Out-of-the-Box Text Analysis for the Digital Humanities

David Hoover
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 ...

### Sunday, 3 June 2018 [DHSI Registration + 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](https://www.parks.ca.gov/?r=39&l=0); we recommend departing early (around 8.00 am) to catch low tide for a better view of the wonderful underside 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](https://www.portrenfewhotel.com/).

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

  A shorter journey to the resplendently beautiful [Butchart Gardens](https://www.butchartgardens.com/) and, if you like, followed by (ahem) a few minutes at the nearby [Church and State Winery](https://www.churchandstatewinery.com/), 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](https://www.saltspringisland.com/). 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](https://oceanriveradventures.com/) 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!

### Psst: Some Suggested Outings

#### Psst: Some Suggested Outings

- Early Class Meeting: 4. [Foundations] DH For Department Chairs and Deans ([Hickman 120, Classroom](https://www.hickman120.ca/))
  
  Further details are available from instructors in mid May to those registered in the class. Registration materials will be available in the classroom.

- **DHSI Registration** ([MacLaurin Building, Room A100](https://www.maclaurinbuilding.ca/))

  After registration, many will wander to [Cadboro Bay](https://cadborobay.ca/) and the pub at [Smuggler's Cove](https://www.smugglerscovepub.com/) OR the other direction to Shelbourne Plaza and [Maude Hunter's Pub](https://www.maudehunterspub.com/) OR even into the city for a [nice meal](https://www.greatcafe.com/).

### 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](https://www.maclaurinbuilding.ca/))

- **8:30 to 10:00** Welcome, Orientation, and Instructor Overview ([MacLaurin A144](https://www.maclaurinbuilding.ca/))
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)

Modelling 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 three-dimensional space, and 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 three-dimensional space, and 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.

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 three-dimensional space, and 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.

Lee Zickel (Case Western Reserve U): "Comfortably Trepid.” Abstract: #myDHi 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
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).

1:30 to 4:00
Classes in Session

▼ 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

Lunch break / Unconference
"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.
Thursday, 7 June 2018

9:00 to Noon

Classes in Session

12:15 to 1:15

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!

1:30 to 4:00

Classes in Session

4:15 to 5:15

Chair: James O’Sullivan

- Defining a Taxonomy of of Abandonment for Online Digital Humanities Projects. Luis Meneses (Electronic Textual Cultures Lab, U Victoria) and Jonathan Martin (King’s College London)
- The Stories We Tell: Representing Gay and Lesbian History through Digital Technologies in the LGCLC Project. Nadine Boulay (Simon Fraser University) and Ewan Matthews (King’s College London)
- Italian Paleography in the Digital Domain. Isabella Magni (Newberry Library)
- Digital Humanities, A Question of Ethics. Negar Basiri (Louisiana State U)
- Writing Poetry in High School. Guadalupe Echegoyen (National Autonomous U Mexico)

6:00 to 7:00

Bring your DHSI nametag and enjoy your first tipple on us!

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?
Joint Reception: DHSI and DLFxDHSI (University Club)
DLFxDHSI Poster/Demo Session

- Mediators as a Colonialist Artifact in Menzies’ Journal. Paula Johanson (U Victoria)
- Camp Edit: the Institute for the Editing of Historical Documents. Nikolaus Wasmoen (Association for Documentary Editing, U Buffalo), Jennifer Steltzer (Association for Documentary Editing, U Virginia), and Cathy Moran Hajo (Association for Documentary Editing, Ramapo C)
- A Digital Archaeology of Life in Cleveland’s Depression-Era Slums. Charlie Harper (Case Western Reserve U) and Jared Bendis (Case Western Reserve U)
- Feminist Pest Control: controlling and not controlling nonhuman pests. Lindsay Garcia (C of William and Mary)
- Legends of the Buddhist Saints. Jonathan S. Walters (Whitman C) and Dana Johnson (Freelance Web Developer)
- Accessibility in Digital Environments Via TEI-Encoded Uncontracted Braille. Gia Alexander (Texas A&M U)
- Translation3point0: Why Literary Translation Data Matters. Katie King (U Washington)
- PoéticaSonora: A Digital Audio Repository Prototype for Latin American Sound Art and Poetry. Aurelio Meza (Concordia U)
- Beauty and the Book: Pre-Raphaelite Artistic Practice Contained. Josie Greenhill (U Victoria)
- Poetic Procedures/Digital Deformances. Corey Sparks (California State U, Chico)
- Miranda, the Folger Shakespeare Library’s new Digital Asset Platform. Meaghan Brown (Folger Shakespeare Library)
- Living Song Project. Quinn Patrick Ankrum (U Cincinnati) and Elizabeth Avery (U Oklahoma)
- Digital Frankenstein Variorum. Rikk Mulligan (Carnegie Mellon U)

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

9:00 to 5:30

- DHSI All Day Workshop Session (click for workshop details and free registration for DHSI participants)
- 53. Building Your Academic Digital Identity (MacLaurin D105, Classroom)

Saturday, 9 June 2018 [DLFxDHSI + DHSI Conference and Colloquium]

9:00 to 5:00

- 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)
- 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)
- 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)
- 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)

9:00 to 5:00

- Building and Analyzing. Chair: Yannis Rammos (New York University)
- 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 (Quemet 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)
Sunday, 10 June 2018 [SINM + DHSI Registration, Workshops]

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 to 9:00</td>
<td>Symposium on Indigenous New Media Registration (MacLaurin A100)</td>
</tr>
<tr>
<td>9:00 to 5:00</td>
<td>DHSI Registration (MacLaurin A100)</td>
</tr>
<tr>
<td>9:00 to 4:00</td>
<td>SINM Sessions</td>
</tr>
<tr>
<td>9:00 to 4:00</td>
<td>DHSI All Day Workshop Sessions (click for workshop details and free registration for DHSI participants)</td>
</tr>
<tr>
<td>9:00 to Noon</td>
<td>DHSI AM Workshop Sessions (click for workshop details and free registration for DHSI participants)</td>
</tr>
<tr>
<td>1:00 to 4:00</td>
<td>DHSI PM Workshop Sessions (click for workshop details and free registration for DHSI participants)</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>Joint Institute Lecture (DHSI and SINM):</td>
</tr>
<tr>
<td></td>
<td>David Gaertner (U British Columbia): &quot;A Landless Territory?: CyberPowWow and the Politics of Indigenous New Media.&quot;</td>
</tr>
<tr>
<td></td>
<td>Chair: Deanna Reder (Simon Fraser U) (MacLaurin A144)</td>
</tr>
</tbody>
</table>

**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]
Tuesday, 12 June 2018

9:00 to Noon

Classes in Session

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)
### Wednesday, 13 June 2018

**1:30 to 4:00**  
DHSI Colloquium Lightning Talk Session 3 (MacLaurin A144)  
Chair: Lindsey Seatter

- 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)
- "Doing DH with Graphic Narratives. John Barber (Washington State U)
- "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)

**4:15 to 5:15**  
DHSI Colloquium Lightning Talk Session 4 (MacLaurin A144)

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

### 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

**4:15 to 5:15**  
DHSI Colloquium Lightning Talk Session 5 (MacLaurin A144)

**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!

### Friday, 15 June 2018

**9:00 to Noon**  
Classes in Session

**12:15 to 1:15**  
Lunch Reception / Course E-Exhibits (MacLaurin A100)
1:30 to 2:30

(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
DHSI 2018 Out-of-the-Box Text Analysis for the Digital Humanities

1. Course Curriculum

2. Hoover, David L. “Using the Wide Spectrum Spreadsheet”

3. Hoover, David L. “Using the Delta Calculation Spreadsheets”

4. Hoover, David L. “Using the Analyze Textual Divisions Spreadsheet”

5. Hoover, David L. “Cluster Analysis, Principal Components Analysis (PCA), and T-testing in Minitab”

6. The instructions above are also available on my website: https://wp.nyu.edu/exceltextanalysis/


DHSI 2017 Curriculum: Out-of-the-Box Text Analysis for the Digital Humanities

**Monday**
10:15-12:00
   - Introductions
   - What is text-analysis good for, anyway? The **Intelligent Archive**
1:30-4:00
   - The Wide Spectrum Spreadsheet

**Tuesday**
9:00-12:00
   - Planning a Project (plan to discuss your own interests)
   - Text Preparation and Editing
   - KWIC, Juxta, Analyze Textual Divisions Spreadsheet, The Apostrophe Project
1:30-4:00
   - MS Excel and Minitab: PCA and Cluster Analysis
   - Some Ready-to-Go Sample Analyses

**Wednesday**
9:00-12:00
   - Minitab: PCA and Cluster Analysis, continued
   - The Wide Spectrum Spreadsheet
1:30-4:00
   - The Delta Spreadsheets

**Thursday**
9:00-12:00
   - Minitab: T-tests, Graphing
1:30-4:00
   - Individualized Help with Projects

**Friday**
9:00-12:00
   - Review, Summing Up
Using the Wide Spectrum Spreadsheet
© David L. Hoover, 2015

Introduction

Zeta and Iota, introduced by John F. Burrows, are measures of textual difference that can be used effectively in authorship attribution investigations and stylistic studies. They are especially interesting as methods for locating an author’s characteristic vocabulary “marker” words that one author uses consistently, but that another author or other authors use much less frequently or not at all. Both measures exclude the extremely common words of the language that have traditionally been one main focus of computational stylistics and authorship attribution. Zeta concentrates on words of moderate frequency, and Iota on relatively rare words.

Hugh Craig has developed an alternative version of Zeta that simultaneously creates sets of marker words and anti-marker words, and I have developed Excel spreadsheets that calculate Zeta, Iota, and CraigZeta (as I will call it). More recently, I have been working with an Excel spreadsheet tool that I call the Wide-Spectrum Spreadsheet that extends CraigZeta so that it can analyze almost the entire vocabulary of a set of texts at once. Operating on a set of training texts of known authorship, it creates a composite score for each word in the analysis by adding the proportion of sections of text by the primary author in which the word occurs to the proportion of sections of texts by the other author in which the word does not occur. It ignores the frequency of the word, concentrating instead on how consistently it occurs. So, for example, assume we have 300 sections of text by our primary author and 400 by our secondary author. Assume further that the word “time” appears in 270 sections by the primary author and 80 sections by the secondary author (thus it is absent from 320 sections). The score for “time” is 270/300 + 320/400, which is equivalent to .90 + .80, or 1.70 (quite a strong distinctiveness score). Possible scores range from 2, for a word that is present in every section by the primary author and absent from every section by the secondary author (thus it is absent from 320 sections), to 0, for a word that is absent from all the sections by the primary author and present in all of the sections by the secondary author. Thus “time” is a word that is highly characteristic of our primary author, with respect to the secondary author. The spreadsheet sorts the words in the analysis on their scores and selects the most distinctive marker words for each author. Finally, it graphs the training texts against the test texts by calculating the percentage of different words (types) in each text section that are marker words for each of the authors.

The Wide-Spectrum Spreadsheet can analyze just about any kind of difference you are interested in. I discuss the simplest case here, involving two authors, but the two “authors” can also be classes of any type. For example, it can be used to contrast a single author to a large number of contemporaries, or to investigate differences in gender, age, genre, nationality, historical period, or political affiliation—indeed, any perceived contrast. It can be used on a single writer, to study the differences between early and late styles. The tool is an excellent difference detector, and it will find differences between almost any two classes. If you pick your contrast unwisely, however, the Zeta scores for the most distinctive words will be weak, and the marker words will be difficult to interpret. (See the references on “The Zeta and Iota Spreadsheet and the Wide-Spectrum Spreadsheet” page for some examples.) This tool requires a substantial amount
of text by the two “authors” to work well. This kind of analysis is too new to have any set
guidelines, but a sensible minimum might be 10 sections of at least 500 words each. Much larger
numbers of longer sections will improve the distinctiveness of the marker words and increase the
accuracy of the analysis. If you have many long texts, you may want to use sections of 10,000
words or more. Fortunately, the automated nature of the sheet makes it easy to test several
different numbers and sizes of sections, and this is always a good practice.

Getting Started

In order to do a Wide-Spectrum analysis with the Wide-Spectrum Spreadsheet, I recommend that
you begin by creating a mnemonically named subdirectory/folder on your computer and copying
WideSpectrumSpreadsheet2015.xlsm to the new directory. (The spreadsheet has been tested on
Macs, but the instructions below assume a PC platform running some version of MS Windows
and MS Excel 2007 or later.) Depending on the version of Excel you have and how it is
configured, you will probably need to enable macros for the spreadsheet to operate. Typically
there will be a warning bar or message that will tell you there are macros and will allow you to
enable them for this spreadsheet. I do NOT recommend that you set Excel to accept macros
automatically because macros are sometimes malicious programs, but the macros in the
distributed spreadsheets are harmless.

This spreadsheet contains a macro that automates and speeds up the process of doing a
Wide-Spectrum analysis, reducing the chance of error, collects the results in a clear and usable
form, and graphs them. Putting all the functions except creating the word lists themselves into
one self-contained package that can operate automatically should allow anyone interested in
textual similarity and difference to perform multiple experiments easily, even if they are not
particularly proficient in computer programming or computational stylistics. On the other hand,
because the macros are open source, those who have the expertise (or are willing to learn it) can
customize and revise them for other purposes or to improve them. This spreadsheet is under
further development, and future enhancements may be available on these pages.

Let’s take an example. Assume the unlikely event that you have found two long poems that are
known to by either Archibald MacLeish or Wallace Stevens. You would like to test which poet is
the most likely to be the author of each poem. To illustrate this kind of problem, we can simulate
it by using actual poetry by MacLeish and Stevens and just pretend we do not know the author of
two of their long poems. Such simulations with known texts are an important way of testing any
new method of authorship attribution. We collect large samples of poetry by both authors (in this
case, I acquired more than 70,000 words of poetry by each author from Literature Online), and
also some long poems (roughly 1,000 words or more) by each poet (these are not included in the
author’s base samples). Wide-Spectrum analysis can help us find sets of characteristic words for
MacLeish and Stevens that will distinguish their independent poems from each other and tell us
which author is likely to be the author of the “anonymous” poems. The analysis will also allow
us to examine the differences between the characteristic words of these poets.

Here, I have seven long poems, three by MacLeish and four by Stevens. I put aside one poem by
each poet as an “unknown,” but include the other five as trial texts. Only if the analysis correctly identifies known texts can we take seriously its attribution of the unknown texts.

**Word Lists**

These instructions assume that you are using Hugh Craig’s Intelligent Archive (with instructions) to create your word frequency lists. The Wide-Spectrum Spreadsheet sheet will accommodate 1,000 sections each by the primary and secondary authors, with room for 100 independent samples by both the primary and secondary author (or group of authors), and up to 100 anonymous texts you want to test. First create the parallel word lists for your primary and secondary authors. You should create a training set that includes only the base samples and you should normally select a block size of 2,000 or more words, longer for longer texts, and set the size of the output to a number large enough to collect the entire vocabulary. In the sample case here, I am using approximately 2,000-word sections and the entire word list of 14,711 words. Copy the word list itself and the data for the primary author out of the Intelligent Archive and paste it into columns A and following of the Author1 sheet, with the word list in column A and the data in B and following. (You can switch active worksheets by clicking on the tabs at the bottom of the window.) Paste the data for the secondary author into columns A and following of the Author2 sheet. Next, you will need to create word lists for any other texts you will be analyzing. These word lists MUST use the same word list that you created for the primary and secondary authors. To do this, create a Text Set that contains the remaining texts, here, the independent poems by MacLeish and Stevens and the “anonymous” poems, in the Intelligent Archive. Next, copy the word list from the Author1 sheet (just the words; in this case, they are in cells A2-A14712), select “Word Frequencies,” and paste the list of words into the “Words to Include in Output:” area. Keep the block size the same, and under “Output Options” select “Inclusion list words unsorted”), and produce your word lists. For all of these word lists, you will need to retain the last two rows in the data, which contain the number of types and tokens in each section, as these are used in the spreadsheet’s calculations. Paste only the lists of frequencies (not including the word list) of any independent texts by your primary author into the Author1Ind sheet, and any independent texts by your secondary author (or authors) into the Author2Ind sheet. (In our sample case, I put the two independent poems by MacLeish into the Author1Ind sheet and the three by Stevens into the Author2Ind sheet; one of the poems by each author is long enough to create two 2,000-word sections.) Then paste the data (again only the frequencies) for your test text(s) into the Test sheet. (In our sample case, I put one “anonymous” poem by each author into the Test sheet.) Finally, enter the names of the two authors into the Wide-Spectrum Spreadsheet sheet in cells E7 and E8 (the spreadsheet has formulas to enter their names elsewhere in the sheet).

**Wide-Spectrum Analysis**

Now run the macro to perform the analysis by clicking on the “Analyze&Graph” button. The macro clears out any existing results before beginning the next analysis, so, before running it, you may want to save any previous results in another spreadsheet for future reference. After clearing out the old data, the macro copies the data out of the Author1, Author2, Author1Ind, Author2Ind,
and Test sheets into the appropriate columns for analysis. It also shrinks most of the columns so that the sheet is easier to read. It sorts our 14,711 rows of data on the Wide-Spectrum score in column A in descending order, so that the strongest marker words for the primary author are at the top and the strongest marker words for the secondary author are at the bottom. It then finds the point at which words switch from being primary author markers to secondary author markers. (Actually, it also finds words with a score of exactly 1 and includes them in both authors’ sets.) It then sorts the secondary authors’ marker words in reverse and records the change-point and enters it into cell F9. Formulas in rows 2-7 of columns H and following calculate the proportion of the different words (types) in each section that are marker words for the primary and secondary authors. It uses the values set in F2-F7 to determine how many of the marker words for each author it tests; for authorship purposes, you should probably use the same number of words for each author. (Note that the sheet initially enters the maximum number of markers for each author in F2 and F3, half of each authors’ markers in F4 and F5, and one fourth of each authors’ markers in F6 and F7, to allow you to easily scale the analysis down. In addition to graphing the results and adding labels, the macro also pastes the word lists upon which the analysis is based and the results of the analysis to the area at the right of the data. The results can be copied into Minitab or another program for graphing if you prefer, or you can use the data to produce line graphs or other kinds of graphs in Excel.

Two other buttons in the sheet set options. “Set/Clear Optional Max F2&F3” allows you to toggle whether to limit the analysis to just 500 markers for each author (as Craig and Kinney do in their work on Shakespeare). “Eliminate/Keep ‘Hapax’” allows you to toggle whether or not to include hapax legomena (words that occur in just one section in the entire corpus). Try the analysis various ways to see the effect on the graph (the distribution version has this option set, limiting analysis to 500 words).

Here is what the graph looks like initially for the data described above with Hapax eliminated, but graphing the full word list:
Obviously, this is impossible to read. To make it clearer, drag the corner of the graph to increase its size. If you like, you can also remove the labels for the primary and secondary texts. Just let your mouse pointer hover over the cloud of Stevens sections until a ‘Series “Stevens Sections” Data Labels’ box pops up. If you click this box, all the Stevens training sections labels will be selected. Pressing Del will delete these labels, while leaving the markers. Then do the same for the MacLeish labels. This makes the graph much easier to read. (I have left all the labels in by default because in smaller analyses you may want them displayed; Excel’s graphing function doesn’t seem to allow the use of a column of text data as labels, so adding them is tricky.) Resizing the graph, deleting the title, deleting the labels for the training texts, moving the legend to a clear spot on the graph, reformatting the Stevens Ind. Section labels to display to the left, rather than right, increasing the font sizes, and dragging a couple of labels further apart to make them more readable yields the graph below. (To move labels, click on any label, selecting all the labels for that group of texts; click on any of the selected labels again to select just that label; then drag the label to make it easier to read; right click on a label, then select “Format Data Labels” to change where it appears relative to its symbol.)

The horizontal axis in the graph shows the percentage of distinct words (types) in each section that are MacLeish marker words and the vertical axis shows the percentage of the distinct words (types) in each section that are Stevens marker words. Thus, for Stevens’s “Notes Toward a Supreme Fiction-2,” about 24% of the word types are MacLeish Marker words, and about 48% are Stevens marker words. The two “unknowns” (“The Woman on the Stair” and “Auroras of Autumn”) are strongly attributed to their correct authors, as are the independent poems for each
author. We might want to examine “First I Will Tell You,” which seems likely to be an uncharacteristic MacLeish poem, but note that even it has a percentage of MacLeish markers about twice that of most of the Stevens sections, while it has a significantly lower proportion of Stevens markers than the Stevens section with the lowest percentage. Finally, note that this graph is based on the 2806 most distinctive MacLeish markers and the 3423 most distinctive Stevens markers. If you want to base your graph on the 500 most distinctive words for each, simply change F2 and F3 to 500, or click the “Set/Clear Optional Max F2&F3” button before clicking the “Analyze & Graph” button. (Do not enter numbers into column G; numbers entered in column F flow automatically into G.) The graph will immediately change to reflect the new choice, as shown below (Note that the graph based on only 500 markers for each author yields more widely scattered author sections, but shows the test texts closer to the author clusters.):

You can gauge roughly what setting is most appropriate by changing F4, F5, F6, and F7 and seeing which number of marker words gives you the largest separation for the known texts (in columns H and following). If you want to avoid any chance of unconsciously slanting your analysis, you can do a graph first without the test texts, select the settings that give the best separation of the training sections and any independent texts, then add the test texts and create a new graph with the same settings.

The sheet also produces a space-optimized graph for each analysis. This graph shows the same information, but it adjusts the starting and ending points of the X and Y axes to eliminate extra white space and give more room for the data itself. Below is the space-optimized version of the analysis based on 500 marker words each:
Finally, in addition to authorship attribution, Wide-Spectrum analysis, is useful for characterizing the vocabularies of two authors with respect to each other. I have copied the 25 most distinctive MacLeish and Stevens words out of the spreadsheet:

MacLeish: surf, mouth, remember, till, slow, open, none, dawn, flesh, done, knees, girl, dusk, mouths, clean, god, girls, sand, stone, stones, wave, strange, eyes, smell, sweet
Stevens: self, eye, being, except, became, form, sense, single, speech, sounds, merely, reality, element, centre, part, clouds, yet, less, nature, central, colors, space, spirit, forms, final

Obviously, MacLeish’s list is much more concrete than Stevens’s. Note also the presence of word-families for each poet, even among the 25 most distinctive words. For MacLeish, mouth/mouths, girl/girls, stone/stones, dawn/dusk, surf/sand/wave; for Stevens, form/forms, centre/central. An analysis of larger sets of marker words can be very revealing. (Some authors also differ a great deal in where in the word frequency spectrum their characteristic words are found. You can examine this by looking at the ranks of the marker words in column F. The marker words for the first author will be at the top of column G, and those for the second author at the bottom.) Remember, however, that these lists are specifically tuned just to this comparison. It is technically possible for a MacLeish marker word with respect to Stevens to be an anti-marker word for MacLeish with respect to another poet. To get a more accurate “absolute” characterization of MacLeish’s vocabulary, you would want to run an analysis with MacLeish versus a large selection of texts by his contemporaries. Many other uses are possible.
Using the Delta Spreadsheets
© David L. Hoover, 2015

Introduction

Delta, introduced by John F. Burrows (2001, 2002, 2003), is a measure of textual difference that can be used effectively in authorship attribution investigations and stylistic studies. It is especially useful in investigations involving relatively large numbers of texts and authors, situations where other methods would be unwieldy. In order to do Delta analysis with the Delta Calculation Spreadsheet, I recommend that you begin by creating a mnemonically named subdirectory/folder on your computer and copy DeltaCalc2015.xls to the new directory/folder. (I’m sorry that I do not know if these spreadsheets will operate on Macs; the instructions below assume a PC platform running some version of MS Windows and MS Excel.) Depending on the version of Excel you have and how it is configured, you will probably need to enable macros for the spreadsheet to operate. Typically there will be a warning bar or message that will tell you there are macros and will allow you to enable them for this spreadsheet. I do NOT recommend that you set Excel to accept macros automatically, because macros are sometimes malicious programs, but the macros in the distributed spreadsheets are harmless. The instructions below are fuller versions of those available within DeltaCalc2015.xls, itself. To access the instructions in the spreadsheet, from the Excel “Tools” menu select “Macro” then “Macros”. Then “Edit” any of the macros and page up to the top for the instructions. I have now included a “DoItAll” button that activates the “DoItAll” macro; please read the explanation of this omnibus macro before running it.

I first wrote a version of these spreadsheets in 2003, and then discussed them in two articles (Hoover 2004a, 2004b), and made them available online. The early versions used Visual Basic macros to automate the process of comparing each text or sample to the samples of the candidate authors. The current versions have been extensively revised to incorporate a simplified Delta formula (Argamon, 2008), that drastically reduces the size of the spreadsheets and improves the speed of the macros. They also automate much more of the process of Delta analysis. The user can now enter a raw word frequency list for a whole corpus of texts and raw word frequency lists for each single text, and the spreadsheet can prepare the word lists for processing and perform a series of Delta analyses based on different numbers of frequent words, collect and analyze the results, and group the data for graphing. Putting all the functions except creating the word lists themselves into one self-contained package that can operate automatically should allow anyone interested in textual difference to perform multiple experiments easily, even if they are not particularly proficient in computer programming or computational stylistics. On the other hand, because the macros are open source, those who have the expertise (or are willing to learn it) can customize and revise them for other purposes or to improve them. The Delta Calculation Spreadsheets are under periodic development, and future enhancements will be available on these pages.

Getting Started

To use the spreadsheets, first collect a corpus of texts that interest you. Delta operates on a set of
primary samples of text by a group of authors and a secondary set of texts or samples, some by authors in the primary set, some by other authors. It was designed to suggest which of the primary authors is likeliest to have written an anonymous text known to be by one of them, but other kinds of investigations are also possible. One limitation of this kind of analysis is that you need enough texts by each author so that you can set aside some of them as a primary sample while also having some independent texts available to test whether the analysis is correctly identifying the samples of known authorship.

Collecting a Corpus

Let’s take a simple example. Assume that you have a novel that was serialized anonymously in a Victorian periodical, and that, further, you have good reason to believe that it was written by either Dickens, Thackeray, Collins, Eliot, Trollope, Meredith, Reade, or Gaskell. Fortunately, there are plenty of e-texts by these authors available, so you collect and prepare as many of these texts for analysis as is practical. Ideally, you would combine three or more texts by each primary author into a single primary authorial sample so that any idiosyncratic texts will not skew the analysis. You then select some additional novels by these same authors, including some that are similar in type and size to the anonymous text, to use as controls. Only if Delta is successful in attributing the known texts to their correct authors can you feel confident in accepting its attribution of the anonymous text. Recently, I have begun using multiple texts by the primary authors rather than combining them. If two or more texts by a single primary author are selected as the most likely authors, that provides a strong argument for a correct result and reveals additional information about the texts by the primary author.

Word Lists

The analysis proceeds as follows. First you produce a word frequency list in descending frequency order for the entire corpus of texts you have gathered, using one of the many programs available, including free tools available on the web, such as KWIC and the tools at the TAPoR project. In this case, you are dealing with large texts, so we’ll assume that you want to use a large number of words in your analysis (see below for more discussion). You place the 5,000 most frequent words of the entire corpus into the Delta Calculation Spreadsheet in column M, of the RawWordlists worksheet, beginning in row 3. (The various worksheets can be accessed by clicking on the tabs at the bottom of the window; you may need to use the arrows at the bottom left to move the tabs for some of the sheets into view.) Note that in the distributed version of the spreadsheet, there is a list of personal pronouns in column G of RawWordlists. These will be deleted from the word list before running the Delta analysis if you tell the spreadsheet to do so (as described below). Removing personal pronouns helps to reduce the effect of differences in point of view, the ratio of dialogue to narration, and differences in the genders of main characters. The hyphens and dashes are included in the list of personal pronouns in the distributed version because some programs (including KWIC), can be set to treat these as part of a word, and therefore as words in their own right under some conditions. Note that you can put any words you want to remove from the analysis in column G; they need not be pronouns. (The pronouns are also stored in column H for possible future use, and column H also includes
common contractions of these pronouns so that they can also be deleted if you like.)

You also produce word frequency lists in descending frequency order for each of the primary authorial samples (maximum 40) and for the samples to be tested (maximum 80). It is important to use the same tool with the same settings to create all of these lists, as different tools or settings can produce quite different results. The word lists for the primary authors are also placed in the RawWordlists worksheet, beginning in column O, with the author in O1, the sample name in O2, the first word in O3, its raw frequency in P3, and the rest of the words arranged below in column O, with their raw frequencies in column P. In the distributed version of the spreadsheet there is some data already entered so that you can see how it should be arranged. (Cells P1, P2 and the corresponding cells to the right of each author and title must remain empty.) The word lists from the texts to be tested are placed in the RawWordlists sheet in the same way in columns CS and following. When you have finished entering your word lists, make sure that columns O through IV in RawWordlists contains nothing except your word lists. CAUTION: If you need to remove unwanted data from the worksheets, simply select what you want to remove and press the “DEL” key or select “Edit” then “Clear” from the menu to clear the cells. Do NOT use “Edit” then “Delete” or otherwise remove whole columns or rows, and do not cut and paste rows or columns. This can cause havoc with the many formulas in the sheets. (It is all right to copy and paste rows or columns, then clear the original.)

The word lists can be entered manually, as described above. However, if your word lists are plain text files with the word in the first column, a tab, then the frequency in the second, you can use the FileImport macro to enter them automatically. This macro reads your list of files from the RawWordLists sheet in columns I through K of rows 17 and following. Simply place the abbreviation for the class of the file in column I (the abbreviations are listed above this area in the sheet), the actual file name in column J (no extension), and the author’s name in column K. The macro assumes your files are produced by KWIC, so that they are in the form “FILENAME.KWC”. If your files have a different extension, either rename them or edit the macro appropriately (there is a comment section near the beginning of FileImport explaining how to do this). The macro also assumes that the word list files are in the currently active directory; you can assure that Excel’s active directory is correct by navigating to the spreadsheet using the “File” menu, then selecting “Open” and opening the correct folder and then opening the spreadsheet. When you have finished entering your word lists, be sure that columns O and following in RawWordLists contains nothing except your word lists.

If you wish, you can produce parallel word lists in another program, such as Hugh Craig’s Intelligent Archive, and place them directly in the “WordlistsPercentagesForAnalysis” sheet, ignoring the “RawWordlists” area. If you do this, you will NOT run the “FileImport” and “FindandRecordFrequencies” macros; be sure to comment them out if you are using the “DoItAll” macro.

**Setting Parameters**

Now you select the DeltaCalculationWorksheet and set some parameters before running the
analysis. In the cell E3, enter “Y” (not case sensitive; no punctuation) if you want the personal pronouns (or whatever words you’ve selected) to be removed from the analysis. If you are unsure, leave this blank initially and run the analysis both with and without them to see which produces the most accurate results on the known texts in your corpus. In cell E4 you can enter a percentage for my automated culling process, which removes words that are very frequent in the entire corpus only by virtue of being extremely frequent in a single text. Although it may seem counterintuitive to remove such distinctive words, and although it occasionally does remove words that are very characteristic of a single author, the words that it normally removes are the names of characters and common place names. So the words “Dombey” and “Chuzzelwit” will disappear, in spite of being distinctive names of Dickens characters, but note that such words are not characteristic throughout Dickens. I have found that a culling percentage from about 60 to 80 percent usually improves the accuracy of the analysis. Again, if you are uncertain about this, leave E4 blank and run the analysis first without culling and then again with culling, or run it with two or three different percentages.

The only other parameters than can be set are the number of words to process and the word list size. Taking first the word list size, set in E6, you should choose a size appropriate for your corpus. With small texts, the 1000 or even the 500 most frequent words may be appropriate, and the smaller the word list, the faster the operation of the program. On the other hand, if you make the list too small, you may not get the most accurate results that are possible. In the hypothetical analysis we are discussing, a large word list is appropriate, so you would probably want to set the word list size at about 4500. (The analysis can only handle 4000 words, but some will be deleted because they do not occur in any of the primary samples and more will be deleted if you remove pronouns and/or cull the word list.) Note that with a large group of large novels, the analysis may take quite a long time, and you will not be able to use Excel for anything else while the analysis is running. One option is to run the analysis when you will be away from your computer for an extended period of time. (I like to run a trial analysis first with a very small word list, say 10 words, just to be sure that I have set everything up correctly.)

The number of words you want included in your analysis is set in E5 (obviously limited by the size of the word list in E6). In most authorship attribution and computational stylistics work in the past, the number of words to be studied was set arbitrarily at the beginning, and, until recently, the number tended to be small, from the 30 to the 100 most frequent words. The logic behind this choice was the reasonable assumption that an authorial wordprint, analogous to a fingerprint, is most likely to exist, if anywhere, among the most frequent words of the language, almost exclusively function words. It was assumed that authors do not and probably cannot regulate the frequencies of such words consciously, so that their relative frequencies result from deeply ingrained linguistic habits that are more likely to be stable across different subject matter, point of view, setting, and so forth. In many analyses from 2001 to the present, using Delta, Cluster Analysis, and PCA, however, I have shown that increasing the size of the word list almost invariably improves the accuracy of authorship attribution methods based on frequent words (Hoover, 2001, 2002, 2003a, 2003b, 2004a, 2004b, 2007, 2008; Hoover and Corns 2004). Because of these studies, the Delta Spreadsheets allow the processing of up to the 4000 most frequent words. In spite of the fact that nearly all of these are content words, the accuracy of
analyses of large texts like novels often improves steadily up through the 2000 most frequent words, and sometimes higher, and often remain at their most accurate levels all the way up to the 4000 most frequent words (at which point the selected words typically account for more than 90% of all the words in the texts). Again, however, the spreadsheets allow you to run several analyses based on different numbers of words so that you can base your conclusions on the number of words that give the most accurate results for the known texts in your analysis.

**Processing the Raw Word Lists with “FindandRecordFrequencies”**

Once you have entered your word lists and set the parameters, you are ready for a Delta analysis. I will describe later a way of automating the process so you can do it with a single click, but will first discuss the required and optional steps separately. Note that they must operate in the order indicated. First, you run FindandRecordFrequencies to turn the raw frequency lists into percentages and enter records with a zero frequency for each word in the whole corpus word list that does not occur in any given sample. This is required so that the statistics can compare the frequencies of each word in every sample. After the lists are processed, the macro runs RemoveWordsNotFoundInMainSet to eliminate words that are not found in any of the primary samples. This is required so that means and standard deviations can be calculated; otherwise, division-by-zero errors are produced and the analysis cannot proceed. The macro then enters the data into the WordlistsForAnalysis sheet and the WordlistsPercentagesComplete sheet. The former is where further processing of the word lists is performed, and the latter stores the full word lists so that they can be reused without running this time-consuming macro again. (To run another analysis on the same data, simply copy the data from WordlistsPercentagesComplete into WordlistsForAnalysis.) FindandRecordFrequencies then checks E3 to see if you have indicated that you want pronouns deleted; if so, it runs DeletePersonalPronouns (described below). This macro, like the others described below, is run from the Excel “Tools” menu by selecting “Macro” then “Macros” then selecting FindandRecordFrequencies and clicking “Run”. Note that the data in WordlistsPercentagesComplete or WordlistsForAnalysis can also be used to run PCA or Cluster Analysis, so that you may want to use Delta to prepare your word lists even in cases where Delta analysis itself is not desirable or not appropriate.

**(Optionally) Deleting Personal Pronouns**

Next you can now decide whether or not to run the DeletePersonalPronouns macro (if you have specified it in E3, FindandRecordFrequencies will have already done it). If you run it, it will delete from the WordlistsForAnalysis sheet the list of personal pronouns (or other words) you’ve listed in the RawWordlists sheet. It will also enter “Y” into E3, so that the results will be properly labeled as coming from an analysis without personal pronouns. Again, you may want to run the analysis twice: once including personal pronouns and once without them.

