum of the proportions of Collins sections in which each word occurs and Besant sections in which it does not. Here answered, the most distinctive Collins word (as in the t-tests), has a Zeta score of 1.8, and is present in 89 of 90 Collins sections and absent from 65 of 77 Besant sections. The most distinctive Besant word is again upon, present in all 77 Besant sections and absent from 25 of 90 Collins sections, with a Zeta of 0.28. Below are the twenty most distinctive Zeta words (those also identified by t-testing in bold):

Besant: upon, fact, presently, therefore, however, everything, real, whole, cannot, though, rich, none, thousand, except, fifty, ago, because, papers, also, twenty

Collins: answered, Mrs, Miss, excuse, suggested, resumed, reminded, doctor, inquired, creature, notice, circumstances, tone, idea, temper, object, sense, feeling, governess, impression

As noted above, Zeta marker words are less frequent than t-tested words. Only two Zeta marker words rank in the top 100 in the novels, compared to 20 of the t-tested words. About 3/4 of the 1000 t-tested marker words are also among the 1000 Zeta markers. Among the 2000 Zeta words are 275 words occurring in only one author; 59 form new single-author families, 27 join existing single-author families, and only 21 form split families.

The Zeta words also effectively detect the change of authorship in Blind Love. In the scatter graph in Fig. 3, the axes show the percentages of all the word types (unique words) in each section that are Besant or Collins marker words (longer texts are divided into 4000-words sections; the labels for even-numbered Collins sections of Blind Love are removed; only a few sections of other novels are included). Note how distinct Besant’s chapters of Blind Love are from Collins’s, though many of them are pulled toward Collins. This graph also includes The Case of Mr. Lucraft (Case in bold), jointly written by Besant and James Rice; it suggests, as has been argued (Boege 1956: 251-65), that Besant did most of the actual writing.

T-tests and Zeta analysis are both effective authorship attribution methods that produce lists of characteristic vocabulary for the authors being compared. Both identify morphological and semantic families of words and uncover extraordinarily consistent patterns and puzzling inconsistencies that suggest new directions for literary and stylistic analysis.

References


The point of departure for the development of the ‘Delta procedure’ (as I call it) was the observation that the methods of comparison and authorial attribution currently employed in computational stylistics are better fitted for ‘closed games’ than for more open ones. The closed games take two forms. Where only two or three writers are eligible candidates for the authorship of a particular text and where that text is of a sufficient length, we are now well equipped to form strong inferences about their rival claims. The classic study of this kind is Mosteller and Wallace (1964). Holmes (2001) offers an excellent recent specimen. Where the real question is whether or not a particular writer (and no other) is the author, we are equally well equipped to test his or her claims. Tweedie et al. (1998) and Burrows and Craig (2001) offer recent specimens of this kind. But in ‘open games’, where we are faced with an anonymous text but have little or no outside evidence to identify the most likely candi-
dates, our current methods must be employed in an exhaustive and possibly fruitless series of iterations.

A reliable means of detecting unique authorial fingerprints (of whose very existence we do not yet have either proof or promise) would clearly be the best way of resolving these open games. But, in its absence, there is room for a simple measure capable of distinguishing the most likely candidates from a large group and also, where no candidate lays a sufficient claim, of indicating that it might be wise to look further afield. For want of such a measure, we are still bound by Bailey’s dictum (1979, p. 7), proposed over 20 years ago and lately put even more strictly by Binongo and Smith (1999, p. 464). We should confine ourselves, they hold, to cases where the choice lies within a narrow range of well-matched sets and we should proceed with only two authors’ texts at a time. But, at least in the initial stages of an inquiry, the ‘Delta procedure’ allows us to shake off these constraints. After employing it to identify the strongest candidates, we can use our current methods to choose among them. The open game is thus transformed into a closed game.

Most of the methods currently employed in computational stylistics rest upon multivariate statistical comparisons between some characteristics of a given specimen and those of an appropriate set of norms. The characteristics, which are used as statistical variables, comprise the relative frequencies of various simple phenomena such as alphabetic characters, strings of characters, whole words, or common grammatical forms. Each of these classes of variables has its advantages and disadvantages, and each has its adherents among scholars in the field. Forsyth and Holmes (1996) offer a reasoned overview. The advantage of working with whole words rests on their accessibility and their meaningfulness. They help us, in particular, to form close and fruitful inferences about the outcome of an inquiry. Whichever class of variables is chosen, it has become customary, in recent years, to allow the particular variables to ‘declare themselves’, thus obviating, as far as possible, the danger of a predetermined outcome. The words used, for example, might be the 100 most common in the database that provides the norms for a particular inquiry. In this sort of work on language, so our researches teach us, a wealth of variables, many of which may be weak discriminators, almost always offer more tenable results than a smaller number of strong ones. Strong features, perhaps, are easily recognized and modified by an author and just as easily adopted by disciples and imitators. At all events, a distinctive ‘stylistic signature’ is usually made up of many tiny strokes.

The multivariate statistical instruments now most used in computational stylistics are designed to elicit subtle trends in complex sets of figures. The results obtained from principal component analysis, for example, are usually rendered in bi-axial or tri-axial graphs where the pattern of the entries allows far-reaching inferences to be drawn. Whenever a specimen is added or removed, the whole pattern alters in a fashion that does not admit strict comparisons between graph and graph. It is as if the ingredients of a mixture were being altered and a fresh state studied each time.
A Measure of Stylistic Difference and Likely Authorship

In experienced hands, such methods yield excellent results. But they are obviously unsuitable for the crude but useful task of ranking many candidates in a single, all-embracing hierarchy, thus singling out the statistically most eligible among them. Even a ranked series of aggregates or means, I told myself, might serve that purpose if a sound basis were available. The path forward from this point was long obscured by the fact that, because the scores for any given specimen on the chosen set of variables diverge in both directions from the norms for the database, an aggregate or mean divergence would comprise an arbitrary mixture of positives and negatives.

Although the differences between positives and negatives—high scores, say, for *the* in this specimen but low ones for *I* and *me*—are most instructive, they are not the heart of the matter. An expression of difference, pure difference, is what we seek. If all the positive and negative divergences were rendered as absolute divergences, their overall aggregate or their mean might be of interest. A ‘delta-score’ is just such a mean divergence. The term ‘Delta’ (best rendered when possible as ‘\(\Delta\)’) was chosen to represent D for Difference and also as a gesture of respect for those heroic pioneers in our field who worked without benefit of computers. Among their various attempts to derive simple expressions of stylistic difference, Udney Yule’s Characteristic K was the most fruitful.

The first step in the procedure is to establish a frequency-hierarchy for the most common words in a large group of suitable texts. The texts are arranged in subsets representing the work of numerous authors appropriate to the task in hand. With texts of a bygone era, it is usual and desirable to standardize spelling and to expand contracted forms of expression to reduce the influence of trivial or accidental variations. (Just such variations were studied by some of the pioneers of stylometry. But when one works with common word-counts, they are merely a distortion.) It has also been our practice, in Newcastle, to tag some of the more common homographic forms to distinguish the different uses of words such as *so* and *that*. When the word-counts have been made, the frequencies are standardized as proportions of each authorial subset so that the larger subsets do not exert an undue influence on the composition or ranking of the hierarchy.

Working on these lines, we formed a database of verse by twenty-five poets of the English Restoration period. These yielded the frequency-hierarchies used for several recent studies of authorship based on principal component analysis. They also yielded the norms for the Delta project. For this project, I have added a further range of texts, all independent of the main set and all of unquestioned if not unquestionable authorship. (It is impossible to be confident of the authorship of every member of a large, mixed set of Restoration poems. But, having gone to reputable sources, I shall stand by the results.)

Table 1, based on a small Microsoft Excel worksheet, offers a ‘closed version’ of the procedure, bringing the top thirty words of the main set of Restoration verse to bear on a simple question. Can we demonstrate that John Milton has a better claim to a selection from *Paradise Lost*—27,154

---

2 Experiments with overall medians yielded less accurate results than those to be described.

3 The present corpus of 540,244 words ranges widely across the work of the following twenty-five poets: Aphra Behn (1640–89) 21,705 words; Alexander Brome (1620–66) 29,539; Samuel Butler (1612–80) 30,932; William Congreve (1670–1729) 30,917; Charles Cotton (1630–87) 12,625; Abraham Cowley (1618–67) 19,272; Sir John Denham (1615–69) 30,092; Charles Sackville, Earl of Dorset (1638–1706) 9,586; John Dryden (1631–1700) 18,238; Thomas D’Urfey (1653–1723) 18,757; Robert Gould (1660–1709) 29,110; Andrew Marvell (1621–78) 23,282; John Milton (1608–74) 18,924; John Oldham (1653–83) 32,462; Katherine Phillips (1631–64) 29,004; Matthew Prior (1664–1721) 32,000; Alexander Radcliffe (floruit 1669–96) 11,889; John Wilmot, Earl of Rochester (1648–80) 12,725; Sir Charles Sedley (1639–1701) 10,304; Elkanah Settle (1648–1724) 24,080; Thomas Shadwell (1642–92) 14,540; Jonathan Swift (1667–1745) 30,974; Nahum Tate (1652–1715) 20,333; Edmund Waller (1606–87) 16,443; Anne Wharton (1659–85) 12,511. Most of the corpus was prepared by John Burrows and Harold Love, assisted by Alexis Antonia and Meredith Sherlock. The Marvell subset was contributed by Christopher Wortham.
### Table 1 Specimen of procedure

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>O</th>
<th>P</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Main set</td>
<td>Milton</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>sum</td>
<td>31.489</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>stdv</td>
<td>0.770</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>World's Infancy</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>sum</td>
<td>31.489</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>World's Infancy</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>World's Infancy</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>World's Infancy</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>World's Infancy</td>
<td>count</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>mean (&quot;delta&quot;)</td>
<td>1.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>stdv</td>
<td>1.163</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
words, made up of 300 lines apiece from each of the twelve books—than to *The World’s Infancy* (1658), Nicholas Billingsley’s 11,111-word versification of the Book of Genesis? Columns A and B show the thirty most common words in descending order of their frequency in the main database. Column C shows their mean frequencies, all represented as percentages of that set, and Column D shows the corresponding standard deviations. Columns F, I, and N show the scores for the whole of Milton’s early verse, for our selection from *Paradise Lost*, and for *The World’s Infancy*, respectively, and Columns G, J, and O give z-scores representing their divergences from the means of the main set. The z-scores are used to obtain cognate figures for all the words in a hierarchy where the original frequencies fall away sharply from top to bottom. The object is to treat all of these words as markers of potentially equal power in highlighting the differences between one style and another. Columns K and P, respectively, show the differences between the z-scores for Milton and *Paradise Lost*, and those for Milton and *The World’s Infancy*.

The next step is to translate the positive and negative measures of difference shown in Columns K and P into absolute differences, as shown in Columns L and Q. By doing so, we obscure some useful stylistic information. But we are now able to derive meaningful totals and means for the whole range of differences. These are shown in L3 and Q3 and in L4 and Q4. A ‘delta-score’, as I propose to term entries like those in L4 and Q4, can be defined as ‘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’. (In the current inquiry, the text-groups are authorial. But that could be altered for tasks of non-authorial classification such as the differentiation of genre or era.) The delta-scores of 1.050 and 1.205 show that *Paradise Lost* is less unlike Milton than is *The World’s Infancy*. Thirty words, of course, are really too few for our purpose, especially when several of them are pronouns of volatile frequency. But even thirty words are enough to show why the differences we wish to add up and average out must be derived from z-scores and not from the original text-percentages. The text-percentages fall away so rapidly as the list extends downward that even sharp differences among lower-order words would be obliterated, in the total, by those from higher in the order.

Although this is a satisfactory outcome, success in a two-horse race does not promise success elsewhere. The addition of further specimens, the complete texts of *Paradise Regained* (15,694 words) and *Samson Agonistes* (12,885 words), each of which behaves as it should, is more encouraging. But the Delta procedure really begins to come into its own when it demonstrates that, although these three of Milton’s poems form no part of our Milton-set, they are less different from it than from any other of twenty-five authorial sets. Table 2 shows how the open, multi-author version of the procedure is used to test *Paradise Lost*.

If Columns A–C, where the output is recorded, are passed over for the moment, Table 2 begins like Table 1. Columns D–G show the upper range of the descending hierarchy of common words, standardized
272

8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42

1
2
3
4
5
6
7


X
Y
Z

Behn
Brome
Butler
Congreve
Cotton
Cowley
Denham
Dorset
Dryden
Durfey
Gould
Marvell
Milton
Oldham
Phillips
Prior
Radcliffe
Rochester
Sedley
Settle
Shadwell
Swift
Tate
Waller
Wharton

MAX
MIN
MEAN
STDEV

A

1.568
1.688
1.502
1.242
1.565
1.316
1.344
1.663
1.393
1.393
1.575
1.367
1.023
1.389
1.828
1.258
1.673
1.697
1.589
1.412
1.433
1.461
1.330
1.523
1.513

1.828
1.023
1.470
0.176
OUTPUT
delta-scores

B

0.560
1.238
0.181
–1.291
0.538
–0.874
–0.716
1.098
–0.435
–0.436
0.596
–0.584
–2.536
–0.456
2.033
–1.201
1.153
1.292
0.674
–0.330
–0.208
–0.049
–0.793
0.300
0.243

delta z-scores

C

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

D

F

G

the
and
of
a
to(i)
in(p)
his
with
to(p)
is
but
he
all
I
it
as
their
her
not
be
you
they
for(p)
by(p)
my
we
from
that(rp)
or
our
thy
was
this
when
are

4.242
3.770
1.821
1.601
1.419
1.358
1.154
1.022
1.014
0.938
0.923
0.803
0.781
0.766
0.766
0.710
0.641
0.623
0.616
0.586
0.580
0.564
0.559
0.555
0.512
0.510
0.500
0.476
0.471
0.460
0.451
0.437
0.426
0.426
0.413

0.630
0.501
0.315
0.430
0.272
0.189
0.323
0.208
0.131
0.312
0.195
0.241
0.193
0.391
0.239
0.224
0.237
0.336
0.174
0.167
0.252
0.234
0.114
0.106
0.370
0.275
0.127
0.228
0.165
0.268
0.247
0.140
0.095
0.105
0.134

DERIVED FROM DATABASE
Word
Mean
SD

E

Table 2 First page of 150-word worksheet (Paradise Lost as test-piece)

4.091
4.165
2.769
0.696
1.289
1.720
1.532
1.484
1.245
0.239
0.696
0.703
0.836
0.700
0.151
0.737
0.795
0.435
0.847
0.401
0.037
0.464
0.000
0.689
0.258
0.265
0.884
0.313
0.906
0.354
0.490
0.250
0.505
0.284
0.063

–0.239
0.789
3.015
–2.103
–0.480
1.916
1.171
2.224
1.761
–2.238
–1.167
–0.413
0.283
–0.171
–2.575
0.119
0.653
–0.560
1.324
–1.109
–2.154
–0.428
–4.905
1.260
–0.687
–0.891
3.019
–0.715
2.636
–0.397
0.158
–1.333
0.820
–1.355
–2.623

Paradise Lost
INPUT
Score
z-score

H
I
Test-piece
150

4.202
3.925
1.783
1.479
1.331
1.120
0.912
0.944
0.986
0.797
0.797
0.792
1.179
1.382
0.733
0.673
0.355
0.299
0.539
0.617
1.133
0.272
0.475
0.479
1.221
0.290
0.396
0.636
0.442
0.290
0.769
0.507
0.304
0.544
0.382

Score
–0.064
0.311
–0.121
–0.283
–0.323
–1.264
–0.747
–0.371
–0.217
–0.452
–0.648
–0.043
2.063
1.574
–0.138
–0.167
–1.206
–0.961
–0.441
0.188
2.196
–1.249
–0.739
–0.717
1.914
–0.800
–0.816
0.699
–0.175
–0.634
1.290
0.500
–1.283
1.117
–0.229

z-score

J
K
Behn
COUNT
SUM
MEAN
STDEV

0.175
0.478
3.136
1.819
0.157
3.181
1.918
2.595
1.978
1.786
0.519
0.370
1.780
1.745
2.437
0.286
1.859
0.402
1.765
1.297
4.350
0.820
4.167
1.977
2.600
0.091
3.835
1.414
2.811
0.236
1.132
1.833
2.103
2.472
2.394

Abs. diff.

150
235.2557
1.568
1.167

L

3.883
4.695
1.229
1.750
1.666
1.198
0.978
0.812
1.026
1.642
1.222
0.897
0.840
1.093
1.290
0.765
0.711
0.200
0.989
1.016
0.620
1.049
0.758
0.569
0.339
1.226
0.318
0.968
0.660
0.951
0.213
0.389
0.552
0.345
0.691

Score

–0.570
1.847
–1.883
0.347
0.908
–0.847
–0.543
–1.006
0.087
2.254
1.536
0.392
0.301
0.836
2.196
0.246
0.296
–1.258
2.135
2.579
0.157
2.071
1.753
0.128
–0.469
2.603
–1.430
2.155
1.145
1.835
–0.961
–0.340
1.316
–0.768
2.078

z-score

M
N
Brome
COUNT
SUM
MEAN
STDEV

0.331
1.058
4.898
2.450
1.387
2.763
1.713
3.229
1.673
4.492
2.703
0.805
0.019
1.007
4.771
0.128
0.357
0.698
0.811
3.688
2.311
2.499
6.658
1.131
0.218
3.494
4.449
2.870
1.490
2.233
1.119
0.993
0.496
0.587
4.701

Abs. diff.

150
253.177
1.688
1.467

O

Lit&Ling 17/3 267-287 Final 12/2/03 12:29 pm Page 272

J. Burrows


means for the frequency of each word in our main database, and the corresponding standard deviations. (In its complete form, the table includes the 150 most common words, ranging down to those that occur about once in every thousand in the main database. The words tagged so as to distinguish homographic forms are accompanied by parenthetic abbreviations: (i) for infinitive, (p) for preposition, (rp) for relative pronoun, and (c) for conjunction.) Columns H and I now provide a site for pasting-in the scores for any chosen test-piece and for the z-scores derived by setting those scores against the means and standard deviations given in Columns F and G. In Columns J–L and M–O, respectively, we have the entries for the first two of our twenty-five authorial sets. (In its complete form, the table continues until it includes them all. The vast arithmetical power of spreadsheets such as Microsoft Excel also allows room for other sets to be added as desired.)

Columns J–L (and each of the corresponding trios that follow) give the standardized score for each word in a particular authorial set, the corresponding z-score, and the absolute difference between each z-score and that of the test-piece. Cells L2–L5, O2–O5 (and those corresponding to them in the pages not shown) sum up the columns beneath, giving a count of the number of entries and the sum, mean, and standard deviation of those entries. As in Table 1, these means are our delta-scores. But, refining on Table 1, they are auto-copied across to Column B where each is listed beside the name of the appropriate author. (The entries marked X, Y, and Z at the foot of Column B allow for the addition of other authors or for the testing of control-sets.)

Cell B3 shows the minimum entry in Column B. Glancing down, we see, in Cell B20, that this is Milton’s delta-score. On this test, then, Paradise Lost differs less from our Milton-set than from any other of our twenty-five authorial sets. With a delta-score of 1.023 on the word-list of 150 (which may be expressed as $\Delta_{150} = 1.023$), Milton therefore has the best claim of these twenty-five poets to the authorship of Paradise Lost. The strength of the result shown in Cell B20 is reinforced in Cell C20, by far the lowest of a fresh set of ‘delta z-scores’ derived from the delta-scores in Column B. Milton’s ‘delta z-score’ ($z_{150} = -2.536$) diverges so far from the other twenty-four that only one case in a hundred of a normal population could be expected to equal or exceed it. Although it would be unwise to assume that twenty-five sizeable authorial sets constitute a fair sample of seventeenth-century English poetry, there is no obvious reason to insist that they do not. The outcome of our further trials is relevant.

By repeatedly entering new sets of scores in Column H, we can apply the procedure to as many test-pieces as we wish, each test being independent of the rest. The companion paper, ‘Questions of authorship’, presents a first group of results. These were obtained by applying the Delta procedure (using the scores for all of the 150 most common words) to sixteen long poems by members of our set of twenty-five poets and to another sixteen by poets from beyond that set. Thirty of the thirty-two long poems behaved as they should. Fifteen of the first sixteen attached
themselves firmly to their true authors. Fifteen of the second sixteen showed only weak affinities for authors within the main set, suggesting that their true authors should be sought elsewhere.

The present paper rests upon an extension of the inquiry. It treats of the results obtained by applying the Delta procedure to 200 poems by members of our set of twenty-five poets. The word-lists used fall into a series ranging from the 150 most common down through 120, 100, 80, and 60, to the 40 most common. (The corresponding analysis of further poems by non-members of the set has not yet been undertaken. But the procedure cannot be employed effectively for this further purpose unless the analysis of members of the set is operating at a high level of success. That tends to restrict its usefulness with putative non-members to longer texts where, as will be seen, the success rate for members of the set rises above 90 per cent.)

The choice of the 200 poems was less carefully designed than one might, with hindsight, have desired. The obvious first step was to examine some much shorter poems than those mentioned above and to include enough of them to reduce the risk of sampling errors. The results were favourable enough to encourage a long series of increments and extensions in which more poems by more authors were gradually added. In terms of their length, the 200 poems fall into five bands. A hundred of them are of 500 words or fewer. (Thirty-three of these range from 103 words to 250 and sixty-seven from 251 to 500). The next forty range up from 500 to 1,000 words (twenty of them lying on either side of 750). The remaining sixty fall into three groups of twenty, ranging upward in successive bands to 1,500, 2,000, and beyond. (The last set of twenty long poems includes the sixteen of the companion paper.)

In the matter of authorship, the process of selection was a struggle against the (admittedly fortunate) constraint that poets do not write to our dictates. With some members of the set, no further authentic pieces were to be had. With others there was a dearth of short poems or of long ones. With some there was an unavoidable sameness, with others an extreme diversity. The final group of 200 comprised between twelve and nineteen texts apiece by twelve of our twenty-five poets and nine more by three others. Except for a different selection from Hudibras, none of them, in whole or part, figured in the original database against which they were to be tested. All but a dozen were whole poems. To round out the set of twenty in the 1,500 band, there are two separate Cantos from Cotton’s Voyage to Ireland in Burlesque and two selections from Paradise Lost. The latter pair replace the one large selection used at the beginning of this paper.

The outcome of the whole battery of tests is summarized in Table 3. (The results are also set out, poem by poem, in the Appendix.) In the successive columns making up the left-hand block of the table, the poems are sorted into five bands according to their length. Each cell in this block shows the number of poems whose true authors attained a given rank (out of twenty-five). Thus, in the first row of data, the true author ranked first out of twenty-five for twenty-seven of the 100 poems ranging up to

---

5 Many of the word-counts derive from texts in the excellent Chadwyck–Healey archive of English poetry, to which my university subscribes. The texts are not used in any other way.
A Measure of Stylistic Difference and Likely Authorship

500 words in length; for eighteen of the forty poems ranging from 501 to 1,000 words; for thirteen of the twenty poems ranging from 1,001 to 1,500 words; and so on up to nineteen of the twenty poems of more than 2,000 words. The corresponding columns of the right-hand block express these figures as percentages.