**Moving the Data to the “DeltaCalculationWorksheet” for Delta Analysis (after Optionally Culling the Data)**

Next you must run CullThenEnterDataCullFullWordSet, which will optionally cull the data of
words that are frequent only in one text and will then enter the data into the DeltaCalculationWorksheet for further analysis. If there is a number in E4, the data will be culled at that percentage, so that any word with a frequency that exceeds that percentage in any single text will be removed from the analysis. If E4 is empty, all the remaining words will be retained and entered into the main sheet. Note that if the number of words entered by this macro is less than the number you have set in E5, this macro will automatically set E5 to the number of words that are available.

Analyzing the Data with “Processit” and “DeltaAnalysis”

The next macros are the ones that actually perform the Delta analysis. The first is Processit, which, along with its sub-macro, DeltaAnalysis, performs a Delta analysis based on the requested number of words. It begins by checking to make sure you haven’t entered a number of words to process that is larger than the number available; if you have, it reduces the number requested to match what is available. The distributed version assumes that you are doing multiple analyses, so it checks to see if the number requested or available is an even multiple of 100, runs the analysis once, and then makes the next number an even multiple of 100 for tidier results. It also assumes an initial analysis of the 4000 most frequent words, and repeats the analysis at set intervals: if there are more than 1000 words processed in the first analysis, it subtracts 500 and reruns the analysis; when the number drops to 1000 or below, it reruns the analysis at 200-word intervals until it runs out of words. Thus, in the distributed version, Processit will run analyses based on the 4000, 3500, 3000, 2500, 2000, 1500, 1000, 800, 600, 400, and 200 most frequent words. The logic of this procedure is that having a sequence of analyses based on a decreasing number of most frequent words allows you to determine the most accurate level of the analysis, which is where you would look for the most likely attribution for your anonymous text. The size of your texts and the characteristics of your authors will obviously affect the results, but several analyses somewhere in the middle of the range typically are equally accurate, suggesting that the results are consistent and that the attribution of the unknown text is likely to be reliable.

You may want to edit this macro so that it operates differently. If so, from the Excel “Tools” menu select “Macro” then “Macros”. Then select Processit and choose “Edit”. (Note that the editor automatically saves any changes you make even if you just close the editing window, so be careful.) For example, if you have smaller texts, you may want to specify 1000 as the number of words to process in E5, 1200 as the word list size in E6, and set Processit to run at intervals of 100 words, so that it produces analyses of the 1000, 900, 800, etc. To do this, change wrds = wrds – 200 to wrds = wrds – 100, or, if you are starting with, say, the 1200 most frequent words, you may want to set the increment for more than 1000 words to 100 as well, so that it doesn’t drop from 1200 to 700 in the first step. To do this, change wrds = wrds – 500 to wrds = wrds – 100.

I suggest that if you edit this macro, you should leave the original lines there, but “comment out” any line you want to change by adding a single quotation mark at the beginning, then create your new version just below it. This will allow you to get back to the original version if you want to later. If you want to run a single analysis rather than a sequence, simply comment out all the lines
following the line wrds = Range(“WdsToProcess”) except for a single instance of Application.Run “DeltaAnalysis”. Note that the macro resets the number of words to process to 4000 after it is finished, so you may want to edit that number for smaller analyses.

**Analyzing the Results with Analyze and CollectInfo**

The last two major macros are normally run only when you are finished with all the sets of words you are interested in. Analyze does some statistical analysis on the results and checks to see how many of the known texts are correctly identified. It collects a summary of your results at the top far right of the DeltaCalculationWorksheet. It also marks each attribution of a known text as correct or indicates the rank of the known author, and marks each text that is anonymous or not by any author in the primary set as “Not in Set”. Texts not by any author in the primary set are normally included so that you can see how the attributions of the anonymous texts compare to attributions known to be false. Remember that Delta shows which author is the likeliest author of the text, and one of them must always be the likeliest (or least unlikely). Finally CollectInfo sorts and formats the results and places them in a block to the right of the results for use in producing a “members, errors, others” graph of the sort in Burrows (2002, 2003).

**Setting up “DoItAll” for Automated Analysis**

Rather than running each of these macros manually from the Excel menu, it usually makes sense to edit the macro DoItAll, which allows you to automate the process flexibly. In the distribution version, this macro looks like this:

Sub DoItAll()
    ‘ Application.Run “FindandRecordFrequencies”
    Application.Run “RemoveWordsNotFoundInMainSet”
    Application.Run “DeletePersonalPronouns”
    Application.Run “CullThenEnterDataCullFullWordSet”
    Application.Run “Processit”
    Application.Run “Analyze”
    ‘ Application.Run “CollectInfo”
End Sub

Note that FindandRecordFrequencies and CollectInfo are both “commented out”. The first can take a very long time to run, and is normally run only once at the beginning of an analysis, so I’ve assumed you’ve run it once already. Just delete the single quotation mark to run it as well. The last is commented out because it is not always used; for example, if you are doing preliminary analyses and do not intend to graph the results, you may not want to run this macro.

Here’s a version of DoItAll that does everything, including running the analysis both with and without personal pronouns and with two different culling percentages:

Sub DoItAll()
    ‘ set the culling percentage to 0 initially
Range(“CullPercentage”) = 0
' CullPercentage is the name of cell E4
' set the sheet so that it does not delete pronouns
' DeletePersPro is the name of cell E3
' this step is required because FindandRecordFrequencies automatically deletes them if
' E3 contains “Y”
Range(“DeletePersPro”) = “”
Application.Run “FindandRecordFrequencies”
Application.Run “RemoveWordsNotFoundInMainSet”
Application.Run “CullThenEnterDataCullFullWordSet”
Application.Run “Processit”
' next run it again, deleting the pronouns
Application.Run “DeletePersonalPronouns”
Application.Run “CullThenEnterDataCullFullWordSet”
Application.Run “Processit”
' next run it again, culling at 80%
Range(“CullPercentage”) = 80
Application.Run “CullThenEnterDataCullFullWordSet”
Application.Run “Processit”
' next run it again, culling at 60%
Range(“CullPercentage”) = 60
Application.Run “CullThenEnterDataCullFullWordSet”
Application.Run “Processit”
' now analyze and collect the data for all of the above analyses
Application.Run “Analyze”
Application.Run “CollectInfo”
End Sub

This sequence of analyses would take a considerable amount of time, and is one you would probably want to run it while you are doing something else, but it has the advantage of giving you a large number of analyses to consider. Assuming you begin with the 4000 most frequent words, for example, this version of DoItAll would produce 44 separate analyses, 11 based on the whole word list, 11 on the list with personal pronouns deleted, 11 with pronouns deleted and culled at 80%, and 11 with pronouns deleted and culled at 60%. If the anonymous text were strongly linked to a single author in all of these analyses, that would constitute a strong argument for its authorship.

**Delta Prime**

Note that I have not given instructions for my proposed variants on Delta (Hoover 2004b), but they are used the same way. I have updated only the two consistently best variants, Delta-Lz and Delta-Oz in 2008, which are available on these pages. DeltaCalc2015.xls contains instructions and comments on the macros.

**References:** for the references in the text above, see the web site.
Using the Analyze Textual Divisions Spreadsheet

© David L. Hoover, 2015

Introduction

This spreadsheet grew out of my annoyance with the tedium and error-prone nature of the process of separating parts of a text for analysis. John F. Burrows’s classic Computation into Criticism (1987) analyzes the dialogue of Jane Austen’s characters, and it would be valuable to follow up this work with examinations of other authors and other kinds of textual variation (see “For Further Reading,” below, for examples). Yet it is extremely tedious and time-consuming to go through a long novel and copy and paste out all the dialogue of all the major characters, saving each character’s dialogue in a separate file for analysis. I have, in the past, used a macro to remove dialogue, but it is a rare electronic text that has all quotation marks paired correctly, and this still does not separate the dialogue of the characters. It is also difficult to know what to do with the speech markers, such as “she incoherently murmured”. They obviously are not part of the dialogue, but if one is considering an analysis of the dialogue versus the narration, they do not seem like narration either. Finally, what frequently happens is that, after doing all of this work, one has forgotten the basis of the earlier decisions or has begun to question them, which leads to further complicated and difficult work.

The Analyze Textual Divisions Spreadsheet is an attempt to improve, simplify, and rationalize the process of analyzing divisions of any kind within a text. It can be used to separate the dialogue of the various characters in a novel or play, to separate the letters of an epistolary novel and label them by writer and addressee, to separate the books, sections, chapters, and so forth of a novel, or to separate the multiple narrators of a text. It can be used to separate any parts of a text that you think may show a difference from some other part. The spreadsheet contains a macro that operates on a very simple set of markup characters that you enter into the original text. Simply put, the macro reads each line of text, labels that line with the appropriate divisions and category labels to allow the text to be sorted by the various divisions and categories, and copies it into a new column. Although the macro removes the markup characters, you’ll obviously want to keep an unmarked copy of your text in reserve.

To use the Analyze Textual Divisions Spreadsheet, I recommend that you begin by creating a mnemonically named subdirectory/folder on your computer and copy AnalyzeTextualDivisions20152015.xls to the new directory/folder. (These spreadsheets should also operate on Macs, but the instructions below assume a PC platform running some version of MS Windows and MS Excel (2003 or later). Depending on the version of Excel you have and how it is configured, you will probably need to enable macros for the spreadsheet to operate. Typically there will be a warning bar or message that will tell you there are macros and will allow you to enable them for this spreadsheet. I do NOT recommend that you set Excel to accept macros automatically, because macros are sometimes malicious programs, but the macros in the distributed spreadsheets are harmless.

Once you have divided your text, you will probably want to analyze the various divisions, genres, categories, etc. Some tools and methods for doing this can be found elsewhere on these pages.
The Markup System (customizable)

I have tried to keep the markup as simple as possible and as flexible as possible while being powerful enough for use on texts with complex structures. Here is the entire set of markup characters:

- `<1>` text division level 1
- `<2>` text division level 2
- `<3>` text division level 3
- `<4>` text division level 4
- `[ ]` Letter writer
- `{ }` Letter addressee
- `/` new speaker (character)
- `\` speech marker
- `>` copy without processing
- `^` special character follows

Four major text divisions are allowed, marked as `<1>`, `<2>`, `<3>`, `<4>`, each followed by the name of the current division. The distributed version was designed with Wilkie Collins’s complex novel *No Name* in mind. That novel is divided into scenes and chapters, but also contains letters and other documents, so I chose “Scene”, “Chapter”, “Letter”, and “Document” as divisions. (A brief marked-up excerpt from this novel is included in the SampleText tab of the spreadsheet.) An epistolary novel might only have “Letter”. Other novels might use “Volume”, “Book”, and “Chapter”; “Narrator” might be used for a multiple-narrator situation. For plays, “Act” and “Scene” would likely be used. The macro can be modified to process texts with other divisions simply by entering the new label in the AnalyzeTextualDivisions2015 macro that runs the analysis. To do this, click on the “View” menu, select “Macros,” then “View Macros” and then “AnalyzeTextualDivisions2015,” then “Edit” and make your changes in the lines below:

```java
div1name = “Scene”
div2name = “Chapter”
div3name = “Letter”
div4name = “Document”
```

For example, for a novel divided into books, you could just change the first line to this: `div1name = “Book”`. Alternatively, you can just run the macro and then change the labels at the top of the sheet to whatever seems appropriate.

**Note: when modifying the macros, it’s a good idea to SAVE YOUR FILE FIRST!**

The four divisions are hierarchical, so that a new `<1>` division ends any of the others, a `<2>` ends any `<2>`, `<3>`, or `<4>`, and so forth. You can handle the dreaded overlapping hierarchies by restarting any division that continues through a “higher” division change. If any division ends without another immediately beginning, you must end it by inserting a blank instance of that
division level. For example, in the No Name sample, the first scene (level 1) is divided into chapters (level 2), and chapter XIII contains a letter (level 3), after which chapter XIII continues. Note that I have entered a <3> division at the end of the letter on a blank line. This begins a new level 3 division, theoretically a letter, but with a blank name, effectively ending the letter (see lines 49-55).

Square brackets enclose the writer of a letter and curly braces enclose the addressee; these can be placed anywhere in the line where a new letter begins. (See the letters in the SampleText tab of the spreadsheet for examples.) The forward slash “/” marks a new speaker (character); for example, within a letter, novel dialog, or drama. The back slash “\” indicates the beginning of a “speech marker,” a word or phrase introducing, characterizing, identifying dialog. So, for example,

“Why not?” John said.
is marked as follows:

/John””Why not?”
\John said.

Here, /John identifies the (new) speaker, and the quotation mark has its normal function of introducing the quotation (you may insert a space before the quotation mark if you like, for ease of reading). The line break ends the quotation, and the backslash indicates the speech marker, which is also attributed to John (on the theory that you might be interested in whether or not the speech of different characters is marked differently). If the speech marker precedes the quotation, it is marked like this:

/John\Then he said,
“Why not?”

Note that a speaker identification must be followed by either dialog or a speech marker, and that the dialog or marker can continue (all three lines below are attributed to Mrs. R):

/Mrs. R””Excuse me,”
\she said,
“how clumsy of me.”

Note the line breaks, which are required for processing, for clarity, and to avoid having to add special characters to end dialog or speech markers. This method retains the quotation mark and, to an extent, the line break, as traditional print formatting devices for dialog. Note that the macro ignores single quotation marks within quotations, and will not correctly process texts with single quotation marks for all quotations; double quotation marks within quotation marks will generate errors. The macro can be modified to make the single quotation mark the default (change the line QuoteCh = “””” to QuoteCh = “””), but you will also have to edit your text to replace apostrophes with another character. Use caution if you do this, as you will probably want the character you
use to replace the apostrophe to be included in a word, and many word list generating programs either assume all single quotation marks are apostrophes and/or do not allow you to choose characters that do not break words.

Two other characters complete the markup. The “>” tells the program to copy the text on the line without further processing and it must appear at the beginning of a line. This is useful for metadata and for any text that is not really part of any of the textual categories you want to analyze. For example, in an epistolary novel, you will usually not want to analyze the indications of addressee, like “To Mrs. Richman” or indications of the location of the writer, and/or the time, like “New Haven, Sunday Morning, 9:00AM”, or the “signature” of the writer. Placing a “>” before them tells the program to copy them without characterizing them. (Conversely, if you believe that location might be significant for your text, you should encode it with one of your four major text divisions.) Finally, “^” is required before any line that begins with a character that Excel treats specially. Aside from the quotation mark (discussed below), the only problem characters I know of are “+”, “-“, and “=” (the hyphen and dash, especially, are problematic, as they can be forced to the beginning of a line by the line breaks required by dialog. The “^” tells the program to put a space before the special character to avoid an error. (Unfortunately, simply adding a space before the offending character in the text itself does not work.)

If you don’t like my selection of characters, or if one or more appear regularly in your text, you can change them. For example, to use “*” to mark special characters, rather than “^”, just replace the macro line SpecialCh = “^” with SpecialCh = “*” and so forth. CAUTION! Be careful to select characters that are not otherwise found in your text, and avoid duplicates. I have used “<1>” for the first major division, for clarity (“<1Letter I” is hard to read), and have also used the closing part of this markup, “>”, to indicate that the line should be copied without processing, but the program removes all three characters of the division marker without processing the final “>”. So you could encode “<1+” or use any other character you like to end the division marker. The same is not true of the other characters.

**Note: when modifying the macros, it’s a good idea to SAVE YOUR FILE FIRST!**

The marked-up text must be saved as Plain Text (ASCII). For example, in MS Word, select the Office Button, then Save As, choose Plain Text, then, under “Options” select “Insert line breaks” and “End lines with” CR/LF. For English text, the “Windows (Default)” encoding is fine; for foreign languages with accented characters and for non-western alphabets, you may need to do some experimentation to find an encoding that works.

**Analysis**

You might want to begin by analyzing the sample from *No Name* that is included in the spreadsheet. Simply select the SampleText sheet by clicking on its tab at the bottom of AnalyzeTextualDivisions2015, then find the Macros tab under View, select “View Macros,” select the “AnalyzeTextualDivisions” marco and select “Run”. You can then try sorting and examining the results to see whether this macro will give you the results you want before
marking up your own text.

The simplest way to analyze your own text is to open AnalyzeTextualDivisions2015.xls, enable macros, if you are asked to, then open the plain-text marked-up file. When the “Text Import Wizard” pops up, select “Fixed width”, then click “Next”. Make sure there are no column breaks, or delete any that appear. IMPORTANT: be sure to change the “Text qualifier:” to {none}. Then click “Finish”.

Note: no one ever said Excel was ideal for text analysis. One of its annoying peculiarities for this purpose is that it treats the quotation mark by default as a text qualifier, so if you import without changing the “Text qualifier”, or if you copy and paste text into Excel, the program is likely to delete any initial quotation marks and truncate any quotation at the end of its first line if it does not end in a quotation mark.

Before running the AnalyzeTextualDivisions macro, you should resize the first column so that your text fits within it, for clarity. Next, find the Macros tab under View, select “View Macros,” select the “AnalyzeTextualDivisions” marco and select “Run”. The macro will analyze the text, then run an error-checking routine that will find many kinds of markup errors (most novels will be analyzed in a minute or less). Resizing the column with the analyzed text will make it easier to find any error messages that the program generates (they are entered in the columns to the right of the analyzed text). It is extremely unlikely that your markup will be completely correct the first time. It will work best to correct any problems in the original text file and run the analysis again (and again), continuing until you have the results you want. Note that the macro removes initial and trailing spaces.

Finally, sort the file in any way you like to isolate the parts of the text in which you are interested. For example, if you are working on differences in the dialog of various characters, you might sort on the Type (narrative, dialogue, marker), then on the Speaker, then on the Line number, to get all of each character’s dialog together in text-order, then select chunks of appropriate size and copy them to separate files for analysis. Alternatively, you can save each character’s dialogue as a single file and then divide each file into equal segments. You can go back to the original form by sorting on line number, and then sort in another way for other purposes.

For Further Reading:

- McKenna, W. and A. Antonia. 1996. “‘A few simple words’ of interior monologue in


Note: All of these instructions assume that you have, or can create, a set of parallel word lists for a group of texts. That is, that there is a single master list of words, and that each text has an entry for each word, whether or not the word is found in that text. One simple method for creating these word lists is to use the Delta Spreadsheet (available, with detailed instructions on this site). The spreadsheet also lets you delete personal pronouns, and/or remove words found in the word list only because they are extremely frequent in one text (see the instructions, and below, for more discussion). A more elegant and much faster method is to use Hugh Craig’s Intelligent Archive (available here with instructions). Once you have created your word lists, you should have data that looks like that below.

### 1. Data Preparation: before you move the data into Minitab for analysis, you need to prepare it a bit.

If you are using Excel 2007 or later, you can prepare the data in Excel. To use the words in your analysis as column labels, first replace any apostrophes with another character; Minitab will not allow apostrophes in labels. Click the grey cell above row 1 (with just “A” in it) to select the entire column, then use Ctrl-F (or Find & Select then Replace, from the Home menu) to find all apostrophes and replace them (I use “^”). Now select the data; in the sample above, click cell A1, then scroll down to row 1003, press and hold the Shift key and click cell K1003. Next, copy the data with Ctrl-C, then click on cell L1, select Paste from the Home menu, then Past Special. When the Paste Special dialogue box comes up, select Transpose (just above the OK button) and then click OK. Your data will now be transposed so that the authors and titles are on the left and the words are along the top. (You can transpose more than 1000 words, but Minitab can only analyze about 1000 at a time in cluster or PCA; t-testing often involves larger numbers of words.) Once your data is transposed, select the data again; here, click cell A1, then scroll down to row 1003, press and hold the Shift key and click cell K1003. Next, copy the data with Ctrl-C, then click on cell L1, select Paste from the Home menu, then Past Special. When the Paste Special dialogue box comes up, select Transpose (just above the OK button) and then click OK. Your data will now be transposed so that the authors and titles are on the left and the words are along the top. (You can transpose more than 1000 words, but Minitab can only analyze about 1000 at a time in cluster or PCA; t-testing often involves larger numbers of words.) Once your data is transposed, select your data again; here, click L1, then press and hold Shift and click ALZ11 and use Ctrl-C to copy the data. Go to Minitab and paste the data into a blank worksheet by clicking the unmarked grey cell below C1 and above row 1 and pasting the data with Ctrl-V. (If you need to create a blank worksheet in Minitab, go to the File menu, select New, then Minitab Worksheet.) Your data should now look like that below.

<table>
<thead>
<tr>
<th>I</th>
<th>A</th>
<th>B</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>M</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>the</td>
<td>5.624896</td>
<td>5.037437</td>
<td>5.571758</td>
<td>4.774714</td>
<td>5.573425</td>
<td>5.509689</td>
<td>5.393801</td>
<td>5.638296</td>
<td>5.179916</td>
<td>5.37596</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>of</td>
<td>2.247523</td>
<td>2.579021</td>
<td>3.253253</td>
<td>2.945589</td>
<td>3.022779</td>
<td>3.946233</td>
<td>3.267213</td>
<td>2.719088</td>
<td>2.947847</td>
<td>2.697771</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>to</td>
<td>2.259782</td>
<td>2.457278</td>
<td>2.857911</td>
<td>2.5927</td>
<td>2.211124</td>
<td>3.208593</td>
<td>3.219104</td>
<td>3.029083</td>
<td>3.271195</td>
<td>2.944985</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>and</td>
<td>2.796215</td>
<td>2.484587</td>
<td>2.517238</td>
<td>2.797949</td>
<td>2.46506</td>
<td>2.653841</td>
<td>2.607895</td>
<td>3.357056</td>
<td>3.802648</td>
<td>3.22261</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>2.664258</td>
<td>2.439544</td>
<td>2.392285</td>
<td>2.254179</td>
<td>2.215983</td>
<td>2.311047</td>
<td>1.978037</td>
<td>2.149754</td>
<td>2.028855</td>
<td>2.106224</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>her</td>
<td>1.29901</td>
<td>2.847637</td>
<td>2.831265</td>
<td>2.471331</td>
<td>1.716419</td>
<td>2.590522</td>
<td>2.210214</td>
<td>1.767495</td>
<td>2.296079</td>
<td>1.831541</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>she</td>
<td>1.001132</td>
<td>2.19467</td>
<td>2.033424</td>
<td>1.465804</td>
<td>2.006095</td>
<td>2.134455</td>
<td>1.747169</td>
<td>1.284067</td>
<td>1.841227</td>
<td>1.285044</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>in</td>
<td>1.621938</td>
<td>1.59686</td>
<td>1.691693</td>
<td>1.898507</td>
<td>1.683502</td>
<td>1.633856</td>
<td>1.545058</td>
<td>1.526081</td>
<td>1.554889</td>
<td>1.637302</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 2. Cluster Analysis
To do a cluster analysis of the data above in Minitab, select the Stat menu, then Multivariate, then Cluster Observations. In the Cluster Observations window that will pop up, first select the words you want to analyze, either by selecting them with your mouse in the box at the left or by typing the column numbers into the “Variables or distance matrix:” box (here, I’ve typed in “C1-C800” to analyze the 800 most frequent words). Next you need to select some options. It’s safe to use the ones shown below, where I’m analyzing the 800 MFW using Ward linkage, and Squared Euclidean distance; you should also select “Standardize variables,” which compensates for the rapid decrease in word frequencies in a word list. I’ve set the number of clusters to 4, the number of different authors in my set. This only changes the color of the lines in the graph; in this trial analysis with known texts, the colors tell me at a glance whether the four authors’ texts group properly. I have also selected Show dendogram, which produces the graph.

You should next click the Customize box so that you can enter a title for the graph and indicate what labels you want to use. The Customize box looks like this:
Make the title informative, so you can keep track of what you did. Select the “Case labels” by clicking in the box next to it and then selecting the correct column from the box on the left, or you can type the label into the “Case labels” box. Here, I’ve typed in “Text,” the label I’ve used for the titles of the texts; if you were interested only in whether the authors cluster properly, and if your titles do not include the authors as mine do, you could also use the authors as labels by entering or selecting “Author.” Now click OK and then OK again and the analysis is performed, in the sample above, it yields this:
Normally, you will want to do several analyses, starting with the largest number of words and working down. In this sample case, I normally start with the 990 most frequent words (which is about the most that Minitab can analyze at once), then do the 900, 800, 700, 600, 500, 400, 300, 200, 100. What you are looking for, and what you will usually find, is a series of analyses at some point that are very similar and which typically are also the most accurate for the known texts. To These are the most reliable results. (In the sample analysis here, the graphs are essentially identical for the 990-200 MFW words.) Note that this graph suggests that Glasgow’s texts show a chronological grouping, with the early texts to the left and the later ones grouping to the right.

Note: I’ve created a little macro for Minitab that runs all of these analyses automatically; it’s called “clall.mac”, and you can download it from the website, or try clallc.mac, if you want to do the same thing but use complete linkage rather than Ward linkage. You will have to put the macro file in the same folder/directory as your Minitab Project, and make sure Minitab is set to look for the file there. To do this, just go to File, select “Open Project”, change the “Files of type:” from “Minitab Project (*.mpj)” to “All files (*.*)” and try to open clall.mac. Minitab will give you an error message, but it will now be set to find the macro file in the correct folder/directory. To run the macro, click on the “Editor” menu at the top and make sure the “Enable Commands” box is checked. Then, at the “MTB >” prompt, type “% clall” or “%clallc” and the macro will run the multiple analyses. This is somewhat convoluted, but it is much faster and simpler than running the ten analyses manually.

3. PCA Analysis

For PCA analysis, prepare the word lists, copy them into Minitab and transpose them as above. If you want to do both Cluster Analysis and PCA, you will only need to prepare and transpose once; both analyses can be done from the same worksheet in Minitab. Select the Stat menu, then Multivariate, then Principal Components. In the Principal Components Analysis window that pops up, first select the words you want to analyze, either by selecting them with your mouse in the box at the left or by typing the column numbers into the “Variables:” box (here, I’ve typed in “C1-C800” to analyze the 800 most frequent words). Next you need to select some options. The “Number of components to compute:” should normally be 2, and you should select Correlation under “Type of matrix.”
Next, click the Graphs button and select “Score plot for first 2 components” and “Loading plot for the first 2 components” and click OK, and click OK again to perform the analysis.
When the graphs appear, the first one, the Score Plot, will plot the relationships between the texts, but it will not have any labels on the dots for the texts. To add these, simply right click on any of the dots, then select Add, and then Data labels. This will bring up the Add Data Labels box below:

In the example shown, I have already selected “Use labels from column:”, have clicked in the empty box next to it, and double-clicked on “Word”, which is the label of the column that contains the titles of the texts. Now when you click the OK button the labels will be added to the graph, giving this:
Often you will want to edit the title of any graph to make it more informative (just double click the title). If you want to change the font size for the labels in a graph (for example, if they are too big or too small), just right click any of the labels and select Edit Data Labels. You can change the font, font size, and the alignment of the text to get the graph to look better, or you can click on a label if you need to move it a bit so it’s not on top of another label. (Be careful with this last operation, as you can seriously mess up your results if you’re not.) You can also right click on the symbol itself if you want to change its size, shape, or color. Here, you’d want to move the labels so they don’t overlap. It often helps to maximize the graph so it fills the entire window for editing.

The second plot, the Loading Plot, displays the words and how they contribute to the positions of the texts in the Score Plot. Together, these two graphs constitute the classic “Burrows” analysis, in which you show both the graph with the texts and the graph with the words that produced the text graph. Because the Loading Plot is based upon the same 800 MFW as the Score Plot of the 10 texts above, the two can be compared with each other. This graph is not very readable because of the large number of words, but it shows that the position of Wharton’s texts on the left of the Score Plot is related to her relatively frequent use of words like near, drawn, usual, continued, and taken (which appear on the left of this graph). Clearly, to study large numbers of words in this way it would be more appropriate to do separate graphs for subsets of the words. (I have moved a few labels slightly to the left to make them more visible; you can also left click on a label and then format the labels by changing to a smaller font for better readability, though you
As with cluster analysis, you will normally want to do several analyses, starting with the largest number of words and working down. In this sample case, I would start with the 990 most frequent words (which is the most that Minitab can usually analyze without an error message), then do the 900, 800, 700, 600, 500, 400, 300, 200, 100. What you are looking for, and what you will usually find, is a series of analyses at some point that are very similar. That should be the most reliable results, and I would do the word cloud only for that number. Note that you should not try to interpret the directions on these graphs. What is important is the positions of the texts relative to each other: the further two texts are apart, the less similar they are. Compare the graphs below, based on the 600 MFW, to the Score Plot above:

Worksheet: Worksheet 3
Though these two graphs may seem quite different, note that they are almost mirror-images of each other, so that the relationships of similarity and difference in the two graphs are essentially the same. It is a curious but normal fact of this kind of analysis that, as in this example, the orientation shifts back and forth several times. All show essentially identical relationships of similarity and difference.

Note: As I did for cluster analysis, I’ve created a little macro for Minitab that runs analyses for the 990, 900, 800, down to 100, automatically; it’s called “PCAall.mac”, and you can download it from the website. PCA is quite slow for large numbers of variables and/or texts, so be prepared for this macro to take a minute or more. You can also try PCA100s.mac, if you want to do analyses for the 900-1000mfw, 800-900mfw, 700-800mfw, down to the 100mfw. This can be a useful way to see if the texts vary in different ways for different parts of the word frequency spectrum. It also has the benefit that each word plot displays only 100 words, and so is more readable than one based on the 990mfw. You will have to put the macro file in the same folder/directory as your Minitab Project, and make sure Minitab is set to look for the file there. To do this, just go to File, select “Open Project”, change the “Files of type:” from “Minitab Project (*.mpj)” to “All files (*.*)” and try to open PCAall.mac. Minitab will give you an error message, but it will now be set to find the macro file in the correct folder/directory. To run the macro, click on the “Editor” menu at the top and make sure the “Enable Commands” box is checked. Then, at the “MTB >” prompt, type “% PCAall” or “% PCA100s” and the macro will run the multiple analyses. This is somewhat convoluted, but it is much faster and simpler than running the ten analyses manually.
4. T-Testing

T-testing is an effective way of selecting a set of words that discriminate between two sets of texts. Typically, this method is used to find words that distinguish two authors, but it can also be used to distinguish between one author and a group of other authors, between groups of authors, between genres, chronological periods, and so forth. In fact, the technique is generally useful whenever one wants to determine differences that could not reasonably have occurred by chance. Any introductory statistics text will explain the calculations, so I will concentrate here on how to do a t-test analysis in Minitab. Let’s say I want to distinguish Wilkie Collins from his less important contemporary, Walter Besant. I could use T-testing to compare the frequencies of words across a large numbers of samples of writing by the two authors and determine what words are very differently distributed. These would be good marker words for the two authors and would be an appropriate set of words to use to test an anonymous text that might be by either of them.

T-testing works best with a large number of text samples for comparison, so I begin with four novels by Besant and three novels by Collins, with a total of more than 300,000 words for each author. I first produce a parallel word frequency list like the ones discussed above for the whole novels. Because t-tests compare frequencies in sections of novels in this case, words that occur in multiple sections of a single novel by one of the authors, but very rarely in any of the other texts by either author, will have statistically significant distributions. For example, George, a character in Besant’s *The Ivory Gate*, is the 574th most frequent word in the entire corpus, occurring 313 times in that novel, once each in two other Besant novels, and just once in one Collins novel. The t-test finds this distribution highly statistically significant, with a p value of .002, which means that this distribution could be expected to occur by chance only two out of a thousand trials (a p value of .05 is the standard cut-off). It would seem highly unwise, however, to suggest that any text containing the word George is by Besant rather than Collins (incidentally, George occurs 102 times in Collins’s *No Name*, which was not included in my samples).

One could do the t-tests and then eliminate proper names and other inappropriate words from the significant words, but it seems more reasonable to remove most of them before doing the tests, and to do this in an automatic way that doesn’t risk unconsciously biasing the analysis. The Delta Spreadsheet optionally does this, so I put the word lists for the whole novels into the spreadsheet and run the macro that removes all words for which a single text supplies more than 90% of the instances (see the instructions on Using the Delta Spreadsheets on these pages). This process occasionally removes a highly distinctive word, but it also removes most proper names and other words that are specific to a single text (*Heathcliff* from *Wuthering Heights* would disappear, as would *telescreen* from 1984.) Once I have culled the words with the macro, I use the new word list as a basis for the word lists for the next process. For example, I enter the culled word list into the RawWordlist sheet for processing the word lists for the sections of texts that are discussed below. However you create it, the t-test requires a set of parallel word lists for all the sections of the texts. If, as here, you have large texts, you can use sections of 2000-4000 words, or even more, so long as you have at least a dozen, but preferably many more sections (there’s a simple program for segmenting texts available at the website above). In the example here, I use 167...
sections of 4000 words each.

For the t-tests, I prepare the data as before and put it into Minitab; the only differences are that I’ve entered data for 3300 words and have added a column that has just two category labels in it, one for each author (I use “C” for Collins and “B” for Besant). Here’s a small section of the Minitab worksheet, which has columns for 3300 words in 167 rows:

<table>
<thead>
<tr>
<th>Word</th>
<th>Auth</th>
</tr>
</thead>
<tbody>
<tr>
<td>past</td>
<td>B</td>
</tr>
<tr>
<td>kising</td>
<td>B</td>
</tr>
<tr>
<td>island</td>
<td>B</td>
</tr>
<tr>
<td>invented</td>
<td>B</td>
</tr>
<tr>
<td>initials</td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>Auth</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
<tr>
<td></td>
<td>B</td>
</tr>
</tbody>
</table>

To do a single t-test, you can use the Minitab menus. Select the Stat menu, then Basic Statistics, then 2-Sample t. In the example above, you’d select Samples in one column, then enter, for example, “initials” (word number 3300) in the “Samples:” box and “Auth” in the “Subscripts:” box, leaving the rest of the defaults and clicking OK. You’ll get a result like this:

Two-sample T for initials
Auth N Mean StDev SE Mean
B 77 0.00100 0.00650 0.00074
C 90 0.00244 0.00922 0.00097

Difference = mu (B) – mu (C)
Estimate for difference: -0.001436
95% CI for difference: (-0.003849, 0.000977)
T-Test of difference = 0 (vs not =): T-Value = -1.18 P-Value = 0.242 DF = 159

Here, “initials,” with a p-value of .242 clearly does not have a significantly different distribution in the two authors. Obviously, doing this manually for 3300 words would be ridiculous, and I actually want to test the 6600 MFW. Fortunately, the process can easily be automated. First type (or cut and paste) the following commands into a file in any text editor or word processor and
save it as a plain text file (I call mine DoAllTtests.mac; the extension must be “.mac”), or simply download a copy here.

NOTE: In the macro below, I have inserted four periods (to simulate a TAB), wherever the text should be indented, and multiple groups of four periods when necessary. This avoids using backspace characters and other other “fancy” methods of indentation. Be sure to replace the periods with spaces before saving the macros.

GMACRO
....DoAllTtests
....#k1 is number of words to process
....#k2 is the column with the two group variables; e.g., B for Besant, C for Collins
....let k1=3300
....let k2=3301
....noecho
....DO k3 = 1:k1
......TwoT Ck3 ck2
....ENDDO
ENDMACRO

This macro will process all 3300 words in the sheet shown above. Before running this macro, select all the text in your session window with the mouse and delete it, so you have a clean window. Note that you will have to put the macro file in the same folder/directory as your Minitab Project, and make sure Minitab is set to look for the file there. To do this, just go to File, select “Open Project”, change the “Files of type:” from “Minitab Project (*.mpj)” to “All files (*.*)” and try to open DoAllTtests.mac. Minitab will give you an error message, but it will now be set to find the macro file in the correct folder/directory. To run the macro, click on the “Editor” menu at the top and make sure the “Enable Commands” box is checked. Then, at the “MTB >” prompt, type “% DoAllTtests” and the macro will run the multiple analyses. This produces 3300 t-tests, but they are not very useful because the words with statistically significant distributions are mixed in with the rest, so I have created an automated process that cleans up and sorts the results. First, select all the t-test results from the session window with your mouse and copy them into an empty Excel worksheet (the copy process in Minitab may take several seconds; be patient). The Excel macro below will remove a lot of useless data and sort the words in descending order of significance. Creating the macro is a bit different in Excel 2007 and earlier versions. In 2007, select Macros from the View menu, then click View Macros and select Edit; in earlier versions, from the Tools menu select “Macro” then Macros then Edit. Then either type in the macro below, or, better, cut and paste it out of this document.