Table 3 Two hundred poems of the late seventeenth century: summary of results of delta-tests

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>150 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>27</td>
<td>18</td>
<td>13</td>
<td>17</td>
<td>94</td>
<td>27.0</td>
<td>45.0</td>
<td>65.0</td>
<td>85.0</td>
<td>95.0</td>
</tr>
<tr>
<td>1st–2nd</td>
<td>40</td>
<td>26</td>
<td>14</td>
<td>18</td>
<td>20</td>
<td>118</td>
<td>40.0</td>
<td>65.0</td>
<td>70.0</td>
<td>90.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>67</td>
<td>32</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>157</td>
<td>67.0</td>
<td>80.0</td>
<td>90.0</td>
<td>100</td>
</tr>
<tr>
<td>06 : 10</td>
<td>15</td>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>15.0</td>
<td>20.0</td>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>120 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>31</td>
<td>16</td>
<td>10</td>
<td>15</td>
<td>19</td>
<td>91</td>
<td>31.0</td>
<td>40.0</td>
<td>50.0</td>
<td>75.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>62</td>
<td>30</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>150</td>
<td>62.0</td>
<td>73.0</td>
<td>90.0</td>
<td>100</td>
</tr>
<tr>
<td>06 : 10</td>
<td>17</td>
<td>9</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>28</td>
<td>17.0</td>
<td>22.5</td>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>13</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>13.0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>100 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>25</td>
<td>17</td>
<td>8</td>
<td>15</td>
<td>16</td>
<td>81</td>
<td>25.0</td>
<td>42.5</td>
<td>40.0</td>
<td>75.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>59</td>
<td>29</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>145</td>
<td>59.0</td>
<td>72.5</td>
<td>90.0</td>
<td>95.0</td>
</tr>
<tr>
<td>06 : 10</td>
<td>18</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>27</td>
<td>18.0</td>
<td>17.5</td>
<td>5.0</td>
<td>5.0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>10.0</td>
<td>10.0</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>11.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>80 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>26</td>
<td>12</td>
<td>9</td>
<td>15</td>
<td>16</td>
<td>78</td>
<td>26.0</td>
<td>30.0</td>
<td>45.0</td>
<td>75.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>66</td>
<td>29</td>
<td>17</td>
<td>19</td>
<td>20</td>
<td>151</td>
<td>66.0</td>
<td>72.5</td>
<td>85.0</td>
<td>95.0</td>
</tr>
<tr>
<td>06 : 10</td>
<td>13</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>13.0</td>
<td>17.5</td>
<td>10.0</td>
<td>5.0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>14.0</td>
<td>5.0</td>
<td>5.0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>4.0</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>60 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>28</td>
<td>7</td>
<td>5</td>
<td>12</td>
<td>17</td>
<td>69</td>
<td>28.0</td>
<td>17.5</td>
<td>25.0</td>
<td>60.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>63</td>
<td>26</td>
<td>17</td>
<td>18</td>
<td>20</td>
<td>144</td>
<td>63.0</td>
<td>65.0</td>
<td>85.0</td>
<td>90.0</td>
</tr>
<tr>
<td>06 : 10</td>
<td>22</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>35</td>
<td>22.0</td>
<td>20.0</td>
<td>15.0</td>
<td>10.0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>9</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
<td>9.0</td>
<td>12.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>5.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1.0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>40 words</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st</td>
<td>23</td>
<td>10</td>
<td>6</td>
<td>8</td>
<td>17</td>
<td>64</td>
<td>23.0</td>
<td>25.0</td>
<td>30.0</td>
<td>40.0</td>
</tr>
<tr>
<td>1st–5th</td>
<td>57</td>
<td>29</td>
<td>14</td>
<td>20</td>
<td>19</td>
<td>139</td>
<td>57.0</td>
<td>72.5</td>
<td>70.0</td>
<td>100</td>
</tr>
<tr>
<td>06 : 10</td>
<td>17</td>
<td>8</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>30</td>
<td>17.0</td>
<td>20.0</td>
<td>20.0</td>
<td>0</td>
</tr>
<tr>
<td>11 : 15</td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>17</td>
<td>13.0</td>
<td>5.0</td>
<td>10.0</td>
<td>0</td>
</tr>
<tr>
<td>16 : 20</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>11</td>
<td>11.0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 : 25</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2.0</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Literary and Linguistic Computing, Vol. 17, No. 3, 2002 275
Studied from head to foot, the table treats the successive word-lists in descending order from 150 to forty. In each set, the table shows the number of poems for which the true author ranked first, first–fifth, sixth–tenth, and so on. In the top set, the number for which the true author ranked first–second is also given.

The most general results lie in the last column of all. Of the whole 200 poems, a vast miscellany, 47 per cent attach themselves to their true authors when the full word-list of 150 is employed. For 59 per cent of them, the true author ranks either first or second out of twenty-five. For almost 79 per cent, the true author ranks among the first five and for 12.5 per cent among the next five. In only 4 per cent does the true author rank below fifteenth. When the word-list of 120 is employed, the results are a little weaker at almost every point. The lower parts of the table show that continued truncation of the word-list produces a continued deterioration in the results. (Trials in which the word-list was truncated from the top instead of the bottom yielded worse results than any of those shown. The words ranking from 76 to 150, for example, yielded some clean hits but many wild misses.)

If we shift the perspective and weigh up the poems in increasing order of length by moving horizontally across the table, the strongest results of all are for the lists of 150 and 120 on the twenty poems of more than 2,000 words. With nineteen of them, the true author ranks first of twenty-five. With the solitary exception, **The Hind and the Panther**, the true author (John Dryden) ranks second. Although the shorter word-lists yield weaker results than these (save for a minor aberration in the list of 40), the progressive lengthening of the poems always makes for increasing accuracy.

A study of so many specimens justifies some strong conclusions. The overall level of success is impressive because the task of identifying the right candidate from a group of twenty-five offers a much higher level of difficulty than the two- or three-author tasks to which we are accustomed. The unfettered operation of chance would lead, after all, to a roughly equal spread over the several ranks. Two hundred trials would yield only eight cases, not ninety-four, in which the true author ranked first out of twenty-five. In forty cases, not 157, the true author would rank between first and fifth. And in forty cases, not three, the true author would rank between twenty-first and twenty-fifth. Only a genuine authorial factor could yield results like those we have.

It is evident that, with texts of 1,500 words or more, the Delta procedure is effective enough to serve as a direct guide to likely authorship. With problem texts of that length, the procedure has much to offer, especially when an extensive word-list is employed. Even though corroborative evidence would usually be sought, that makes it a worthwhile addition to our scholarly armoury.

It is also evident that, even with much shorter texts, the Delta procedure is useful in two less direct ways. It makes a basis for selecting a likely group of candidates. The word-lists of 100, 80, and 60 all include the true author among the top five candidates for 85 per cent or more of texts
A Measure of Stylistic Difference and Likely Authorship

of above 1,000 words. That supports the use of such Delta trials as a prelude to tests in the ‘closed forms’ to which I have previously referred. Even more usefully, perhaps, with shorter texts, the Delta procedure helps (as the old song says) to ‘ee-lim-inate the negative’. If a putative author does not rank within the top ten of twenty-five candidates, one might demand extremely strong external evidence in his or her favour before discountenancing the doubt so cast. We are thus offered useful negative evidence for cases where it is appropriate. This negative evidence also offers strong general support for the ancient but no longer unchallenged belief that the concept of authorial signatures is well-founded.

How are we to choose the most useful from among these word-lists? Even with the shortest poems, the list of forty yields the least accurate results and should therefore be abandoned. For all groups except the shortest, where some small irregularities appear, the pattern of results improves with each extension of the list. Now, as was noted above, large sets of variables usually yield the most accurate stylistic signatures, possibly as a reflection of the wealth of information they incorporate. But one would scarcely have expected that principle to extend to the point where the list of 150 words still yields some of the best results for poems of as little as 100 words in length.

To move beyond this point, it is necessary to go behind Table 3 and inspect the full authorial hierarchies for the various poems. As the data derived from very short poems become too sparse to be reliable, a statistical artefact intervenes. Of these twenty-five poets, John Milton has the most constrained and strongly delineated stylistic repertoire and, accordingly, moves furthest from the common patterns of the language. Far more often than any of the others, his scores diverge below the norms derived from our main database. (Of the top 100 words, he lies below the norm for seventy-two!) The fact that this makes for many negative $z$-scores is obliterated when we treat all the scores as absolute. But the fact that many of them are strong divergences produces an unusual pattern of absolute differences. That pattern, as it happens, coincides with one of a very different origin. With very short poems, many of our most common words do not occur at all. Here, again, therefore, the scores often diverge below the norm. Here, again, the divergences yield an unusual pattern—a pattern not unlike Milton’s. When the list of forty words is applied to the thirty-three shortest poems, which range in length from 103 to 250 words, Milton’s performance is entirely unremarkable. Although he is not the author of any of them, he ranks first for one and between first and fifth for seven. But when the list of 100 words is applied to the same thirty-three texts, he ranks first for seven and between first and fifth for fourteen. With the shortest poem of all, a little song of Congreve’s running to only 103 words, the list of forty words puts Rochester at the head of the field. But Milton, who does not rank among the top five candidates, sweeps into first place when the list of 100 is employed. Milton’s spurious claim to the authorship of the shortest poems grows even more pervasive when the lists are extended to 120 and 150.
The tests also give Rochester many short poems to whose authorship he has no claim. In this case, however, the progressive extension of the wordlist works differently. With the list of forty words, Rochester ranks first for eight of these thirty-three short poems and between first and fifth for nineteen. With the list of 100 words, he ranks first for six and, once more, between first and fifth for nineteen. The reason this time is not that the scores are unusual. It is rather that they are so consistently normal—so characteristic of the period—that they are set in low relief. So long as there is a sparsity of information, this pattern of low relief allows many short poems to show a statistical affinity for Rochester. But the longest word-lists, with their richer information, put paid to this false effect. The better delineated frequency-profiles of the longer texts, likewise, are not vulnerable in this way.

Given the weight of misleading influences like these, we must consider how it is possible for the procedure to yield so many accurate results for the thirty-three shortest poems. Even with the list of forty, the true authors rank first for six poems and between first and fifth for seventeen, a score outmatched only by Rochester. With the list of 100, the true authors rank first for eight poems and between first and fifth for twenty-two, outmatching both Rochester and Milton. No other poet of the twenty-five surpasses the constraints of chance.

A poem of only 209 words, To Alexis, On his saying, I lov’d a Man that talk’d much, is correctly assigned to Aphra Behn by all six of our wordlists. The poem’s 209 word-tokens represent 120 word-types. Of these, only sixty-five word-types and 136 word-tokens lie within the ambit of our 150 most common words. Eighty-five spaces in the hierarchy are not occupied and a further forty spaces each contain a solitary member. Such a list as this does not constitute a frequency-profile in the usual sense of the term. And yet, so the result implies, it best matches the frequency-profile for Aphra Behn’s authorial subset in the main database. The absolute z-scores on which the Delta procedure operates have the effect, it seems, of presenting the lower part of this pattern of divergences in almost binary terms. Provided a sufficient number of the occupied spaces in the poem’s profile match those where Behn’s authorial frequency-profile shows positive divergences and provided a sufficient number of the blank spaces match those where she diverges below the norm, she must (and does) emerge as the top-ranking candidate. This interpretation of the effect complies with the rather hit-or-miss character of the results obtained from the shorter word-lists: a binary hierarchy lacks the subtlety of a true frequency-profile and can easily go amiss.

A more detailed scrutiny of the results reveals some of the limitations of the Delta procedure. The poem-by-poem record shown in the Appendix makes it clear that this is far from a level playing field. The procedure yields a very high success rate with Samuel Butler, whose poems are longer than most, and Katherine Phillips (Orinda), whose style is both idiosyncratic and extremely homogeneous. Many of the worst results can be attributed to the fact that, in one way or another, a given poem or group of poems is uncharacteristic of its author. The three love
poems that open Oldham’s set, the elegy on Katharine Kingscote, and the epistle to Madam L. E. are all remote from the strong vein of satire for which he is best known and which rightly predominates in his authorial subset. The weak results for Robert Gould’s short poems reflect the regrettable fact that his authorial subset is a badly skewed sample of his work. Knowing him only for his long satires, I did not include any of his very diverse short pieces in the original database. The weak results that open Cowley’s set are the product of a conscious and (unduly?) successful experiment. The first five of his nineteen poems are the work of his youth, written as much as half a century before the poems for which he is best known. My experience has been that, although authors’ styles change over the years, they do not change beyond recognition. This more usual situation is represented here by Waller’s Of the Danger his Majesty . . . Escaped, which is also a work from the beginning of a long career. But in Cowley’s case, the signs of change are so strong as to affect our overall result. For a fine specimen of ‘stylochronometry’, see Forsyth (1999).

Apart from the very short poems considered earlier, those cases where the procedure gives poor results on poems of more than 500 words represent bad matches—recognizably bad matches—between specimen and authorial subset. They mostly arise from a difficulty encountered by everyone who works in computational stylistics—the fact that authors work at times in very uncharacteristic literary genres. Whereas procedures such as principal component analysis can often overcome the aberrations that arise in this way by absorbing them in the lesser vectors of their output, the single vector of the Delta procedure has no such cushion. With the Delta procedure, an aberrant set of scores is expressed as a lower ranking for the author in question. This being so, it is surprising that the procedure is as resilient as the results demonstrate.

In cases where we are testing the claims of acknowledged candidates, it would usually be possible to foresee this genre-difficulty through our knowledge of their writings. A misleading ranking could be given its due and further tests could be undertaken. But if the Delta procedure is to be of real value in open cases, another answer must be found. Table 4 takes us back to the difference between ‘inside’ and ‘outside’ candidates and suggests an answer to this difficulty.

The companion paper, ‘Questions of authorship’, contrasts the results of the Delta procedure for thirty-two long poems, sixteen by members of our set of twenty-five and sixteen by outsiders. It shows that, when the 150 word-list is employed, the true authors rank first for fifteen of the sixteen poems by insiders and that, in the sixteenth case, the true author ranks second. The sixteen poems by outsiders must, by definition, be ‘least unlike’ some member of the set of twenty-five authors. They have nowhere else to go. But ‘least unlike’ need not be ‘much like’. The average delta-score for these sixteen is 1.332 as opposed to 1.097 for the sixteen ‘insiders’.

Table 4 takes eight poems from each set to reach inside these averages and show how insiders differ from outsiders. (The insiders occupy the
Table 4  Sixteen long poems: summary of results

<table>
<thead>
<tr>
<th></th>
<th>Cowley</th>
<th>Waller Instructions to Painter</th>
<th>Dryden Absal. and Achit.</th>
<th>Dryden Hind and Panther</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Davidides</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6812 (sel.) delta</td>
<td>2606 delta</td>
<td>7824 delta</td>
<td>19896 delta</td>
</tr>
<tr>
<td>List score</td>
<td>1.006 1.274</td>
<td>0.993 2.136</td>
<td>0.861 1.625</td>
<td></td>
</tr>
<tr>
<td>z-score LU score</td>
<td>1.006 2.209</td>
<td>2.209 2.136</td>
<td>2.209 2.136</td>
<td></td>
</tr>
<tr>
<td>List score</td>
<td>1.024 1.393</td>
<td>0.998 2.119</td>
<td>0.858 1.798</td>
<td></td>
</tr>
<tr>
<td>z-score LU score</td>
<td>2.486 1.910</td>
<td>2.119 2.115</td>
<td>1.798 1.884</td>
<td></td>
</tr>
<tr>
<td>List score</td>
<td>1.030 1.456</td>
<td>0.965 2.115</td>
<td>0.843 1.884</td>
<td></td>
</tr>
<tr>
<td>z-score LU score</td>
<td>2.195 1.584</td>
<td>2.115 2.115</td>
<td>1.884 1.884</td>
<td></td>
</tr>
<tr>
<td>List score</td>
<td>0.934 1.400</td>
<td>0.966 1.876</td>
<td>0.886 1.853</td>
<td></td>
</tr>
<tr>
<td>z-score LU score</td>
<td>2.137 1.643</td>
<td>1.876 1.876</td>
<td>1.853 1.853</td>
<td></td>
</tr>
<tr>
<td>List score</td>
<td>0.938 1.459</td>
<td>0.966 1.643</td>
<td>0.815 1.683</td>
<td></td>
</tr>
<tr>
<td>z-score LU score</td>
<td>1.932 1.559</td>
<td>1.643 1.643</td>
<td>1.683 1.683</td>
<td></td>
</tr>
</tbody>
</table>

|                  | Oldham Satyr 2   | Gould The Presbytery | Durfrey The Malecontent | Swift Verses on the Death |
|                  | 2210 delta       | 4492 delta           | 7817 delta              | 3206 delta               |
| List score       | 1.215 1.024      | 1.204 1.024          | 1.284 2.777             |                         |
| z-score LU score | 3.600 3.002      | 3.003 3.002          | 2.777 2.777             |                         |
| List score       | 1.237 1.203      | 1.207 1.207          | 1.298 2.662             |                         |
| z-score LU score | 2.553 1.911      | 1.568 1.568          | 2.662 2.662             |                         |
| List score       | 1.206 1.082      | 1.704 1.704          | 1.351 2.478             |                         |
| z-score LU score | 2.378 1.704      | 2.573 2.573          | 2.478 2.478             |                         |
| List score       | 1.170 1.052      | 1.418 1.418          | 1.266 2.245             |                         |
| z-score LU score | 2.544 1.418      | 2.491 2.491          | 2.245 2.245             |                         |

|                  | Fletcher Purple Island | Davenant Gundihert | Wild Iter Boreale | Wase Divination |
|                  | 5933 (sel.) delta      | 5167 (sel.) delta  | 3321 delta        | 2156 delta       |
| List score       | 1.210 1.302           | 1.324 1.324        | 1.362 1.362       | 1.374 1.374     |
| z-score LU score | 1.822 2.185           | 1.558 1.558        | 1.322 1.322       | 1.222 1.222    |
| List score       | 1.205 1.306           | 1.314 1.314        | 1.347 1.347       | 1.368 1.368     |
| z-score LU score | 1.720 1.786           | 1.788 1.788        | 1.619 1.619       | 1.368 1.368    |
| List score       | 1.218 1.315           | 1.297 1.297        | 1.293 1.293       | 1.368 1.368     |
| z-score LU score | 1.733 1.505           | 2.024 2.024        | 1.368 1.368       | 1.368 1.368    |
| List score       | 1.202 1.317           | 1.246 1.246        | 1.286 1.286       | 1.549 1.549     |
| z-score LU score | 1.513 1.508           | 1.855 1.855        | 1.549 1.549       | 1.549 1.549    |
| List score       | 1.172 1.263           | 1.142 1.142        | 1.374 1.374       | 1.222 1.222     |
| z-score LU score | 1.659 2.033           | 2.235 2.235        | 1.222 1.222       | 1.222 1.222    |

|                  | Pordage Heyrick Duke | Blackmore King Arthur |
|                  | The Medal Revers’d | The New Atlantis Paris to Helena | 6986 (sel.) delta |
|                  | 3103 delta          | 8797 (sel.) delta      | 3892 delta       |
| List score       | 1.542 1.162         | 1.279 1.665           | 1.399 1.526      |
| z-score LU score | 1.799 1.707         | 1.665 1.665           | 1.526 1.526      |
| List score       | 1.466 1.183         | 1.260 1.950           | 1.495 1.517      |
| z-score LU score | 2.114 1.694         | 1.950 1.950           | 1.517 1.517      |
| List score       | 1.413 1.210         | 1.269 1.882           | 1.589 1.487      |
| z-score LU score | 2.491 1.622         | 1.882 1.882           | 1.487 1.487      |
| List score       | 1.446 1.230         | 1.294 1.936           | 1.636 1.312      |
| z-score LU score | 2.503 1.438         | 1.936 1.936           | 1.312 1.312      |
| List score       | 1.466 1.119         | 1.160 2.057           | 1.603 1.330      |
| z-score LU score | 2.201 1.542         | 2.057 2.057           | 1.330 1.330      |

A = Waller A = Dryden A = Swift
B = Cowley B = Durfrey B = Swift 2nd
C = Dryden 2nd

A Measure of Stylistic Difference and Likely Authorship

top half of the page.) The table includes the two cases, Dryden’s *The Hind and the Panther* and Davenant’s *Gondibert*, that give most difficulty. For each poem, the table shows the delta-score on each of the top five word-lists, the corresponding delta z-score from a set of twenty-five, and in the column headed ‘LU’ (for ‘least unlike’), a set of codes for the top-ranking candidates. Cowley’s *Davideis*, the first entry, shows delta-scores approximating to 1.00, delta z-scores ranging down to almost −2.5, and a consistent top-ranking for Cowley.

A close study of Table 4 shows several contrasting tendencies, none of which is absolute, in the two sets of eight poems. The delta-scores for the insiders run lower than those for the outsiders. The delta z-scores run much more strongly into the negative. (Their special value is to offset the fact that delta-scores rise rapidly for shorter texts. The delta-scores shown here, averaging 1.075 for insiders and 1.318 for outsiders, are consistent with the length of these poems.) For the insiders, the lowest delta-scores and, accordingly, the most strongly negative of the delta z-scores tend to emerge from the longer word-lists, where the information used is richer. And the insiders tend to sit consistently with a single candidate. When all of these tendencies are weighed up, none of the poems by ‘outsiders’ can easily be taken for the work of any member of the set of twenty-five poets. In the other set, the overall case for each of the true authors is impressive for all but *The Hind and the Panther*. The particular reasons why that poem breaks the pattern are examined in ‘Questions of authorship’ and a second round of testing rectifies the first.

‘Questions of authorship’ includes a bar-graph in which the delta z-scores for my original set of thirty-two long poems on the 150 word-list shows that a threshold of −1.9 neatly separates the two sub-sets of sixteen poems save for the two exceptions I have mentioned. (In a normal population, a z-score of −1.9 separates around 3 per cent of cases from the remainder. Two exceptions out of thirty-two are by no means unexpected.) The possibility of using just such a bar-graph as a grid for testing specimens of unknown authorship was not to be ignored.

But the addition of only four more long poems by ‘insiders’ to round out the new group of twenty poems of above 2,000 words yielded another exception. John Oldham’s pompous 2,237-word eulogy *Upon the Works of Ben. Johnson* does register as his on the longer word-lists and, after a fleeting affinity with Dryden, returns to him on the shorter word-lists. But the delta z-score for the list of 150 is only −1.679, far below the threshold of −1.9. A poem that opens with the invocation ‘Great Thou! Whom it is a crime almost to dare to praise’ is likely to differ in many ways from the characteristic work of this harsh and argumentative satirist. Although it is reassuring to find that, even here, the Delta procedure does identify the true author, the delta z-score is a sharp reminder that the system for distinguishing between insiders and outsiders is not foolproof. It behoves us, as always, to remember that, by relying on statistical analysis, even in this simple form, we are dealing in probabilities and not in absolutes.

With this necessary proviso and without forgetting the other limitations we have observed, it is clear that the Delta procedure satisfies the
purposes enunciated at the beginning of this paper. Even with very short
texts, it is more successful than might have been expected at the ‘open
game’ of picking out the most likely set of candidates from a large group.
By comparison with the results described (my own among them) in the
extensive trials of Forsyth and Holmes (1996), it is extremely accurate in
singling out the true author of texts of more than 1,500 words in length.
(That claim is strengthened by the fact that it is much more difficult to
identify the true author in a field of twenty-five candidates than in a
comparison of two candidates or three.) And it shows promise as a means
of indicating that the true author of a given text may lie beyond a current
set of candidates, a task we have not hitherto accomplished. How is it that
such a primitive statistical instrument can satisfy these purposes? The
answer must lie, I believe, in areas where we are still extremely ignorant—
in the communicative resilience of the language and the astonishing force
of human individuality.

Acknowledgements
The research on which this paper is based has been generously supported
by the Australian Research Council and the University of Newcastle. I am
also indebted to Harold Love of Monash University and to my colleagues
in the Centre for Literary and Linguistic Computing at Newcastle.