NOTE: In the macros below, I have inserted four periods (to simulate a TAB), wherever the text should be indented, and multiple groups of four periods when necessary. This avoids using backspace characters and other other “fancy” methods of indentation. Be sure to replace the periods with spaces before saving the macros.
Sub cleanupttestresults()
  ....Dim drow, dcol As Integer
  ....Application.Run “cleanuperrors”
  ....Application.Run “trimoutput”
  ....Application.Run “formatandsort”
End Sub

Sub formatandsort()
  Range(“B1:B1”) = “T-value =”
  ....Range(“D1:D1”) = “P-value =”
  ....Range(“C1:C1”).Formula = “=value(MID(A2,FIND(“”T-V””,A2)+10,FIND(“”P-V””,A2)-FIND(“”T-V””,A2)-12))”
  ....Range(“E1:E1”).Formula = “=value(MID(A2,FIND(“”P-V””,A2)+10,FIND(“”DF””,A2)-FIND(“”P-V””,A2)-12))”
  ....Range(Cells(1, 2), Cells(20000, 5)).Select
  ....Selection.FillDown
  ....Columns(“A:E”).Select
  ....Selection.Copy
  ....Range(“G1”).Select
  ....Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks _
  ........:=False, Transpose:=False
  ....Application.CutCopyMode = False
  ....Selection.Sort Key1:=Range(“K1”), Order1:=xlAscending, Key2:=Range(“I1”) _
  ........, Order2:=xlDescending, Header:=xlNo, OrderCustom:=1, MatchCase:=False _
  ........, Orientation:=xlTopToBottom, DataOption1:=xlSortTextAsNumbers, _
  ........DataOption2:=xlSortTextAsNumbers
  ....Columns(“A:F”).Select
  ....Range(“F1”).Activate
  ....Selection.Delete Shift:=xlToLeft
  ....Range(“A1”).Select
  ....Selection.End(xlDown).Select
  ....ActiveCell.Offset(1, 0).Activate
  ....drow = ActiveCell.Row
  ....dcol = ActiveCell.Column
  ....Range(Cells(drow, dcol), Cells(drow + 10000 – drow, dcol + 4)).Select
  ....Selection.Delete Shift:=xlUp
End Sub

Sub cleanuperrors()
  ....With Range(“A1:A65536”)
  .........Set c = .Find(“* ERROR *”, LookIn:=xlValues, LookAt:=xlPart, _
  ........SearchOrder:=xlByRows)
  .........If Not c Is Nothing Then
  ............Do
  

When you are ready to process the data, select the worksheet that contains your t-test results, which should look something like this:
Now run the macro by going to Macros in the View menu, then View Macros, then select cleanupttestresults and click Run (in older versions of excel, go to the Tools menu, select Macro then Macros then select cleanupttestresults and click “Run”). This macro may take several minutes (during which Excel will not be usable for any other purpose), but it will produce something that looks like the following:
The list is sorted by p value, then by T-value, so that you can select whatever p value you would like. In the example here, there are 535 words significant at the .001 level, with a chance of less than one in a thousand of occurring in this pattern by chance. For purposes of discriminating the authors, these very distinctive words may be superior to a larger set of less significantly distinctive words, but you may also want to look at the larger list of more than 1500 words that are significant at the .05 level, especially if you are interested in studying the kind of vocabulary that differentiates the two authors.

The positive T-values are for words more frequent in the first author (whichever one is alphabetically first in column 3301, row 1), and the negative are for the other author. You can separate the two groups of words by sorting on the T-value. Then, if you put them into separate sets of columns, you can sort on the p value as the first level, then the T-value, to put the words in order of distinctiveness for each author (be sure to sort the positive ones from largest to smallest T-value and the negative ones from smallest to largest). Here are the 20 most distinctive words among the 3300 tested for Besant and Collins:

|   | A         | B   | C    | D    | E    | F    | G    | H    | I    | J    | J    | K    |
|---|-----------|-----|------|------|------|------|------|------|------|------|------|------|------|
| 1 | upon, Auth| T-value = 19.17 | P-value = 0 | answered, Auth | T-value = -18.6 | P-value = 0 |
| 2 | all, Auth  | T-value = 17.56 | P-value = 0 | to, Auth | T-value = -17.56 | P-value = 0 |
| 3 | but, Auth  | T-value = 15.92 | P-value = 0 | had, Auth | T-value = -13.04 | P-value = 0 |
| 4 | them, Auth | T-value = 15.71 | P-value = 0 | miss, Auth | T-value = -12.5 | P-value = 0 |
| 5 | and, Auth  | T-value = 12.19 | P-value = 0 | on, Auth | T-value = -11.86 | P-value = 0 |
| 6 | not, Auth  | T-value = 11.97 | P-value = 0 | asked, Auth | T-value = -11.56 | P-value = 0 |
| 7 | or, Auth   | T-value = 11.75 | P-value = 0 | in, Auth | T-value = -10.38 | P-value = 0 |
| 8 | very, Auth | T-value = 11.68 | P-value = 0 | miss, Auth | T-value = -10.37 | P-value = 0 |
| 9 | so, Auth   | T-value = 11.65 | P-value = 0 | mind, Auth | T-value = -9.52 | P-value = 0 |
| 10 | because, Auth | T-value = 10.91 | P-value = 0 | suggested, Auth | T-value = -9.05 | P-value = 0 |
| 11 | great, Auth | T-value = 10.83 | P-value = 0 | person, Auth | T-value = -8.87 | P-value = 0 |
| 12 | thing, Auth | T-value = 10.17 | P-value = 0 | resumed, Auth | T-value = -8.75 | P-value = 0 |
| 13 | things, Auth | T-value = 10.17 | P-value = 0 | excuse, Auth | T-value = -8.69 | P-value = 0 |
| 14 | much, Auth | T-value = 10.08 | P-value = 0 | left, Auth | T-value = -8.67 | P-value = 0 |
| 15 | every, Auth | T-value = 9.98 | P-value = 0 | at, Auth | T-value = -8.4 | P-value = 0 |
| 16 | there, Auth | T-value = 9.86 | P-value = 0 | reminded, Auth | T-value = -8.25 | P-value = 0 |
| 17 | man, Auth  | T-value = 9.75 | P-value = 0 | creature, Auth | T-value = -8.04 | P-value = 0 |
| 18 | everything, Auth | T-value = 9.73 | P-value = 0 | inquired, Auth | T-value = -7.97 | P-value = 0 |
| 19 | is, Auth   | T-value = 9.72 | P-value = 0 | reply, Auth | T-value = -7.95 | P-value = 0 |
| 20 | well, Auth | T-value = 9.48 | P-value = 0 | when, Auth | T-value = -7.68 | P-value = 0 |

Besides being very good for discriminating these two authors, the lists of words have other interesting characteristics. For example, notice thing and things in Besant’s list and answered, asked, suggested, resumed, and inquired in Collins’s list. These are just the beginning of many families of morphologically or semantically related words that may be of interest stylistically or thematically. Putting the entire list back together and sorting it alphabetically will show many families of related words, and will let you see at a glance whether all members of a family are markers for the same author. For example, belong, belonged, belonging, and belongs are all Besant marker words, while discovered, discoveries, discovering, and discovery are all Collins marker words.

In my study of Collins and Besant, I created two lists of 3300 words, ran the t-tests on both lists, then combined the words with statistically significant distributions into one large list of about
1700 words p < .05, for further study.

For Further Reading:

- For extensive use of Cluster Analysis, see also my “Multivariate Analysis and the Study of Style Variation,” *Literary and Linguistic Computing* 18: 341-60.
Computer-assisted textual analysis has a long, rich history, despite the fact that, as has often been noted, it has not been widely adopted in contemporary literary studies. Instead of debating the causes for this neglect, I will concentrate here on computational methods that can be of use in many different kinds of literary research (for two contrasting views, see Ramsay; Hoover, “End”). I would argue that almost any literary study can benefit from at least some modest and basic kinds of computer assistance. For example, it would seem perverse not to use an available digital text of a work for searching for a vaguely remembered passage that is important for an argument or for locating every significant example of a word or phrase, and studying a concordance remains an effective method for understanding a text. In these cases, the computer is valuable despite the fact that one could perform the activities without it. When the collection of texts is larger or the items to be investigated occur more frequently, however, it becomes impossible to perform the work without a computer (imagine studying personal pronouns in one hundred Victorian novels). Many kinds of evidence produced by statistical methods are simply not accessible without a computer. I will argue with John Burrows that “computer-assisted textual analysis can be of value in many different sorts of literary inquiry, helping to resolve some questions, to carry others forward, and to open entirely new ones” (“Textual Analysis”).

Producing electronic texts and locating and accessing data within them are simple but vital functions the computer can perform, but the computer’s greatest strengths are in storing, counting, comparing, sorting, and performing statistical analysis. This makes computer-assisted textual analysis especially appropriate and effective for investigating textual differences and similarities, either in an exploratory way (examining the novels of an author or a group of authors for unexpected similarities or differences) or in a more directed investigation (studying the shared vocabulary of gothic novels). Some of the many kinds of investigations and questions that can be approached through textual analysis are the following:

- Testing a hunch, hypothesis, or thesis about an author, text, passage, genre, or period. Was Shakespeare’s vocabulary really unusually large? (Apparently not, according to Elliott and Valenza). Did Milton’s use of visual imagery change when he became blind? Do authors’ styles change when they dictate rather than write their texts by hand? What textual characteristics, if any, define the gothic novel?
- Testing the claims of an unsatisfying critical work or supporting and building on a compelling critical work. For example, Wallace Stevens’s “The Snow Man” is not really a “noun-heavy” poem, as Jerome McGann and Lisa Samuels claim (Hoover, “Hot-Air Textuality” 81–87).
- Investigating how and the extent to which authors differentiate the voices of characters or narrators in a novel or play or correspondents in an epistolary novel (Burrows, Computation; Rybicki; Stewart; McKenna and Antonia; Hoover, “Evidence”).
- Investigating how perceived radical shifts in style in a text are accomplished (Hoover, “Multivariate Analysis”).
- Studying how an author’s style changes over time or how genres develop and decay
(Hoover, “Corpus Stylistics”; Stamou; Martindale; Pennebaker and Stone; Garrard, Maloney, Hodges, and Patterson; Craig, “Jonsonian Chronology”; Burrows “Computers”; Moretti).

- Investigating the history of an important word, concept, or group of words or concepts over a long time span (Algee-Hewitt).
- Studying the effects of genre conventions on the language of texts or characterizing or refining how genres are defined (Biber).
- Answering questions like the following: “Do playwrights from different classes, with different education, or brought up in different places write differently? . . . Which playwrights are the most diverse stylistically across their various works? Which show the widest variation across the characters?” (Craig and Kinney 14).
- Exploring the characteristic vocabulary of an author, genre, period, group of texts, a single text, or a part of a text or the similarities and differences between the vocabularies of various writers, genres, periods, or groups of texts (Burrows, “Textual Analysis”; Hoover, “Quantitative Analysis,” “Corpus Stylistics,” “Searching”).
- Studying the extent to which and how gender, sexual orientation, race, nationality, and age of authors are reflected in the language of their texts (Koppel, Argamon, and Shimoni).
- Assessing how similar imitations, pastiches, completions, continuations, prequels, and sequels of texts written by other authors are to the original texts (Hoover, “Authorial Style”; Sigelman and Jacoby; Burrows, “Who Wrote Shamela?,” “Englishing”; Rybicki).
- Investigating questions of authorship attribution (Holmes, Robertson, and Paez; Craig and Kinney; Hoover and Hess; Forsyth, Holmes, and Tse; Juola “Authorship”; Grieve).
- Exploring thematic language in texts (Fortier).

Before discussing a few kinds of analysis more fully, the problems of planning a project and collecting digital texts must be addressed. “Planning” may take a very loose form at first for an exploratory project based on a hunch, yet even a hunch has implications for what texts and what methods will be appropriate, and more explicit planning will eventually become necessary to avoid the wasted effort of a poorly conceived study. Conversely, any project may need to take new directions in response to the availability of texts and computational tools, and even well-defined and carefully planned projects often uncover promising and unexpected avenues of exploration and sometimes fail to produce significant results. Thus flexibility is an important virtue in computer-assisted textual analysis, and testing a project on a subset of texts or methods can avoid wasted effort.

Any investigation must begin with a preliminary list of texts and some idea of method, and one good way of preparing is to study previous computational work addressing similar questions or methods.1 Those new to computational approaches may also benefit from one of the increasingly common university courses on digital humanities, humanities computing, or text analysis or from shorter, specialized workshops, such as those offered at the Digital Humanities Summer Institute.

Once a preliminary set of texts to be investigated has been identified, the first question is whether those texts are available in digital form.2 The temptation to begin with a Web search
should be resisted. Though many electronic texts can be found through such a search, thousands more cannot. The most critical factor determining where to look for an electronic text and whether it is likely to be found at all is its copyright date. In the United States, the crucial date is 1923: texts published earlier are very likely out of copyright. (The term of copyright is different in other countries; in the European Union, for example, copyright generally extends seventy years from the death of the author.)

For texts likely to be out of copyright, The Online Books Page (J. Mark Ockerbloom) and Alex Catalogue of Electronic Texts (Morgan) are especially valuable. Both list texts available at many sites, including most of those available at Project Gutenberg (the oldest electronic text collection), though it is worthwhile searching Gutenberg itself as well, since texts are added continually. Users of Google Books can limit searches to books for which the full text is available or search by author or title. The University of Oxford Text Archive contains many high-quality texts, some still in copyright but available by permission, and The Internet Archive contains a huge number of electronic texts of extremely variable quality, most of which cannot be found by a Web search.

Many university libraries have their own digital collections and even more subscribe to services like Early English Books Online, Literature Online, Eighteenth Century Collections Online, or Orlando: Women’s Writing in the British Isles from the Beginnings to the Present, and many others, most of which are accessible only through a library search. Many also have librarians specializing in digital resources, and, because some electronic resources are not widely known outside their specific subject area, subject librarians are another valuable resource, as are the subject-specific pages of resources on library Web sites.

If these searches fail, a general Web search may locate specialized collections hosting texts or versions of texts not available elsewhere, such as academic sites devoted to a historical period, like The Victorian Web (Landow); to a geographic area, like Documenting the American South; to special interests, like A Celebration of Women Writers (M. Mark Ockerbloom) or The Brown University Women Writers Project; or to extraordinary individual efforts, like The Wilkie Collins Pages (Lewis) and the Henry James site The Ladder (Dover).

Unfortunately, finding the electronic texts is only the first step: they vary so much in nature and quality that it is worthwhile to compare the available versions before selecting one. Consider Henry James’s The Awkward Age (1899), available from Project Gutenberg, The Ladder, Google Books, and The Internet Archive, among other sites. The Gutenberg text, as is usually true, contains no information about its print source, although a comparison shows that it is the revised New York Edition of 1908.3 Dover’s text matches the first British edition. A Google Books advanced search (with “awkward age” as title and “james” as author) finds one copy of the 1899 British edition and two copies of the New York Edition (one from 1908, one from 1922), but there are links to five other versions available at books.google.com at The Internet Archive, three of the New York Edition (two from 1908 and one printed later) and two of the first American edition, one much better than the other. The Internet Archive also has four independent versions. No two of these electronic texts are identical, and the best edition to select will depend on what kind of analysis will be performed and what texts, if any, will be compared with this novel.

If the original version is the most appropriate, the first American edition is probably the
best choice. The Internet Archive version and the Google Books version have competing strengths and drawbacks. The optical character recognition (OCR) used to digitize the Internet Archive version seems slightly more accurate, and the entire text can be downloaded at once, but it has hundreds of line-end hyphens. The hyphenation has been corrected in the Google Books version, but it seems to have more errors and can only be copied and pasted into a document a few pages at a time. If British spellings are appropriate, Dover’s excellent first British edition at The Ladder is the obvious choice. If the New York Edition is more appropriate, the Project Gutenberg text, with far fewer errors than any of the others, is the obvious choice, unless James’s spaced contractions (e.g., “could n’t,” “they ’re,” “I ’m”) are of interest, in which case, the Google Books version taken from a later printing seems the most accurate.

For texts still in copyright, the search should probably begin at the library, rather than on the Web, since many authors or their estates vigorously protect their copyrights and frequently force Web sites to remove works, especially novels. Literature Online, an expensive but widely held resource, contains a huge number of texts in English from AD 600 to the present. Many of these are still in copyright, including scholarly editions (but not the most recent ones) of works that are out of copyright, similar editions of the collected or complete poems of modern and contemporary poets, collections of modern drama, and national and regional literature (modern fiction, much of it still in print, is not well represented, though there is a large selection of novels in English by African writers). As noted above, The University of Oxford Text Archive also has some texts still in copyright, including novels.

For texts not available in digital form, an electronic text can be created by scanning and OCR. Unfortunately, it is not entirely clear that this is legal for texts in copyright. Although I am not a lawyer and cannot give any legal advice on this subject, creating an electronic text of a work in copyright seems defensible under the “fair use” exception of United States copyright laws (17 USC, sec. 107), provided that the electronic text is not sold or distributed in any way. This view is supported by the fact that both Literature Online and The University of Oxford Text Archive allow authorized users to download copyrighted materials with restrictions on their use and by the exceptions to the prohibition on copying and disseminating copies of such materials for libraries and archives (17 USC, sec. 108). One respected and detailed source for information on copyright is Stanford University Libraries’ Copyright and Fair Use.

There is too great a variety of hardware and software available for scanning and OCR to permit a detailed discussion here. Many university libraries, IT departments, and computer labs have expertise and equipment and may be able to help. The process is not difficult, however, and even inexpensive scanners (under $100) typically come bundled with an OCR program, so that no one who wants to produce electronic texts from printed texts should feel intimidated. Yet even the most accurate OCR produces errors. Many programs boast impressive accuracy rates of 98% or higher, but these must be taken with a grain of salt, and the accuracy is reduced by complex formats, multiple fonts, yellowed paper, stains, underlining, and marginal comments. Even at 98% accuracy, scanning and performing OCR on a typical novel produces several thousand errors. Many can be found using a spell-checker or grammar checker, but the effort is certainly not trivial, and the entire process of scanning, checking, and proofreading a novel can easily occupy many hours of tedious labor. Clearly, scanning and OCR should normally be reserved for small numbers of texts that will be used extensively or for texts out of copyright that
will be made available online.

However digital texts are acquired, they almost invariably require some editing. (A copy of the original, unedited text should be kept for reference purposes and for extracting passages to be quoted.) Normally textual analysis is performed only on text actually written by the author, so that introductions, prefaces, footnotes, tables of contents, title pages, indexes, appendixes, quotations, running heads, epigraphs, part and chapter numbers and titles, like “Chapter II,” and any other material not by the author should be removed. Occasionally, even some of the author’s own words, such as prefaces, explanatory footnotes, or poems, are so different in genre that they should be removed. Any header and licensing information (as in Project Gutenberg texts) and other similar markup should be removed.

Some typographic elements may need to be addressed. For example, to prevent dashes from being treated as hyphens by some text-analysis software, spaces may need to be added before and after them. In most electronic texts, apostrophes and opening and closing single quotation marks are identical; this is especially problematic for dialect forms, scare quotes, quotation within quotation, and dialogue marked with single quotation marks. It may be necessary to examine every apostrophe and single quotation mark and perhaps delete each one that is not an apostrophe or replace it with a double quotation mark or acute accent. Literature written before about 1800 presents additional problems, such as variant spellings, frequent and variable editorial intervention, and a high proportion of anonymous texts and texts of doubtful authorship.

The more detailed the analysis, the more important these editing processes are. They may not be feasible for large collections of texts and may not be necessary for analyses in which precise word frequencies are not at issue. Fortunately, most methods of textual analysis will not be severely affected unless there are a great many errors. It may be wise, therefore, to clear up just the most important problems and then perform some preliminary analysis to test whether the analysis seems likely to be effective before spending a great deal of time cleaning up the texts.

Given the wide variety of literary studies for which textual analysis is appropriate and the many methods that exist for pursuing literary studies, it would be impossible to discuss even a small sample of them in detail here. Rather, I will discuss processes that are common to a large number of methods, suggest some resources for learning about methods, and then discuss a few methods in more detail.

For most literary study, the smallest unit of analysis is the word. There is some evidence that letter sequences and information about parts of speech sometimes work better than words for authorship attribution (Clement and Sharp), but words have the advantage of being meaningful in themselves and in their significance to larger issues like theme, characterization, plot, gender, race, and ideology. Many tools for generating word-frequency lists and concordances exist, including good free programs that can be downloaded: AntConc (Anthony), KWIC Concordance for Windows (Tsukamoto), and Conc. Online collections of tools often allow the user to upload texts to Web-based tools that do not require installation; these seem especially valuable for exploratory work (TAPoR). Finally, there are inexpensive, commercially available programs like WordSmith Tools (Scott), MonoConc Pro (Barlow), and Concordance (Watt), which tend to be more powerful, versatile, and comprehensive than the free programs. These programs and others have various strengths and weaknesses, but they all typically produce word-frequency lists.
in alphabetic or descending frequency order and concordances that list designated words along with a substantial amount of context. Word lists and concordances are very useful exploratory tools, and concordances can also be used to test hunches about how specific words are used in a text. Many of these programs can also statistically compare word frequencies among several texts and can show which texts have unusual frequencies of words of interest. Many can also generate lists of collocations (words that occur repeatedly near each other); an examination of collocations can be especially useful for thematic studies. Note that many text-analysis programs can only process plain text files; if a word-processing program is used for editing and cleanup of the text, it will probably be necessary to save the file as plain text.

Most textual analysis begins with word-frequency lists and compares the frequencies of words across two or more texts. This comparison requires a parallel word-frequency list, consisting of the words listed in descending frequency order for the entire group of texts and the relative frequency of each word in each text, including zero frequencies for texts in which the word does not occur. Unfortunately, not many simple, easy-to-use programs for producing this kind of list are available. I know of only three that can handle large numbers of texts: WordSmith Tools (Scott), The Intelligent Archive, and my own The Parallel Wordlist Spreadsheet.

The parallel word lists are processed with general purpose statistical programs, a variety of which are frequently installed in computer labs, and many IT departments offer instructional sessions on using these programs. I normally use Minitab, which is relatively easy to learn, has a good graphing function and an excellent help function, and is inexpensive enough that most users can afford to purchase a copy. The most frequently used statistical techniques are principal components analysis (PCA) and cluster analysis, but discriminant analysis and other techniques have also been used. Many of the essays cited here give some information on how to perform statistical analysis of word lists, and there are detailed instructions for doing PCA and cluster analysis in Minitab on my The Excel Text-Analysis Pages. PCA Online allows the user to experiment with PCA on Shakespeare’s plays without learning a statistical program.

As a demonstration of how PCA and cluster analysis work, consider figure 1, a cluster analysis of ten texts by Walter Besant (five novels and five stories) and sixteen novels by Wilkie Collins, based on the one hundred most frequent words of the entire set (the last two digits of the date of publication precedes each abbreviated title; all are from the eighteen hundreds). These texts were collected for a study of Besant’s completion of Collins’s unfinished novel Blind Love, which shows that, despite Besant’s use of the extensive notes Collins provided, the point at which Besant takes over is very clearly marked (Hoover, “Authorial Style”). Cluster analysis compares the frequencies of all one hundred of the most frequent words simultaneously, determines which two texts are most similar to each other in how they use these words, and joins them into a cluster, then proceeds to find the next most similar pair or group of texts until all the texts are joined in a single cluster. The more similar the frequencies of the one hundred most frequent words are in two or more texts, the closer to the left those texts form a cluster. Thus figure 1 shows that “72 PoorF” and “75 LawLady” are much more similar to each other than are “95 Quarantine” and “93 Shrinking.” Similarly, “84 Dorothy” and “82 RevoltMan” are much more similar to each other than they are to the eight texts in the cluster below them. Finally, the ten Besant texts at the top of the graph are much more similar to each other than they are to the sixteen Collins texts at the bottom. Clearly, even the frequencies of the one hundred most
frequent words very distinctly separate the styles of these two authors. Equally clearly, Collins’s
texts are more similar to each other than are Besant’s, possibly because of the great variation in
length in Besant’s texts. Furthermore, Collins’s texts show some tendency to group by date of
publication: the five texts in the bottom cluster were all written after 1880 (the clustering by date
becomes more accurate when larger numbers of words are analyzed).

Figure 1. Cluster analysis of texts by Walter Besant and Wilkie Collins based on the one hundred
most frequent words.

The results of PCA based on the same texts and the same words can be seen in figure 2.
Instead of clustering similar texts together on the basis of the frequencies of the one hundred
most frequent words, PCA compresses as much of the information about the frequencies of the
one hundred most frequent words as possible into a small number of unrelated new variables, or components. The values for the two most important of these variables are then used to locate each text on a two-dimensional graph, with the first component on the horizontal axis and the second on the vertical axis. Figure 2 shows that the first component, which accounts for almost 33% of the variation in the frequencies of the words, is capturing authorship, with all the Collins texts to the right and all the Besant texts to the left. This means that many words are more frequent in all of the texts by Besant than in those by Collins and vice versa. The second component has no clear interpretation, though it is suggestive that later texts by both authors tend to appear toward the top of the graph. (The wider scattering of the texts by Besant reflects the same greater variation among his texts than among those by Collins that is evident in figure 1.)

Figure 2. PCA analysis of texts by Walter Besant and Wilkie Collins based on the one hundred most frequent words.

PCA and cluster analysis are valuable for both exploratory work and in-depth analyses, but they have different strengths and weaknesses. Cluster analysis has the benefit of giving unequivocal results, while PCA graphs are more dependent on judgment, especially where the texts being compared do not separate as clearly as these. But PCA has one great advantage: using the same data as in figure 2, PCA can produce a graph like that in figure 3, in which the words are graphed onto the same two dimensions as the texts, so that it is immediately apparent which words are disproportionately rare or frequent in which texts.
The words most favored by Besant over Collins (on the far left) are but, would, one, man, a, all, or, so, and about, while those most favored by Collins over Besant (on the far right) are to, in, on, had, left, letter, time, and back. Even a cursory examination of figure 3 uncovers other interesting characteristics of the vocabularies of these two authors that suggest further directions for research (see Hoover, “Authorial Style”). For example, Besant favors negatives like no, not, nothing; forms of to be (be, is, are, was, were; only been and am are about equally favored); and the third-person plural pronouns they, them, and their. Collins favors the feminine pronouns she and her, the titles Mrs. and Miss, and the noun lady (together suggesting more emphasis on women); the first-person singular pronouns I, me, my; and several content words—house, room, letter, time, lady, way, first, and looked—compared with only man and know for Besant. Other methods can locate characteristic vocabulary, and PCA graphs quickly become unreadable as more words are analyzed, but the ability to produce graphs like those in figure 2 and figure 3 from a single set of data has made this kind of analysis among the most frequently used in computational studies.

As the incipient chronological differentiation in the texts above suggests, these same techniques can be used to study an author’s stylistic development by treating early and late periods as different authors. Both cluster analysis and PCA easily distinguish the early from the late Henry James (see also Hoover, “Corpus Stylistics”). As can be seen in figure 4, the
clustering of his twenty-one major novels matches their chronology extremely closely, except for the unusual late novel, The Outcry (adapted from a 1909 play). This graph does more than dramatically demonstrate the development of James’s style, however. It also casts doubt on the widely held notion that the late style is a result of James’s adoption of dictation because of wrist pain in 1897, during the composition of What Maisie Knew. There is certainly no sign of any radical transformation of James’s style in 1897.

Figure 4. Cluster analysis capturing the chronology of twenty-two Henry James novels based on the five hundred most frequent words.

My final example of computer-assisted textual analysis is an exploratory study of differences in poetic vocabulary among a group of twenty-six male and female American poets born between 1911 and 1943.9 This discussion will not pretend to settle the complex and contentious debate about the existence of feminine writing and will make no global claims about gender theory. Rather, it will demonstrate that textual analysis can produce provocative results that point toward areas where more research is needed, and will argue that interesting results are
the norm for such an analysis.

Burrows has shown that it is relatively easy to distinguish male and female writers of the seventeenth and eighteenth centuries even using only a subset of the 150 most frequent words of the texts but that it becomes progressively more difficult with more recent writers (“Computers,” “Textual Analysis”). A 2002 study (Koppel, Argamon, and Shimoni), however, has shown that it remains possible, using more sophisticated methods, to identify the gender of both fiction and nonfiction documents in the British National Corpus (mostly written between 1974 and 1993) at a rate of about 80%.

Here I will use a method that focuses not on the most frequent words of texts that have been the province of so much textual analysis but on words that are neither very common nor very rare. The goal is not primarily to show that authors can be identified by gender on the basis of their characteristic words but rather to explore the vocabularies of the poets. The method is a modification of Burrows’s Zeta (“Who Wrote Shamela?,” “All the Way Through”) developed by Craig (Craig and Kinney) that I call Craig Zeta. This simple method divides two sets of texts into approximately equal-sized sections and compares how many sections for each author contain each word, ignoring the frequencies of words and concentrating on their consistency of appearance across the sections. The sets can be selected on the basis of any perceived contrast, but here the contrast is texts by women versus those by men. Combining the ratio of the sections by women in which each word occurs with the ratio of the sections by men from which it is absent yields a single measure of distinctiveness that ranges from two (words found in every women’s section and absent from every men’s section) to zero (vice versa). Sorting the words on this composite score produces two lists of marker words, one favored by these women and avoided by these men, and one favored by these men and avoided by these women.

Testing fourteen additional poets, seven men and seven women, with these marker words produces the result shown in figure 5, where the vertical axis shows the proportion of all the different words in each text that are among the five hundred most distinctive male marker words, and the horizontal axis shows the proportion of all the different words in each text that are among the five hundred most distinctive female marker words. Despite the limitations of this exploratory study, it does a remarkably good job, correctly identifying the genders of twenty of the twenty-five new sections of poetry by poets who played no part in the selection of the words (the errors are in bold type). These same words produce a similar result for seven male and seven female contemporary novelists, which is further evidence that the method is capturing some kind of genuine difference. Consider now the one hundred most distinctive male and female marker words, shown in table 1 (the lists are identified by gender only in the note below each table, so that readers who want to can try identifying which is which—an informal preliminary survey suggests that most readers can).
Figure 5. Scatter graph of male and female poets based on five hundred male and five hundred female marker words.
<table>
<thead>
<tr>
<th>Mother's</th>
<th>Sign</th>
<th>Breathe</th>
<th>Cross</th>
<th>Hidden</th>
<th>Powder</th>
</tr>
</thead>
<tbody>
<tr>
<td>Father's</td>
<td>Delicate</td>
<td>Velvet</td>
<td>Song</td>
<td>Flies</td>
<td>Paris</td>
</tr>
<tr>
<td>Skin</td>
<td>Fist</td>
<td>Slit</td>
<td>Memory</td>
<td>Future</td>
<td>Exact</td>
</tr>
<tr>
<td>Table</td>
<td>Hotel</td>
<td>Bite</td>
<td>Souls</td>
<td>Trail</td>
<td>People</td>
</tr>
<tr>
<td>Lovers</td>
<td>Heels</td>
<td>Thorns</td>
<td>Dancing</td>
<td>Cell</td>
<td>Winter</td>
</tr>
<tr>
<td>Eggs</td>
<td>Sea</td>
<td>Doorway</td>
<td>Dirty</td>
<td>Needs</td>
<td>Heavy</td>
</tr>
<tr>
<td>Mother</td>
<td>Door</td>
<td>Stems</td>
<td>Darkness</td>
<td>Justice</td>
<td>Rest</td>
</tr>
<tr>
<td>Father</td>
<td>Round</td>
<td>Lesson</td>
<td>Mean</td>
<td>Surround</td>
<td>Given</td>
</tr>
<tr>
<td>Hot</td>
<td>Itself</td>
<td>Interior</td>
<td>Dog</td>
<td>Roll</td>
<td>West</td>
</tr>
<tr>
<td>Breath</td>
<td>Wet</td>
<td>Rushes</td>
<td>Ate</td>
<td>Faith</td>
<td>Heaven</td>
</tr>
<tr>
<td>Next</td>
<td>Gold</td>
<td>Proper</td>
<td>Flow</td>
<td>Beer</td>
<td>Moved</td>
</tr>
<tr>
<td>Grow</td>
<td>Stones</td>
<td>Angle</td>
<td>Whether</td>
<td>Goodbye</td>
<td>Lonely</td>
</tr>
<tr>
<td>Grew</td>
<td>Gave</td>
<td>Children</td>
<td>Many</td>
<td>Radio</td>
<td>Clouds</td>
</tr>
<tr>
<td>Secret</td>
<td>Wait</td>
<td>Eye</td>
<td>Town</td>
<td>Vanished</td>
<td>Happy</td>
</tr>
<tr>
<td>Five</td>
<td>Almost</td>
<td>Tell</td>
<td>Die</td>
<td>Prayers</td>
<td>Hours</td>
</tr>
<tr>
<td>I'd</td>
<td>Mirror</td>
<td>Want</td>
<td>Others</td>
<td>Flute</td>
<td>Fool</td>
</tr>
<tr>
<td>Watched</td>
<td>Later</td>
<td>Lie</td>
<td>Road</td>
<td>Confusion</td>
<td>Try</td>
</tr>
<tr>
<td>Chair</td>
<td>Afternoon</td>
<td>Turned</td>
<td>Fly</td>
<td>Kicked</td>
<td>Laughter</td>
</tr>
<tr>
<td>Lines</td>
<td>Kept</td>
<td>Arms</td>
<td>Below</td>
<td>Dust</td>
<td>Ancient</td>
</tr>
<tr>
<td>Shining</td>
<td>Shoulder</td>
<td>Three</td>
<td>Friends</td>
<td>Few</td>
<td>Stir</td>
</tr>
<tr>
<td>Watches</td>
<td>Thighs</td>
<td>Found</td>
<td>Hell</td>
<td>Bird</td>
<td>Spinning</td>
</tr>
<tr>
<td>Ear</td>
<td>Somewhere</td>
<td>Slowly</td>
<td>Spirit</td>
<td>Dance</td>
<td>Emptiness</td>
</tr>
<tr>
<td>Rare</td>
<td>Touched</td>
<td>Watch</td>
<td>Either</td>
<td>Seen</td>
<td>Lands</td>
</tr>
<tr>
<td>Wore</td>
<td>Trying</td>
<td>Done</td>
<td>Hill</td>
<td>Less</td>
<td>Exist</td>
</tr>
<tr>
<td>Sealed</td>
<td>Cream</td>
<td>Floor</td>
<td>Living</td>
<td>Beat</td>
<td>Fade</td>
</tr>
<tr>
<td>Wool</td>
<td>Angels</td>
<td>Pale</td>
<td>Caught</td>
<td>Wrong</td>
<td>Drifting</td>
</tr>
<tr>
<td>Spin</td>
<td>Orange</td>
<td>Making</td>
<td>Care</td>
<td>Sang</td>
<td>Thousands</td>
</tr>
<tr>
<td>Going</td>
<td>Blew</td>
<td>Beside</td>
<td>Falls</td>
<td>Understand</td>
<td>Police</td>
</tr>
<tr>
<td>Lay</td>
<td>Spent</td>
<td>Doors</td>
<td>Miles</td>
<td>Dogs</td>
<td>Hawk</td>
</tr>
<tr>
<td>Sat</td>
<td>Remembered</td>
<td>Break</td>
<td>Bad</td>
<td>Whom</td>
<td>Everybody</td>
</tr>
<tr>
<td>Fruit</td>
<td>Bowl</td>
<td>Corner</td>
<td>Sing</td>
<td>Lust</td>
<td>Worms</td>
</tr>
<tr>
<td>Instead</td>
<td>Carefully</td>
<td>Closed</td>
<td>Hate</td>
<td>York</td>
<td>Rolling</td>
</tr>
<tr>
<td>Leaving</td>
<td>Fair</td>
<td>Closed</td>
<td>Strength</td>
<td>Screaming</td>
<td></td>
</tr>
<tr>
<td>Language</td>
<td>Wrists</td>
<td></td>
<td>That's</td>
<td>Freedom</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Lists of the one hundred most distinctive gender marker words in thirteen female and thirteen male poets.

The most distinctive female and male marker words can be distributed variously, so long as there is a great difference between the two genders. Relatively common words like mother are found in twenty women’s sections but only eleven men’s; some less frequent words like cross are
found in sixteen men’s sections but only three women’s; others, still less frequent, like spin, are found in nine women’s sections but no men’s sections. Female markers like children and mirrors and male markers like beer and lust seem almost stereotypical, but there are also surprises, like the female marker fist and the male markers song and dancing. Studying a concordance of the entire set of texts is an excellent way to examine these words in context. The fact that Sara Teasdale, H.D., and Edna St. Vincent Millay, whose texts cluster with the men’s, were born about twenty to thirty years before the poets on which the words are based also seems worth investigating, as do some large, distinctive clusters of related words, shown in table 2, which are drawn from among the five hundred most distinctive male and five hundred most distinctive female markers. Any study of the vocabulary of male and female poets would benefit from larger numbers of poets and larger samples, and many other configurations that address different contrasts are possible (e.g., nationality or historical period).

Cluster 500 Women’s Markers 500 Men’s Markers

Family

mother's, father's, mother, father, children, ancestral, aunt, baby, birth, child, child’s, cousins, daughters, family, generations, uncles

Religion

altar, nuns, praying

faith, heaven, hell, prayers, souls, spirit, Christ, gods, myth, paradise, religion, spirits, temple

Houses/Furniture

table, chair, door, doorway, floor, doors, bathroom, bedroom, carpet, ceiling, cellar, chimney, closet, cupboards, gates, hotel, kitchen, palace, rooms, rug

attic, buildings, ruins, shack, temple, wall

Song/Dance

danced

song, dancing, sing, dance, sang, dancer, music, singer, singing, sings

Personal Pronouns

he’ll, I'd, mine, ourselves, she'd, she's, they'd, you’d, you’re, yourself

Table 2. Clusters of related words drawn from the five hundred male and five hundred female marker words.

Examples and suggestions could be multiplied almost indefinitely, but I hope to have provided a general idea of the varieties, the challenges, and the benefits of computer-assisted textual analysis and of the opportunities it provides for a wide range of literary studies. Textual
analysis can help the literary scholar in the relatively simple but important tasks of collecting, organizing, and evaluating examples and evidence that are relevant to a more traditional study. It can act as a kind of discovery procedure for revealing previously unnoticed trends and suggesting productive and original questions in open-ended, exploratory work. It can inform much more specific and directed kinds of research that test a hypothesis, hunch, thesis, or critical claim. It can provide access to detailed and precise kinds of evidence that would otherwise be impractical to assemble or completely unavailable. It can also help establish or revise an author’s canon by removing spurious works, by adding previously unknown works, or by suggesting or confirming the chronological relations among an author’s works. Computer-assisted textual analysis is neither a panacea nor a substitute for sound literary judgment, but its ability to refine, support, and augment that judgment makes it an important analytic method for literary studies in the digital age.

Notes

1. The best resources for this kind of research are the journals Literary and Linguistic Computing and Computers and the Humanities (through 2004), which specialize in computational approaches. Computational work is, however, increasingly appearing in other journals, such as Eighteenth-Century Studies, Ben Jonson Journal, Milton Quarterly, Modern Language Review, Style, Victorian Periodicals Review, and Early Modern Literary Studies.

2. Increasingly sophisticated online resources built around collections of texts provide another opportunity for textual analysis (see, e.g., Cooney, Roe, and Olsen’s essay in this volume). Most of Cather’s texts, for example, can be analyzed online through The Willa Cather Archive (Jewell). The Brown University Women Writers Project allows users to perform textual analysis on a large collection of writings by women, though it requires an institutional subscription or an inexpensive license. ARTFL has a huge collection of French texts with tools for analysis, and MONK offers sophisticated tools that operate on some large, publicly available collections of texts.

3. One good way to examine differences among texts is with Juxta, a collation tool that can be used to compare and evaluate two versions of a text.

4. Faulkner’s novels, for example, are not generally available, and an earlier online electronic version of Light in August is no longer available. The full text of The Sound and the Fury has also recently been removed from an online scholarly edition (Stoicheff et al.).

5. Also available are several free online OCR tools of varying quality and ease of use, and some older versions of Microsoft Office include document imaging, which can scan paper documents and perform OCR.

6. Juola’s JGAAP is a simple but powerful and versatile suite of authorship methods, best used in conjunction with his “Authorship Attribution,” which discusses many of the methods it implements.
7. Lee has a good list of software with commentary.

8. In WordSmith this function, located under the Detailed Consistency tab of Wordlist, operates on special word-list files produced earlier (the View Column Totals option must also be selected). The Web sites for my The Parallel Wordlist Spreadsheet and The Intelligent Archive provide detailed instructions.

9. The texts come from Literature Online; to simplify the analysis and avoid problems of samples of different sizes, I truncated each poet’s sample at about eight thousand words. Obviously, an analysis based on only eight thousand words by each of twenty-six authors must be considered preliminary.

10. For a more detailed explanation of this method applied to the vocabularies of Wilkie Collins and Stephen Crane, see Hoover, “Authorial Style.”

Works Cited


Not Unless You Ask Nicely: The Interpretative Nexus Between Analysis and Information

J. F. BURROWS
University of Newcastle, Australia

Abstract

Computer-based evidence, especially when it incorporates statistical analysis, is too often regarded with special deference or special scepticism. It is better assessed upon its merits, like any other application of inductive logic. The nexus between the inductive process and the information available is studied in two paradigmatic attempts to interpret sets of statistically based distinctions between different texts.

1. Introduction

In recent years, as people have become more accustomed to computers, they have become less inclined to pass off the effects of human incompetence as ’computer error’. But the notion that the machine is something more than an instrument survives in subler forms. Thus, in disputes about the computer-assisted testing of authorship, the fact that a computer has been used (especially but not only when its output is statistical) seems to colour many people’s attitudes to evidence and to standards of proof. Computer-based evidence is sometimes put forward as if it deserved special credence: it is sometimes received, conversely, as if it must meet special standards or else be swept aside as worthless. In both cases, the analogy of fingerprints has often been put to unreasonable use.

At one time, the analogy was mostly used to bolster the imprudent hopes attached to new ways of testing authorship. But no one has yet identified a stylistic attribute as idiosyncratic or as durable as human fingerprints. Nothing in the nature of the case suggests that anyone will ever do so: for, though writers have their tricks of expression and though some of these become habitual, they are not absolute identifiers because they are not often peculiar to one writer, because they do not always persist over the length of a writer’s career, and because many of them are easily copied. Nowadays the analogy is mostly used to express satisfaction at the supposed failure of those earlier hopes. It is used in this way by literary scholars who believe (and perhaps by some who choose to claim) that computers can somehow offer and should be required to yield incontrovertible evidence. But, in response to their demand for absolute proof, one must note that that is a rarity in any field and that even fingerprints are unlikely to provide it. In response to the claim that the computer has failed in cases where the data were too sparse or the methods of analysis defective, one must insist that it is for us to use our instruments effectively. Even a clear fingerprint has nothing to tell us unless its presence in a certain place is revealing and unless a good set of its owner’s prints is available for comparison.

Readers of this journal will recognize the forms of argument I have glanced at. If I proceed without particularizing them, it is because I wish to get behind the immediacies of controversy and to consider some basic principles. My point of departure is that, as the files of this journal also show, we no longer find it difficult to draw clear and objective distinctions between one text and another: data ranging from hapax legomena to differences in the rate at which hitherto unused words accrue and, again, to differences of transitional probability (the likelihood that a given word will be followed by one rather than another of its possible successors) have been employed for that purpose; and the best of the methods used have withstood close scrutiny. Such distinctions can even be derived, as I shall show, from tables of comparative frequencies for the most common words (tables of comparative frequencies follow on after the Appendix). But, not being absolute identifiers, these distinctions cannot be interpreted as inter-authorial, intra-authorial, or otherwise unless some information can be brought to bear. Whatever instruments we employ and whatever form our data may take, data are not transmuted into evidence until a question can be formulated in a manner fit for testing. And, in formulating our questions, we must rely on some knowledge of the case.

The two main cases to be examined are chosen as paradigms of the most difficult and the simplest forms of attribution-problem. In the former, we shall proceed as if we had no information except a set of frequency tables. In the latter, we shall examine a (supposedly) doubtful text on the premiss that it was written by one or other of two known authors. Most attribution-problems fall short of these extremes. Even at worst, we would rarely be denied the opportunity to inspect the text itself and to form rough inferences for testing. Even at best, we would rarely be asked merely to determine whether the author was Smith or Jones: the ghostly presence of A. N. Other usually clouds that question.

2. The Descriptive Analysis of Data

Let us begin by specifying some ground rules. I am presented with four frequency tables. Table 1.1 shows raw frequencies for the fifty most common words of a certain Text A. Tables 1.2, 1.3, and 1.4 show raw frequencies for the same words in three other texts. I am asked to compare the four tables, with special reference to Text A, and to comment on the outcome. I am to proceed as if the tables alone and not the texts were...
available for my scrutiny. And if, as here, I know the identity of the texts, I must take no advantage of that knowledge.

Before I can reasonably attempt to analyse such tables, I must make two stipulations: that the rows lie in an authentic hierarchy and that the columns be truly comparable. I also prefer but need not demand to know whether the forms of any words were modified for counting. In the present case, I am assured, first, that these are fifty of the fifty-two most common word-types of Text A, the names of two people having been excluded. (Provided their frequencies are high enough to bear analysis, the number of word-types tabulated is a matter of convenience: the top fifty, whatever they may be, make up about half of all the word-tokens in most texts written in English. The hierarchy is that of Text A and not that of all four texts combined because I am asked to give particular attention to Text A.) I am assured, secondly, that all but the last column of each table covers 8,000 words of text: since the last column of each table covers an unequal residue, the figures for every column need to be standardized (as percentages or rates per thousand) before they can be compared. And I am assured, thirdly, that the words have been counted as they stand in the four texts without any separation of homographic forms and without any expansion of contracted forms like 'I'd' and 'don't'. The relevant effects are twofold. When all forms of a word like 'that' are left as one, some nice distinctions can be blurred. When contracted forms are left as they stand in the text, a broad distinction between more and less colloquial texts is preserved but the frequencies for 'not', for the personal pronouns, and for some auxiliary verbs are distorted.

From a study of these tables, an experienced observer would soon infer that all four texts were likely to be novels. A text running to sixteen or more segments of 8,000 words in which the personal pronouns 'I' and 'you', 'he' and 'she', and the verb 'said' rank high among the most common words of all is unlikely to be couched in any other literary form (This line of thought is supported by some of the variations that occur, across the sets of columns, in the frequencies of these and other words: it is not difficult to see where, from segment to segment, each text fluctuates between a predominance of narrative and a predominance of dialogue.) Other literary forms show equally distinctive patterns. A play would almost certainly be much shorter and would be likely to run high on the same pronouns but not on 'said'. A treatise might well be as long as a novel; but it would almost certainly run to lower frequencies for the pronouns and much lower frequencies for 'said'; and it would be most unlikely to make much use of 'don't'. A set of personal letters might run to any length; but the first- and third-person pronouns would usually rank far higher than the second, marking the fact that (except in tutelary letters of advice or close enquiry) those to whom letters are addressed are not often their main topic.

On comparing the four parts of the table, our observer would also draw some clear inferences about differences among the four texts. 'She' and 'her', for example, occur even more often in Tables 1.2 and 1.3 than in 1.1 and 1.4. Perhaps the protagonist (or else the character most discussed) in Texts B and C is female: perhaps the reverse applies in Texts A and D. In Tables 1.2 and 1.4, where 'don't' is especially frequent, 'am', 'are', and 'will' are comparatively uncommon. It seems likely that Texts B and D are comparatively colloquial and that these and other verbs mostly appear there in contracted forms. The consequent effects upon the frequencies of the pronouns are less easily visible but the effect upon 'not' may explain the contrast with its high frequency in Text C. This last set of inferences could be tested by examining a parallel set of tables in which all the contracted forms were separated into their constituents. All these and many similar inferences could be refined by calculating means and standard deviations for this row or that. And they could be given real statistical weight by introducing a test of distribution between populations like Student's $t$-test.

But all such inferences rest upon outside knowledge. They could not be drawn by a person who was ignorant of the ways the language changes from one literary form to another and of the different levels of formality that usually prevail in narrative and dialogue. To deny ourselves the advantage of such knowledge, let us proceed for a time as if even the word-list were stripped away and we were left with bare tables of frequencies.

Our last remaining piece of outside knowledge is that the figures for each text occupy specific columns in the frequency tables. That information could be withheld by presenting the data as a single table of standardized frequencies—a compound, for example, of the forty-four columns for Texts A and B—in which the columns lay in random order. Even under those conditions, the form of principal components analysis I am about to describe would establish discrete clusters for what were in fact (though we would not know it) the entries for Texts A and B respectively. Once the clusters were established, their statistical cohesion could be assessed by testing for co-variance. But, in order to move from statistical differences back towards any form of literary interpretation, it would be necessary to know which columns actually belonged to which text.

To compare the data for Texts A and B, we must standardize the raw frequencies of Tables 1.1 and 1.2 and unite the two tables. The rows of the joint table are then correlated, each with every other, using the Pearson product-moment method. If we were allowed to see the word-list itself, the resulting matrix would show the varying degrees of concomitant frequency among these word-types in these texts. Since we are not, we are confined to the abstract observation that far more rows than chance would dictate are linked by high correlation coefficients, whether positive or negative. The underlying pattern of relationships can be made more intelligible by using any of the statistical methods that extract the 'principal components' of such matrices. (My own practice is to employ 'eigen analysis', using the MINITAB package: the full sequence of commands is given in an appendix.) The most powerful of the principal components (the first two or three 'eigenvectors') can then be portrayed as the axes of one or more biaxial graphs. Given access to the word-list, the entries in such 'word-plots' can be labelled as in Graph 1. In a
final statistical step, which is entirely independent of our access to the word-list, the ‘eigen matrix’ is multiplied through the original table of standardized frequencies for each text. The results can then be shown in graphs like Graph 2.

The salient feature of Graph 2 is that it distinguishes sharply between the sixteen entries derived from Table 1.1 and the twenty-eight derived from Table 1.2. Three of the forty-four lie some way towards the other set but the overall result is plain: the use of this statistical procedure yields a clear distinction between Texts A and B. It should be noted, however, that the clusters are separated only on the horizontal axis, which represents the most powerful eigenvector. On the vertical axis, representing the second eigenvector, the larger cluster ranges more widely than the smaller: in some significant respects, therefore, Text B is more diverse than Text A. But what Graph 2 does not show is whether these two texts are the work of a single author.

Let us now reinstate our access to the word-list so that, drawing on Graph 1, we can see which word-types contribute most to the distinction between Texts A and B. Since Graph 2 is a product of Graph 1, its easterly entries reflect comparatively high frequencies, in Text A, for those words that lie towards the east of Graph 1 and comparatively low frequencies for those that lie towards the west: the reverse applies to Text B. Among the most easterly words of all, the majority are first and second-person pronouns and present and future inflections of auxiliary verbs. ‘She’, ‘her’, and several past tense verbs are among the most westerly words. Such a contrast between the two texts may rest upon simple differences in content and temporal perspective. But other factors cast doubt on that neat interpretation. The south-easterly location of ‘I’ and ‘you’, along with ‘said’, suggests that Text A is more given to dialogue than Text B. That possibility, in turn, is complicated by the fact that the frequencies for contracted word-forms were counted as they stood. Since all five of the easterly auxiliaries and two of those five pronouns are often contracted, their presence, in their uncontracted state, at the easterly extremity of the graph suggests that Text A contains fewer colloquial contractions than Text B. This possibility (which is in keeping with the location of ‘don’t’) might simply reinforce the idea that Text A runs high on dialogue: but it might mark, instead, a difference between the stylistic habits of two authors. As the
complexities increase, one notes the locations of 'me', 'my', 'your', and 'her' (none of which takes a contracted form). For they show that, though the original neat interpretation may need to be modified, it cannot simply be dismissed.

The words that lie at the northern and southern extremities of Graph 1 mark a secondary contrast between the more descriptive or disquisitory and the more dialogic parts of each text. In this respect, as was noted earlier, Text B is more diverse than Text A. The presence, at the northern extremity, of most of the early entries for both texts is a strong indication that both texts begin with long passages of scene-setting in which 'which' unites with the articles and the most common prepositions. At other identifiable stages, both texts show high frequencies for 'which' and 'if', 'it' and 'you' and 'me', some of the staple words of dialogue. But, despite its interest for the literary critic, the north-south axis of Graphs 1 and 2 sheds no light whatever on the authorship of these texts.

When Texts A and C are compared, in precisely the same fashion, the distinction between them is clearer still. The unoccupied territory between the clusters in Graph 4 is broader than it was in Graph 2. Not a single entry approaches the other cluster. And, as is common when the distinction between two such clusters is strongly marked, the clusters lie not in a horizontal but rather in a diagonal opposition to each other: here (as was not the case in Graph 2) both of the main eigenvectors contribute strongly to the distinction between the texts. The most striking feature of Graph 3, the accompanying word-plot, is that its clusters are all-pervasive. Of our fifty common words, only 'not' lies in the broad, slanting no man's land that separates the clusters: only 'not' and the handful of other words that lie along the edges of no man's land contribute nothing to the overall distinction between Texts A and C. The many words that do contribute include a higher proportion than before of connectives and intensifiers, the sorts of words where authors do not easily change their habits.

Graph 3  Texts A and C in 8,000-word segments (word-plot for the fifty most common word-types of Text A)

Graph 4  Texts A and C in 8,000-word segments (based on the fifty most common word-types of Text A)
from text to text. But the more ad hoc words also contribute. Those of a dialogic cast lie in a diagonal band across the south-eastern corner: Text A, once more, runs higher on dialogue than its opposite number. The feminine pronouns and those auxiliary verbs that reflect an emphasis on the past run towards Text C. The isolated location of the three masculine pronouns points to a strong local emphasis on a male figure, unusual in these texts but pronounced in both the opening and closing phases of Text A.

Graphs 5 and 6 show that Texts A and D are comparatively difficult to distinguish from each other. In Graph 6, the boundaries between the clusters are blurred; an entry from each text penetrates deep into the other cluster; and an outlying pair, one from each text, make unexpected partners. In Graph 5, moreover, the words are spread more uniformly than they were in the previous word-plots. There is, nevertheless, an imperfect separation of clusters along a diagonal from south-west to north-east. Returning to Graph 6 with that diagonal in mind, one finds that it separates about half of the entries for Text A from all the other entries for either text. The fact that the smaller group includes seven successive entries (A5–A11) from the middle of Text A and only one other (A2) suggests that this pattern may not be a chance-effect. One notes, moreover, that a closely corresponding set of entries (D6–D12, along with D2) lies at the eastern edge of the

---

**Graph 5**  Texts A and D in 8,000-word segments (word-plot for the fifty most common word-types of Text A)

**Graph 6**  Texts A and D in 8,000-word segments (based on the fifty most common word-types of Text A)
larger group. And the last entry for each text (A16 and D18) lies far to the west of all its fellows. Once seen, this curious matching of partners in a pattern of displaced replication pervades Graph 6. But though the pattern can be seen, it cannot be interpreted without more information than we have.

From an analysis of the four frequency tables, we have been able to distinguish Text A more or less sharply from each of the other three. From a study of the relationships between word-plots and text-plots, we have been able to draw various inferences, some tolerably firm, some highly tentative, about the stylistic bases of those distinctions. But only one such inference does anything to suggest which, if any, of the distinctions may reflect a difference in authorship I allude, of course, to the fact that Graph 3 shows a much more pervasive distinction between the word-frequencies of Texts A and C than is evident in the word-plots treating of Texts A and B or A and D. Yet no less sharp and pervasive distinctions might be encountered, say, in a comparison between texts by one author but in different genres. (In the present case, however, the frequency tables give us some reason to believe that all four texts were novels.)

Since I do know the true identity of the four texts, this exercise has been factitious. That is not reflected, I hope, in my having made observations that would have escaped me if I had not known the identity of the texts. It has certainly made itself felt, on the other hand, in a persistent and inhibiting sense of doubt about things I did observe: would I have noticed this or that if I had not known the truth? Wearing a blindfold is nothing like being blind. Yet the exercise meets its purpose if it shows the conceptual gap between analysis and interpretation; if it shows how that gap can be spanned only by information; and if it shows how even a little information can be put to use. Where the information is of kinds we scarcely know we have (such as the likelihood that ‘said’ will occur more often in most novels than in other forms of literature), we can easily delude ourselves that we are getting answers from the bare statistics. But the principle is clear: where we do not have information enough to formulate specific questions about the authorship or provenance of the texts, we cannot reasonably hope to go beyond a descriptive analysis of the data, interesting as that may be.

Texts A, B, C, and D are the 1877 version of James’s The American; The Portrait of a Lady; Mansfield Park; and the 1907 revision of The American. Anyone who, knowing this, wishes to review the argument so far may care to know a little more. In narrative as well as dialogue, James makes less use of contrived forms in his earlier work than in his later: a contrast that often serves as an inter-authorial differentia in intra-authorial in James. The displacement in the matching of the numbered pairs of entries in Graph 6 arises from the fact that the revised version of The American is almost 9,000 words longer than the original: thus, though D1 and A1 are a true pair, D17 and D18 more nearly correspond to A16 than does D16. The two outlying entries, A14 and D15, cover the original and revised versions of the long episode where the old servant, Mrs Bread, tells Newman how her former master, the Marquis de Bellegarde, came to his death.

3. The Predictive Testing of Evidence

In the second of these paradigmatic cases, I am informed that Text A is undoubtedly the work of either Jane Austen or Henry James and am asked to determine which. (As before, I am not to know that Text A is the 1877 version of The American.) I am granted my two previous stipulations about the structure of the frequency tables and I am allowed to ask for additional tables. I am still not allowed to see any of the texts themselves.

Our previous comparisons were encumbered by the fact that the contracted word-forms had been left unmodified and also by the likelihood that some texts contained a higher proportion of dialogue than others. The effect was that the frequency tables did not afford proper comparisons between like and like. For our second trial, therefore, as in most statistical analyses of written texts, the contracted forms are expanded. ‘I’m’ counts as ‘I’ and ‘am’, ‘don’t’ as ‘do’ and ‘not’, and so on. (In most versions of standard English, this procedure is complicated only by the need to treat ‘cannot’, by analogy with ‘can’t’, as ‘can’ and ‘not’. Dialectal variants, which are not in issue here, produce some interesting problems.) As for the relationship between dialogue and narrative, a distinction based on the crude but tolerably objective determinant that that which is not marked in the original text as dialogue should be treated as narrative yields the following proportions for the texts previously compared: The American (1877), 47.6% dialogue; The American (1907), 47.2%; The Portrait of a Lady, 38.2%; and Mansfield Park, 40.2%. For our second trial, the word-frequencies tabulated are not for each text as a whole but for the narrative component of each text. (Here again, ‘narrative’ is taken simply as ‘non-dialogue’. An attempt to distinguish more subtly between the different facets of narrative, though rewarding in itself, would require some highly subjective decisions and complicate our paradigm.) Since this component makes up roughly half of each text, the number of columns can best be sustained by reducing the size of each segment from 8,000 to 4,000 words.

A further change is that the common homographic forms have been separated out. Eleven of the fifty words in the new set of frequency tables are affected. Both the infinitive and the prepositional forms of ‘to’ occur often enough to qualify for the word-list. ‘That’ qualifies both as a conjunction and as a relative pronoun but not as a demonstrative. ‘Which’ qualifies only as a relative pronoun, ‘so’ as an adverb of degree, and ‘no’ as an adjective. ‘In’, ‘for’, ‘on’, and ‘by’ qualify as prepositions but not as adverbial particles. (The use of ‘of’, ‘from’, and ‘with’ as adverbial particles is so rare that the distinction between preposition and particle has not been drawn.)

The most appropriate word-list for testing a doubtful case against the work of two known authors is one that amalgamates the most common words of both authors without being influenced by the doubtful case itself.
But it would be a counsel of heroism or folly to think of gathering the most common words of narrative in the collected works of Jane Austen and Henry James. Of James’s novels, I am conﬁned at present to The Portrait of a Lady and The Awkward Age. (Even if I were not privy to the true identity of Text A, I would recognize, I hope, that it is inadvisable to include any of those texts, like The American and Daisy Miller, which were heavily revised by James. To make an inadvertent use of two versions of the same work would be most misleading.) Having only these two narratives to work with, I decide to balance them with three of Jane Austen’s, adding Emma and Pride and Prejudice to Mansﬁeld Park. That gives over 191,000 words of James’s narratives and almost 237,000 of Jane Austen’s.

To give each author an equal inﬂuence on the composition of the word-list, we begin by standardizing the frequencies of the ﬁfty most common words of each narrative. Then, by ranking the standardized scores for all those words that lie within the top ﬁfty of either of James’s narratives, we can select the top ﬁfty overall. We likewise arrive at our list for Jane Austen by taking the most common ﬁfty, overall, from the top ﬁfty for each of her three narratives. Our ﬁnal joint-list, the basis of Tables 2.1–2.6, is then constructed, in the same fashion, by selecting the combined top ﬁfty from his list and hers.

We might now attempt a series of descriptive analyses like those undertaken in Section 2. But that would carry us away from our real question and waste the advantage of knowing that Text A is by one of two named authors. It is better to take a further preparatory step and try to establish whether a ‘signiﬁcant’ proportion—a point taken up a little later—of the ﬁfty word-types act as authorial discriminators, distinguishing the two narratives known to be by James from the three known to be by Jane Austen. If that can be done, we can, in other words, align Text A against a select group of words whose frequencies have independently been shown to discriminate between the two authors. This carries us beyond statistical description and allows us to test the hypothesis that Text A is the work of whichever of the authors we choose to specify.

To establish such a set of ‘discriminating word-types’, we take each word in turn, comparing each row of Tables 2.2 and 2.3 with the corresponding row of Tables 2.4, 2.5, and 2.6. (Here, too, the raw frequencies shown in the tables must ﬁrst be standardized: note that, in Tables 2.4 and 2.5, I have avoided the danger of ‘rounding up’ a sparse residue by creating ‘oversized’ last columns in D23 and E19.) For such comparisons between two groups, one uses distribution tests like the t-test, which treats of divergences from means, and the Mann-Whitney test, which treats of divergences from medians. The principle on which both tests rely is that scores characteristic of two populations, each fairly homogeneous, can sometimes be distinguished from the pattern characteristic of a single heterogeneous population: does the height of all 12-year-olds range across a single spectrum or is it possible to distinguish separate spectra for boys and girls? The simplest cases, which hardly require testing, are those where the scores for the two populations do not overlap and those where they are completely intermingled. The tests come into their own when the scores show an incipient distinction, whose ‘signiﬁcance’ needs to be assessed. On most occasions where it is appropriate to use them, the two tests yield not dissimilar results. But, since each test is better ﬁtted to cope with a different kind of aberration in the pattern of the scores, it seems prudent to employ both: in deciding whether an outcome is ‘signiﬁcant’, one has the beneﬁt of both ﬁndings.

It is obviously desirable that the two populations should contribute more or less equally to a comparison of this kind. How best to employ forty-seven columns from two of James’s narratives with ﬁfty-eight columns from three of Jane Austen’s? How best to overcome the imbalance, in James’s set, between the thirty-four columns of Table 2.2 and the thirteen of Table 2.3? My decision was to make a random selection of twelve columns apiece from Tables 2.2 and 2.3 and of eight apiece from Tables 2.4, 2.5, and 2.6. The resulting twenty-four columns for each author offer the parity desired while amounting to a sufﬁcient number of scores to allow the effective operation of the distribution tests. The columns yielded by the process of selection are as follows: from Table 2.2, columns 1, 4, 11, 13, 15, 20–2, 25, 30, 32–3; from Table 2.3, all except column 3; from Table 2.4, columns 5, 7, 10–11, 14, 16, 21–2; from Table 2.5, columns 8–10, 12, 15–17, 19; and from Table 2.6, columns 2, 8–10, 13–16.

When they are used to compare the scores for two known groups on any given variable, the essential function of the distribution tests is to assess the probability that the two groups do not differ from each other. The observer must decide upon an acceptable level of ‘statistical signiﬁcance’ and determine, in that light, whether the results for the several variables surpass that level often enough to make it unlikely that mere chance is a sufﬁcient explanation of the outcome. A probability of 0.05 or one chance in twenty is often chosen as the point at which the results are deemed ‘signiﬁcant’ enough to be pursued. A probability of 0.01 or one chance in 1,000 is usually regarded as ‘very highly signiﬁcant’. Whatever the chosen level, it must be met on a sufﬁcient number of occasions to show that the overall outcome is unlikely to be the effect of chance. Thus, though a probability of one in 100 is commonly regarded as 99/1 against, it is also 1/99 for. Provided the calculation is accurate, any such event will certainly occur—once, on average, in 100—and no ‘signiﬁcance’ attaches to it.

But when our two groups of twenty-four columns are compared, row by row for each of the ﬁfty variables, about half of the ﬁfty results satisfy both distribution tests at levels of probability beyond one chance in 100: if chance alone were at work, we could conduct twice as many trials before even one result might be expected to attain that level. An even more striking feature of Table 3, where these results are shown, is that more than twenty of the ﬁfty words satisfy both tests at levels beyond one chance in 1,000 and that thirteen do so at levels beyond one chance in 10,000. The fact that extremely high levels of probability occur here in such abundance is powerful evidence that our two populations differ, and differ most consistently. The consistency


97
of the difference can also be demonstrated in another way. Of the twenty-two words omitted from Table 3 because they do not operate as discriminators at the chosen level of significance, only five show 'weakly significant' results while seventeen satisfy neither of the distribution tests. Where one might have expected a body of middling results tapering off to either of the extremes, one finds two sets of extreme results with little in the middle. Such a pattern of results, in other words, is not one where several differentiae are in lively counterplay. It reflects one overriding differential, affecting most columns when it operates at all. The precise identity of such differentiae can never be declared with certainty; what looks like James versus Jane Austen might conceivably reflect some other feature of these narratives. But, in the face of Table 3, anyone not *parti pris* would presumably prefer to maintain that a difference of authorship is the most likely interpretation than to accept the burden of finding a better interpretation.

The list shown in Table 3 might reasonably have ended after 'with', the twenty-second word, reflecting the fact that 'from' is the first word not to satisfy both tests at a level beyond 0.001. Or, having excluded 'from', one might have gone on to the 0.01 threshold, adding the next three words. I have chosen in fact to go two steps still further down the hierarchy, including 'little' and 'him', which satisfy one but not both of the distribution tests at 0.01. 'Little' is the highest ranking lexical word to offer a significant result. 'Him' completes the set of masculine pronouns, all three occurring significantly more often in James. (A positive t-score in the fourth column marks the first group entered for analysis, a negative t-score the second: in the present case, James's columns were entered before Jane Austen's.)

The third-person plural pronouns run higher in Jane Austen and the two indefinite articles run very high in James. But, to a greater extent than when we were matching text against text and to a much greater extent than when we were matching James against James, the strongest results are found among the connectives (prepositions included), the deictics, and the words that express emphasis and limitation. Although the details vary when different pairs of authors are compared, these are usually the sorts of words that make relatively stable authorial discriminators.

The fact that most of the t-scores in Table 3, especially the strongest of them, are negative suggests that most of the word-frequencies for Jane Austen's work are more mutually consistent than are James's set. Yet (assuming our randomly chosen sets are representative) that is ostensibly an unexpected outcome because three of her narratives are matched against only two of his. In actuality, the outcome is in keeping with James's remarkable capacity for giving each of his narratives its own peculiar 'note'. In saying that, I am not overlooking the great length of James's literary career and the great stylistic changes that it brought. Whereas Jane Austen's narratives fall into two groups—the early novels and those she wrote at Chawton—James's differ considerably even within a given phase of his career. Within its limits, Graph 8 supports this judgement: the entries for Jane Austen are intermingled, while those for James lie in discrete clusters.

But the main function of Graphs 7 and 8 is to show the outcome of a predictive comparison, based on the twenty-eight words of Table 3, between the seventeen segments of the narrative of Text A and the two sets of twenty-four segments selected from the other five narratives. We know that Text A is the work of one or other of these authors. If the differences revealed by

Graph 7 Five narratives by James and Austen (selected segments) and Text A (word-plot for the twenty-eight most common 'discriminating word-types' of the five narratives)
the distribution tests are truly inter-authorial, the entries for Text A should align themselves with one set or the other even though Text A had no part in the choice of the word-list. If no stylistic determinants were at work among the common function words, the sixty-five entries in Graph 8 might arrange themselves in any of a great multitude of patterns. Various possible stylistic determinants might yield particular arrays. But only a genuine inter-authorial differentia or an astonishing freak of chance could produce the pattern shown in Graph 8 or another of those few patterns that would satisfy our prediction, distinguishing sharply between the two main sets of twenty-four entries and aligning the seventeen entries for Text A with one or other of those sets.

In Graph 7, from which Graph 8 derives, all those words that bore positive t-scores in Table 3 lie at the eastern or Jamesian end: the opposite is true of those that bore negative t-scores. Having now assigned Text A to James, we may also care to note which words distinguish it from James's other narratives. On the face of it, the main differentia is the set of masculine pronouns: it would be fair to infer that this narrative gives more attention than the others to a male protagonist—and even, therefore, to propose The American as a plausible Text A. But in order to pursue intra-authorial questions of that kind, it would be desirable to remove Jane Austen's narratives from consideration and start afresh in a more closely focused comparison of James with James. Such a comparison, which might for example be between early James and late, would undoubtedly yield other significant differentiae to put beside this group of pronouns. (If that were not so, the attempt, in the first part of this article, to distinguish between Texts A and D would not have been successful.) But intra-authorial questions lie beyond our present brief.

The conclusion that each and every segment of Text A is James's rather than Jane Austen's, as illustrated in Graph 8, may be expressed instead as a series of probabilities by using the method of co-variance to discriminate between the two sets of twenty-four segments and by using the result as the basis of a separate prediction for each segment of Text A. The conclusion can also be independently verified by excluding the two sets of twenty-four segments and, after replacing them by the two sets of 'residual segments', proceeding as before. The outcome, as shown in Graph 9, is slightly blemished by the fact that there is only one residual segment of The Awkward Age. But the enormous corroborative power of Graph 9 rests upon the fact that not one of the segments included in it has influenced the choice of the significant discriminators upon which it rests. We have, if I may be permitted to repeat it, derived a set of statistically significant discriminators from a comparison of two sets of twenty-four randomly selected segments from five narratives. We have set those forty-eight segments aside and made a fresh comparison between the two residual sets of segments of those same narratives and the seventeen segments of Text A. The pattern of entries in Graph 9 itself and its resemblance to Graph 8 bear out the same conclusions: that our method of analysis has yielded an inter-authorial distinction between these two sets of narratives and that it justifies us in assigning Text A to James.