References
Knowles, F. E., and Smith, J. (eds), Advances in Computer-aided Literary
and Linguistic Research: Proceedings of the Fifth International Symposium on
Computers in Literary and Linguistic Research. Birmingham: John Goodman,
pp. 1–15.
Binongo, J. N. G. and Smith, M. W. A. (1999). The application of principal com-
seventeenth-century poems: a computational approach. The Seventeenth
Century, 16: 259–82.
Forsyth, R. S. (1999). Stylochronometry with substrings, or: a poet young and
Christiana, attributed to John Milton: a statistical investigation. Literary and

## Appendix

### Table A1  Two hundred long poems of the late seventeenth century: list of texts with delta-ranks

<table>
<thead>
<tr>
<th>Word-list</th>
<th>150</th>
<th>120</th>
<th>100</th>
<th>80</th>
<th>60</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Author's rank (ex 25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Behn
- a 248 1 1 3 3 3 3 3 A Pindaric to Mr. P. who sings finely
- b 500 7 3 7 5 5 5 On the Author of that Excellent Book ... The Way to Health [etc.]
- c 559 7 5 10 10 11 8 To the Honorable Sir Francis Fane, on his Excellent Play, The Sacrifice
- d 263 12 15 14 15 12 17 A Satyr on Doctor Dryden
- e 295 12 10 20 22 19 24 To Alexis in Answer to his Poem against Fruation. Ode
- f 209 2 1 1 1 1 1 To Alexis, On his saying, I lov'd a Man that talk'd much
- g 273 16 11 9 11 9 15 To Amintas, Upon reading the Lives of some of the Romans
- h 416 6 3 3 6 4 3 On the first discovery of falseness in Amintas
- i 454 3 2 3 3 4 5 On the death of E. Waller, Esq.
- j 161 3 3 3 2 4 4 Verses design'd by Mrs. A. Behn, to be sent to a fair lady ... Left unfinish'd
- k 199 5 2 1 1 1 5 On a Pin that hurt Aminta’s Eye
- l 282 11 3 4 4 7 3 A Letter to the Earl of Kiddare, dissuading him from marrying Moll Howard
- m 818 1 2 1 2 2 2 On Desire. A Pindarick
- n 761 7 6 6 3 2 1 A Pindaric Poem to the Reverend Doctor Burnet
- o 16419 1 1 2 2 3 2 A Voyage to the Island of Love

#### Brome
- a 379 1 1 1 1 1 1 The Answer (Stay, stay, prate no more)
- b 437 1 1 1 1 1 1 The Leveller
- c 366 1 1 1 1 1 1 The Polititian
- d 428 1 1 1 1 1 1 On Sir G. B., his defeat
- e 453 3 4 4 3 4 4 On a Butcher’s Dog
- f 325 1 2 2 1 1 1 Palinode
- g 436 1 3 9 8 7 10 To a Potting Priest upon a quarrel
- h 364 5 6 7 5 5 4 To the Meritoriously Honorable Lord Chief Justice of the Kings bench
- i 479 7 10 16 8 1 13 Upon the miscarry of Letters betwixt his Friend and him
- j 376 2 2 3 1 2 1 To his friend Mr. I. W. on his translation of a romance
- k 241 1 1 1 1 2 7 Upon the Kings imprisonment
- l 693 1 1 1 1 2 5 The Answer (Did I not know thee friend)
- m 645 4 9 11 6 6 9 Upon the Death of that Reverend and learned Divine, Mr. Josias Shute
- n 761 4 4 7 11 14 1 The Satyr of Money
- o 767 2 3 3 3 4 9 To C. S. Esquire (Dear Charles, I am thus far come)
- p 1385 2 2 2 2 2 2 The Answer (My Friend, in troth, I am glad to hear)

#### Butler
- a 1217 1 1 1 1 2 1 Satyr upon Plagiaries
- b 1401 1 1 1 3 2 2 Satyr upon the Weakness and Misery of Man
- c 1528 1 1 1 1 2 3 Satyr upon the Licentious Age of King Charles the 2d
- d 1720 1 1 1 1 1 2 1 Upon a Hypocritical Nonconformist
- e 1950 1 1 1 1 1 1 Satyr upon the Imperfections and Abuse of Human Learning, Pts i and ii, 1–72
- f 1621 1 1 1 1 2 1 Hudsbris, the third part, Canto ii, 1–266
- g 615 1 1 1 1 1 1 Satyr upon Gaming
- h 604 1 1 1 1 4 6 A Panegyric upon Sir John Denham’s Recovery from his Madness
- i 705 1 1 3 4 6 3 Satyr upon Drunkenness
- j 962 1 1 1 1 1 2 Satyr upon Marriage
- k 813 1 1 1 1 1 1 Upon Philip Nye’s Thanksgiving Beard
- l 844 1 1 1 1 2 1 Upon Modern Critics
Table A1 (continued)

<table>
<thead>
<tr>
<th>Length</th>
<th>Author’s rank (ex 25)</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>120 100 80 60 40</td>
</tr>
</tbody>
</table>

**Congreve**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Author’s rank</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>203</td>
<td>2 2 4 2 2</td>
<td>Paraphrase upon Horace. Ode xix., Lib. 1</td>
</tr>
<tr>
<td>b</td>
<td>247</td>
<td>1 1 1 1 2</td>
<td>The Reconciliation</td>
</tr>
<tr>
<td>c</td>
<td>313</td>
<td>17 18 16 4 10 12</td>
<td>To Mr. Dryden on his Translation of Persius</td>
</tr>
<tr>
<td>d</td>
<td>260</td>
<td>11 12 9 12 6 15</td>
<td>Prologue to Pyrrhus, King of Epirus</td>
</tr>
<tr>
<td>e</td>
<td>400</td>
<td>2 2 5 3 3 1</td>
<td>Doris</td>
</tr>
<tr>
<td>f</td>
<td>935</td>
<td>2 2 4 5 11 14</td>
<td>Of Pleasing. An Epistle to Sir Richard Temple</td>
</tr>
<tr>
<td>g</td>
<td>133</td>
<td>2 2 2 3 6 8</td>
<td>Song (As Amoret and Thyris, lay)</td>
</tr>
<tr>
<td>h</td>
<td>337</td>
<td>3 4 8 9 7 6</td>
<td>Buxom Joan of Lynnas’s Love for a jolly Sailer</td>
</tr>
<tr>
<td>i</td>
<td>103</td>
<td>3 3 3 4 4 5</td>
<td>The Decay, a Song</td>
</tr>
<tr>
<td>j</td>
<td>321</td>
<td>2 2 3 3 4 2</td>
<td>A Poem in Praise of the Author</td>
</tr>
<tr>
<td>k</td>
<td>403</td>
<td>5 7 11 10 10 16</td>
<td>Prologue, written for Mr. Haynes</td>
</tr>
<tr>
<td>l</td>
<td>617</td>
<td>3 3 2 3 2 2</td>
<td>A Letter from Mr. Congreve to the Lord Viscount Cobham</td>
</tr>
<tr>
<td>m</td>
<td>1349</td>
<td>2 2 2 1 1 2</td>
<td>The Tears of Amaryllis and Amyntas, a pastoral</td>
</tr>
<tr>
<td>n</td>
<td>1376</td>
<td>1 1 1 1 2 2</td>
<td>An Impossible Thing</td>
</tr>
</tbody>
</table>

**Cotton**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Author’s rank</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>174</td>
<td>5 11 7 5 10 13</td>
<td>Elegy</td>
</tr>
<tr>
<td>b</td>
<td>212</td>
<td>16 20 19 11 14 3</td>
<td>Advice</td>
</tr>
<tr>
<td>c</td>
<td>168</td>
<td>1 1 2 1 1 2</td>
<td>Epigramme de Monsieur des-Portes</td>
</tr>
<tr>
<td>d</td>
<td>219</td>
<td>5 11 4 5 6 13</td>
<td>Ode de Monsieur Racan</td>
</tr>
<tr>
<td>e</td>
<td>292</td>
<td>1 1 1 2 1 4</td>
<td>To my dear and most worthy friend, Mr. Isaac Walton</td>
</tr>
<tr>
<td>f</td>
<td>351</td>
<td>3 3 1 2 2 2</td>
<td>The World, Ode</td>
</tr>
<tr>
<td>g</td>
<td>345</td>
<td>1 1 1 1 1 1</td>
<td>A Journey into the Peak</td>
</tr>
<tr>
<td>h</td>
<td>1959</td>
<td>1 1 1 1 1 3</td>
<td>Burlesque. Upon the Great Frost</td>
</tr>
<tr>
<td>i</td>
<td>1830</td>
<td>2 3 3 4 5 5</td>
<td>A Voyage to Ireland in Burlesque. Canto 2</td>
</tr>
<tr>
<td>j</td>
<td>1543</td>
<td>1 1 1 1 1 1</td>
<td>A Voyage to Ireland in Burlesque. Canto 3</td>
</tr>
<tr>
<td>k</td>
<td>1537</td>
<td>1 1 1 1 1 3</td>
<td>Epistle to Sir Clifford Clifton</td>
</tr>
<tr>
<td>l</td>
<td>789</td>
<td>3 7 7 6 8 3</td>
<td>On Tobacco</td>
</tr>
<tr>
<td>m</td>
<td>816</td>
<td>1 3 4 3 4 4</td>
<td>Eclogue</td>
</tr>
<tr>
<td>n</td>
<td>664</td>
<td>4 5 3 6 7 7</td>
<td>On Marriott</td>
</tr>
</tbody>
</table>

**Cowley**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Author’s rank</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>393</td>
<td>4 6 6 4 3 2</td>
<td>On his Majesties returne out of Scotland</td>
</tr>
<tr>
<td>b</td>
<td>443</td>
<td>9 12 16 12 10 16</td>
<td>A Poetical Revenge</td>
</tr>
<tr>
<td>c</td>
<td>256</td>
<td>21 22 20 22 15 18</td>
<td>To his very much honour’d Godfather, Master A. B.</td>
</tr>
<tr>
<td>d</td>
<td>294</td>
<td>23 21 19 20 12 20</td>
<td>An Elegie on the death of Mrs Anne Whitfield</td>
</tr>
<tr>
<td>e</td>
<td>217</td>
<td>22 17 17 19 19 20</td>
<td>An Answer to an Invitation to Cambridge</td>
</tr>
<tr>
<td>f</td>
<td>469</td>
<td>2 1 1 1 1 2</td>
<td>A Discourse By way of Vision Concerning the Government of O. Cromwell</td>
</tr>
<tr>
<td>g</td>
<td>432</td>
<td>3 10 12 3 4 4</td>
<td>It is a Truth so certain and so clear</td>
</tr>
<tr>
<td>h</td>
<td>312</td>
<td>2 1 2 1 1 1</td>
<td>When, Lo, e’re the last words were fully spoke</td>
</tr>
<tr>
<td>i</td>
<td>482</td>
<td>1 1 1 1 1 1</td>
<td>A paraphrase on an Ode in Horace’s third Book</td>
</tr>
<tr>
<td>j</td>
<td>323</td>
<td>10 9 12 12 9 12</td>
<td>The Motto</td>
</tr>
<tr>
<td>k</td>
<td>328</td>
<td>1 1 3 3 11 9</td>
<td>Life and Fame</td>
</tr>
<tr>
<td>l</td>
<td>357</td>
<td>8 14 16 12 13 17</td>
<td>Life</td>
</tr>
<tr>
<td>m</td>
<td>933</td>
<td>1 1 1 1 1 1 1</td>
<td>The First Nemaean Ode of Pindar</td>
</tr>
<tr>
<td>n</td>
<td>772</td>
<td>2 1 1 1 2 2</td>
<td>Hymn. To Light</td>
</tr>
<tr>
<td>o</td>
<td>1291</td>
<td>1 1 1 1 2 2</td>
<td>The Second Olympique Ode of Pindar</td>
</tr>
<tr>
<td>p</td>
<td>1182</td>
<td>9 10 9 11 7 14</td>
<td>The Complaint</td>
</tr>
<tr>
<td>q</td>
<td>1315</td>
<td>1 1 2 2 1 1</td>
<td>To the Royal Society</td>
</tr>
<tr>
<td>r</td>
<td>1897</td>
<td>1 1 1 1 1 1</td>
<td>The Plagues of Egypt, II, 1–262</td>
</tr>
<tr>
<td>s</td>
<td>6812</td>
<td>1 1 1 1 1 1</td>
<td>Davideis, Book ii</td>
</tr>
</tbody>
</table>
### Table A1 (continued)

<table>
<thead>
<tr>
<th>Word-list</th>
<th>150</th>
<th>120</th>
<th>100</th>
<th>80</th>
<th>60</th>
<th>40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gould</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>221</td>
<td>18</td>
<td>20</td>
<td>20</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>b</td>
<td>219</td>
<td>4</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>c</td>
<td>216</td>
<td>14</td>
<td>10</td>
<td>19</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>d</td>
<td>219</td>
<td>15</td>
<td>16</td>
<td>16</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>e</td>
<td>176</td>
<td>19</td>
<td>22</td>
<td>21</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>f</td>
<td>184</td>
<td>3</td>
<td>6</td>
<td>10</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>612</td>
<td>8</td>
<td>10</td>
<td>13</td>
<td>16</td>
<td>21</td>
</tr>
<tr>
<td>h</td>
<td>616</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>i</td>
<td>651</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>j</td>
<td>904</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>k</td>
<td>786</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>l</td>
<td>908</td>
<td>7</td>
<td>12</td>
<td>15</td>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>m</td>
<td>602</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>n</td>
<td>4019</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>3988</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>p</td>
<td>1995</td>
<td>1</td>
<td>4</td>
<td>9</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>q</td>
<td>4037</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>r</td>
<td>4492</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Oldham</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>568</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>b</td>
<td>577</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>c</td>
<td>652</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>d</td>
<td>887</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>266</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>f</td>
<td>1254</td>
<td>4</td>
<td>8</td>
<td>14</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>g</td>
<td>3795</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>h</td>
<td>1945</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>i</td>
<td>1726</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>j</td>
<td>2534</td>
<td>4</td>
<td>8</td>
<td>14</td>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>k</td>
<td>3795</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>l</td>
<td>2327</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>m</td>
<td>1503</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>n</td>
<td>2210</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>3378</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>p</td>
<td>1298</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Phillips</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>229</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>b</td>
<td>242</td>
<td>6</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>181</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>297</td>
<td>4</td>
<td>11</td>
<td>14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>e</td>
<td>238</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>344</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>358</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>h</td>
<td>415</td>
<td>2</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>i</td>
<td>190</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>j</td>
<td>445</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>k</td>
<td>299</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>l</td>
<td>356</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>m</td>
<td>667</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>n</td>
<td>554</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>565</td>
<td>7</td>
<td>9</td>
<td>9</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>p</td>
<td>1210</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