4. Boundary Conditions: Putative and Real

The strength of these particular conclusions does not justify complacent generalizations. Let us consider some practical difficulties before turning to the more general question of the relationship between samples and populations. Perhaps James and Jane Austen are unusually easy to separate: having succeeded with chalk and cheese, we should try chalk and chalk. Perhaps the method of analysis requires large texts, like
Graph 9 Five narratives by James and Austen (residual segments) and Text A (based on the twenty-eight most common ‘discriminating word-types’ of the five narratives) A, Text A; B, Portrait of a Lady; C, The Awkward Age; D, Mansfield Park; E, Emma; F, Pride and Prejudice

novels, which offer the statistical luxury of many segments of several thousand words apiece. Perhaps it is too easy to establish inter-authorial differentiae when both authors are working, as here, in much the same literary form. The only way to meet these doubts and others like them is to assess them as they arise, framing precise questions and choosing appropriate sets of data.

Graph 10 treats of three varieties of chalk—or Millstone Grit. There may well be other groups of three authors whose styles happen to resemble each other more closely than do those of the Bronté sisters; but I know of no other trio who might, from their whole background, be expected to write more like each other. By taking the ‘histories’—the first-person retrospective narratives that each of them embeds in her novels, as when Rochester tells Jane of his past life—as my database, I have tried to restrict the possible effect of genre-difference. The difference is not quite eliminated because the histories in Anne’s novels are recorded in letters and journals, whereas those in her sisters’ novels are recounted orally by one character to another. But the striking feature of Graph 10, which rests upon fifteen or 20,000 words from each of the three sisters, is that the three sets can be so clearly distinguished. To predict which sister was the author of a doubtful text, we need only modify the procedures described earlier in such a way as to place three rather than two candidates upon an equal footing.
Graph 11  Scott and Byron: selected letters in 500-word segments (based on the twenty most common words of the corpus). S, Scott—letters; B, Byron—letters

By treating of brief, non-literary texts, Graph 11 does much to allay two other kinds of doubt. The texts are small sets of personal letters by two near-contemporaries, running to about 8,000 words for Scott and 6,000 for Byron. To minimize the effect of genre-difference (which often manifests itself, in letters, as a reflection of the sort of person who is addressed), I have chosen letters addressed to the authors’ distinguished wellwishers, Lady Abercorn and Lady Melbourne respectively. Each set is broken into segments of only 500 words and the word-list is truncated to the twenty most common words of the two sets. (The truncation is necessary to ensure that even the less common of the common words occur often enough to admit statistical analysis. As a rule of thumb, I find that the distribution tests have room to work properly when the mean raw frequency of a given word in either set of texts is over five. In such cases, the corresponding frequency for the other set of texts can safely lie at any level.) In another trial of this same comparison between Scott and Byron, when the segments were reduced to 400 words and the word-list to the top fifteen, the separation of the two clusters began to weaken. But Graph 11, based on segments of 500 words and using the twenty most common words, shows a clear distinction between the two sets of letters. In two ‘forensic’ enquiries, using texts of as little as 400 words apiece, I have been able to distinguish between sets of anonymous letters and control-sets of signed letters and to test the result by a predictive analysis of other letters known to have been written by the alleged author of each anonymous set. My impression is that the narrow linguistic repertoires of ordinary people make their ‘signatures’ easier to identify than those of more gifted writers.

By putting a multiplicity of genres beside each other, Graphs 12 and 13 extend the putative limits of this kind of analysis in yet another way. After excluding Scott’s dialectal narratives, I have drawn comparisons in which the diversity of his work and the almost equal diversity of Byron’s are represented. (To exclude Scott’s work in broad Scots is to lose much of his best writing: to include it, however, would afford a spurious advantage to my attempt to assess any inter-authorial distinction that might transcend the shared characteristics of each genre. Specimens of their work as lyrical poets are omitted only because they have not yet been prepared for analysis.)

Graph 12, which is based upon the sixty most common words, shows a complex pattern in which genre transcends authorship. Its most obvious feature is a separation between prose and verse, the only exception being the entry for Scott’s prose-play, The House of Aspen, which lies with the entries for the verse-dramas of both writers towards the western end of the graph. The narrative poems of both writers lie to the north-west, the most northerly being long excerpts from their first-person retrospective narratives, while the three that lie nearer to the entries for drama are from Don Juan. (The almost dialogic give and take between narrator and reader in that work lends it a semblance of the patterns of dialogue that unite the dramatic writings of both authors.) The only entry for Byron’s prose sets his letters beside Scott’s towards the southerly edge of the graph. The slightly more formal cast of Scott’s fictive letters—they are those of Alan and Darsie in the first 100 pages of Redgauntlet—sets them with his autobiographical memoir and his first-person retrospective narratives in the south-east. The two entries for Scott’s third-person narratives—‘The Highland Widow’ and ‘The Fortunes of Martin Waldeck’—lie high up at the eastern extremity of Graph 12.

The word-list for Graph 13 is derived by using the t-test and the Mann–Whitney test to select, from among the sixty words used for Graph 12, those that yield the strongest inter-authorial differences. Twenty of the sixty
discriminate at levels beyond 0.01 (or one in 100), ten at levels beyond 0.001 (or one in 1,000). Counting from the top, these ten strongest (the basis of Graph 13) include 'but', 'not', 'all', and 'no' on Byron's side; 'of' and 'the' on Scott's; and 'be', the demonstrative 'that', 'what', and 'or' on Byron's. The implied habits of emphasis in Byron and of impersonal amplification in Scott are not unexpected. As for Graph 13 itself, traces of the differences of genre and of the broader division between prose and verse can still be seen. But, though the entry for Byron's letters reaches to the edge of Scott's cluster, the graph shows a clear inter-authoral separation even among the entries for those genres in which both writers work. Given the heterogeneity of the texts, that is a strong result. It even implies that, at least in circumstances where a direct comparison can be undertaken, it may be possible to distinguish the work of some pairs of writers no matter what kind of documents are involved.

That may seem an unduly cautious inference from a result in which a distinction of much that kind has been so plainly drawn. Yet even in its application to these two writers, leaving other pairs aside, our result rests only upon samples. Conclusions about a whole population can legitimately be drawn from small samples if it is fair to assume that the population is tolerably homogeneous. But, though we are making worthwhile progress, the statistical analysis of literary and other
texts does not yet afford firm empirical guidance on such matters as the degree of homogeneity that may reasonably expected in an author's œuvre.

Even if such guidance were available or if, instead, we bypassed the problems of sampling by analysing everything ever written by a given pair of authors, we could show at most that they never did write like each other and not that they could never do so. In resting my case upon one of the great commonplace of inductive logic, I look back towards my introductory comments. Given the information that makes nice questions possible and a nice enough formulation of those questions, literary statisticians can undoubtedly help to identify the authors of doubtful texts. They will usually rely on the analysis of stylistic resemblances, differences, and concomitant variations to adduce evidence of an inductive cast. Like anyone else who does that, from nomadic hunters to astronomers, they will sometimes be able to present a compelling argument but they will never be entitled to claim certainty. Whether or not they make use of computers in gathering their data, they need ask only one concession from their colleagues: that their evidence should not be treated with special deference or special scepticism but that it be taken, case by case, upon its merits.

Acknowledgements

My research could not have been conducted without generous support from the Australian Research Council and the University of Newcastle, NSW. The computer package employed for the analysis forms part of MINTAB (University of Pennsylvania). I owe more personal debts to several colleagues, especially Alexis Antony, David Hoole, Sandra Britz, Nicole Cox, and John Lambert for their work in the preparation and analysis of the texts. N. M. McLaren, N. Collins George, C. S. Wallace, and Annette Dobson have given valuable statistical advice at different times. The responsibility for any errors and misjudgements remains my own.

Notes

2. Despite a subsequent revision of the machine-readable versions of the texts, the totals given for Mansfield Park and Pride and Prejudice differ little from those set out in my Computation into Criticism (Oxford, Clarendon Press, 1987), p. xv. In Emma, however, the total number of words assigned to the narrative is now reduced by 500. The main changes are in passages marked by 'Emma said' but not actually spoken aloud to any other character. The new figures tally with our forthcoming electronic version of Chapman's edition of The Complete Works of Jane Austen (Oxford, Oxford University Press, 1992).

Appendix

Annotated List of 'Mintab' Commands

When applied to standardized equivalents of the appropriate parts of Tables 1 and 2, the following set of MINTAB commands will produce rough drafts of Graphs 1–9

MTB> oh 0  # allow continuous scrolling
      # through loops
MTB> outfile 'table out'  # name output file
MTB> read 'table doc' c1-c16  # read tabulated data into
      # columns
MTB> let k1 = no. of rows  # i.e. no. of word-types in
      # 'table doc'
MTB> let k2 = no. of columns  # i.e. no. of 'texts', authors, etc.,
      # in 'table doc'
MTB> execute 'rowprop.mtb'  # use first stored loop described
      # below to convert text-
      # percentages of 'table doc.' to
      # decimal fractions of the total
      # of entries in each row
MTB> let k4 = 400 + k1  # prerequisite for executing
      # 'eigenloop mtb'
MTB> set c204  # set labels for word-types. e.g
      # 1 50
MTB> set c304  # set labels for text or segments
      # e.g. 1 44
MTB> set c305  # set further group-labels. e.g
      # 16(1) 28(2)
MTB> execute 'eigenloop mtb'  # use second stored loop
      # described below to
      # carry out correlation and eigen
      # analysis and make rough lines-
      # plots of results
MTB> stop  # (unless making use of following
      # optional step)

# Optional Extra Step, to precede 'stop'
# To facilitate direct comparisons between graphs, it is often useful
# to rotate a graph upon either or both of its axes.

MTB> let c201 = -1 * c201  # rotate horizontal axis of 'word-
      # plot'
MTB> let c301 = -1 * c301  # rotate horizontal axis of 'text-
      # plot'
      # likewise c202 and c302 for
      # vertical rotation

Commands in First Stored Loop, 'rowprop mtb'

The functions of this loop are: to standardize 'text-percentages' (the raw frequencies shown in each column expressed as percentages of the amount of text covered by that column) by converting them into decimal fractions of the sum of each row; and to transpose the rows and columns so that data for word-types, not texts, can be correlated. The simpler process of correlating text-columns was used in my Computation into Criticism (1987).

rsum c1-c2 k100  # calculate divisor for use in
      # conversion

copy c1-c2 k1 m1  # columns and rows are now

transpose m1 m2

copy m2 c1-c1k1  # transposed

let k3 = 1
noecho
store
let ck3 = ck3/c100(k3)  # convert the first column to
      # decimal fractions

let k3 = k3 + 1  # allow progress to second and
      # later columns

end

print k3  # print the commands k1 times

execute the commands k1 times

print c1 c1  # show that the procedure is
      # complete

Commands in Second Stored Loop, 'eigenloop mtb'

The functions of this loop are: to derive a Pearson product-moment correlation matrix from the data; to extract, scale, and compound the


103
principal components of the matrix, and to print rough plots of the results and prepare them for more accurate plotting.

echo copy c1-cck1 m3
correlate c1-cck1 m4
eigen m4 c151 m5

let k10 = sum(c151)
let c152 = c151/k10
let c153 = 100*c152
let c155 = 100*c153
round c153 c155
let c157 = c153/100
write c151, c153

# make vector plot for first two
# vectors
# make similar plots for other
# combinations of vectors

# compound two earlier matrices
to give new matrix in which data
# for texts are fitted to word-data

# make data-plot now shown as m3, for
# later use
# correlation matrix now in m4
# eigenvectors now in m5.
# eigenvalues m c151

# make data-plot now shown as m3, for
# later use
# correlation matrix now in m4
# eigenvectors now in m5.
# eigenvalues m c151

# scale values of first new vector
# by multiplying by k2 (number of
texts of corresponding step
above), and by multiplying
# results by 10 for pure
# convenience
# likewise for second new vector
# likewise for third new vector
# print new vectors and their
# author- or text-labels
# make text-plot for first two new
# vectors
# make similar plots for other
# combinations of vectors

Table 1.1 Text A: raw frequencies for the fifty most common words of the whole text (the columns represent successive segments of 8,000 words: the last segment comprises 11,946 words).

<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
<th>A9</th>
<th>A10</th>
<th>A11</th>
<th>A12</th>
<th>A13</th>
<th>A14</th>
<th>A15</th>
<th>A16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>of</td>
<td>be</td>
<td>an</td>
<td>in</td>
<td>that</td>
<td>for</td>
<td>it</td>
<td>as</td>
<td>on</td>
<td>by</td>
<td>not</td>
<td>to</td>
<td>if</td>
<td>with</td>
</tr>
<tr>
<td>2</td>
<td>to</td>
<td>be</td>
<td>in</td>
<td>be</td>
<td>of</td>
<td>on</td>
<td>in</td>
<td>it</td>
<td>be</td>
<td>on</td>
<td>on</td>
<td>on</td>
<td>to</td>
<td>in</td>
<td>of</td>
</tr>
<tr>
<td>3</td>
<td>be</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>be</td>
<td>be</td>
<td>be</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>of</td>
</tr>
<tr>
<td>4</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>5</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>6</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>7</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>8</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>9</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
<tr>
<td>10</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>of</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>of</td>
<td>be</td>
<td>be</td>
</tr>
</tbody>
</table>

104

## Table 1.4 Text. raw frequencies for the fifty most common words of Text A (the columns represent successive segments of 8,000 words; the last segment comprises 4,886 words)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>22,482</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>17,815</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>15,741</td>
</tr>
<tr>
<td>4</td>
<td>and</td>
<td>12,397</td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>10,991</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>10,503</td>
</tr>
<tr>
<td>7</td>
<td>for</td>
<td>9,765</td>
</tr>
<tr>
<td>8</td>
<td>that</td>
<td>9,605</td>
</tr>
<tr>
<td>9</td>
<td>it</td>
<td>9,179</td>
</tr>
<tr>
<td>10</td>
<td>be</td>
<td>8,871</td>
</tr>
<tr>
<td>11</td>
<td>is</td>
<td>8,576</td>
</tr>
<tr>
<td>12</td>
<td>with</td>
<td>8,373</td>
</tr>
<tr>
<td>13</td>
<td>on</td>
<td>8,204</td>
</tr>
<tr>
<td>14</td>
<td>as</td>
<td>8,132</td>
</tr>
<tr>
<td>15</td>
<td>this</td>
<td>7,976</td>
</tr>
<tr>
<td>16</td>
<td>in which</td>
<td>7,460</td>
</tr>
<tr>
<td>17</td>
<td>and from</td>
<td>7,428</td>
</tr>
<tr>
<td>18</td>
<td>and as</td>
<td>7,116</td>
</tr>
<tr>
<td>19</td>
<td>a proposition</td>
<td>6,913</td>
</tr>
<tr>
<td>20</td>
<td>with which</td>
<td>6,891</td>
</tr>
</tbody>
</table>

## Table 2.1 Text A: raw frequencies for the most common words of narrative in the main set of texts (the columns represent successive segments of 4,000 words of narrative: the last column comprises 5,142 words)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>the</td>
<td>22,482</td>
</tr>
<tr>
<td>2</td>
<td>of</td>
<td>17,815</td>
</tr>
<tr>
<td>3</td>
<td>to</td>
<td>15,741</td>
</tr>
<tr>
<td>4</td>
<td>and</td>
<td>12,397</td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>10,991</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>10,503</td>
</tr>
<tr>
<td>7</td>
<td>for</td>
<td>9,765</td>
</tr>
<tr>
<td>8</td>
<td>that</td>
<td>9,605</td>
</tr>
<tr>
<td>9</td>
<td>it</td>
<td>9,179</td>
</tr>
<tr>
<td>10</td>
<td>be</td>
<td>8,871</td>
</tr>
<tr>
<td>11</td>
<td>is</td>
<td>8,576</td>
</tr>
<tr>
<td>12</td>
<td>with</td>
<td>8,373</td>
</tr>
<tr>
<td>13</td>
<td>on</td>
<td>8,204</td>
</tr>
<tr>
<td>14</td>
<td>as</td>
<td>8,132</td>
</tr>
<tr>
<td>15</td>
<td>this</td>
<td>7,976</td>
</tr>
<tr>
<td>16</td>
<td>in which</td>
<td>7,460</td>
</tr>
<tr>
<td>17</td>
<td>and from</td>
<td>7,428</td>
</tr>
<tr>
<td>18</td>
<td>and as</td>
<td>7,116</td>
</tr>
<tr>
<td>19</td>
<td>a proposition</td>
<td>6,913</td>
</tr>
<tr>
<td>20</td>
<td>with which</td>
<td>6,891</td>
</tr>
</tbody>
</table>

### Abbreviations
- (i) = infinitive, (p) = preposition, (c) = conjunction, (rp) = relative pronoun; (av d) = adverb of degree, (adj) = adjective
Table 2.4 Mansfield Park: raw frequencies for the fifty most common words of narrative in the main set of texts (the columns represent successive segments of 4,000 words of narrative: the last column comprises 6,186 words)

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1684452</td>
</tr>
<tr>
<td>of</td>
<td>474812</td>
</tr>
<tr>
<td>and</td>
<td>1235236</td>
</tr>
<tr>
<td>she</td>
<td>255678</td>
</tr>
<tr>
<td>he</td>
<td>135467</td>
</tr>
<tr>
<td>it</td>
<td>112234</td>
</tr>
<tr>
<td>his</td>
<td>234567</td>
</tr>
<tr>
<td>this</td>
<td>456789</td>
</tr>
<tr>
<td>you</td>
<td>789012</td>
</tr>
<tr>
<td>my</td>
<td>123456</td>
</tr>
<tr>
<td>our</td>
<td>345678</td>
</tr>
<tr>
<td>our</td>
<td>456789</td>
</tr>
<tr>
<td>but</td>
<td>567890</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>all</td>
<td>789012</td>
</tr>
<tr>
<td>all</td>
<td>890123</td>
</tr>
<tr>
<td>all</td>
<td>901234</td>
</tr>
<tr>
<td>all</td>
<td>012345</td>
</tr>
<tr>
<td>all</td>
<td>123456</td>
</tr>
<tr>
<td>all</td>
<td>234567</td>
</tr>
<tr>
<td>all</td>
<td>345678</td>
</tr>
<tr>
<td>all</td>
<td>456789</td>
</tr>
<tr>
<td>all</td>
<td>567890</td>
</tr>
<tr>
<td>and</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
<tr>
<td>and</td>
<td>345678</td>
</tr>
<tr>
<td>and</td>
<td>456789</td>
</tr>
<tr>
<td>and</td>
<td>567890</td>
</tr>
<tr>
<td>all</td>
<td>678901</td>
</tr>
<tr>
<td>and</td>
<td>789012</td>
</tr>
<tr>
<td>and</td>
<td>890123</td>
</tr>
<tr>
<td>and</td>
<td>901234</td>
</tr>
<tr>
<td>and</td>
<td>012345</td>
</tr>
<tr>
<td>and</td>
<td>123456</td>
</tr>
<tr>
<td>and</td>
<td>234567</td>
</tr>
</tbody>
</table>
Table 2.6  Pride and Prejudice: raw frequencies for the fifty most common words of narrative in the main set of texts (the columns represent separate segments of 4,000 words of narrative; the last column comprises 4,007 words)

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1876157</td>
</tr>
<tr>
<td>and</td>
<td>1486101</td>
</tr>
<tr>
<td>be</td>
<td>769059</td>
</tr>
<tr>
<td>of</td>
<td>4369155</td>
</tr>
<tr>
<td>as</td>
<td>10610811</td>
</tr>
<tr>
<td>in</td>
<td>848093</td>
</tr>
<tr>
<td>to</td>
<td>9596518</td>
</tr>
<tr>
<td>that</td>
<td>6273115</td>
</tr>
<tr>
<td>10</td>
<td>482499</td>
</tr>
<tr>
<td>11</td>
<td>394419</td>
</tr>
<tr>
<td>12</td>
<td>323413</td>
</tr>
<tr>
<td>13</td>
<td>292312</td>
</tr>
<tr>
<td>14</td>
<td>282211</td>
</tr>
<tr>
<td>15</td>
<td>262511</td>
</tr>
<tr>
<td>16</td>
<td>232411</td>
</tr>
<tr>
<td>17</td>
<td>212211</td>
</tr>
<tr>
<td>18</td>
<td>202211</td>
</tr>
<tr>
<td>19</td>
<td>192211</td>
</tr>
<tr>
<td>20</td>
<td>192211</td>
</tr>
<tr>
<td>21</td>
<td>192211</td>
</tr>
<tr>
<td>22</td>
<td>192211</td>
</tr>
<tr>
<td>23</td>
<td>192211</td>
</tr>
<tr>
<td>24</td>
<td>192211</td>
</tr>
<tr>
<td>25</td>
<td>192211</td>
</tr>
<tr>
<td>26</td>
<td>192211</td>
</tr>
<tr>
<td>27</td>
<td>192211</td>
</tr>
<tr>
<td>28</td>
<td>192211</td>
</tr>
<tr>
<td>29</td>
<td>192211</td>
</tr>
<tr>
<td>30</td>
<td>192211</td>
</tr>
<tr>
<td>31</td>
<td>192211</td>
</tr>
<tr>
<td>32</td>
<td>192211</td>
</tr>
<tr>
<td>33</td>
<td>192211</td>
</tr>
<tr>
<td>34</td>
<td>192211</td>
</tr>
<tr>
<td>35</td>
<td>192211</td>
</tr>
<tr>
<td>36</td>
<td>192211</td>
</tr>
<tr>
<td>37</td>
<td>192211</td>
</tr>
<tr>
<td>38</td>
<td>192211</td>
</tr>
<tr>
<td>39</td>
<td>192211</td>
</tr>
<tr>
<td>40</td>
<td>192211</td>
</tr>
<tr>
<td>41</td>
<td>192211</td>
</tr>
<tr>
<td>42</td>
<td>192211</td>
</tr>
<tr>
<td>43</td>
<td>192211</td>
</tr>
<tr>
<td>44</td>
<td>192211</td>
</tr>
<tr>
<td>45</td>
<td>192211</td>
</tr>
<tr>
<td>46</td>
<td>192211</td>
</tr>
<tr>
<td>47</td>
<td>192211</td>
</tr>
<tr>
<td>48</td>
<td>192211</td>
</tr>
<tr>
<td>49</td>
<td>192211</td>
</tr>
<tr>
<td>50</td>
<td>192211</td>
</tr>
</tbody>
</table>

For abbreviations see Table 2.1

Table 3. Five narratives by Henry James and Jane Austen: most significant scores on distribution tests for fifty most common words

<table>
<thead>
<tr>
<th>Rk.</th>
<th>t</th>
<th>df</th>
<th>Threshold</th>
<th>t p</th>
<th>W p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.01</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>0.10</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>3</td>
<td>0.25</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>4</td>
<td>0.50</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>10</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

The first three columns identify the word itself and its rank in the orginal hierarchy of frequencies. The fourth column gives the t-score and the next two match that score against the appropriate level of the Mann–Whitney test. The two last columns summarize the outcome as levels of probability deriving from the t-test and the Mann–Whitney test respectively.

1. Introduction

Authorial style can be studied in many ways, using any of the methods or approaches of stylistics. Here, I concentrate on the distinctiveness of authorial style, mainly in the area of lexis, using the methods of computational stylistics. This approach will emphasize the comparative nature of style, while at the same time providing an opportunity to explore the styles of four authors in a context in which the subject matter, plot, characters, and themes remain constant for each pair of authors. This exploration will also suggest further directions for investigating authorial style, both in these two cases and more generally.

Authorial style is chiefly linguistic, though a description of the styles of some authors might also invite attention to punctuation and other graphological features or to illustrations. In some cases, even the page size and layout and the physical characteristics of the text, such as the weight and colour of the paper, may be considered stylistic features. An author’s characteristic philosophy, world view, themes, tones, topics, and characters are all expressed linguistically, as are the more obviously linguistic elements of grammar, lexis, morphology, phonology, figures of speech, cohesion, and collocation (for an excellent checklist of features, see Leech and Short 1981: 75-82). Style, and especially prose style, is also distributed and patterned. Although there are obviously quite local and striking stylistic effects, it is difficult to know how to take a truly isolated effect: When Dickens opens *A Tale of Two Cities* with ‘It was the best of times,’ consider the effect of following it with ‘it was the worst of times,’ and the importance of the pattern of opposites that loads the first half of his sentence. Style is also essentially, if not always explicitly, comparative. Any remark on a stylistic characteristic implies a comparison, even if it does not state one. To comment on the ‘staccato effect of Hemingway’s short sentences’ implicitly asserts that his sentences are shorter than ‘normal.’ Otherwise, how would the effect arise? The vexed question of the appropriate norm for any given text remains vexed, but the widespread availability of electronic texts now allows a much wider range of defensible answers.

The linguistic, distributed, patterned, and comparative nature of style lends itself naturally to computational methods. While I do not suggest that stylistic analysis must be computational, I hope to show here that computational methods can reveal stylistic features and characterize authorial style in ways that would be practically impossible in any other way, and that statistical analysis is ‘an essential and important tool in stylistic description’ (Leech and Short 1981: 71). I concentrate here on words, for reasons of both convenience and principle. Words are easily identifiable and countable, compared with figures of speech or syntactic patterns. They are also very frequent, which makes computational methods both necessary and appropriate. In addition, unlike sequences of letters, which some recent work suggests may more effectively differentiate authors (Clement and Sharp 2003), words are clearly, if not unproblematically, meaningful.

Although “word” seems an intuitively simple concept, various definitions are defensible. For my purposes, a word (type) is a unique sequence of alphanumeric characters not broken by a space or by any punctuation except the hyphen or the apostrophe. (A type is a unique form, a token is an instance of a type: the previous sentence contains two tokens of the type “by”.) This definition treats contractions as single words. In some cases, one might be particularly interested in modal verbs or negatives, and so might want to separate contractions; however, in the late 19th
and early 20\textsuperscript{th} centuries (the period of the texts I will be analyzing here), contracted forms were increasing in frequency, so that they seem likely to be stylistically meaningful. My definition also treats hyphenated words as single words, which seems the simplest choice, though in some analyses one might well want to analyze the elements of such words separately. Unfortunately, my definition does not distinguish homographs, such as the verb and noun meanings of bear, but the corpora I analyze are too large to make such distinctions feasible, and I will be analyzing very large numbers of words, so that these relatively rare words are unlikely to be problematic. Finally, in some cases, one might want to treat inflected or variant forms of a word as equivalent (go/went/gone, book/books), but, as we will see, different forms of a word often behave quite differently, so that it seems safer, as well as much simpler, to treat them as different (see Sinclair 2003, Task 18, for a discussion of the different behaviour of eye and eyes).

Most of the methods of computational stylistics have their origins in the field of authorship attribution, where the focus is on classification, on searching for the linguistic equivalent of the fingerprint. These methods capitalize on the linguistic, distributed, patterned, and comparative nature of style. There are more cases of unsolved literary attribution than is often realized; in the Victorian era, for example, many novels and even more periodical fiction was published anonymously, including many texts by canonical authors. My own practice, however, is to apply the methods of literary attribution to the analysis and exploration of individual authorial style. Here I focus on Wilkie Collins’s \textit{Blind Love} (1890, 1900, 2003) and Steven Crane’s \textit{The O’Ruddy} (1903, 1971) two novels left unfinished at the authors’ deaths and completed, respectively, by Walter Besant and Robert Barr.

2. \textit{Blind Love}

Wilkie Collins, friend of Charles Dickens, inventor of the detective novel, and one of the most widely read and successful of Victorian novelists, did not live to complete his final novel, \textit{Blind Love}. When he realized that he would not be able to meet the demanding schedule of the serial publication of the novel, he asked Walter Besant, another prolific, though less important and now little-read novelist, to finish it for him, and furnished an extensive synopsis of the part of the novel he had not finished. In this case, we have letters, manuscripts, and other records that indicate Collins had completed a long prologue and chapters 1-48 of the novel, and that Besant wrote chapters 49-64 and an epilogue (referred to below as chapter 65) (Collins 2003).

2.1 The Most Frequent (Function) Words: Cluster Analysis

Among the most widely used variables for authorship attribution are the frequencies of the 30-50 most frequent function words. The popularity of these words as authorship indicators is based largely on the reasonable assumption that an author’s use of such high-frequency and low-content words is likely to be habitual and ingrained, and that authors are unlikely to alter their use of such words intentionally. These words are extremely frequent and, unlike nouns, adjectives, and verbs, are relatively free from the effects of theme, setting, tone, and other literary characteristics that often vary widely among an author’s works. As Fig. 1 shows, the frequencies of the 50 most frequent function words easily distinguish Collins from Besant.
The Cluster Analysis that produces Fig. 1 begins with a list of the 50 function words that are the most frequent in all of the texts combined (the, to, of, and, i, in, a, you, he, her, was, that, it, she, my, his, had, me, on, with, at, as, for, is, have, him, be, not, which, by, this, what, if, your, from, no, but, there, will, we, are, an, were, been, they, who, out, do, so, or). It works in such a way that the further to the left that two texts or groups of texts join into one cluster, the more similar they are in terms of the frequencies of all 50 of the words considered simultaneously (see https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html for detailed instructions for doing Cluster Analysis). Thus, Collins’s novels are generally more similar to each other in their use of these 50 words than Besant’s are, and Collins’s three latest novels (at the bottom of the graph) are least like his other novels. Although the frequency of words like the, to, and of do not seem likely to be very interesting to a student of style, John F. Burrows (1987)
has shown that frequent function words can yield significant stylistic insights. Nevertheless, recent work (Hoover 2001, 2004, 2007) has shown that, however reasonable the assumption that the most frequent function words are the most appropriate words to use for authorship attribution, in practice, increasing the size of the word list to include all of the 600-1,200 of the most frequent words, regardless of type, almost invariably increases the accuracy of an analysis of novel-sized texts. For the set above, for example, an analysis based on the 990 MFW, which together account for more than 80 percent of all the words of the texts, not only keeps the two authors distinct, it also groups the Collins novels neatly into an early group, 1860-76, and a late group, 1879-88. (Minitab (2005), my preferred statistical software, has a practical limit of about 990 words, which accounts for the odd-seeming use of 990 words.) Fortunately, computational stylistics provides methods for identifying the characteristic vocabulary of writers that extend beyond the most frequent words of the text, and *Blind Love* provides an opportunity to examine more closely what such methods can teach us about the styles of Collins and Besant (on Cluster Analysis and chronological style, see Hoover 2007).

### 2.2 Student’s T-test

Student’s t-test is a well-studied method for testing whether or not any difference between two groups is likely to have arisen by chance. (Any introductory statistics text will provide a full explanation; for a classic discussion of the use of t-tests in authorship and stylistics, see Burrows 1992.) In this case, we want to identify words that are used so differently in the writings of Wilkie Collins and Walter Besant that those differences are extremely unlikely to be a result of chance, and therefore are very likely to indicate real stylistic differences between the writers. Other variables, such as sentence length, word length, the frequencies of various word classes, syntactic units, or any other feature that can be counted, can be tested using this method, but I want to concentrate here on words as an obviously meaningful category, and the other variables just mentioned would require a prohibitive amount of manual identification. I begin by collecting substantial samples of writing by the two authors, in this case, four third-person novels by Besant (1882-1893) and three by Collins (1883-1885):

- **Besant:** *The Revolt of Man, In Luck at Last, The Ivory Gate, The Rebel Queen*, about 308,000 words.
- **Collins:** *Heart and Science, I Say No, The Evil Genius*, about 358,000 words

The corpus just described provides a kind of norm, here controlled for point of view, genre, and date; as noted above, Collins’s novels show signs of chronological differences in style, and genre effects are often stronger than those of authorship.

I first create a word frequency list for these seven texts combined, then delete words that occur only once or twice in the entire set, reducing the word list from more than 18,000 words to about 9,000. The t-test favours relatively frequent words, so these rare words, which could hardly be called ‘characteristic,’ can safely be deleted; their distributions will not be statistically significant. I also delete personal pronouns, which are closely related to the number and gender of characters; given that the goal is to examine *Blind Love*, this seems appropriate, though there are so few pronouns that they are unlikely to have a noticeable effect on the analysis. Next, I delete all words for which a single text accounts for more than 90 percent of the occurrences;
Hoover Authorial Style

these are almost exclusively proper nouns, typically the names of characters and places. Although the character names an author selects clearly have stylistic significance, removing these words prevents the analysis from inappropriately treating a high frequency of, say, Tom in two novels as evidence that they are by the same author. Finally, I also delete words that appear only in Collins or Besant; the t-test cannot be calculated when a variable is absent from either group. (I will examine later a method of identifying characteristic vocabulary that includes words like these.) The process just described leaves about 6,600 words upon which to perform t-tests.

To produce strong results, the t-test needs a large number of measurements, so I divide the seven novels into 167 sections of about 4,000 words each, 90 sections for Collins and 77 for Besant, collect the frequencies of the 6,600 words in each section, and perform a t-test for every word. Finally, I sort the results on the p value that results, so that I can select the most distinctive words for further analysis. The normal practice is to retain any variables for which p < .05; that is, here, words with distributions that have a probability of less than 5 percent of occurring by chance. In practice, however, I often select a smaller group of even better marker words with p < .01, or even p < .001. In this case, among the 6,600 words are about 1,700 words with p < .05, more than 1,000 with p < .01, and more than 500 with p < .001. In all of these groups, the words that Collins favours outnumber those that Besant favours; nearly 1,000 of the 1,700 are Collins words (see https://files.nyu.edu/dh3/public/ClusterAnalysis-PCA-T-testingInMinitab.html for detailed instructions for doing multiple t-tests).

2.2.1 Student’s T-test: Distinguishing Besant and Collins

This is a large number of very distinctive marker words for Collins and Besant, and, as Fig. 2 shows, the 993 most distinctive marker words, about 495 for each author, clearly separate new texts by Besant and Collins that had no part in the creation of the list of words. This graph is based on six texts for each author, an additional novel and five short stories for Besant and six additional novels for Collins (I began with the 1,000 most distinctive words, then deleted those that did not occur in any of the new texts). To keep the graph readable, I divided the novels into 10,000 word sections and retained only half the sections for each novel. Including short stories and first-person novels intentionally makes the task more difficult; nevertheless, the marker words derived from the authors’ third-person novels are obviously more generally characteristic of the authors, and in this case, at least, limiting the original corpus to third-person novels seems to have been an unnecessary precaution.
2.2.2 Student’s T-test and the Authorship of the Chapters of *Blind Love*

With the chapters of *Blind Love* added to this analysis (with the texts divided into 1,000 word sections to better match the length of the chapters), the change in authorship after chapter forty-eight is starkly apparent, as Fig. 3 shows, in spite of Besant’s use of the long synopsis Collins provided. The graph is based on the sums of the frequencies of the 500 most distinctive Besant words and the 500 most distinctive Collins words in each section. (Only a few sections of the novels are shown, and the frequencies of the Collins words are multiplied by -1 to make the graph easier to read. A cluster graph is very crowded with so many texts, but shows essentially the same results.)
Fig. 3 1,000-Word Sections of Besant, Collins, and Blind Love: The 1,000 Most Distinctive T-Tested Words
2.2.3 Student’s T-test and Characteristic Vocabulary

These marker words are not simply useful in distinguishing the authors from each other, they can also help to characterize the authors’ styles, though it is important to remember that the t-tested words are selected in such a way as to bring out the differences between Collins and Besant. (The method can be modified by testing one author against a group of other authors, which produces a more comprehensive characterization.) 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

The relative banality of the Besant’s marker words compared to Collins’s is quite striking. Most Besant words are frequent function words: 12 of the 20 rank in the top 100 most frequent words in the entire corpus; only 1, everything, ranks above 200, and the average rank is 106. In the Collins list, only 8 words rank in the top 100, 7 rank above 300, and the average rank is 262. Also striking is the heavy concentration of Collins words related to speech presentation (answered, asked, inquired, resumed, suggested; possibly also reply and reminded). This concentration is partly a result of a somewhat higher proportion of dialogue in Collins than in Besant—as measured very crudely by calculating the ratio of quotation marks to the total number of words—but that difference is small enough that further research seems likely to discover more interesting causes. The presence of added, begged, declared, exclaimed, explained, expressed, muttered, rejoined, and said as likely speech markers among the other Collins marker words, compared with only gasped, groaned, murmured, replied, and stammered for Besant, suggests that the two authors present speech in very different ways. A brief examination of the beginnings of Collins’s The Evil Genius and Besant’s In Luck at Last suggests that, while both authors frequently leave out speech markers entirely, Besant is more likely than Collins to replace a reporting word with a phrase like ‘he turned a beaming and smiling face upon the assistant,’ or ‘Mr. James’s cheek flushed.’ Here, as elsewhere below, I can only suggest further avenues of investigation.

Although these forty most distinctive words are fascinating in themselves, it is more instructive to sort all 1,700 of the p < .05 words alphabetically, along with the scores that show which author favours each word. It is immediately apparent that these distinctive words tend to group in morphologically related families, and that each family strongly tends to be favoured by one of the authors, as can already be seen by the presence of thing, things, and everything among the twenty most distinctive Besant words. These are joined by anything and nothing further down his list (something is not distinctive for either author). In addition, every and everything are joined by everybody and everywhere, and anything is joined by any and anywhere, so that the families are related to each other. The same could be said of nothing and not, which are joined by never, no, nobody, none, and nor. (Not all of these families are equally cohesive; I have tried to err on the side of inclusiveness because it gives more chances for the group to fail to cohere.) Finally, much is joined by more, moreover, most, and mostly. For Collins, answered is joined by
answer, answering, and unanswerable, and five more of his twenty most distinctive words are joined by two other words: ask, asked, asks; inquired, inquiries, inquiry; leave, leaving, left; person, personally, persons; suggest, suggested, suggestion (there are some additional two-word families, including one, reply and replied, in which the second is a Besant marker word).

Among the 1,700 distinctive words, about 600 form morphologically related groups, nearly 400 for Collins and more than 200 for Besant. Only about 175 words form groups that contain members from both authors; these form 73 groups that fall into several intriguing patterns. One large pattern shows us that Collins uses more contractions, so that, for example, while did, and does are Besant words, didn’t, doesn’t, and don’t are Collins words. The same is true for must, need, should, and would and their negative contractions. This pattern is one reason for the corresponding frequency of not in Besant, though it does not explain his distinctive use of the other negative words. In another interesting pattern, partly semantic and partly morphological, the singular and possessive forms of brother, friend, sister, and son are Collins words and the plural forms are Besant words. The singular Collins vs plural Besant pattern is continued in thirty-two more nouns; the only exception is that troubles is a Collins word and trouble is a Besant word. Obviously, some of these words, including trouble, are complicated by the fact that some of their occurrences are verbs. Some other plurals are also Collins markers and some singulars are Besant markers, but in these cases all forms of the words are Collins or Besant markers. Another strong pattern is Collins’s use of the -ing forms of verbs and Besant the 3rd singular present forms, and Besant’s plurals and verb forms in -s obviously overlap and reinforce each other. A more thorough examination of some of the groups would clearly require part-of-speech disambiguation, but automated tagging is still too inaccurate for my taste, and manual tagging of such a large corpus is impractical.

One of the most surprising patterns is that all nineteen of the distinctive words that refer to cardinal numbers are strong Besant markers: all of the numbers from one to ten, twelve, eighteen, twenty, twenty-one, five-and-twenty, thirty, forty, fifty, and sixty. These are joined by hundred, thousand, and thousands, and perhaps one might add dozen, half, and quarter. The pattern is reversed to some extent in the ordinal numbers; fourth is a Besant marker, but first and second are Collins markers. Naturally, this pattern is related to the preference for plural nouns, which are one of the most common word classes to follow numbers, and which less frequently follow ordinals. Other less consistent patterns exist, but authors’ characteristic words are obviously highly patterned, and many more families could be found or made larger by including words that are more frequent in either author, but not significant at the p < .05 level. The existence of morphological and semantic families among authors’ characteristic words may not seem particularly surprising, but so little research has been done into the nature and character of literary vocabulary that it is difficult to know what should seem surprising, and it is unlikely that the number or extent of such patterns would be discovered without a computational analysis. The strong patterns for plurals and number words cry out for more investigation, and further examination of the lists of distinctive words would undoubtedly reveal other semantic and possibly even phonological families, and families of families (for more on word-families, see Hoover 2007).

2.2.4 Student’s T-test and the Density of Authorial Style

Another way that t-tested words can be used to investigate style is simply by highlighting them.
in a text. Doing so reinforces the distributed nature of style and visually emphasizes just how densely patterned style is. In the passages below, I have highlighted just the 500 most distinctive marker words for each author:

This has made a race of **men** **quick** to **fight** and careless of life, since, willy nilly, they **went** **daily** in peril; and **many** families there are whose **men**, until a **hundred** years ago, **never** **knew** what it was to die in their beds.

*Dorothy Forster*: 44 words, 15 **Besant markers**, 0 Collins

Confronted **by** the **serious** responsibility that he had undertaken, he **justified** what he had **said** to me. **Still** pale, **still** **distressed**, he was now **nevertheless** master of himself. I **turned** to the **door** to **leave** him alone with the Prisoner. She called me **back**. *The Legacy of Cain*: 44 words, 12 **Collins markers**, 0 Besant

The density of marker words obviously varies throughout each text, but passages as densely marked as these and without any marker words for the opposing author are not difficult to find. The high density of words with distributions in the two authors that are statistically significant (for these 1000 words, the weakest p value is .016) helps to make comprehensible the fact that many readers can recognize the style of an author they know well, even in a short passage they have never read before (see also Hoover 2007, 2008).

### 2.3 Zeta Analysis

One drawback of the t-test as a method of characterizing an author’s vocabulary is that it privileges high-frequency words: words that occur in many of the sections of one author’s texts but fewer sections of the other. These words clearly characterize the authors with respect to each other, but words like *are, back, by, said, there,* and *this* are not usually very interesting stylistically. Another drawback is that, as noted above, it cannot measure the importance of words that do not occur at all in one of the authors. Other statistical methods can cope with zero frequencies, but here I would like to present Zeta, a simpler, less technical, and more pragmatic method that is very effective in identifying characteristic vocabulary without requiring formal statistical testing. Zeta was invented by John F. Burrows’s Zeta (Burrows 2006), but the specific form I will be using here was developed by his colleague Hugh Craig (Craig and Kinney, forthcoming; for an automated spreadsheet that performs a Craig’s Zeta analysis, along with detailed instructions, see [https://files.nyu.edu/dh3/public/TheZeta&IotaSpreadsheet.html](https://files.nyu.edu/dh3/public/TheZeta&IotaSpreadsheet.html)). Like the t-test, Zeta identifies words that are much more frequent in one author than in one or more other authors, but Zeta words are substantially less frequent than those identified by the t-test (for more on original Zeta, see Hoover 2007, 2008).

Calculating Zeta is very simple. Here I use the same seven novels as were used for the t-test and prepare the word list in the same way, retaining only words with a minimum frequency of three, and removing words for which a single text supplies more than 90 percent of the examples. For this analysis, I have not deleted the personal pronouns and have retained words that are present only in Besant or Collins. Next, I again divide the seven novels into 167
sections, and then compare how consistently each word appears in the texts of each author. This is done by dividing the number of sections by the first author in which the word appears by the total number of sections for that author, then dividing the number of sections by the second author in which the word does not appear by the total number of sections for that author. (This method ignores the word’s frequency in each section.) The two ratios are added together to get a composite measure of affinity: a score for how much one author favours the word plus how much the other avoids it. A word that appeared in all Collins segments and no Besant sections would have a score of 2.0; conversely, one that appeared in no Collins segments and all of Besant’s would have a score of 0.0. In practice, such words are vanishingly rare, and in the present analysis, answered is the most distinctive Collins word, as it was in the t-tests, with a Zeta score of 1.8. It is present in eighty-nine of the ninety Collins sections and is absent from sixty-five of the seventy-seven Besant sections. Likewise, upon is again the most distinctive Besant word, present in all seventy-seven Besant sections and absent from twenty-five of the ninety Collins sections, with a score of 0.28.

2.3.1 Zeta and Characteristic Vocabulary

A comparison of the twenty most distinctive Zeta words with those that the t-test identifies is instructive. In the lists of Zeta words below, those that are also identified by the t-test are in bold type:

**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

Because Collins’s t-tested marker words are much less frequent than Besant’s, the overlap is much greater for his Zeta words than for Besant’s. Of the twenty t-tested marker words for both authors that ranked in the top 100 in frequency in the corpus, only two are left (Mrs and Miss), and the average ranks for the Zeta words are 445 for Besant and 473 for Collins (the average ranks for the t-tested markers were 106 and 261, respectively). For Collins, 372 of the 500 most distinctive t-test words are also among the 500 most distinctive Zeta words; for Besant, the figure is 361. Of the 500 Collins Zeta words, 37 are not found in Besant; of the 500 Besant Zeta words, 50 are not found in Collins.

An examination of all 2000 Zeta words reveals 153 Besant-only words and 122 Collins-only words. One of the benefits of the inclusion of these words is that they augment the morphological families discussed above. Of these 275 words, 59 form new families favoured by a single author; another 27 join the single-author families of the t-tested words. Only 21 words join into families split between the authors. Among the most interesting additions are the following new number words for Besant: sixteen, twenty-first, one-and-twenty, two-and-twenty, seventy-five, millions, and possibly multitude, quantity, and several. Most of these number words collocate with o’clock, pounds, thousands, and years, depending on the size of the number, suggesting that money and the passage of time are strong thematic elements in Besant’s fiction.
Again, I can only scratch the surface. Another Besant word-family, *somewhat*, is augmented by the Besant-only word *somewhere*. The Collins-only words *alluding, consultation, and excitable* augment the *allude, alluded, allusion, allusions* family, the *consult, consulted, consulting* family, and the *excite, excited, exciting* family.

One final Besant-only word that deserves a brief comment is *thou*. As noted above, I retained the personal pronouns in this Zeta analysis; *herself, myself, and yourself* are Collins marker words and *themselves, ourselves, ours, and thou* are Besant marker words; note that the pattern of the singular for Collins and the plural for Besant recurs, and no forms of the simple personal pronouns are distinctive. All of these words except *ours* (p = .051) and *thou* (not present in Collins, so that the t-test is impossible), would have been significant at the p < .05 level if they had been included in the t-test analysis. There are, however, too few of them to have had a discernable effect on the analyses presented above, and an argument could be made that they should be excluded in spite of their distinctiveness because the kind of preferences they mark are intimately linked to the nature of the novel, which was determined by Collins.

### 2.3.2 Zeta and the Density of Authorial Style

Highlighting the 1,000 most distinctive Zeta words for each author rather than the 500 most distinctive t-tested words in the same passages reveals a similar density of marker words, though fewer function words and more meaningful words are highlighted:

This has made a race of *men quick to fight* and careless of life, since, willy nilly, they *went daily* in peril; and many families there are whose *men*, until a hundred years ago, never *knew* what it was to *die* in their beds. *Dorothy Forster*

44 words, 14 Besant markers, 0 Collins

**Confronted** by the *serious responsibility* that he had undertaken, he justified what he had said to me. Still pale, still *distressed*, he was now nevertheless master of himself. I *turned* to the *door* to *leave* him *alone* with the Prisoner. She called me *back.* *The Legacy of Cain*

44 words, 11 Collins markers, 1 Besant

### 2.3.4 Zeta and the Authorship of the Chapters of *Blind Love*

Graphing the frequencies of the Zeta words as the t-tested words were graphed in Fig. 3 gives similar results. Rather than duplicating them, Fig. 4 presents a scatter graph in which the vertical axis is the percentage of all the word types (unique words) in the section that are Besant marker words (longer texts are divided into 4000-words sections) and the horizontal axis is the percentage of all word types in the section that are Collins marker words. To make the graph more readable, I have removed the labels for the even-numbered sections two to forty-four of *Blind Love*, and have included only a few sections of other novels.
Note how distinct Besant’s chapters of *Blind Love* are from Collins’s. Clearly Zeta marker words are capturing distinctive stylistic vocabulary characteristics. I have also included an additional text in this analysis, *The Case of Mr. Lucraft*, a long story jointly written by Besant and James Rice, just to see what would happen (the segments of this story are indicated as *Case* in bold type). Fig. 4 suggests that, as has been argued, Besant did most of the actual writing of the large number of co-written stories and novels that the pair produced from 1872 to 1882, a period during which Besant published nothing on his own (Boege 1956: 251-65). Further research would likely clear up this question, but probably neither Besant nor Rice is important enough to make such research very compelling from a literary point of view.

3. *The O’Ruddy*

I turn next to a discussion of Stephen Crane’s posthumously published novel *The O’Ruddy* (1903), completed by Robert Barr. Though Crane was only twenty-eight and lacked the stature of Collins, he was more highly regarded than Barr, who, like Besant, was a popular novelist whose works are now little read. I was initially drawn to this novel by a fascinating short book (O’Donnell 1970) that investigates its authorship, using a small group of variables that includes sentences, words, clauses, clause types, metaphor, parts of speech, and punctuation. O’Donnell
reports that, at the time of Crane’s death, he had written a manuscript of 65,000 words, of which only chapter twenty-four is known to exist, but that Barr at one point claimed to have written three-fourths of what came to be a thirty-three-chapter novel of about 100,000 words. Given the uncertainty about the relative contributions of Crane and Barr, O’Donnell analyses the entire novel paragraph-by-paragraph (apparently limiting himself to paragraphs of about 100 words or longer). His analysis identifies chapters one to twenty-four as Crane’s, chapters twenty-six to thirty-three as Barr’s, and twenty-five as a borderline case. To test his findings using frequent words and without preconceptions, I first divided the novel into thirty-seven sections of 2,500 words each (about the size of the average chapter) and performed several cluster analyses of the novel and some other texts by Barr and Crane based on the most frequent words. Sections one to twenty-four and twenty-five to thirty-seven cluster separately and consistently over the entire range of analyses, which suggests, as O’Donnell concludes (1970: 76), that the change of authorship was abrupt, without much joint authorship, except perhaps for chapter twenty-five. When I divided the novel into chapters, however, I found that Cluster Analysis divides the novel consistently after chapter twenty-five, not chapter twenty-four.

These findings prompted me to further research, and a reference in A Stephen Crane Encyclopedia (Wertheim 1997: 251-54), led me in turn to the University of Virginia’s authoritative edition of Crane’s works. There I discovered that a complete manuscript of the novel exists for chapters one to twenty-five (first recorded in 1969, and unknown to O’Donnell), and that Crane’s text was not extensively revised (Crane 1971: xv-lxxiv, 271-98). An independent comparison of chapters seventeen and eighteen of the original text and the Virginia edition (based on the manuscript) confirms that there are only a handful of differences on a typical page, many of which are matters of hyphenation (compound words that are written as one word in the 1903 edition are often hyphenated in the Virginia edition). I might have wished that this information had not come to light until after I had completed my analysis, but it was encouraging that my initial attempts were accurate, even more accurate than O’Donnell’s much more labour-intensive methods. (One often overlooked benefit of computational work is that it frequently pushes the analyst toward further research.) The case now appears remarkably similar to that of Blind Love, and a further comparison of the results for the two novels is instructive.

3.1 The Most Frequent (Function) Words: Cluster Analysis

After assembling a large set of texts for Barr and Crane, I performed a Cluster Analysis without including The O’Ruddy and discovered that Crane and Barr are much more difficult to distinguish than Collins and Besant. All analyses based on the 200 to the 990 MFW attribute Crane’s Active Service to Barr, and all those based on the 500-990 MFW also attribute Crane’s Third Violet and The Second Generation to Barr. Here, the only analyses with just one error are ones based on the 300, 100, and 70 MFW, with errors for three different texts, one by Barr and two by Crane. The variability of the analyses based on small numbers of words and the close similarity of the analyses based on the 500-990 MFW both suggest that Crane’s and Barr’s lexical styles are quite similar.

When the The O’Ruddy is included, typically from one to three Crane texts are attributed to Barr in the most accurate analyses, but Crane’s chapters of The O’Ruddy almost universally cluster together, though within the same large cluster that contains Barr’s texts, as shown in Fig. 5:
Note that, though it confirms the division point in the novel, this graph suggests that all of The O’Ruddy is more similar to Barr than to Crane, presumably reflecting the fact that this first person comic novel is considered very uncharacteristic of Crane. Barr himself described it in a letter as ‘different from anything he ever wrote before’ (quoted in Crane 1932: 18; see also Wertheim 1997: 254; O’Donnell 1970: 33-35; Sorrentino 2008: 61).

3.2 Student’s T-test, Zeta, and the Authorship of the Chapters of The O’Ruddy

To make my analysis of The O’Ruddy comparable to the analysis of Blind Love, I collected a similar-sized corpus of texts by Crane and Barr, about 315,000 words each. In this case, because Crane only wrote a few relatively short novels, many of the texts are stories or novellas. I created
word lists and collected the t-tested words, just as for Collins and Besant. In this case, 1,691 words \( p < .05 \), 769 for Barr and 922 for Crane, and 1,000 Zeta words for each author. Analyses based on the two sets of words are remarkably similar, and they confirm the clear distinction between Crane’s and Barr’s chapters of *The O’Ruddy*. A graph based on the t-tested words is shown in Fig. 6. Note that Barr’s chapters are much more similar to his other texts than Crane’s chapters are to his other texts, showing just how uncharacteristic this novel is for Crane.

![Graph showing word percent of types](image)

Fig. 6 Barr, Crane, and *The O’Ruddy*: 1691 T-tested Words \( p < .05 \)

### 3.2.1 Student’s T-test and Characteristic Vocabulary

There is only space here for a brief examination of the vocabulary differences in Crane and Barr, but consider the 30 most distinctive marker words for Crane and Barr.

**Barr:** have, than, will, for, is, my, so, your, has, may, or, before, am, although, not, shall, own, on, I, that, me, put, should, taken, our, myself, cannot, are, therefore, next

**Crane:** suddenly, they, a, upon, presently, great, turned, began, at, manner, arose, toward, gesture, moved, faces, finally, were, like, went, seemed, eyes, their, black, near, fro, he, obliged, kind, rage, stared

Barr’s list, like Besant’s is quite banal and much less revealing than Crane’s. Both lists also
Hoover Authorial Style

again suggest word-families. For example, my, your, I, our, and myself show Barr’s preference for first and second person pronouns, a preference that is continued by mine, yours, yourself, you, and we, among his other marker words. Crane’s greater use of contractions (not evident here), however, places we’ll and we’ve among his marker words. It may also help to explain the presence of the auxiliary verbs in Barr’s list above, by reducing the frequency of the full forms in Crane, though the presence of seven lexical verbs in Crane’s list and only two in Barr’s suggests that contraction is not the full explanation. Crane favours the third person pronouns, as they, their, and he suggest, and as is confirmed by the presence of his, them, ‘em, and it’ll among his other marker words (note the reinforcement of the contractions). Crane and Barr both typically write in the third person, yet first and second person pronouns and contractions all suggest dialogue, so that more research would be needed to explain these patterns. The presence of faces, eyes, and stared, along with black, suggest that Crane is more interested in the visual, and this is confirmed by the presence of stare, staring, face, faced, apparent, appear, appeared, appearing, glance, glanced, glances, look, looked, looks, resemble, resembled, scene, scenes, shade, shadows, shadowy, shine, shining, and shine, among his other marker words. Among Barr’s marker words are only evident, evidently, recognize, recognized, notice, and noticed, a list that, besides being much shorter, is also less clearly visual. Finally, the contrast between Barr’s is, am, and are and Crane’s were is completed by was among Crane’s other maker words, and Crane’s other past tense verbs above, turned, began, arose, moved, went, seemed, and stared, compared with only taken and possibly put for Barr, suggest a more general preference for past tense.

Among the 1,691 t-tested markers are about 300 morphological families, containing about 730 words: about 130 Crane families, 120 Barr families, but only 50 families with at least one marker word for both Barr and Crane. Among the mixed families, twenty involve a past tense and a present tense verb; seventeen of the past tense verbs are Crane markers and only three are Barr markers, confirming Crane’s preference for the past tense. Furthermore, although past tense verbs are about equally prevalent among the Barr and Crane families of words, among the words not in families, about nineteen percent of Crane’s words are past tense verbs, compared with only about nine percent for Barr. Five of the mixed families involve a contraction for Crane, and a few other minor patterns exist, so that more than half of the mixed families exhibit further patterns. Even families that are not fully consistent are often partially patterned, as with we versus we’ve and we’re, above. Consider the do family: do, does, and doesn’t are Barr markers, and don’t and didn’t are Crane markers, so that the pattern of present tense and full forms for Barr and past tense and contractions for Crane are still visible in spite of the partial exceptions of doesn’t and don’t. Forms of any, every, and some present another interesting mixed pattern (Barr markers in bold type):

<table>
<thead>
<tr>
<th>any</th>
<th>anyone</th>
<th>anybody</th>
<th>anyhow</th>
<th>anyway</th>
</tr>
</thead>
<tbody>
<tr>
<td>every</td>
<td>everyone</td>
<td>everybody</td>
<td>everywhere</td>
<td></td>
</tr>
<tr>
<td>some</td>
<td>someone</td>
<td>somebody</td>
<td>sometimes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>somewhat</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Within this somewhat confused situation it is clear that the -one vs -body words form a strong pattern of the kind that is often important for authorship attribution, and the absence of any of the -thing forms as marker words is intriguing.
Finally, there are interesting semantic groups of words among the Barr and Crane markers that are not morphologically based. The word black within Crane’s thirty most distinctive markers is joined by all the other colour words among the marker words: blue, gray, grey, green, pink, purple, red, white, and yellow. Crane also monopolizes the semantic field of swearing, with cursed, curses, damn, damned, hell, swearing, and swore among his marker words (many of Crane’s texts were originally bowdlerized). Barr, in turn, monopolizes the field of time, with day, days, hour, minutes, moment’s, moments, month, year, and years among his marker words. Finally, Crane overwhelmingly favours words related to speed or quickness, with the following among his marker words: abrupt, abruptly, flurry, gallop, galloped, galloping, hastily, hurried, immediately, presently, prompt, ran, rush, rushed, shortly, sudden, suddenly, swift, and swiftly. For Barr, in contrast, there are only runs, speedily, and suddenness. Again, I can only point in the directions that this kind of analysis makes possible, and further examination of the word lists would undoubtedly uncover additional patterns of interest.

4. Conclusion

Computational stylistics is neither a substitute for nor an alternative to other kinds of stylistic and literary analysis. Rather, it provides additional tools that can serve as discovery procedures and can augment and enhance other methods. Computational stylistics provides a practical way of elucidating the extensive and richly-patterned nature of style, of coping with the dense, yet widely-distributed linguistic phenomena that are or constitute the stylistic features that both make an authorial style distinctive and characterize its nature. Cluster Analysis effectively measures the similarity and difference of authors, texts, or parts of texts by taking into account large numbers of features simultaneously, emphasizing and facilitating the comparisons upon which style so intimately depends. T-tests and Zeta analysis are also effective methods of measuring difference and similarity, but they go beyond or beneath this to provide access to patterns too subtle, extensive, and numerous to be readily accessible to traditional stylistic analysis, especially where the texts in question are whole sets of novels. They allow the analyst to uncover morphological and semantic families of words and to investigate the relationship between those families and larger issues of authorial style. They can uncover extraordinarily consistent patterns and puzzling inconsistencies and aberrations that point the way toward more minute and searching literary and stylistic analysis, while also providing a measure of objectivity and consistency that can bolster and support other kinds of argument and analysis. Computational stylistics has earned a place in the expanding stylistics toolbox, and is an especially effective method for analyzing authorial style.
References


Hoover Authorial Style 20

203.


Digital humanities, like most fields of scholarly inquiry, constituted itself through a long accretion of revolutionary insight, territorial rivalry, paradigmatic rupture, and social convergence. But the field is unusual in that it has often pointed both to a founder and to a moment of creation. The founder is Roberto Busa, an Italian Jesuit priest who in the late 1940s undertook the production of an automatically generated concordance to the works of Thomas Aquinas using a computer. The founding moment was the creation of a radically transformed, reordered, disassembled, and reassembled version of one of the world’s most influential philosophies:

00596 in veniale peccatum non cadat; ut sic hoc verbum habemus non determinatum, sed confusum praesens importet
-003(3SN)3.3.2.b.ex/56

00597 intellegit profectum scientiae christi quantum ad experientiam secundum novam conversionem ad sensibile praesens,
-S 003(3SN)14.1.3e.ra4/4

00598 ita quot apprehenditur ut possibile adipisce, apprehenditur ut jam quodammodo praesens: et ideo spec delectionem
-003(3SN)26.1.2.ra3/8

00599 operationibus: quia illud quod certudinaliter quasi praesens tenemus per intellectum, dicimur sentire, vel videre;
-003(3Sn)26.1.5.co/11 (Index 65129)

Undertaking such transformations for the purpose of humanistic inquiry would eventually come to be called “text analysis,” and in literary study, computational text analysis has been used to study problems related to style and
authorship for nearly sixty years. As the field has matured, it has incorporated elements of some of the most advanced forms of technical endeavor, including natural language processing, statistical computing, corpus linguistics, data mining, and artificial intelligence. It is easily the most quantitative approach to the study of literature, arguably the oldest form of digital literary study, and, in the opinion of many, the most scientific form of literary investigation.

But "algorithmic criticism"—criticism derived from algorithmic manipulation of text—either does not exist or exists only in nascent form. 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 largely concerned with the interpretative analysis of written cultural artifacts. Texts are browsed, searched, and disseminated by all but the most hardened Luddites in literary study, but rarely are they transformed algorithmically as a means of gaining entry to the deliberately and self-consciously subjective act of critical interpretation. Even text analysis practitioners avoid bringing the hermeneutical freedom of criticism to the "outputted" text. Bold statements, strong readings, and broad generalizations (to say nothing of radical misreadings, anarchic accusations, and agonistic paratextual revolts) are rare, if not entirely absent from the literature of the field, where the emphasis is far more often placed on methodology and the limitations it imposes.

It is perhaps not surprising that text analysis would begin this way. Busa's own revolution was firmly rooted in the philological traditions to which modern criticism was largely a reaction. Reflecting on the creation of the Index some forty years after the fact, Busa offered the following motivations:

I realized first that a philological and lexicographical inquiry into the verbal system of an author has to precede and prepare for a doctrinal interpretation of his works. Each writer expresses his conceptual system in and through his verbal system, with the consequence that the reader who masters this verbal system, using his own conceptual system, has to get an insight into the writer's conceptual system. The reader should not simply attach to the words he reads the significance they have in his mind, but should try to find out what significance they had in the author's mind. ("Annals" 83)

Such ideas would not have seemed unusual to nineteenth-century biblical scholars, for whom meaning was something both knowable and recoverable through careful, scientific analysis of language, genre, textual recension, and historical context. Nor would it, with some rephrasing, have been a radical proposition either for Thomas himself or for the Dominican friars who produced the first concordance (to the Vulgate) in the thirteenth century. How-
ever, we do no injustice to Busa’s achievement in noting that the contemporary critical ethos regards Busa’s central methodological tenets as grossly naive. Modern criticism, increasingly skeptical of authorial intention as a normative principle and linguistic meaning as a stable entity, has largely abandoned the idea that we could ever keep from reading ourselves into the reading of an author and is no longer concerned with attempting to avoid this conundrum.

But even in Busa’s highly conventional methodological project, with its atomized fragmentation of a divine text, we can discern the enormous liberating power of the computer. In the original formation of Thomas’s text, “presence” was a vague leitmotif. But on page 65,129 of the algorithmically transformed text, “presence” is that toward which every formation tends, the central feature of every utterance, and the pattern that orders all that surrounds it. We encounter “ut sic hoc” and “ut possibile,” but the transformed text does not permit us to complete those thoughts. Even Busa would have had to concede that the effect is not the immediate apprehension of knowledge, but instead what the Russian Formalists called ostranenie—the estrangement and defamiliarization of textuality. One might suppose that being able to see texts in such strange and unfamiliar ways would give such procedures an important place in the critical revolution the Russian Formalists ignited—which is to say, the movement that ultimately gave rise to the hermeneutical philosophies that would supplant Busa’s own methodology.

But text analysis would take a much more conservative path. Again and again in the literature of text analysis, we see a movement back toward the hermeneutics of Busa, with the analogy of science being put forth as the highest aspiration of digital literary study. For Roseanne Potter, writing in the late 1980s, “the principled use of technology and criticism” necessarily entailed criticism becoming “absolutely comfortable with scientific methods” (91–92). Her hope, shared by many in the field, was that the crossover might create a criticism “suffused with humanistic values,” but there was never a suggestion that the “scientific methods” of algorithmic manipulation might need to establish comfort with the humanities. After all, it was the humanities that required deliverance from the bitter malady that had overtaken modern criticism: “In our own day, professors of literature indulge in what John Ellis (1974) somewhat mockingly called ‘wise eclecticism’—a general tendency to believe that if you can compose an interesting argument to support a position, any well-argued assertion is as valid as the next one. A scientific literary criticism would not permit some of the most widespread of literary critical practices” (93). Those not openly engaged in the hermeneutics of “anything goes”—historians old or new—were presented with the settling logic of truth and falsehood proposed by computational analysis:
This is not to deny the historical, social, and cultural context of literature (Bakhtin, 1981), and of language itself (Halliday, 1978). Nor can one overlook the very rich and subtle elaborations of literary theory in the forty years since Barthes published *Le degré zéro de l'écriture* (1953). In point of fact, most of these elaborations have the technical status of hypothesis, since they have not been confirmed empirically in terms of the data which they propose to describe—literary texts. This is where computer techniques and computer data come into their own. (Fortier 376)

Susan Hockey, in a book intended not only to survey the field of humanities computing but also to “explain the intellectual rationale for electronic text technology in the humanities,” later offered a vision of the role of the computer in literary study to which most contemporary text analysis practitioners fully subscribe:

Computers can assist with the study of literature in a variety of ways, some more successful than others. . . . Computer-based tools are especially good for comparative work, and here some simple statistical tools can help to reinforce the interpretation of the material. These studies are particularly suitable for testing hypotheses or for verifying intuition. They can provide concrete evidence to support or refute hypotheses or interpretations which have in the past been based on human reading and the somewhat serendipitous noting of interesting features. (66)

It is not difficult to see why a contemporary criticism temperamentally and philosophically at peace with intuition and serendipity would choose to ignore the corrective tendencies of the computer against the deficiencies of “human reading.” Text analysis arises to assist the critic, but only if the critic agrees to operate within the regime of scientific methodology with its “refutations” of hypotheses.

Perhaps the boldest expression of these ideas comes from a 2008 editorial in the *Boston Globe* titled “Measure for Measure.” In it, literary critic Jonathan Gottschall describes the field of literary studies itself as “moribund, aimless, and increasingly irrelevant to the concerns not only of the ‘outside world,’ but also to the world inside the ivory tower.” The solution is one that even C. P. Snow would have found provocative:

I think there is a clear solution to this problem. Literary studies should become more like the sciences. Literature professors should apply science’s research methods, its theories, its statistical tools, and its insistence on hypothesis and proof. Instead of philosophical despair about the possibility of knowledge, they should embrace science’s spirit of intellectual optimism. If they do, liter-
ary studies can be transformed into a discipline in which real understanding of literature and the human experience builds up along with all of the words.

This proposal may distress many of my colleagues, who may worry that adopting scientific methods would reduce literary study to a branch of the sciences. But if we are wise, we can admit that the sciences are doing many things better than we are, and gain from studying their successes, without abandoning the things that make literature special.

Gottschall offers no suggestions for how we might retain those things that make humanistic discourse itself “special.” He admits to being not overly fond of what he presumes to be the main outlines of that discourse (the “beauty myth,” the death of the author, the primacy of social and cultural influences in the constitution of identity, and the sexism of the Western canon), but his main concern is that such notions have become the unexamined ground truths of contemporary criticism. This in itself is hardly objectionable; it is difficult to imagine a healthy episteme that does not constantly question even its most cherished assumptions. But that these ideas were themselves the product of decades of humanistic reflection and debate, that they supplanted other ideas that had come to be regarded as similarly uncontroversial, and that they provide a powerful counterexample to the “philosophical despair about the possibility of knowledge” against which he inveighs seems not to lessen Gottschall’s faith in final answers. Only the methodologies of science and the rigor of computation can render unexamined assumptions “falsifiable.”

Even Franco Moretti, whose outlook on literary study is assuredly quite different from Gottschall’s, shows strong evidence of embracing this faith in the falsifiable: “I began this chapter by saying that quantitative data are useful because they are independent of interpretation; then, that they are challenging because they often demand an interpretation that transcends the quantitative realm; now, most radically, we see them falsify existing theoretical explanations, and ask for a theory, not so much of ‘the’ novel, but of a whole family of novelistic forms. A theory—of diversity” (Graphs 30). Moretti is right to be excited about what he is doing. It is breathtaking to see his graphs, maps, and trees challenging accepted notions about the nineteenth-century novel. But one wonders why it is necessary to speak of these insights as proceeding from that which is “independent of interpretation” and which leads to the “falsification” of ideas obtained through more conventional humanistic means. It is as if everything under discussion is a rhetorical object except the “data.” The data is presented to us—in all of these cases—not as something that is also in need of interpretation, but as Dr. Johnson’s stone hurtling through the space of our limited vision.
The procedure that Busa used to transform Thomas into an alternative text is, like most text-analytical procedures, algorithmic in the strictest sense. If science has repeatedly suggested itself as the most appropriate metaphor, it is undoubtedly because such algorithms are embedded in activities that appear to have the character of experiment. Busa, in the first instance, had formed an hypothesis concerning the importance of certain concepts in the work. He then sought to determine the parameters (in the form of suitable definitions and abstractions) for an experiment that could adjudicate the viability of this hypothesis. The experiment moved through the target environment (the text) with the inexorability of a scientific instrument creating observable effects at every turn. The observations were then used to confirm the hypothesis with which he began.

Some literary-critical problems clearly find comfort within such a framework. Authorship attribution, for example, seeks definitive answers to empirical questions concerning whether or not a work is by a particular author. Programs designed to adjudicate such questions can often be organized scientifically with hypotheses, control groups, validation routines, and reproducible methods. The same is true for any text analysis procedure that endeavors to expose the bare empirical facts of a text (often a necessary prelude to textual criticism and analytical bibliography). Hermeneutically, such investigations rely upon a variety of philosophical positivism in which the accumulation of verified, falsifiable facts forms the basis for interpretative judgment. In these particular discursive fields, the veracity of statements like “The tenth letter of The Federalist was written by James Madison” or “The 1597 quarto edition of Romeo and Juliet is a memorial reconstruction” are understood to hinge more or less entirely on the support of concrete textual evidence. One might challenge the interpretation of the facts, or even the factual nature of the evidence, but from a rhetorical standpoint, facts are what permit or deny judgment.

For most forms of critical endeavor, however, appeals to “the facts” prove far less useful. Consider, for example, Miriam Wallace’s discussion of subjectivity in Virginia Woolf’s novel The Waves:

In this essay I want to resituate The Waves as complexly formulating and re-formulating subjectivity through its playful formal style and elision of corporeal materiality. The Waves models an alternative subjectivity that exceeds the dominant (white, male, heterosexual) individual western subject through its stylistic usage of metaphor and metonymy. . . . Focusing on the narrative construction of subjectivity reveals the pertinence of The Waves for current feminist reconfigurations of the feminine subject. This focus links the novel’s visionary limitations to the historic moment of Modernism. (295–96)
An Algorithmic Criticism

Wallace frames her discourse as a “resitation” of Woolf’s novel within several larger fields of critical discourse. This will presumably involve the marshaling of evidence and the annunciation of claims. It may even involve offering various “facts” in support of her conclusions. But hermeneutically, literary critical arguments of this sort do not stand in the same relationship to facts, claims, and evidence as the more empirical forms of inquiry. There is no experiment that can verify the idea that Woolf’s “playful formal style” reformulates subjectivity or that her “elision of corporeal materiality” exceeds the dominant Western subject. There is no control group that can contain “current feminist reconfigurations.” And surely there is no metric by which we may quantify “pertinence” either for Woolf or for the author’s own judgment.