A Measure of Stylistic Difference and Likely Authorship
Table A1  (continued)

<table>
<thead>
<tr>
<th>Word-list</th>
<th>Length</th>
<th>Author's rank (ex 25)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prior</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>303</td>
<td>4</td>
</tr>
<tr>
<td>b</td>
<td>336</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>541</td>
<td>4</td>
</tr>
<tr>
<td>d</td>
<td>364</td>
<td>15</td>
</tr>
<tr>
<td>e</td>
<td>192</td>
<td>4</td>
</tr>
<tr>
<td>f</td>
<td>290</td>
<td>3</td>
</tr>
<tr>
<td>g</td>
<td>321</td>
<td>3</td>
</tr>
<tr>
<td>h</td>
<td>156</td>
<td>3</td>
</tr>
<tr>
<td>i</td>
<td>209</td>
<td>3</td>
</tr>
<tr>
<td>j</td>
<td>245</td>
<td>5</td>
</tr>
<tr>
<td>k</td>
<td>242</td>
<td>7</td>
</tr>
<tr>
<td>l</td>
<td>149</td>
<td>15</td>
</tr>
<tr>
<td>m</td>
<td>933</td>
<td>2</td>
</tr>
<tr>
<td>n</td>
<td>1112</td>
<td>4</td>
</tr>
<tr>
<td>o</td>
<td>1115</td>
<td>7</td>
</tr>
<tr>
<td>p</td>
<td>1615</td>
<td>5</td>
</tr>
<tr>
<td>q</td>
<td>2697</td>
<td>1</td>
</tr>
<tr>
<td>r</td>
<td>6033</td>
<td>1</td>
</tr>
<tr>
<td><strong>Swift</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>310</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>316</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>406</td>
<td>7</td>
</tr>
<tr>
<td>d</td>
<td>489</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>354</td>
<td>2</td>
</tr>
<tr>
<td>f</td>
<td>535</td>
<td>2</td>
</tr>
<tr>
<td>g</td>
<td>408</td>
<td>6</td>
</tr>
<tr>
<td>h</td>
<td>366</td>
<td>8</td>
</tr>
<tr>
<td>i</td>
<td>894</td>
<td>1</td>
</tr>
<tr>
<td>j</td>
<td>823</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>899</td>
<td>1</td>
</tr>
<tr>
<td>l</td>
<td>1023</td>
<td>3</td>
</tr>
<tr>
<td>m</td>
<td>1210</td>
<td>5</td>
</tr>
<tr>
<td>n</td>
<td>1243</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>1093</td>
<td>1</td>
</tr>
<tr>
<td>p</td>
<td>1433</td>
<td>1</td>
</tr>
<tr>
<td>q</td>
<td>3381</td>
<td>1</td>
</tr>
<tr>
<td>r</td>
<td>3206</td>
<td>1</td>
</tr>
<tr>
<td><strong>Waller</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>321</td>
<td>3</td>
</tr>
<tr>
<td>b</td>
<td>554</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>388</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>327</td>
<td>4</td>
</tr>
<tr>
<td>e</td>
<td>247</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>263</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>202</td>
<td>1</td>
</tr>
<tr>
<td>h</td>
<td>276</td>
<td>6</td>
</tr>
<tr>
<td>i</td>
<td>330</td>
<td>14</td>
</tr>
<tr>
<td>j</td>
<td>440</td>
<td>1</td>
</tr>
<tr>
<td>k</td>
<td>336</td>
<td>6</td>
</tr>
<tr>
<td>l</td>
<td>422</td>
<td>1</td>
</tr>
</tbody>
</table>
Table A1  (continued)

<table>
<thead>
<tr>
<th>Length</th>
<th>Author’s rank (ex 25)</th>
<th>Word-list</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>150</td>
<td>120</td>
</tr>
<tr>
<td>m</td>
<td>1777</td>
<td>3</td>
</tr>
<tr>
<td>n</td>
<td>1322</td>
<td>1</td>
</tr>
<tr>
<td>o</td>
<td>2606</td>
<td>1</td>
</tr>
<tr>
<td>Others</td>
<td></td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>1170</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1650</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>7824</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>19896</td>
<td>2</td>
</tr>
<tr>
<td>e</td>
<td>1926</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>1867</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>15694</td>
<td>1</td>
</tr>
<tr>
<td>h</td>
<td>12885</td>
<td>1</td>
</tr>
<tr>
<td>i</td>
<td>7817</td>
<td>1</td>
</tr>
</tbody>
</table>
Rolling stylometry

Maciej Eder
Polish Academy of Sciences, Institute of Polish Language, Kraków, Poland
Pedagogical University of Kraków, Poland

Correspondence: Maciej Eder, Polish Academy of Sciences, Institute of Polish Language, al. Mickiewicza 31, 31-120 Kraków, Poland
E-mail: maciejeder@gmail.com

Abstract

This paper introduces a new stylometric method that combines supervised machine-learning classification with the idea of sequential analysis. Unlike standard procedures, aimed at assessing style differentiation between discrete text samples, the new method, supported with compact visualization, tries to look inside a text represented as a set of linearly sliced chunks, in order to test their stylistic consistency. Three flavors of the method have been introduced: (i) Rolling SVM, relying on the support vector machines classifier, (ii) Rolling NSC, based on the nearest shrunken centroids method, and (iii) Rolling Delta, using the classic Burrowsian measure of similarity. The technique is primarily intended to assess mixed authorship; however, it can be also used as a magnifying glass to inspect works with unclear stylometric signal. To demonstrate its applicability, three different examples of collaborative work have been briefly discussed: (i) the 13th-century French allegorical poem Roman de la Rose, (ii) a 15th-century translation of the Bible into Polish known as Queen Sophia’s Bible, and (iii) The Inheritors, a novel collaboratively written by Joseph Conrad and Ford Madox Ford in 1901.

1 Introduction

In classical approaches to stylometry—be it authorship attribution, genre recognition or, say, a distant-reading classification of hundreds of novels—the goal is to compute a measure of similarity between the texts in a corpus, in order to discover hidden patterns or regularities. In authorship attribution, it involves extracting the authorial profile from a disputed text, followed by a procedure of identifying the best match in a set of “candidates”; in stylometry beyond attribution, it is aimed at finding groups of stylistically similar works. Even if the input texts are split into samples, the basic high-level stylometric unit is a literary work in its entirety. The hypothesis that consecutive sections of a given text might reveal linear development of certain stylistic features is still a relatively new perspective in this field.

Sequential analysis is a very attractive way of assessing linear phenomena; it is widely used in signal processing, econometrics, weather forecasting, electroencephalography, and so forth. It relies on the general assumption that the sequential order of elements is as important as the elements themselves: if a given series of events is nonrandom, the next element in the series should be—to some extent—modelable and hence predictable. Since natural languages are linear by definition, it was only a matter of time before the methods of sequential analysis were adopted to linguistics. In a study on language as a probabilistic phenomenon, Herdan introduces a classical distinction between two domains of quantitative linguistics: “language in the mass” vs. “language in the line” (Herdan, 1966: 423). While the former
category refers to the popular “bag-of-words” approach, the latter emphasizes the importance of—among other things—the preceding context of analyzed words.

Interestingly, the intuition that language is a modelable sequence of nonrandom units was verbalized as early as in 1913, in the fundamental study introducing Markov chains: to test mathematical assumptions of his new method, Markov used sequences of letters from Eugene Onegin by Alexander Pushkin (Markov, 2006 [1913]; Petruszewycz, 1981). Even if this particular contribution to linguistics was symbolic rather than significant, Markov’s studies had a great impact on theoretical foundations of sequential methods. In particular, the concept of the moving (sliding) window needs to be mentioned here.

This relatively simple idea revolutionized the way in which linear phenomena could be assessed: supposing that a sequence of events (referred to as time series) consists of $N$ elements, the goal is to measure mathematical properties of a subset of $k$ consecutive elements extracted from the beginning of the sequence, and then to move such a “window” of the size $k$ through the entire time series until the position $x = N - k$ is reached. In consequence, one obtains insight into particular segments of the dataset in their development. It allows to detect periodic regularities in the time series on the one hand, and possible disturbances or local idiosyncrasies on the other. The idea of moving window will be the key concept in the present paper.

Sequential methods, including time series analysis, spectral analysis, and Markov models, have been widely used in natural language processing (NLP), including part-of-speech tagging, parsing, speech recognition and synthesis, to name but a few applications. In the field of stylometry, autoregression models proved effective to assess versification (Pawłowski, 1999; Pawłowski and Eder, 2001). Markov chains—at character level—have been introduced to authorship attribution (Khmelev and Tweedie, 2001). Some elements of sequential analysis have been implicitly used in several attribution studies involving word $n$-grams, character $n$-grams, or POS $n$-grams as discriminative features (Stamatatos, 2009; Koppel et al., 2009; Eder, 2011; Hirst and Feiguina, 2007), since $n$-grams are in fact sequentially ordered series of nonrandom units.

The concept of moving window has been extended by van Dalen-Oskam and van Zundert in their study on the medieval Dutch Arthurian epic poem entitled Roman van Walewein (van Dalen-Oskam and van Zundert, 2007). Instead of analyzing sequences of 3 or so letters, the authors used a very high-level moving window of the length of several hundreds words. The aim of such an approach was to generate a series of virtual subsamples from the Walewein in order to test their stylistic consistency throughout the whole text. Other notable approaches to visualize stylistic shifts using moving windows include a paper on Middle-Dutch rhyme words (Kestemont, 2010), on three disputed English prose texts (Burrows, 2010), and on The Tutor’s Story by Kingsley and Mallet approached with $t$-tests (Hoover, 2011).

Similar studies have now been made possible by the recently introduced Rolling Delta method, available in the R package ‘Stylo’ (Eder et al., 2013). The technique has been applied to examine collaborative works by Joseph Conrad and Ford Madox Ford (Rybicki et al., 2014), and used in a benchmark study on Dickens (Tabata, 2014). In this method, the standard windowing procedure is run throughout a reference corpus: a representative centroid for each reference text that consists of the mean relative frequency for each of the $N$ words in the windows extracted from the text is calculated. Next, the test text is also divided into windows and a distance measure (in this case, Burrows’s Delta) between each text window and each reference centroid is computed. The results are visualized using a set
of curves—one for each reference text. The final step involves identifying, for each window, the lowest line, i.e. the most similar reference text. Whenever a takeover (line crossing) occurs, the respective window is assumed to reveal a stylistic change.

Regardless of the level of mathematical complication of the aforementioned sequential stylometric techniques, the basic underlying concept they share is quite simple. Namely, the goal is to split an input text into several subsequent samples (windows) and to contrast them one by one against the reference corpus. It is crucial, however, to keep the original order of the analyzed samples.

Arguably, a generalization of the above idea is straightforward. Since the windowing procedure can be used as a framework for simple similarity measures, such as $t$-tests, it will be also, by extension, applicable to any machine-learning classifier. The present study is aimed at discussing such a generalization and at introducing a robust method of rolling classification. The method will use the concept of moving window in combination with standard supervised classification techniques.

The paper is divided into two parts. In the first part, theoretical assumptions of the new method are discussed; they concern sampling issues (moving window), classification (three different machine-learning techniques) and visualization. In the second part, exemplary applications of the rolling method will be presented. The chosen examples include (i) the 13th-century French allegorical poem *Roman de la Rose*, (ii) a 15th-century translation of the Bible into Polish known as *Queen Sophia’s Bible*, and (iii) *The Inheritors*, a novel collaboratively written by Joseph Conrad and Ford Madox Ford in 1901.