The hermeneutical implications of these absences invoke ancient suspicions toward rhetoric, and in particular, toward the rhetorical office of inventio: the sophistic process of seeking truth through the dialectical interplay of trust, emotion, logic, and tradition, which has, since the seventeenth century, contended with the promises of empiricism (Bold 543–44). In some sense, humanistic discourse seems to lack methodology; it cannot describe the ground rules of engagement, the precise means of verification, or even the parameters of its subject matter. Yet as Gadamer pointed out in Truth and Method:

The hermeneutic phenomenon is basically not a problem of method at all. It is not concerned with a method of understanding by means of which texts are subjected to scientific investigation like all other objects of experience. It is not concerned primarily with amassing verified knowledge, such as would satisfy the methodological ideal of science—yet it too is concerned with knowledge and with truth. . . . But what kind of knowledge and what kind of truth? (544)

Gadamer’s question is not easily answered, but we may say that from a purely cultural standpoint, literary criticism operates at a register in which understanding, knowledge, and truth occur outside of the narrower denotative realm in which scientific statements are made. It is not merely the case that literary criticism is concerned with something other than the amassing of verified knowledge. Literary criticism operates within a hermeneutical framework in which the specifically scientific meaning of fact, metric, verification, and evidence simply do not apply. The “facts” of Woolf—however we choose to construe this term—are not the principal objects of study in literary criticism, and “evidence” stands as a metaphor for the delicate building blocks of rhetorical persuasion. We “measure” only to establish webs of interrelation and influence. “Verification” occurs in a social community of scholars whose agreement or disagreement is almost never put forth without qualification.
All of this leaves the project of text analysis in a difficult position. For even if we are willing to concede the general utility of computational methods for the project of humanistic inquiry, we must nonetheless contend with a fundamental disjunction between literary-critical method and computational method. The logic that underlies computation, though not scientific in the strict sense of the term, conforms easily to the methodologies of science. Computers are, as Hickey noted, good at counting, measuring, and (in a limited sense) verifying data, and we judge the tractability of data by the degree to which it can serve the requirements of these procedures. When it comes to literary criticism, however, we find that the "data" is almost entirely intractable from the standpoint of the computational rubric. Paper-based textual artifacts must either be transformed from a continuous field into some more quantized form (i.e., digitized), or accompanied, as in the case of markup, with an elaborate scaffolding by which the vagaries of continuity can be flattened and consistently recorded. We accept the compromises inherent in such transformations in order to reap the benefits of speed, automation, and scale that computational representations afford. But the situation is considerably more complicated in the case of the analysis that is undertaken with these objects. Not a single statement in Wallace's précis, and indeed very few of the statements one encounters in literary critical discourse, can be treated in this way. No extant computer can draw the conclusions that Wallace does by analyzing the links between "the novel's visionary limitations" and "the historic moment of Modernism"—particularly since the Modernism being invoked here is not a matter of shifting consumer prices or birth statistics. Literary-critical interpretation is not just a qualitative matter; it is also an insistently subjective manner of engagement.

Given the essential properties of computation, we might conclude that text analysis is precisely designed to frame literary-critical problems in terms of something analogous to consumer prices and birth statistics, and in general text analysis has chosen low-level linguistic phenomena as its primary object of study. Doing so would seem to demand that we assume the methodological posture of computational linguistics, with its (entirely appropriate) claims toward scientific rigor. According to this hermeneutical vision, text analysis is simply incapable of forming the sorts of conclusions that lie outside of a relatively narrow range of propositions.

It is not at all uncommon to encounter explicit statements of such interpretative limitation in text-analytical scholarship. John Burrows and D. H. Craig's use of principal component analysis for comparing Romantic and Renaissance tragedy—a masterful work of text-analytical scholarship by any measure—is typical in how it commits itself to an essentially scientific vision
of permissible conclusion. The goal of the study is to elucidate the stylistic differences between the two periods of drama—one widely considered to have produced some of the greatest works in English, and another that is almost universally regarded as one of the low points of English literary drama. They draw a number of conclusions from their use of sophisticated statistical clustering methods, but in the end they confidently state that the sort of insight offered by George Steiner, who felt that the loss of a “redemptive worldview” had rendered Romantic tragedy an impossibility, is “well beyond the ambit of present computational stylistics” (Burrows and Craig 64).

For an algorithmic criticism to emerge, it would have to come to a philosophical decision concerning statements like these. But the question is less about agreement or disagreement, and more about a willingness to inquire into the hermeneutical foundations that make such statements seem necessary. The computer is certainly incapable of offering “the shift to a redemptive worldview” as a solution to the problem at hand; it is wholly incapable of inferring this from the data. But is it likewise the case that computational results—the data and visualizations that the computer generates when it seeks to quantize and measure textual phenomena—cannot be used to engage in the sort of discussion that might lead one to such a conclusion?

It is useful to put the question this way, because in doing so we refocus the hermeneutical problem away from the nature and limits of computation (which is mostly a matter of methodology) and move it toward consideration of the nature of the discourse in which text analysis bids participation. Burrows and Craig’s statement of limitation is valid if we consider computational stylistics to be essentially a scientific pursuit, because within this hermeneutical framework it makes sense to frame conclusions in terms of what the data “allows.” But in literary criticism—and here I am thinking of ordinary “paper based” literary criticism—conclusions are evaluated not in terms of what propositions the data allows, but in terms of the nature and depth of the discussions that result. The scientist is right to say that the plural of anecdote is not data, but in literary criticism an abundance of anecdote is precisely what allows discussion and debate to move forward.

Wallace’s essay concerns what many consider to be Virginia Woolf’s most experimental work. The novel consists of a series of monologues that trace the lives of six friends from early childhood to old age, with each monologue (beginning always with “Susan said” or “Bernard said”) telling the characters’ stories at seven distinct stages of their lives. They speak about different things and have different perspectives on the world, but they all speak in roughly the same manner, and do so from childhood to adulthood—employing, as one critic puts it, “the same kind of sentence rhythms and similar kinds of image
patterns” throughout (Rosenthal 144). Some critics have suggested that there are differences that lie along the axis of gender or along a rift separating the more social characters from the more solitary ones, but in the end one has the sense of an overall unity running against the perspectival conceit that frames the narrative.

It is natural for a Modernist critic to pursue patterns of difference amid this apparent unity, in part because, as Wallace points out, subjectivity is a major concern for “the historic moment of Modernism.” 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?

It is tempting for the text analysis practitioner to view this as a problem to be solved—as if the question were rhetorically equivalent to “Who wrote Federalist 10?” The category error arises because we mistake questions about the properties of objects with questions about the phenomenal experience of observers. We may say that Woolf’s novel “is” something or that it “does” something, but what we mean to capture is some far less concrete interpretative possibility connected with the experience of reading. We may ask “What does it mean?” but in the context of critical discourse this is often an elliptical way of saying “Can I interpret (or read) it this way?”

It is reasonable to imagine tools that can adjudicate questions about the properties of objects. Tools that can adjudicate the hermeneutical parameters of human reading experiences—tools that can tell you whether an interpretation is permissible—stretch considerably beyond the most ambitious fantasies of artificial intelligence. Calling computational tools “limited” because they cannot do this makes it sound as if they might one day evolve this capability, but it is not clear that human intelligence can make this determination objectively or consistently. We read and interpret, and we urge others to accept our readings and interpretations. Were we to strike upon a reading or interpretation so unambiguous as to remove the hermeneutical questions that arise, we would cease to refer to the activity as reading and interpretation. That we might refer to such uncontested statements as “facts” hardly bespeaks their superiority over less certain judgments.

If text analysis is to participate in literary critical endeavor in some manner beyond fact-checking, it must endeavor to assist the critic in the unfolding of interpretative possibilities. We might say that its purpose should be to generate further “evidence,” though we do well to bracket the association that term holds in the context of less methodologically certain pursuits. The evidence we seek is not definitive, but suggestive of grander arguments and schemes. The “problem” (to bracket another term) with Woolf’s novel is that
An Algorithmic Criticism
despite evidence of a unified style, one suspects that we can read and interpret it using a set of underlying distinctions. We can uncover those distinctions by reading carefully. We can also uncover them using a computer.

It is possible—and indeed an easy matter—to use a computer to transform Woolf's novel into lists of tokens in which each list represents the words spoken by the characters ordered from most distinctive to least distinctive term. *Tf - idf*, one of the classic formulas from the field of information retrieval, endeavors to generate lists of distinctive terms for each document in a corpus. We might therefore conceive of Woolf's novel as a "corpus" of separate documents (each speaker's monologue representing a separate document), and use the formula to factor the presence of a word in a particular speaker's vocabulary against the presence of that word in the other speakers' vocabularies.

Criticism drifts into the language of mathematics. Let *tf* equal the number of times a word occurs within a single document. So, for example, if the word "a" occurred 194 times in one of the monologues, the value of *tf* would be 194. A term frequency list is therefore the set of *tf* values for each term within that speaker's vocabulary. Such lists are not without utility for certain applications, but they tend to follow patterns that are of limited usefulness for our purposes. Since the highest-frequency terms in a given document are almost always particles ("the" can account for as much as 7 percent of a corpus vocabulary), and the lower-frequency words are almost always single-instance words (or "hapax legomena," as they are referred to in the field), we often end up with a list of words that is better at demonstrating the general properties of word distribution in a natural language than it is at showing us the distinctive vocabulary of an author.

If, however, we modulate the term frequency based on how ubiquitous the term is in the overall set of speakers, we can diminish the importance of terms that occur widely in the other speakers (like particles) and raise the importance of terms that are peculiar to a speaker. *Tf - idf* accomplishes this using the notion of an inverse document frequency:

\[ tf - idf = tf \cdot \left( \frac{N}{df} \right) \]

Let *N* equal the total number of documents and let *df* equal the number of documents in which the target term appears. We have six speakers. If the term occurs only in one speaker, we multiply *tf* by six over one; if it occurs in all speakers, we multiply it by six over six. Thus, a word that occurs 194 times, but in all documents, is multiplied by a factor of one (six over six). A word that occurs in one document, but nowhere else, is multiplied by a factor of six (six over one).
Here are the first twenty-five lines of output from a program designed to apply the tf – idf formula to the character of Louis:

<table>
<thead>
<tr>
<th>Weight</th>
<th>Term</th>
<th>Weight</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.917438</td>
<td>mr</td>
<td>4.2756658</td>
<td>disorder</td>
</tr>
<tr>
<td>5.7286577</td>
<td>western</td>
<td>3.9164972</td>
<td>accent</td>
</tr>
<tr>
<td>5.5176187</td>
<td>nile</td>
<td>3.7602086</td>
<td>beaten</td>
</tr>
<tr>
<td>5.0021615</td>
<td>australian</td>
<td>3.7602086</td>
<td>bobbing</td>
</tr>
<tr>
<td>5.0021615</td>
<td>beast</td>
<td>3.7602086</td>
<td>custard</td>
</tr>
<tr>
<td>5.0021615</td>
<td>grained</td>
<td>3.7602086</td>
<td>discord</td>
</tr>
<tr>
<td>5.0021615</td>
<td>thou</td>
<td>3.7602086</td>
<td>eating-shop</td>
</tr>
<tr>
<td>5.0021615</td>
<td>wilt</td>
<td>3.7602086</td>
<td>england</td>
</tr>
<tr>
<td>4.675485</td>
<td>pitchers</td>
<td>3.7602086</td>
<td>eyres</td>
</tr>
<tr>
<td>4.675485</td>
<td>steel</td>
<td>3.7602086</td>
<td>four-thirty</td>
</tr>
<tr>
<td>4.2756658</td>
<td>attempt</td>
<td>3.7602086</td>
<td>ham</td>
</tr>
<tr>
<td>4.2756658</td>
<td>average</td>
<td>3.7602086</td>
<td>lesson</td>
</tr>
<tr>
<td>4.2756658</td>
<td>clerks</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Few readers of The Waves would fail to see some emergence of pattern in this list. Many have noted that Louis seems obsessed with Egypt and the Nile. The list indicates that such terms are indeed distinctive to Louis, but the second most distinctive term in his vocabulary is the word “western.” Louis is also very conscious of his accent and his nationality (he is Australian; all the other characters are English), and yet the fact that “accent” is a distinctive term for Louis would seem to indicate that the other characters aren’t similarly concerned with the way he talks. Further analysis revealed that only one other character (Neville) mentions it. Louis is likewise the only character in the novel to speak of “England.”

This list is a paratext that now stands alongside the other, impressing itself upon it and upon our own sense of what is meaningful. Does this “western” stand against Louis’s “east”? Returning to the text, but with our focus narrowed and reframed, we discover that Louis’s fondness for the words “western,” “wilt,” and “thou” comes from his repetition of a famous sixteenth-century poem: “Western wind, when wilt thou blow? / The small rain down can rain. / Christ, if my love were in my arms, / And I in my bed again” (Waves 203). Woolf quotes the poem again in the nearly contemporaneous second series of The Common Reader (“How Should One Read a Book?”), noting, “Who when they read these four lines stops to ask who wrote them, or conjures up the thought of Donne’s house or Sidney’s secretary; or enmeshes them in the intricacy of the past and the succession of generations? The poet is always our contemporary” (265).
### An Algorithmic Criticism

<table>
<thead>
<tr>
<th>Bernard</th>
<th>Louis</th>
<th>Neville</th>
</tr>
</thead>
<tbody>
<tr>
<td>thinks</td>
<td>mr</td>
<td>catullus</td>
</tr>
<tr>
<td>letter</td>
<td>western</td>
<td>doomed</td>
</tr>
<tr>
<td>curiosity</td>
<td>nile</td>
<td>immittagble</td>
</tr>
<tr>
<td>moffat</td>
<td>accent</td>
<td>marvel</td>
</tr>
<tr>
<td>final</td>
<td>beaten</td>
<td>papers</td>
</tr>
<tr>
<td>important</td>
<td>beasting</td>
<td>bookcase</td>
</tr>
<tr>
<td>low</td>
<td>grained</td>
<td>bored</td>
</tr>
<tr>
<td>simple</td>
<td>thou</td>
<td>camel</td>
</tr>
<tr>
<td>canopy</td>
<td>discord</td>
<td>стир</td>
</tr>
<tr>
<td>getting</td>
<td>eating-shop</td>
<td>detect</td>
</tr>
<tr>
<td>hoot</td>
<td>pitchers</td>
<td>expose</td>
</tr>
<tr>
<td>hums</td>
<td>complex</td>
<td>admirable</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Jinny</th>
<th>Rhoda</th>
<th>Susan</th>
</tr>
</thead>
<tbody>
<tr>
<td>tunnel</td>
<td>cabinet</td>
<td>setter</td>
</tr>
<tr>
<td>prepared</td>
<td>coach</td>
<td>washing</td>
</tr>
<tr>
<td>melancholy</td>
<td>crag</td>
<td>apron</td>
</tr>
<tr>
<td>billowing</td>
<td>dazzle</td>
<td>pear</td>
</tr>
<tr>
<td>fiery</td>
<td>defly</td>
<td>seasons</td>
</tr>
<tr>
<td>game</td>
<td>equipped</td>
<td>squirrel</td>
</tr>
<tr>
<td>native</td>
<td>eyebrows</td>
<td>window-pane</td>
</tr>
<tr>
<td>peers</td>
<td>felled</td>
<td>window-bulk</td>
</tr>
<tr>
<td>quicker</td>
<td>frightened</td>
<td>window</td>
</tr>
<tr>
<td>victory</td>
<td>gage</td>
<td>kitchen</td>
</tr>
<tr>
<td>band</td>
<td>jump</td>
<td>baby</td>
</tr>
<tr>
<td>banners</td>
<td>lockets</td>
<td>Betty</td>
</tr>
</tbody>
</table>

Similar convergences appear in the other lists (see above). For Jinny, whose relationships with men form the liminal background of her narrative, words like “billowing” (a sexually charged word almost always used in reference to her skirts), “fiery,” “victory,” and “dazzle” appear in the top twenty-five. For Bernard, the aspiring novelist who some say is modeled on Woolf herself, the top word is “thinks.” Susan becomes a housewife and frequently invokes the virtues of a pastoral life in the country; nearly every word in her vocabulary seems directly related to the domestic. For Neville, the brilliant unrequited lover of Percival (a mutual friend of all the characters who dies while serving in India), the word “doomed” is in second place.

We might begin to wonder how vocabulary plays out along the gender axis. For example, we might modify the $tf - idf$ program so that it gives us lists of words that are spoken (but shared) only by the women in the novel and another that lists words spoken only by the men. When we do that, we find that the women possess fourteen words in common:
The men have ninety words in common:

- boys
- everybody
- united
- wheel
- hundred
- alas
- power
- language
- became
- poet
- observe
- central
- beak
- story
- course
- friend
- waste
- ease
- shoes
- bowl
- rushes
- lambert
- breath
- soften
- million
- coarse
- stockings
- pirouetting
- cotton
- wash
- antlers
- diamonds

These are provocative results, but the provocation is as much about our sense of what we are doing (the hermeneutical question) as it is about how we are doing it (the methodological question).

We might want to say that the purpose of these procedures is to confirm or deny the “serendipitous reading” of literary critics. Is Louis obsessed with his accent? Yes; the data confirms that he is. Critics who have argued for a deep structure of difference among the characters—one perhaps aligned along the gender axis—might also feel as if the program vindicates their impressions. Is there a gender divide? Yes; the characters are divided along the gender axis by a factor of 6.4285 to 1.

To level such arguments, however, is to turn the hermeneutical question back into a methodological one. To speak of the procedure as “verifying” some other finding is to beg questions of the procedure itself. And here, we are on somewhat shaky ground. The formula $tf - idf$ “works” in the context of information retrieval because it appears to match our general expectations. When we undertake a search for the term “baseball” with a search engine, we want to rule out passing references in favor of documents that are substantially about this topic. If we get back relevant hits, we could say that the $tf - idf$ formula has done its job. In the case of Woolf, we might say that we are getting back results that conform to our general expectations of what distinguishes the characters. But $tf - idf$ itself has no more claim to
truth value than any ordinary reading procedure. Manning and Schütze, in their magisterial work on statistical natural language processing, note that the "the family of \(tf-idf\) weighting schemes is sometimes criticized because it is not directly derived from a mathematical model of term distribution or relevancy" (544). The full version of the formula (the one used to generate the results above) includes a \(\log\) function and an addition:

\[
tf - idf = 1 + tf \cdot \log \left( \frac{N}{df} \right)
\]

The main purpose of these additions is not to bring the results into closer conformity with "reality," but merely to render the weighting numbers more sensible to the analyst. The logarithm dampens the function so that one term isn't a full six times more important than another; the +1 keeps the end of the curve from trailing off into negative territory.

Some text-analytical procedures do rely on empirical facts about language (or on statistical and mathematical laws in general). But even when they do, we often find ourselves unable to point to the truth of the procedure as the basis for judgment. We might say that this is because literary criticism is insufficiently scientific. We might even long for a "scientific literary criticism." We would do better to recognize that a scientific literary criticism would cease to be criticism.

It is no longer controversial to point out that science involves interpretation, rhetoric, social construction, and politics—as if this exposure of science's hidden humanism could somehow discredit the achievements of one of the world's greatest epistemological tools. No serious scientist could ever deny that interpretation, disagreement, and debate is at the core of the scientific method. But science differs significantly from the humanities in that it seeks singular answers to the problems under discussion. However far ranging a scientific debate might be, however varied the interpretations being offered, the assumption remains that there is a singular answer (or a singular set of answers) to the question at hand. Literary criticism has no such assumption. In the humanities the fecundity of any particular discussion is often judged precisely by the degree to which it offers ramified solutions to the problem at hand. We are not trying to solve Woolf. We are trying to ensure that discussion of The Waves continues.

Critics often use the word "pattern" to describe what they're putting forth, and that word aptly connotes the fundamental nature of the data upon which literary insight relies. The understanding promised by the critical act arises not from a presentation of facts, but from the elaboration of a gestalt, and it
rightfully includes the vague reference, the conjectured similitude, the ironic
twist, and the dramatic turn. In the spirit of *inventio*, the critic freely employs
the rhetorical tactics of conjecture—not so that a given matter might be
definitely settled, but in order that the matter might become richer, deeper,
and ever more complicated. The proper response to the conundrum posed
by Steiner’s “redemptive worldview” is not the scientific imperative toward
verification and falsification, but the humanistic propensity toward disagree-
ment and elaboration.

If algorithmic criticism is to have a central hermeneutical tenet, it is this:
that the narrowing constraints of computational logic—the irreducible ten-
dency of the computer toward enumeration, measurement, and verification—
is fully compatible with the goals of criticism set forth above. For while it is
possible, and in some cases useful, to confine algorithmic procedures to the
scientific realm, such procedures can be made to conform to the method-
ological project of *inventio* without transforming the nature of computation
or limiting the rhetorical range of critical inquiry. This is possible because
critical reading practices already contain elements of the algorithmic.

Any reading of a text that is not a recapitulation of that text relies on a
heuristic of radical transformation. The critic who endeavors to put forth a
“reading” puts forth not the text, but a new text in which the data has been
paraphrased, elaborated, selected, truncated, and transduced. This basic prop-
erty of critical methodology is evident not only in the act of “close reading”
but also in the more ambitious project of thematic exegesis. In the classroom
one encounters the professor instructing his or her students to turn to page
254, and then to page 16, and finally to page 400. They are told to consider
just the male characters, or just the female ones, or to pay attention to the
adjectives, the rhyme scheme, images of water, or the moment in which Nora
Helmer confronts her husband. The interpreter will set a novel against the
background of the Jacobite Rebellion, or a play amid the historical location
of the theater. He or she will view the text through the lens of Marxism, or
psychoanalysis, or existentialism, or postmodernism. In every case, what is
being read is not the “original” text, but a text transformed and transduced
into an alternative vision, in which, as Wittgenstein put it, we “see an aspect”
that further enables discussion and debate.

It is not that such matters as redemptive worldviews and Marxist read-
ings of texts can be arrived at algorithmically, but simply that algorithmic
transformation can provide the alternative visions that give rise to such
readings. The computer does this in a particularly useful way by carrying
out transformations in a rigidly holistic manner. It is one thing to notice
patterns of vocabulary, variations in line length, or images of darkness and
light; it is another thing to employ a machine that can unerringly discover every instance of such features across a massive corpus of literary texts and then present those features in a visual format entirely foreign to the original organization in which these features appear. Or rather, it is the same thing at a different scale and with expanded powers of observation. It is in such results that the critic seeks not facts, but patterns. And from pattern the critic may move to the grander rhetorical formations that constitute critical reading.

It might still make sense to speak of certain matters being "beyond the ambit of present computational stylistics." Research in text analysis continues to seek new ways to isolate features and present novel forms of organization. But the ambit of these ways and forms need not be constrained by a hermeneutics that disallows the connotative and analogical methods of criticism. Algorithmic criticism would have to retain the commitment to methodological rigor demanded by its tools, but the emphasis would be less on maintaining a correspondence or a fitness between method and goal and more on the need to present methods in a fully transparent manner. It would not be averse to the idea of reproducibility, but it would perhaps be even more committed to the notion of "hackability." For just as one might undertake a feminist reading of a text by transporting a set of heuristics from one critical context to another, so might the algorithmic critic undertake a particular type of reading by transforming a procedure that has been defined in terms of that most modern text, the computer program.

Algorithmic criticism undoubtedly requires a revolution of sorts, but that revolution is not one of new procedures and methods in contradistinction to the old ones. Algorithmic criticism seeks a new kind of audience for text analysis—one that is less concerned with fitness of method and the determination of interpretative boundaries, and one more concerned with evaluating the robustness of the discussion that a particular procedure announces. Such an audience exists, of course, and has existed for the better part of a century in the general community of literary critics from which text analysis has often found itself exiled. For this reason, text analysis practitioners should view the possibility of such a revolution as both welcome and liberating—not a critique of their methods, but a bold vote of confidence in the possibilities they hold.
Collaborative authorship in the twelfth century: A stylometric study of Hildegard of Bingen and Guibert of Gembloux

Mike Kestemont
Institute for the Study of Literature in the Low Countries & CLiPS Computational Linguistics Group, University of Antwerp, Belgium

Sara Moens and Jeroen Deploige
History Department, Ghent University, Belgium

Abstract

Hildegard of Bingen (1098–1179) is one of the most influential female authors of the Middle Ages. From the point of view of computational stylistics, the oeuvre attributed to Hildegard is fascinating. Hildegard dictated her texts to secretaries in Latin, a language of which she did not master all grammatical subtleties. She therefore allowed her scribes to correct her spelling and grammar. Especially Hildegard’s last collaborator, Guibert of Gembloux, seems to have considerably reworked her works during his secretaryship. Whereas her other scribes were only allowed to make superficial linguistic changes, Hildegard would have permitted Guibert to render her language stylistically more elegant. In this article, we focus on two shorter texts: the *Visio ad Guibertum missa* and *Visio de Sancto Martino*, both of which Hildegard allegedly authored during Guibert’s secretaryship. We analyze a corpus containing the letter collections of Hildegard, Guibert, and Bernard of Clairvaux using a number of common stylometric techniques. We discuss our results in the light of the Synergy Hypothesis, suggesting that texts resulting from collaboration can display a style markedly different from that of the collaborating authors. Finally, we demonstrate that Guibert must have reworked the disputed visionary texts allegedly authored by Hildegard to such an extent that style-oriented computational procedures attribute the texts to Guibert.

1 Introduction

Since the end of the 1960s, literary studies have seen a clear shift of focus from the analysis of authorial intentions to reader-oriented criticism. The repudiation of the modern idea of autonomous authorship has perhaps gone furthest in medieval studies, with the rise, since the late 1980s, of Material Philology (Nichols, 1997). Medievalists have become increasingly aware of the importance of manuscript culture in their understanding of texts: medieval texts should not primarily be studied, it is argued, as abstract entities resulting from authorial ambitions, but rather as tangible objects, materialized in...
specific manuscript contexts. Every material manifestation of a text is unique, because the acts of copying and compiling nearly always resulted in textual changes—from minor changes in orthography to complete rewritings. Our modern post-romantic conception of authorship therefore seems profoundly anachronistic with respect to the Middle Ages (Cerquiglini, 1999, p. 8–10). Yet, even if medieval culture did not share our present-day view on the significance of original authorship, the Middle Ages have known many respected and authoritative individuals who were recognized by their contemporaries and posterior readers as producers of very specific literary works. Some kind of correlation even existed between the degree to which texts were susceptible to alterations and the religious and intellectual authority of their authors (Deploige, 2005).

This did not mean, however, that such recognized authors were necessarily acting individually in the process of conceiving their treatises or narratives—quite the contrary. Writing in the Middle Ages meant entering into a dialogue with a long line of predecessors, whether through citations, paraphrasing, or allusions. In the actual process of literary composition too, medieval authors only seldom worked alone. A ‘new’ text could be the result of drafts on wax tablets copied by professional scribes, of processes of dictation and subsequent correction, etc. A twelfth-century authority like the Cistercian abbot Bernard of Clairvaux (1090–1153), one of the most prolific and influential medieval authors, is known to have been surrounded by a team of secretaries. For his sermons and letters in particular, he was assisted by a number of collaborators to whom he could dictate his messages or who were asked to produce texts in accordance with his own views. Some of his collaborators were even trained in imitating his writing style, thus facilitating Bernard’s work of final editing or correcting (Leclercq, 1962; 1987, pp. 147–52). In the case of the remarkably few medieval female authors known to us, the role of secretaries and collaborators is even more intricate. Women writers like the German nuns Hildegard of Bingen (1098–1179) or Elizabeth of Schönau (1129–1165) were considered unlearned and incapable of independently writing down their visionary experiences, even if these were ‘divinely inspired’. These women therefore had to be assisted by male collaborators, often also serving as their spiritual directors. The precise nature and implications of such cross-gender collaborations remain a topic of scholarly debate.

The immediate incentive for the present article is the preparation of a new critical edition of two lesser known texts attributed to Hildegard of Bingen, supposedly dating from the last years of her life: the *Visio de Sancto Martino*, which is conceived as a letter addressed to the worshippers of Saint Martin, and the *Visio ad Guibertum missa*, containing spiritual advice to an anonymous monk-priest, generally identified as her last secretary, Guibert of Gembloux (1124–1213) (Deploige and Moens, forthcoming). Among the few scholars who paid attention to these texts, there is still no consensus as to the extent to which they should be attributed to either Hildegard herself or to her collaborator Guibert. As neither traditional stylistic analysis nor contextual historical research has so far been able to resolve the problem, we will approach this issue through a stylometric analysis. We will focus on three research questions.

First, does stylometry allow for an authorial differentiation between the writings of twelfth-century Latin authors, belonging to highly similar intellectual circles? To answer this question, we will investigate the letter collections or *epistolaria* of Hildegard of Bingen, her secretary Guibert of Gembloux, and their famous contemporary, Bernard of Clairvaux. Our aim is to assess to what extent we can distinguish stylistic profiles for these authors, despite the marked *variance* within medieval manuscript culture (Cerquiglini, 1999), as well as the fact that these authors, like many of their contemporaries, were often assisted by secretaries. Next, we wish to analyze in more detail to what extent we can discern in Hildegard’s epistolary work, the influence of her last secretary, Guibert of Gembloux. Did her style undergo detectable stylistic changes under the editorial assistance of Guibert, or does the same homogeneous authorial voice appear throughout her epistolary work? Finally, we will assess the complex question to which author we should attribute, at least on
stylistic grounds, the *visiones* at stake in this article. In answering these research questions, we do not aim to develop novel stylometric techniques. The originality of this research is to be found in our application of a number of well-established techniques to assess their feasibility when dealing with medieval Latin texts, a textual tradition that until now has only rarely received attention in computational authorship attribution. Before addressing these issues, we will first briefly introduce the state of research with respect to the so-called *Mittarbeiter* problem in the Hildegard scholarship.

2 ‘Uneducated in the Art of Grammar’

The Benedictine nun Hildegard of Bingen was one of the most productive female authors of the Middle Ages (Newman, 1998). After a youth as anchoress at the abbey of the monks of Disibodenberg in the Rhineland near Mainz, she ended up as abbess of her own convent at the nearby Rupertsberg. Her extensive oeuvre includes genres as diverse as visionary books, letters, hagiographical texts, treatises on monastic life, musical compositions, and some works on physics and medical healing. Considered a true prophetess, receiving revelations and admonitions from God, she enjoyed a special status, even in the highest ecclesiastical milieux. Her extensive circle of correspondents, comprising, among others, popes and the emperor, testifies to her prophetic reputation. She was therefore able to gain an authority unprecedented for a woman, enabling her to even criticize the male clergy of her time. Among the first to approve her visionary gift was Bernard of Clairvaux, in a letter answering her request for support. Her female authorship was built on her recognition as a mouthpiece of God, which caused her to present herself during her entire life as a poor and uneducated woman—uneducated precisely because she was a woman (Deploige, 1998). In one of her *vitae*, her biographer Guibert of Gembloux specifies that she was ‘uneducated as to her schooling in the art of grammar’ (Derolez, 1988–1989, p. 377). Her status, both as a woman and an allegedly unlearned prophetess who may not have had the same type of schooling as young monks, meant that throughout her life Hildegard had to be assisted by secretaries (Ferrante, 1998).

Her first and principal secretary was Volmar of Disibodenberg, who remained her close associate until his death in 1173. He assisted in the redaction of the majority of her works. As we can learn from a famous miniature in the now lost manuscript (henceforth MS) Wiesbaden, Landesbibliothek, 1, dating from the end of her life, Hildegard dictated and wrote drafts on wax tablets, which were subsequently copied on parchment and linguistically ‘polished’ in accordance with the rules of grammar (Fig. 1). In addition, several Rupertsberg nuns must have aided their abbess as scribes during this period, given the number of known manuscripts produced in Rupertsberg under Hildegard’s supervision (Embach, 2003, p. 76, 128–9, 160, 184–5; Herwegen, 1904, p. 302–8). After Volmar’s death, Hildegard had to complete her last major visionary cycle, the *Liber divinorum operum* (‘Book of the Divine Works’), with more occasional assistance by a number of different collaborators from her immediate circle of spiritual acquaintances (Herwegen, 1904, p. 308–15). At the very end of her life, however, she was unexpectedly joined by Guibert, a monk from the abbey of Gembloux in Brabant (nowadays Belgium). Himself a fervent letter writer and hagiographer (Moens, 2010), he served as her secretary from 1177 until her death in 1179 (Delehaye, 1889; Ferrante, 1998, p. 122–30).

While even the authenticity of her female authorship had not always gone uncontested, until the seminal work by Schrader and Fürhrkötter (1956), a lot of scholarly efforts have been concerned with the precise role of Hildegard’s secretaries. Just as for other female writers working under the direction of father confessors (Coakley, 2006), the question has been raised to what extent Hildegard’s secretaries interfered with the final versions of her works, possibly generating male, clerical interpretations rather than original female viewpoints. Following the pioneering research by Herwegen (1904), most specialists now agree that the role of Hildegard’s collaborators was restricted to minor grammatical and stylistic alterations. Generally speaking, they had to copy her words *verbatim* unless they received Hildegard’s

It is generally assumed, however, that Hildegard must have granted a somewhat greater liberty to Guibert, who only entered into her life when she was already at the very advanced age of 79. Although their involvement was short, Guibert nevertheless had a significant impact on Hildegard’s literary
legacy. For example, he may have assisted her as one of the correctors in the final redaction of the Liber divinorum operum, of which MS Ghent, University Library, 241 (Fig. 9), can be considered the autograph copy most true to Hildegard’s own words (Derolez and Dronke, 1996, pp. xci–xciv). He also aided her in both the writing and compilation of portions of her epistolarium. On the basis of manuscript evidence, content, and dating, we can distinguish in Hildegard’s letter collection a part that must have been written and compiled with the help of Volmar and another group of letters that must have been written or transmitted under Guibert’s supervision.1 Last but not least, Guibert is also thought to have directed the compilation of the so-called Riesenkodex (MS Wiesbaden, Landesbibliothek, 2), the manuscript in which, by the end of her life, Hildegard had collected all the authorized versions of her works (Van Acker, 1989, pp. 129–34).

3 Two Suspect Visions

The Visio de sancto Martino (‘Vision of Saint Martin’) and Visio ad Guibertum missa (‘Vision sent to Guibert’), which are at stake in this article, cannot be found in the Riesenkodex. They are only preserved in three manuscripts that can be linked to the abbey of Gembloux and Guibert’s own oeuvre.2 Therefore, both texts are traditionally not included in the core of Hildegard’s canon (Schrader and Führkötter, 1956, p. 182; Embach, 2003, p. 469). Whereas the titles in the manuscripts (Fig. 2), as well as Guibert’s accompanying letters, firmly attribute these visiones to Hildegard, there are good reasons to suspect that Guibert must have been extensively involved in their final redaction. The figure of Saint Martin for instance—the main topic of the Visio de sancto Martino—is entirely absent from Hildegard’s oeuvre. Guibert, on the other hand, developed a lifelong fascination for this saint and devoted nearly half of his life to spreading his cult. The Visio ad Guibertum missa discusses the role of the priest as well as the topic of literary collaboration, both issues of direct relevance to Guibert. Moreover, the end of the latter text contains a passage of particular interest in which Hildegard grants Guibert the exceptional right to revise her texts more fundamentally than simply at the level of style and grammar:

When you correct [the Visio de sancto Martino] and the other works, in the emending of which your love kindly supports my deficiency, you should keep to this rule: that adding, subtracting, and changing nothing, you apply your skill only to make corrections where the order or the rules of correct Latin are violated. Or if you prefer—and this is something I have conceded in this letter beyond my normal practice—you need not hesitate to clothe the whole sequence of the vision in a more becoming garment of speech, preserving the true sense in every part. For even as foods nourishing in themselves do not appeal to the appetite unless they are seasoned somehow, so writings, although full of salutary advice, displease ears accustomed to an urbane style if they are not recommended by some color of eloquence (translated by Newman, 1987, p. 23).