Since the goal of the paper is to introduce a new technique rather than to perform an extensive benchmark, the presented applications are not case studies in a strict sense. In the first case, there is no reference corpus available, the second example is a highly collaborative work at multiple levels—probably too complex to be reliably solved—while the third application’s results substantially depend on the input parameters. Obviously this is not a good material for a systematic benchmark. The chosen examples, however, aim at showing the research questions that can be assessed using the new technique; they are also used to point out a few methodological issues that need to be solved in the future.

2 Method

2.1 Sampling

Supposing there is a work written collaboratively—i.e. a text in which some authorial takeovers are suspected to have happened—the procedure should start with chunking the text into consecutive samples. It is true that any already-existing sections (chapters, acts, cantos, and so forth) can be used as natural chunk delimiters: this solution is obviously preferable when external evidence suggests such a character of authorial collaboration. However, sequential methods show their real power when the samples are distributed evenly throughout the input dataset. In such a case, the dataset shows the properties of a time series, and thus it can be analyzed using dedicated tools, e.g. one can estimate the autocorrelation function to see if there are any cyclic regularities. Also, being a representation of actual timeline, the internal development of textual units can be easily visualized and reliably compared. For this reason, it is better to split the text in question into equal-size blocks of $N$ words (tokens). The desired sample size is a key parameter here, and it
needs to be decided arbitrarily. On the one hand, keeping this parameter small increases the resolution of sampling—an essential factor to pinpoint stylistic breaks—on the other hand, however, below a certain sample size, classification methods become blind. It has been shown that minimal sample length for authorship attribution is roughly 5,000 words, depending on the language and the genre examined (Eder, 2013).

The resolution of sampling can be substantially improved when the concept of moving window is applied. Its inherent feature is that it allows sample overlapping, so that some observations in a dataset can be re-used several times. In the case of rolling stylometry, it means that a sequence of tokens from the end of sample A will re-appear in the middle of sample B, and at the beginning of sample C. The idea of overlapping is shown in Fig. 1, where different types of sampling—with and without overlap—are represented. The only difference is that in classical approaches, e.g. in Markov models, the length of the moving window rarely exceeds 2-3 units (letters or syllables), and consequently the sample overlap is also very small. Stylometric windows, on the other hand, rely on extremely wide windows of hundreds of words, and thus their overlap needs to be augmented accordingly. In the exemplary applications discussed below, the window size of 5,000 words has been used, with an overlap of 4,500 words. Arguably, other combinations of the window size and the overlap parameters will lead to (slightly) different final results. This issue needs to be examined systematically, not only taking into account chunking options, but also a few style-markers (POS-tags, letter n-grams, and so forth), feature normalization algorithms, as well as alternative classifiers. Such a controlled benchmark experiment exceeds substantially the scope of this paper.

Fig. 1: Chunking a text into subsequent samples using a windowing procedure: in the above three variants, the same sample size of k elements is used in combination with different sample overlap (denoted by d).
2.2 Classification

The observation that the chunks produced by the windowing procedure can serve as regular text samples for authorship attribution is quite obvious. In short: instead of attributing a given text in its entirety, the goal is to perform an independent similarity test for each chunk, and then to inspect the results as a sequence of ordered stylistic signals.

Arguably, any classification method can be combined with the above framework. In the present study, however, three supervised classification techniques known for their high accuracy will be used. These are support vector machines (SVM), nearest shrunken centroids (NSC), and Delta in its classical Burrowsian flavor. These classifiers have been thoroughly tested in authorship attribution (Koppel et al., 2009; Burrows, 2002a; Hoover, 2004; Jockers and Witten, 2010; etc.).

Even if they rely on substantially different mathematical kernels, SVM, NSC and Delta use exactly the same corpus setup to carry out the classification. Namely, a number of representative samples for each class is expected to constitute a reference set (training set), while the remaining samples, including anonymous ones, go to a test set. Next, each sample from the test set is checked against the training set in order to identify the most similar authorial profile (i.e. the best matched training class). Unlike SVM and NSC, however, Delta does not combine individual training samples into averaged profiles for each class. This means that for a two-class problem Delta produces a ranking of the most similar training samples, beginning from the best match, e.g. \{A, A_0, B, A_b, B_z, B_y, \ldots\}, while a standard classifier provides a ranking of composite classes: \{A_{abc}, B_{xyz}\} or \{B_{xyz}, A_{abc}\} (i.e. the most probable class followed by the less probable one). This feature of Delta will be used below (see section 3.3) to visualize the first, the second and the third hint of the classifier at a time, even for two-class problems. However, such a peculiarity does not affect the general setup of the experiment, which is shared across the methods. In rolling stylometry, the same general setup is used as well. The only difference is that the test set contains a single work to be chunked automatically into equal-sized segments.

Worth mentioning is the fact that multidimensional methods are very sensitive to the number of features used. In stylometry, the commonly accepted type of features (style-markers) is frequencies of the most frequent words (MFWs), even if there is no consensus how many MFWs should be used. The rolling techniques discussed in this paper can be used either with MFWs, or with alternative style-markers, such as character n-grams, POS tags, or even manually selected content words. The number of features can be customized too. To keep things simple, the experiments discussed below were carried out using MFWs, with three different frequency ranges of 100, 500 and 1000 MFW (one range at a time).

2.3 Visualization

The final stage of the analysis involves a graphical representation of stylistic changes throughout a chunked text. It is true that standard classification techniques do not need any visualization (they provide a list of assigned classes, which is informative enough per se). However, one cannot deny an escapable explanatory power of graphs, trees and diagrams. For that reason, a simple home-brew graph has been added.

The design was based on the assumption that simplicity improves visual informativeness. Thus, to keep the plot clean, any redundant information have been removed. The goal was to
emphasize visually the most likely candidate—i.e. the actual answer of the classifier—and to keep less probable candidates slightly in the shadow. To this end, horizontal stripes colored according to the assigned classes were used, the primary stripe bolded. In pure form, this idea is implemented in the Delta variant of the method (Fig. 6–8). For each chunk, the first, the second and the third ranked candidates are represented by an appropriate (tiny) segment of the bottom stripe, the middle stripe, and the top one, respectively. When classification results are consistent across a number of chunks, the stripes tend to be unicolored rather than patchwork-like.

Simple as it is, the visualization does not provide a good cue as to how reliable the output is—this is due to inherent limitation of the Delta method. In the case of SVM and NSC, however, the output has been slightly enhanced: it seemed reasonable to take advantage of the final probability scores these methods optionally provide. In standard SVM, class assignment is made according to decision values, which can be either negative or positive: the higher they are, the more robust the assignment is. For instance, if the decision values for a given sample are as high as 0.93 in favor of the class A, and -0.93 of the class B, the sample is attributed to A with a high degree of certainty. If the values are 0.01 and -0.01, respectively, then the sample is still attributed to A, but such a classification is not very robust. Now, if one normalizes the decision values, they can be used to control the width of the plotted stripes (Fig. 2–3). In consequence, a stripe wide in its bottom part stands for a robust classification, and the other way around: the more a given segment is overtaken by the secondary (i.e. top) stripe, the more dubious the method. Since the normalized values always add up to 1, the adjusted widths of the two stripes make them look like a single band, gently waving up and down the reference line.

In NSC, regular final probabilities were used. Adopting them to control the width of the stripes turned out to be straightforward (Fig. 4-5). At first glance, the stripes produced by NSC and SVM are significantly different: the former flow gently along the x axis, the latter are more solid, with some ragged areas. This is because NSC is generally less hesitant in classification than SVM. One might say, NSC is devoted to the phrase “let your statement be, ‘Yes, yes’ or ‘No, no’; anything beyond these is of evil” (Matthew 5.37). However, even if sometimes too self-confident, NSC is still one of the best classifiers for stylometry (Jockers and Witten, 2010).

To show the above variants of the rolling technique in action, three exemplary applications are discussed in the following sections. For each case, a different flavor of the method is used: Rolling SVM, Rolling NSC, and Rolling Delta, respectively.

3. Exemplary applications

3.1 Roman de la Rose

This 13th-century French poem, styled as an allegorical dream vision, is a perfect material to test sequential stylometric methods. There is a consensus among scholars that the first part of the poem has been written by Guillaume de Lorris around 1230, and the second part has been completed by Jean de Meun about the year 1275. More importantly, unlike the Middle-Dutch poem Roman van Walewein, which has also been written by two authors in the 13th century, the takeover point in the Roman de la Rose is well known: Guillaume de Lorris is the author of the opening 4,058 lines (ca. 50,000 words), and the second part by Jean de Meun’s consists of 17,724 lines (ca. 218,000 words). This knowledge is supported by the text itself.
Namely, Jean de Meun explicitly writes about the collaborative authorship of the poem, he indicates the name of his predecessor as well as his own name, and he points out the last lines written by Guillaume de Lorris.

The research hypothesis is quite simple in this case: a given classification method is to be called effective if it captures the authorial takeover. Also, one can expect that an effective method reveals a clean segment for the first author followed by an equally clean segment for the second one. Any contamination within the two parts should be considered suspicious.

There is a fly in the ointment, though. Straightforward at first glance, this case contains a non-trivial issue, since there are no extant texts written by the first author of the *Roman de la Rose* that could serve as a comparison corpus. In their seminal study on *Walewein*, the authors faced exactly the same problem (van Dalen and van Zundert, 2007). The proposed solution was rather tricky: two fragments from *Walewein* itself were copied to form the reference corpus. One sample was transplanted from the beginning of the poem, the other from its final part.

![Fig. 2: “Roman de la Rose” assessed using Rolling SVM and 100 MFWs. The level of certainty of the classification is indicated by the thickness of the bottom stripe. The commonly-accepted division into two parts of the poem is marked with the vertical dashed line “b”.

The same method—with the same caveat—can be used to assess *Roman de la Rose*: 10,000 words (roughly 1,000 lines) from the beginning, and the same amount of data from the middle (words from the range of 113,000-123,000) have been copied to serve as a surrogate for the training corpus. To carry on the analysis, the edition by Marteau, slightly outdated but open-accessibly available at the Gutenberg Project website, has been used (Marteau, 1878). The results obtained using Rolling SVM and 100 MFWs are shown in Fig. 2-3. The first observation is that the takeover—marked with the dashed vertical line “b”—has been more or less precisely recognized. Also, a vast majority of the chunks have been robustly attributed to their actual authors; certainly, the sections that were transplanted to the training corpus, delimited with the vertical lines “a” and “c-d”, are correctly recognized as well. In the remaining sections, the attribution accuracy depends on the analyzed author: the part by Jean de Meun reveals a more consistent authorial signal, while Guillaume de Lorris’s chunks are significantly cluttered in the middle of his share.
There are at least three possible explanations of these authorial inconsistencies. Firstly (and pessimistically), the method is not precise enough. Secondly, the transplanted sample from the beginning of the poem did not contain enough information about the authorial profile of Guillaume de Lorris. Last but not least—horribile dictu—the second author did some corrections in the Guillaume’s passages. Wrongly attributed chunks at the end of the poem are also interesting: they seem to show that Jean de Meun put less stylistic effort (in the sense that the authorial signal is weak) when his work was about to be finished. It should be emphasized here that very similar local misclassifications were noticed using the remaining methods: Rolling NSC and Rolling Delta.

3.2 Queen Sophia’s Bible

The first known translation of the Bible into Polish was completed in the year 1455. It was commissioned by Sophia of Halshany, wife of Władysław II Jagiello, King of Poland and Grand Duke of Lithuania. The translation was not intended for liturgical purposes, neither was it known outside the royal court. The text was handwritten on parchment in two large codices. The second volume has been disintegrated quite early to provide parchment for book covers (a few folios have survived), and the first volume, containing a good share of the Old Testament, was lost during the World War II, probably damaged. Luckily enough, however, a black-and-white facsimile of the extant manuscript was published in 1930.

The Queen Sophia’s Bible is a very interesting example of a collaborative (at many levels) work. It is translated directly from Czech rather than from the Vulgate, but it is difficult to decide which of the several variants of the Czech Bible was used (Deptuchowa, 2008: 9-12; Wanicowa, 2009: 76-81). Textual similarities suggest that the beginning passages were translated from one of the Bibles of the oldest 1st redaction, while the next passages were probably rendered after the 2nd redaction—it should be emphasized, however, that despite differences, all the Czech redactions are ultimately derived from the same translation (Kyas, 1997). When it comes to textual source, then, Queen Sophia’s Bible is a translatorial composite: the original Hebrew/Greek Bible was translated into Latin (the Vulgate), then into Czech, then it was modernized (2nd Czech redaction), and then rendered in Polish.

Closer examination of the codex (or its extant facsimile, to be precise) reveals another level of collaboration: five different scribal hands can be distinguished quite easily. The scribal takeovers can be also noticed at the level of orthography. Moreover, the sections for particular scribes seem to be different at stylistic level as well: the quality of translation in the sections written by the 2nd and 5th scribal hands are considered to be smooth, while the 1st hand is claimed to be “very literal” and generally of poor quality (Urbańczyk, 1933). Thus, it has been hypothesized that several translators were involved in rendering the Biblical text or, alternatively, that the scribes were at the same time translators.

All the above variables taken together make Queen Sophia’s Bible a multifaceted collaborative work, in which the translatorial, authorial, and scribal signals are heavily mixed. Certainly, telling these signals apart is unrealistic. What seems feasible, however, is corroborating the hypothesis that the scribal takeovers are correlated with stylistic (i.e. translatorial) transitions. From stylometric point of view, translators are known to be (almost) invisible. However, even if this is the author of the original that is usually stylistically predominant in a translated text, a few successful translatorial attributions show that such a presence-in-a-shadow can also be pinpointed (Burrows, 2002b; Rybicki, 2012; Rybicki and
Heydel, 2013). In the case of *Queen Sophia’s Bible*, the voice of the original Hebrew authors is covered by so many layers that the risk of its predominance is lower than in usual translations, and thus any stylistic changes in the Polish text might be, with a reasonable probability, attributed to the translator rather than to the original author. Two issues are to be resolved, though. Firstly, the scribal signal has to be neutralized in order not to interfere with the stylistic one. Secondly, a comparison corpus has to be compiled.

It is a truth commonly known that in medieval manuscripts, orthographic variants are innumerable and they highly depend on scribes’ preferences, education, particular handwriting school, and so forth. Certainly, the same applies to the Old-Polish language and to *Queen Sophia’s Bible*. Being a strong discriminator, orthography can be used to tell the scribes apart (Thaisen, 2012), but at the same time it weakens—or interferes with—the stylistic signal. It has been shown, however, that the impact of scribes’ voices can be neutralized by dealing with a corpus in modernized transcription rather than in transliteration (Kestemont and van Dalen, 2009). Since *Queen Sophia’s Bible*, along with a dozen of other 15th-century Polish texts, has been recently edited in a form of a transliterated and transcribed parallel corpus (Twardzik, 2006), the suggested solution could be immediately applied.

When it comes to the second issue: a standard reference corpus cannot be compiled for the same reasons as in *Roman de la Rose*, but, similarly, the takeover points are known very well. Thus, it was possible to reliably transplant a few samples into the training corpus. Segments of 10,000 words (one segment per class) have been transplanted from the parts by the 1st, 2nd, 4th and 5th scribe, i.e. roughly 50%, 50%, 25% and 10% of the material by these scribes, respectively. The 3rd scribe’s contribution (mere 2,000 words in total) is too short to allow any sample extraction. Certainly, since one training class is missing, this short segment will be by definition misclassified.

The results for Rolling NSC are shown in Fig. 4–5; handwriting changes are marked with vertical dashed lines. Regardless of the number of MFWs tested, the classifier detects a few style breaks in the dataset. The most valuable result, however, is that the stylistic breaks take place in parallel with scribal takeovers (with some minute shifts). This is a strong evidence in favor of the “many translators involved” hypothesis. Interestingly, the third stylistic break took place still in the 2nd scribe’s section (i.e. between the markers “a” and “b”), as if a newly hired translator was forced to work with his predecessor’s scribe for a while. Alternative explanations of this particular takeover are also possible, though: it might be a stylistic break in the original, e.g. the point where the 1st Czech redaction has been replaced with the new-fashioned 2nd redaction, it might be a break in the Latin pre-original, and so forth: one should remember that this case is stylometrically far too complex to make any conclusive statements.
Stylistic inconsistencies in the second half of the text should also be commented on. This part, written by the 5th scribe and recognized to be stylistically more or less consistent, contains some apparently misclassified segments. In Fig. 4, representing the results for 100 MFWs, some of the chunks seem to be partially eclipsed, and some are entirely overtaken by the 4th scribe’s class. In Fig. 5, for 500 MFWs, an even less probable attribution to the 1st scribe appears, along with a lengthy section robustly yet wrongly attributed to the 4th scribe. This can be interpreted as a weaker stylistic voice in the chunks in question (or yet another translator involved), but at the same time it is a meaningful caveat: when the number of classes is limited and the open-set case cannot be ruled out, false positives might appear.

Fig. 3: “Queen Sophia’s Bible” assessed using Rolling NSC and 100 MFWs.

Fig. 4: “Queen Sophia’s Bible” assessed using Rolling NSC and 500 MFWs.

Close reading of the text itself provides a convincing explanation of the wrongly attributed final chunks. Unlike the already-discussed misclassifications, which are rather accidental than
systematic, the final chunks are robustly attributed to the 4th scribe regardless of the method used (Rolling SVM, Rolling NSC, Rolling Delta). The reason of the apparent blindness of the method turned out to be embarrassingly trivial. Namely, at the end of the critically edited text, all the extant fragments from the second volume of Queen’s Sophia’s Bible are collected. As a result of corpus setup error, these concatenated fragments produced a fake stylistic signal, marked in Fig. 5 with the vertical dashed line “e”.

3.3 Conrad vs. Ford revisited

The last example, aimed at introducing Burrows’s Delta as a rolling classifier, is a replication of the experiment conducted by Rybicki, Kestemont and Hoover, reported in their study on Joseph Conrad’s and Ford Madox Ford’s collaboration (Rybicki et al., 2014). Having scrutinized three novels that were written collaboratively by Conrad and Ford, the authors of the paper conclude: “The decisive domination of Ford’s style over Conrad’s in The Inheritors and The Nature of a Crime is interesting, as it seems to have survived Conrad’s extensive editing that is confirmed by biographical evidence” (ibid., 429). To test the above claim, the same reference corpus and the same test text have been used. Namely, The Inheritors has been compared with six novels by Conrad (The Nigger of the Narcissus, 1897; Heart of Darkness, 1898-99; Lord Jim, 1900; Chance, 1913; Under Western Eyes, 1911; Victory, 1915) and six novels by Ford (The Benefactor, 1905; Privy Seal, 1907; An English Girl, 1907; Mr. Apollo, 1908; Ring for Nancy, 1913, The Good Soldier, 1915).

![Fig. 5: “The Inheritors” by Conrad/Ford assessed using Rolling Delta and 1,000 MFWs. The bottom stripe indicates the first ranked candidate (i.e. the most probable), then comes the second and the third suggested class.](image)

The results for a very long vector of 1,000 MFWs confirm the findings of the previous study (Fig. 6), in which 1,000 MFWs were used as well. Apparently, it suggests that The Inheritors indeed were mostly written by Ford, with some final sections contributed by Conrad. However, when the MFW parameter is lowered, the picture of the collaboration changes substantially. For 500 MFWs, Conrad becomes quite visible, especially in the background (the second and the third suggestion of the classifier). When the number of MFWs is further
reduced to 100 or so, Conrad boldly comes out from the shadow to take a leading role in the duo’s collaboration.

Very similar results were obtained using two other classifiers: Rolling SVM and Rolling NSC. Thus, such an ambiguous outcome can be interpreted as valid results, or as a palimpsest, through which the nature of a collaboration can be seen. One author (Ford) seems to have been responsible for setting the plot and drafting the chapters—which is visible in long vectors of mostly content words examined—while the other author (Conrad) seems to have done the actual word weaving (or heavy editing), the trace of which has been left in the function words’ usage. In any case, further benchmarks are needed to corroborate the palimpsestic hypothesis.

**Fig. 6:** “The Inheritors” by Conrad/Ford assessed using Rolling Delta and 500 MFWs.

**Fig. 7:** “The Inheritors” by Conrad/Ford assessed using Rolling Delta and 100 MFWs.
4 Conclusions

The sequential method as introduced above seems to be an attractive addition to the existing stylometric toolbox. Unlike standard procedures, aimed at assessing style differentiation between discrete text samples, it tries to look inside a text represented as a set of linearly sliced chunks, in order to test their stylistic consistency. Supported by compact visualization, and available, so far, in three variants—Rolling SVM, Rolling NSC and Rolling Delta—the technique is designed to assess mixed authorship. However, it can be also used as a magnifying glass to inspect works that behave strangely when analyzed using classical stylometric methods: if, say, *Night and Day* by Virginia Woolf does not cluster together with other works by the same author, it might be interesting to see which parts of the novel are more, and which less similar to Woolf’s authorial voice.

However, no matter how promising the rolling method is, it has also some drawbacks, at least at the current state of development. The most dangerous one is the risk of accidental false positives, as could be observed in the case of *Roman the la Rose* and *Queen Sophia’s Bible*. Arguably, this is due to an inherent characteristic of any written text rather than a pitfall of the method itself: even if sliced into considerably long samples, some local authorial idiosyncrasies cannot be avoided. In this regard, the modernists, and generally the authors experimenting with style, will probably be more difficult to pinpoint than stylistically consistent realistic literature.

In standard supervised machine-learning, the stage of attributing usually is, or at least should be, accompanied by a cross-validation procedure. This issue, as an important addition to the rolling classification techniques, will be discussed in a sequel of this paper.

References


Appendix

The rolling stylometry technique as introduced above is supported by the R package ‘stylo’ (ver. 0.5.8), through the function `rolling.classify()`. However, since the function is not supplemented by GUI yet, it might look non-intuitive. This appendix provides a concise step-by-step howto explaining its usage. To reproduce the experiments discussed in this document, one should use the following procedure:

1. Install the package stylo in the version >= 0.5.8. In R shell, type:

   ```r
   install.packages("stylo")
   ```

2. Create a new folder: it will serve as a working space for your experiment. Create two subfolders named “test_set” and “reference_set” (all file names are case sensitive!). Put your disputed text into the “test_set”, put the remaining texts into the “reference_set”. The setup for the Conrad/Ford case was as follows:
The complete setup for the *Roman de la Rose* case can be downloaded from the Computational Stylistic Group website <https://sites.google.com/site/computationalstylistics/>.

3. Load the library ‘stylo’, set your working directory:

```r
library(stylo)
setwd("path/to/the/folder/containing/two/subcorpora")
```

4. Optional but important: read the help page for the function `rolling.classify()`:

```r
help(rolling.classify)
```

5. Run the function `rolling.classify()`, use as many arguments as needed:

```r
# Fig. 2
rolling.classify(write.png.file = TRUE, classification.method = "svm", mfw=100, training.set.sampling = "normal.sampling", slice.size = 5000, slice.overlap = 4500)
```

```r
# Fig. 3
rolling.classify(write.png.file = TRUE, classification.method = "svm", mfw=500, training.set.sampling = "normal.sampling", slice.size = 10000, slice.overlap = 9000)
```
The vertical dashed line that divides the part by Guillaume de Lorris and Jean de Meun is produced by adding the word “xmilestone” into the input text, after the line 4,058 (i.e. after ca. 50,000 words). One can add as many milestones as needed; they will be reproduced in the final plot and labelled automatically using lowercase roman letters.

```
# Fig. 4
rolling.classify(write.png.file = TRUE, classification.method = "nsc", mfw=100, training.set.sampling = "normal.sampling", slice.size = 5000, slice.overlap = 4500)

# Fig. 5
rolling.classify(write.png.file = TRUE, classification.method = "nsc", mfw=500, training.set.sampling = "normal.sampling", slice.size = 5000, slice.overlap = 4500)

# Fig. 6
rolling.classify(write.png.file = TRUE, classification.method = "delta", mfw=1000)

# Fig. 7
rolling.classify(write.png.file = TRUE, classification.method = "delta", mfw=500)

# Fig. 8
rolling.classify(write.png.file = TRUE, classification.method = "delta", mfw=100)
```

Black-and-white variants of the above plots can be produced using an additional parameter:

```
# Fig. 8a, black-and-white version
rolling.classify(write.png.file = TRUE, classification.method = "delta", mfw=100, colors.on.graphs = "greyscale")
```

The source code of the package ‘stylo’, including the newly-added function `rolling.classify()`, can be downloaded from the GitHub repository: <https://github.com/computationalstylistics/stylo>.
COLOR FIGURES:

Fig. 2

Fig. 3
Fig. 4

Fig. 5
Fig. 6

Fig. 7