With this statement, Hildegard allegedly granted Guibert editorial privileges that she had not allowed any other previous collaborator. The passage also prompted scholars to have a closer look at the authorship, style, and content of these visionary texts. Already in his 1882 edition, Pitra voiced doubts with respect to Hildegard’s alleged authorship. He stated that Guibert, if not their original author altogether, must at least have reworked the texts profoundly. Pitra based his verdict on a number of syntactical features, on metaphors which he considered typical of Guibert, and on the extensive insertion of Biblical quotations (Pitra, 1882, p. 370–1, 375). Herwegen remained more cautious: although he accepted that Guibert had refined the texts stylistically, he still discerned Hildegard’s authorial voice shimmering through Guibert’s multiple corrections. He recognized Hildegard’s genius in the overall structure of the visions and in some typically Hildegardian vocabulary. He also rejected Pitra’s assertion that the numerous Biblical quotations could only have been inserted by Guibert (Herwegen, 1904, p. 394–6).
Newman recently stated that the Visio ad Guibertum missa was ‘written by Guibert in Hildegard’s persona’ (Newman, 1987, p. 24), although Van Acker (1989, p. 130) and Coakley (2006, p. 61) continued to consider Hildegard as the text’s author and Guibert as a mere stylistic reviser.

These assertions concerning the authorship of the visiones seem to have been predominantly based on subjective appreciations of style and content and the arguments used in this debate remain, at best, intuitive. The appearance of a new critical edition of the visiones once more put the question of their authorship at the forefront: should the texts be regarded as Hildegardian or pseudo-Hildegardian? Stylometric methods may provide a more objective basis for disentangling the issue and to re-assess the nature of Guibert’s secretaryship.

4 Corpus Preparation

For the present study, Brepols Publishers generously provided a digital corpus containing the nearly complete works of Hildegard, Guibert, and Bernard of Clairvaux. We obtained these texts in raw format, corresponding to the way they are included in the Brepols electronic Library of Latin Texts, on the basis of modern critical editions. Fortunately, these editions are all based on manuscripts that were compiled under the supervision of the original authors or at least in their close vicinity, so that we do not have to worry about major scribal interventions. The fact that all three authors in our corpus have been productive letter writers rendered their epistolaria an attractive point of departure. Moreover, the two short visionary texts of dubious origin that are at issue in this article are mostly comparable with Hildegard’s letters with respect to length, topics, and manuscript tradition. Obviously, we restricted our authors’ letter collections to the letters they wrote themselves, leaving aside the letters that were merely addressed to them and that were usually contained in the same manuscripts (Constable, 1976). For Bernard, this resulted in a sub-corpus of 166,063 words and for Guibert of 124,580 words. Hildegard’s letter collection contained 109,633 words, 82,154 of which are contained in the part compiled with the help of her first secretary.
Volmar, while the remaining 27,479 words constitute the letters that, as discussed earlier, have most probably been edited in some way by Guibert.5

Medieval Latin is characterized by unstable orthography. As even a single scribe often used different spellings for the same word, modern editors already tend to silently normalize morphological variants. We have normalized the orthography in our corpus even further via lemmatization, a useful procedure in stylometry for medieval texts (Kestemont et al., 2010). The texts were first tokenized using the Natural Language Toolkit (Bird et al., 2009). The coordinating conjunction –que (‘and’) was not realized as a separate word in medieval Latin, but it was appended to the preceding word (e.g. terra aquaque, ‘land and water’). To automatically isolate the clitic, we have stripped the suffix (‘que’) from every word that did not occur in a list of words proposed by Schinke et al. (1996, p. 180–1).6 We have also split up the medieval contraction of the reflexive pronoun se and the idiomatic reinforcement ipsum in seipsum (or teipsum, teipsam, etc.).

A number of specific character combinations were freely interchangeable in medieval Latin, such as ph for f, v for u, oe or ae for e (or for ë, the so-called ‘e caudata’) (Rigg, 1996). We have therefore lifted the difference between v and u, as well as between ae, oe, and e, by substituting all vs for us and all aes and oes for es. For the substitution of ae and oe by e, this actually meant that we were sometimes forced to erase the distinction between grammatically important morphemes (e.g. between the male vocative singular domine and the female nominative plural dominæ). Yet, this was unavoidable, as a good deal of the aes and oes in our corpus were already contracted to es, making it nearly impossible to automatically normalize them the other way round. Subsequently, we checked whether the surface tokens in our corpus were present in a large and representative word list from the Perseus Project (Tufts University). When a token was not, we used a permutation algorithm to generate plausible spelling variants for it. If one of these newly generated forms was contained in the word list, the original form was replaced by its newly generated counterpart. To generate these variants, we constructed an array with all possible variations for the consecutive character groups. Next, we combined these options through the Cartesian product in the matrix by means of a permutation algorithm (Kestemont et al., 2010). Table 1 lists the series of common alternative character combinations we have considered, loosely based on Riggs (1996).7 An example matrix for a word like chirographum would be: {[c], [h | Ø], [i | y], [r], [o], [g], [r], [a], [ph | f], [u], [m]}. All unique, alternative word spellings that can be generated on the basis of the matrix are: chirographum, chirographus, chirographum, chirographum, cirografum, cirografum, cirografum, and cyrografum.

Finally, we automatically annotated the tokens with lemmas using the medieval Index Thomisticus Treebank (IT-TB: Passarotti and Dell’Orletta, 2010) as training material (ca. 170,000 tokens; ca. 9,000 sentences).8 For the lemmatization of our corpus we have used Morfette (Chrupala et al., 2008). Unlike other popular lemmatization tools, such as TreeTagger (Schmid, 1994), Morfette also lemmatizes input tokens that the tagger did not already encounter verbatim in the training data. Morfette considers pairs of input tokens and lemmas in the training material. From these pairs it learns ‘shortest edit scripts’ or ways to transform tokens into their lemmas using character insertions, deletions, and replacements. An annotated sample from the Visio ad Guibertum missa is listed as an example (Table 2), illustrating how this procedure did not manage to identify all lemmas correctly. Especially content words that are not typical of Thomas Aquinas’s scholastic vocabulary were not always recognized. For the function words used in our analyses (see below), this problem was fortunately hardly an issue.

Table 1 Interchangeable medieval Latin character combinations allowed in our permutation algorithm

<table>
<thead>
<tr>
<th>a</th>
<th>e</th>
<th>i</th>
<th>o</th>
<th>u</th>
</tr>
</thead>
<tbody>
<tr>
<td>v</td>
<td>f</td>
<td>j</td>
<td>u</td>
<td>y</td>
</tr>
<tr>
<td>c</td>
<td>h</td>
<td>k</td>
<td>l</td>
<td>m</td>
</tr>
<tr>
<td>n</td>
<td>p</td>
<td>q</td>
<td>r</td>
<td>s</td>
</tr>
<tr>
<td>t</td>
<td>u</td>
<td>v</td>
<td>w</td>
<td>x</td>
</tr>
<tr>
<td>z</td>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
</tr>
<tr>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
<td>I</td>
</tr>
<tr>
<td>J</td>
<td>K</td>
<td>L</td>
<td>M</td>
<td>N</td>
</tr>
<tr>
<td>O</td>
<td>P</td>
<td>Q</td>
<td>R</td>
<td>S</td>
</tr>
<tr>
<td>T</td>
<td>U</td>
<td>V</td>
<td>W</td>
<td>X</td>
</tr>
<tr>
<td>Y</td>
<td>Z</td>
<td>a</td>
<td>b</td>
<td>c</td>
</tr>
</tbody>
</table>

---

6. We have also split up the medieval contraction of the reflexive pronoun se and the idiomatic reinforcement ipsum in seipsum (or teipsum, teipsam, etc.).

7. An example matrix for a word like chirographum would be: {[c], [h | Ø], [i | y], [r], [o], [g], [r], [a], [ph | f], [u], [m]}. All unique, alternative word spellings that can be generated on the basis of the matrix are: chirographum, chirographus, chirographum, chirographum, cirografum, cirografum, cirografum, and cyrografum.

8. For the lemmatization of our corpus we have used Morfette (Chrupala et al., 2008). Unlike other popular lemmatization tools, such as TreeTagger (Schmid, 1994), Morfette also lemmatizes input tokens that the tagger did not already encounter verbatim in the training data. Morfette considers pairs of input tokens and lemmas in the training material. From these pairs it learns ‘shortest edit scripts’ or ways to transform tokens into their lemmas using character insertions, deletions, and replacements. An annotated sample from the Visio ad Guibertum missa is listed as an example (Table 2), illustrating how this procedure did not manage to identify all lemmas correctly. Especially content words that are not typical of Thomas Aquinas’s scholastic vocabulary were not always recognized. For the function words used in our analyses (see below), this problem was fortunately hardly an issue.
Table 2 Example of lemmatization based on Morfette

<table>
<thead>
<tr>
<th>Original</th>
<th>Lemma</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>in</td>
<td>in</td>
<td>‘in’</td>
</tr>
<tr>
<td>uisionem</td>
<td>uisio</td>
<td>‘vision’</td>
</tr>
<tr>
<td>anime</td>
<td>anima</td>
<td>‘soul’</td>
</tr>
<tr>
<td>mee</td>
<td>meus</td>
<td>‘my’</td>
</tr>
<tr>
<td>,</td>
<td>/</td>
<td>/</td>
</tr>
<tr>
<td>uidi</td>
<td>uideo</td>
<td>‘I see’</td>
</tr>
<tr>
<td>ingentem</td>
<td>ingentem</td>
<td><em>not recognized</em> [ingens = ‘gigantic’]</td>
</tr>
<tr>
<td>rutilantis</td>
<td>rutilo</td>
<td>‘glow’</td>
</tr>
<tr>
<td>ignis</td>
<td>ignis</td>
<td>‘fire’</td>
</tr>
<tr>
<td>nubem</td>
<td>nubem</td>
<td><em>not recognized</em> [nubes = ‘cloud’]</td>
</tr>
</tbody>
</table>

Translation: ‘In a vision of my soul, I saw a gigantic cloud of glowing fire.’

5 Feature Selection

Today’s stylometry has become an umbrella term for a still growing number of techniques for authorship analysis. Each of these has been the subject of both criticism and praise, making it hard to discern a consensus on best practice in this field. For this research too, we had to balance the pros and cons of a number of tried and tested methodologies. Recent studies still tend to agree on the undeniable methodological advantages of using function words in authorship attribution (Binongo, 2003, p. 11). An author’s use of function words is said, for instance, to be relatively unaffected by a text’s topic or genre. (Dis-)similarities between texts regarding function words are therefore to a certain extent content-independent and can be more easily associated with authorship than e.g. content words or other topic-specific stylistics (Juola, 2006, p. 264–5). Numerous empirical studies have effectively demonstrated that analyses of the high-frequency strata of function words yield reliable indications about a text’s authorship (Koppel et al., 2009, p. 11–12; Stamatatos, 2009, p. 540–1). In this research, we have therefore restricted our analyses to function words, using a number of approved methods—many of them implemented in the publicly available script suite ‘Stylometry with R’ (Eder et al., 2013).

Preliminary analyses showed that the upper tail of the frequency spectrum in our corpus still contained a good deal of content-rich lemmas. Among the ca. 200 most frequent lemmas in our entire corpus, listed in Table 3, we came across multiple topic-specific nouns like deus, dominus, sanctus, … and verbs like facio, uideo, uiuo, … The inclusion of such lemmas obviously reflects the corpus’s fairly specific, religious semantics. It is also related, however, to the simple fact that a highly inflected language like Latin with its many declensions makes less use of function words than weakly inflected languages like English. A third explanatory factor might be the fact that we worked with the frequencies of lemmas instead of surface forms. It thus seemed advisable to remove these content words from our data tables.

The content-rich words we chose to remove are marked by a hashtag (#) in Table 3. The words followed by an asterisk (*) in the same Table 3 are non-reflexive personal pronouns, which are also often culled in stylometry to avoid the intrusion of genre-related or topic-specific features. Naturally, a collection of letters will contain more instances of the second-person pronouns tu/vos (‘you’) or tuus/vester (‘your’) than a saint’s life. In our analyses, we have deleted this kind of pronoun. Just as in Table 2, one can still distinguish a certain number of wrongly lemmatized tokens in Table 3. The surface form sui, for example, often seems to have remained unchanged, whereas it should have been transformed into suus. This particular error, however, is neutralized by our elimination of non-reflexive personal pronouns. In sum, our culling of the lemmas in Table 3 resulted in 65 function words with which to form the basis for the actual analyses.

It should be noted, however, that character n-grams might have been an attractive additional feature type for our research, as these have often been shown to be excellent features in authorship attribution (Koppel et al., 2009, p. 12–13; Stamatatos, 2009, p. 541–2). This method, which does not require any kind of normalization or lemmatization, segments texts into consecutive, partially overlapping groups of n characters—the word ‘bigram’ for instance contains the bigrams ‘_b’, ‘bi’, ‘ig’, ‘gr’, ‘ra’, ‘am’, ‘m_’. Contrary to a word-level approach, character n-grams are also sensitive to stylistic information below the word level, like case endings or other grammatical morphemes that are
not realized as separate words (Rybicki and Eder, 2011, p. 320). Latin, for instance, is a heavily inflected language that makes use of affixes to mark the grammatical functions of words—‘by iron, not by sword’ being for example ‘ferro non gladio’ (Sapir, 1921, ch. VI). Therefore, it would have made sense to additionally study the character n-grams in the corpus.

However, one runs into the aforementioned problem that historical languages are characterized by unstable orthography (Piotrowski, 2012). Although Latin spelling variation seems to have been less pronounced than in vernacular medieval languages, it does constitute a serious issue. When comparing two texts written by the same author, surviving in manuscripts with a strongly divergent orthography, stylometric methods may detect artificially large differences. Conversely, and likewise due to scribal interference, texts of non-identical authorial provenance may show artificial similarities when they survive in manuscripts with a similar orthographical profile. In medieval manuscripts, we might even find inconsistent word spellings for the same words throughout the same text (Rigg, 1996). This ultimately implies that an approach based on character n-grams is unadvisable for medieval Latin (cf. Kestemont and Van Dalen Oskam, 2008).

Unfortunately, this means that our approach based on lemmatization cannot take into account stylistic subtleties below the word level (e.g.

| Table 3 Most frequent lemmas in the corpus (# = content words; * = non-reflexive pronouns) |
|---------------------------------------------|---------------------------------------------|---------------------------------------------|---------------------------------------------|
| et   | e   | quoniam | caritas | consilium | contra |
| qui  | uel | uerbum  | uenio   | rex      | pono  |
| in   | #possum | aut  | quasi   | dum      | amicus |
| #sum | pro  | idem   | scilicet | talis    | honor |
| non | quam | super  | causa   | ceterus  | nomen |
| #tu* | #uester* | #terra | manus   | caro     | uelut |
| #is* | autem | #ulo   | iustitia | #fides   | ante  |
| #ego* | multus | nunc | #modus | #res     | #ta   |
| #deus | habeo | iam    | #primus | #parus   | #judicium |
| ad   | ne  | #uita  | semper  | #pax     | apud  |
| hic  | sanctus | ac  | #audio  | salus    | quantum |
| sed  | etiam | #nam  | #mundus | siue     | #lex   |
| ut   | enim | #cor  | #debeo | #salus   | #fidelis |
| de   | #noster* | #do  | uiiuo   | #eternus | #sol   |
| #suus* | #uerus | #solus | #cado   | #inuenio | #celestis |
| #ille* | #uideo | unde | inter   | #frater | #potior |
| a    | sicut | quidem | o      | #uir    | uidelicet |
| cum  | alius | tam  | #diligo | magis    | tun   |
| quod | ita  | propter | #voluntas | #fors    | #angelus |
| ipse | tamen | #quidam | #gloria  | #us      | #diluinus |
| #tuus* | #filius | #bonus | quoque  | #certus  | #summus |
| #omnis | #spiritus | ergo | #atque  | #loquor  | #ideo |
| si   | #christus | #tempus | #aliqui | #uoX     | #prior |
| #sui* | #bonum | sine  | #malum  | #iustus  | #populus |
| per  | #ecclesia | nisi | #mens   | post     | #episcopus |
| #facio | #opus | #unus | #oculus | #misericordia | #similis |
| #homo | xque | #dies | #nihil | #celum  | #os   |
| #dico | sic | #nullus | #secundum | #adhuc | #nous |
| quia | #magnus | ubi  | #pars   | #domus  | #tamant |
| #dominus | #iste* | #corpus | #mors   | #uis    | #ui |
| #meus* | #anima | #locus | #peccatum | #beatus | licet |
| nec  | #pater | #uirtus | #scio  | #quomodo | #predico |
| #quis | #gratia | #totus | #hildegars | #ueritas | #fratres |

Collaborative authorship in the twelfth century
indicative versus subjunctive mood, as expressed in case endings). However, we will demonstrate that our method is still able to harvest sufficient stylistic information from the texts. Indirectly, our results will therefore even serve to emphasize how much grammatical information is in fact still expressed by isolated function words in medieval Latin.

6 Testing Principal Components Analysis

The first stylometric technique we adopt is principal components analysis (PCA), a procedure derived from multivariate statistics and commonly used to reduce the dimensionality of a data set (Binongo, 2003). By combining the original variables of a data table into new, uncorrelated compound variables or ‘principal components’, PCA is able to summarize large and complex data sets into insightful lower-dimensional scatterplots. When applied to the frequencies of high-frequency items in texts, this technique often successfully reveals the authorial structure in a data set. PCA’s good performance in authorship attribution is due to the fact that it explicitly tries to model correlations between word frequencies. Especially the frequencies of function words show complex correlations that are related to stylistic, arguably authorial choices between small sets of alternative options. A mere visual inspection of the samples’ positions in PCA scatterplots often shows that samples written by the same author will cluster, whereas groups of samples written by distinct authors lie further apart.

Because of the considerable size of the epistolaria in the corpus, we could start with a large sample size of 10,000 lemmatized words per sample. Recent research has demonstrated that the accuracy of most authorship attribution techniques is likely to increase when larger samples are taken (Eder, 2010; Luyckx and Daelemans, 2011). Our selection of the epistolaria of exactly three authors—Hildegard of Bingen, Guibert of Gembloux and Bernard of Clairvaux—respects the fact that it is theoretically unadvisable to include more than three authors in a PCA, especially when the discussion of the results is restricted to the two first Principal Components (PCS) (Binongo and Smith 1999, p. 464). As is customary since Burrows (1987), our PCA is based on the correlation matrix, appropriately scaling the original word frequencies.

Fig. 3 shows the scatterplot that results from our first experiment. Each author’s samples are visualized as black letter combinations: the first letter of the author’s name is followed by a digit, indicating the sample’s indexed position in the respective epistolaria. G_EP-4, for instance, is the fourth sample of 10,000 lemmatized words taken from Guibert’s epistolium. At this stage, we are restricting Hildegard’s epistolium to the letters that are not associated in any way with Guibert’s secretaryship. Fig. 3 displays a remarkably clear authorial separation of the samples. Guibert’s samples (G_EP) are invariably positioned to the left. Finally, Bernard’s samples (B_EP) form a tight cluster of samples in the lower-right half of the plot. The density of this last cluster thus points at a clear stylistic unity, despite the fact that, as noted earlier, Bernard must have been assisted in his epistolary work by a true personal chancellery consisting of at least five different collaborators (Leclercq, 1987, p. 147–52).

Additionally, the plot in Fig. 3 contains a series of high-frequency items in light grey, the ‘component loadings’, visualizing how strongly the 65 lemmas have contributed to the creation of the PCS. If a word can, for instance, be found to the far left of the scatterplot, this demonstrates that it is relatively more frequent in samples with a similar position in the plot. Our first scatterplot thus shows that the use of et (‘and’) and a (‘from’) is surprisingly typical of Guibert’s writings, whereas the use of the preposition in (‘in’) is very characteristic of the Hildegard samples. In comparison, the use of the lemmas non or si seems to be relatively more typical of Bernard’s writing. The scatterplot does not reveal any anomalies and it is safe to assume that the high-frequency grammatical lemmas argue in favor of a clear stylistic differentiation between our authors.

The remarkable stylistic differences with respect to a number of specific lemmas used by our authors can be highlighted in another way. The boxplots in Fig. 4 visualize information about the absolute
frequencies (medians, quartiles, etc.) for three interesting function words—*in*, *et*, and *non*—in samples of 2,000 words. In boxplot (a) concerning the use of *in*, the primary column refers to the counts in Hildegard; in the second boxplot (b) dealing with *et*, the left column concerns Guibert; and in boxplot (c), with the results for *non*, Bernard’s results are displayed in first column. The secondary column in all three boxplots refers to the material by the two other authors, e.g. Guibert and Bernard in boxplot (a). These boxplots indeed reveal unmistakable differences between the respective *epistolaria* with respect to the frequency of these important function words. Interestingly, these differences coincide with stylistic observations that have been made in traditional philological research. Given the visionary discourse developed in much of her writings—even in her letters—it is not surprising to come across an intensive use of the preposition *in* in Hildegard’s letters. She repeatedly sees things *in* divine visions; she continuously searches the allegorical meanings buried *in* the multitude of details that she discovers in her visions (Dronke, 1998). Guibert’s writings are especially notorious for their all too inflated and artificial style, and Guibert’s wearisome tendency to compose extremely long
sentences, full of coordinating conjunctions (see also Derolez, 1988, p. v and ix). Bernard’s frequent use of non can be related to the didactic nature of his epistolary expositions in which he very often relies on an antithetical style to illustrate his thoughts (Mohrmann, 1958; Pranger, 2011, p. 222).

7 Testing Delta

For our PCA displayed in Fig. 3, we have been working with extremely generous sample sizes of 10,000 lemmas each. Because the ultimate goal of this article remains the attribution of the Visio ad Guibertum missa and the Visio de Sancto Martino of which the authorship seems very questionable, the problem of sample size needs to be put forward (Eder, 2010; Luyckx and Daelemans, 2011): while the first disputed visio at stake in this article still contains 7,489 lemmas, the latter only counts 3,301 words. The scatterplots in Fig. 5a and b show the results of the same procedure as in Fig. 3 but using sample sizes of 5,000 and 1,000 lemmas, respectively. This clearly illustrates the decrease in discriminatory performance of our PCA when we reduce the sample size in our experiments. Fig. 5b demonstrates that the authorial discrimination becomes less powerful, in particular between Guibert and Bernard in the vertical component.

To what extent will we be able to rely on PCA for a fairly solid attribution of a text, like the Visio de

![Boxplot for "in"](a) Boxplot for "in"

![Boxplot for "et"](b) Boxplot for "et"

Fig. 4. (a–c) Boxplots of the absolute frequencies of in, et, and non in epistolary samples of 2,000 lemmas
Sancto Martino, of only ca. 3,000 words? Although the scatterplots in the previous section demonstrate the general validity of the stylometric approach for our corpus, it makes sense to apply a second attribution technique to our corpus to validate the outcome of the PCA more precisely. Because it is unfeasible to generate new scatterplots for every small change in parameter settings like e.g. sample size in our experiments, we additionally apply Burrows’s Delta (2002) to the *epistolaria*.

In its traditional implementation, Delta offers a similarity metric to determine the authorship of anonymous works. Based on the frequencies of a small set of high-frequency items, Delta computes the stylistic distance between an unknown sample and a set of samples written by a series of candidate authors. It will attribute the anonymous sample to the author of the (single) sample in the data set to which it is closest in style according to the metric. As such, Delta uses a ‘nearest neighbor’ reasoning (Argamon, 2008). We can apply a ‘leave-one-out validation’ with Delta as follows. We can temporarily treat each sample in our collection as anonymous. Next, we can have Delta attribute the anonymized sample to one of the candidate authors and check whether the suggested attribution is successful or not. If at the end of this procedure, we divide the number of correct attributions by the total number of samples in the data set, we get a percentage that offers a useful approximation of the general effectiveness of our technique, should it, for instance, be applied to real-world samples of unknown provenance.

Fig. 6 shows the result of this leave-one-out validation for various sample sizes (multiples of 100 lemmas, ranging from 500 to 4,000). It is obvious that larger sample sizes invariably lead to higher accuracies in cross-validation. Yet, whereas the initial accuracies are fairly low (even < 85%), the attribution success quickly rises above the psychological barrier of 95% (sample sizes > 1,500 lemmas) and becomes entirely flawless when dealing with sample sizes of ca. 3,000 lemmas or more. For a text counting 3,301 lemmas, like the *Visio de sancto Martino*, we might well reach an attribution accuracy of about 99%. Moreover, because these numbers are in line with earlier reports concerning modern languages (Eder, 2010; Luyckx and Daelemans, 2011), Fig. 6 again demonstrates that even a highly inflected language like Latin contains a satisfying amount of useful stylistic information in its grammatical lemmas alone.

By now, we can assume that, when applied cautiously, PCA should offer enough solid ground to make conjectures about the authorship of the visions in the corpus traditionally attributed to Hildegard. Following a nearest neighbor reasoning (Argamon, 2008), we can plot unseen, anonymous texts together with the works of established authorial origin and investigate to which of the authorial clusters the unseen work is most similar in style. However, before moving on to the analysis of the visions, we have first tested this attribution procedure. In the PCA scatterplot in Fig. 7, we have added a new, ‘anonymous’ sample (amounting to 3,706

**Fig. 4. Continued**
lemmas) by author ‘X’ to equal-sized samples from the aforementioned epistolaria. The new sample turns out to be stylistically much more similar to Bernard’s samples than to those by Hildegard or Guibert. Should this sample have been truly anonymous, the analysis would have offered firm grounds for conjecture that the text from which the sample is derived is actually authored by Bernard of Clairvaux. In this specific case, this reasoning would have led to a historically sound attribution, as the anonymous text we have questioned is in reality the *Sermo in festo sancti Martini*, written by Bernard around 1150. An interesting fact about this example is that even though the topic and genre of this text are perhaps quite different from the epistolary material of our candidate authors (viz. a sermon about the aforementioned Saint Martin), it is clear that our PCA procedure allows for solid conclusions. Although one should perhaps not always expect such clear-cut stylistic, authorial differentiation in historical corpora, this promising example clearly illustrates the benefits of the present methodology for (future) research.

![PCA plot](image-url)

**Fig. 5** (a and b) PCAs with reduced sample sizes (5,000 and 1,000 lemmas/sample)
**8 Guibert’s Secretaryship: Synergy and Beyond?**

As discussed earlier, we have discerned two groups of letters in Hildegard’s *epistolarium*: one that must have originated at the time when Volmar was still Hildegard’s secretary and that bears no potential traces of Guibert’s interference, and another containing the letters that are likely to have been revised by Guibert. If we confront samples of 5,000 lemmas from both portions, labeled here H_EPNG and H_EPG, respectively, in a PCA, we get the result in Fig. 8.

We notice that the first, horizontal PC captures an impressive 37% of the original variation in our data and primarily relates to the stylistic differentiation between Guibert’s own letter collections (G_EP) and the anterior portion of Hildegard’s *epistolarium* (H_EPNG). Interestingly, we see that the second PC in the right half of the plot (still capturing 9.4% of the original variation) discriminates between Hildegard’s non-Guibertian letters and her letters that can be associated with Guibert’s secretaryship.

---

**Fig. 5 Continued**
These results thus suggest that there do indeed exist stylistic differences between the oldest portion of Hildegard’s *epistolarium* and the letters in which we expected to discern Guibert’s editorial fingerprints. They also confirm what can be deduced from the surviving manuscript evidence. The so-called autograph copy of the *Liber divinorum operum* mentioned earlier offers unique insight into the way in which Hildegard’s collaborators must have edited her texts under her supervision (Derolez, 1972). Fig. 9, showing a number of lines from the randomly selected page 370 of MS Ghent, University Library, 241, makes it clear that it was the function words in particular that were often altered by Hildegard’s correctors; *tam* being erased, *quod* being replaced by *ut* or *quia*, *ad* being added, *et cetera*. A collaborator—especially Guibert, who is known to have had a great deal of freedom in his editorial work—may thus have had a notable impact on Hildegard’s stylistic profile.

However, in Fig. 8, we see that the samples from Hildegard’s *epistolarium* that bear the influence of Guibert’s interference do not seek the company of Guibert’s own writings in the scatterplot. After all, they continue to be somewhat more similar to Hildegard’s style. This result is reminiscent of the Synergy Hypothesis, recently discussed by Pennebaker (2011). Pennebaker puts forward three hypotheses concerning the stylistic effect of collaborations between different authors. Such projects can produce a language that is (1) similar to the one produced by a single person writing alone, (2) the average of the two writers, or (3) unlike either of one of the styles that the collaborating authors would produce on their own. Based on exploratory research on the Federalist papers and Beatles songs, Pennebaker ultimately argues in favor of the latter, so-called ‘Synergy view’ on collaborative authorship, not refuting however the possibility that one of the collaborating authors might have remained more influential with respect to the end product (cf. Petrie et al., 2008). This Synergy Hypothesis thus might be applicable to a certain extent to the Hildegard–Guibert ‘collaboration’, where the result of the creative process does not fit in with the other letter samples written by Hildegard or Guibert individually, although the result is somewhat more similar to Hildegard.
More can be learned about the stylistic dichotomy in Hildegard’s *epistolarium* by applying a Mann–Whitney test to the lemmas occurring at least twice in 4,000 lemma samples. Here, we temporarily leave the realm of high-frequency lemmas and venture into the lower-frequency strata of the lexical spectrum. Hence, this test will not particularly emphasize the discriminatory power of high-frequency lemmas, as was the case with our other tests (Kilgariff, 2001). Fig. 10 contrasts the words that were predominantly used in the Hildegard’s letters written under Volmar’s secretaryship with those that become typical when Guibert took over the editorial work in the preservation of her letters. The lemmas have been ranked and plotted according to the $U$ test statistic obtained for each lemma. Fig. 10 learns how the use of the relative pronoun *qui* (‘who’) for instance only becomes prominent in letters edited by Guibert, who is indeed notorious for constructing eloquent but complex sentences with a lot of embedded relative clauses. Moreover, this latter group of letters is also characterized by a
more dry and stereotypical ecclesiastical vocabulary (omnipotens, sanctus, spiritus, verus, . . .), whereas the letters not influenced by Guibert betray a more direct and lively narrative style (sed, tunc, nunc, dico, ergo, deinde, . . .), possibly more true to Hildegard’s own preferred way of expressing herself. We might thus be inclined to agree with Newman (1987, p. 24) when she stated: ‘Purists can at least rejoice that the collaboration [between Guibert and Hildegard] began only after the seer’s major works were completed’. From the methodological point of view, these results also show that the discriminatory effects in lower-frequency strata correspond with the stylistic dichotomy present in the high-frequency vocabulary, thus corroborating the performance of the latter methodology.

Fig. 8 PCA of the epistolarium of Guibert, of the letters of Hildegard transmitted without Guibert’s editorial assistance, and of the Guibertian letters in Hildegard’s epistolarium (5,000 lemmas/sample)

Fig. 9 MS Ghent, University Library, 241, p. 370 (detail). Reproduced with permission
Let us finally turn to the original incentive for the present article, namely, the authorship discussion concerning two texts of dubious provenance: the relatively short *Visio de Sancto Martino* about Saint Martin (3,301 lemmas) and the somewhat longer *Visio ad Guibertum missa* (7,492 lemmas). Fig. 11 offers the result of three PCAs in which we have confronted both ‘*dubia*’ (hence D_MART and D_MISSA) with the previously discussed epistolary collections, again using the same 65 lemmas and a sample size of 3,301 lemmas. Fig. 11a considers all texts by all authors; Fig. 11b excludes Bernard’s texts; Fig. 11c only considers Guibert’s *epistolarium* and the ‘anonymous’ visionary texts.

All subplots in Fig. 11 clearly show that both visions tightly cluster with Guibert’s *epistolarium*, instead of with Hildegard’s. This effect is perhaps least prominent in Fig. 11a, where D_MART and D_MISSA display modest similarities to some of the epistolary samples from the portion of Hildegard’s *epistolarium* that was revised by Guibert. In all three plots, however, the visions are generally speaking far more similar to Guibert’s writings than to Hildegard’s. Significantly, most samples resulting from the combined authorial voices of Hildegard and Guibert again do not display any significant rapprochement to the *epistolaria* of the individual authors. These observations seem to reinforce the Synergy Hypothesis. Moreover, the visions’ quasi-

---

**Fig. 10** Results of Mann–Whitney test (*U* statistic) comparing the vocabulary in Hildegard’s *epistolarium* before and during Guibert’s secretaryship

---

**Before Guibert**

- sed
- nunc
- dico
- in
- hic
- deinde
- ergo
- non
- ubi
- populus
- rectus
- uenio
- mysticus
- amo
- quare
- quomodo
- mens
- surgis
- quod
- uale
- interdum
- semetipse

**With Guibert**

- omnipotens
- e
- qui
- uerus
- licet
- summus
- efficio
- precipio
- designo
- a
- per
- uel
- imito
- sanclus
- indumentum
- numquam
- iesus
- cesso
- semper
- possum
- solus

Mann-Whitney *U*
random position in the final subplot (Fig. 11c) reveals no pronounced stylistic differences with Guibert’s letters, regarding the high-frequency lemmas analyzed. They invariably cluster with Guibert’s epistolary oeuvre, making him a much more plausible author than Hildegard—at the very least, from a stylistic point of view.

An important, yet inconspicuous, last feature of Fig. 11a is that it includes the *Sermo in festo sancti Martini*, even though it can hardly be spotted among Bernard’s other samples. This sermon deals, just like the *Visio de Sancto Martino*, with Saint Martin. Both texts were even clearly influenced by the same late Antique hagiographical narratives concerning this saint, namely, the works of his first hagiographers Sulpicius Severus (c. 363–425) and Gregory of Tours (538–594). It is interesting to note that despite their interwovenness within the same intertextual tradition, they are still clearly distinguished and therefore demonstrate that topic-related stylistics hardly interferes with the author-related differences. The visionary texts under investigation thus betray Guibert’s stylistic influence to such an advanced extent that we could wonder whether we should not entirely attribute these texts to Guibert, instead of arguing for any form of ‘synergetical collaboration’, as was still possible for the portion of the *epistolarium* over which both Hildegard and Guibert labored.

Fig. 11 PCAs including the *Visio de Sancto Martino* and the *Visio ad Guibertum missa* (continued)
9 Conclusions

It is obvious that the experiments reported in this article only touch the tip of the iceberg of the research on Hildegard's complicated authorship, to say nothing of the exciting, broader topic of twelfth-century Latin writing. As stated in our Introduction, individuality and authorship remain complex issues when it comes to medieval literature. Even an authoritative and highly idiosyncratic author like Bernard of Clairvaux is known to have been assisted by a team of collaborators. It is moreover clear that medieval scribes often gradually introduced errors and deviations when successively copying exemplars, thus possibly altering the original authors’ style in the surviving copies of texts. Nevertheless, we hope to have demonstrated that these issues do not need to imply that stylometry, when applied cautiously, cannot yield valid research results in the field of medieval philology.

First we showed that authorial discrimination was possible in the corpus studied. Although samples had to be big enough to yield correct attributions, stylometric methods were generally able to model the overall differences in writing style. This suggests that superficial interference from scribes (or even later editors) can be by-passed to a certain extent, for instance through lemmatization. Interestingly, we obtained satisfying results with a word-level approach, notwithstanding the fact that Latin is a highly inflected language. Although other strategies might increase attribution accuracies in the future, this shows that even in highly inflected...
languages, plenty of stylistic information can already be harvested at the word-level.

In the course of our research, we have also touched on collaborative authorship, an issue that recently has raised considerable interest in stylometry (Reynolds et al., 2012). Our methodology enabled us to discover clear stylistic differences in Hildegard of Bingen’s epistolary work between those letters for which she had relied on the modest assistance of her first collaborator Volmar and the letters that have been compiled and copy-edited by Guibert of Gembloux. Interestingly, the letter samples influenced by the collaboration between Hildegard and Guibert formed an isolated cluster that did not display advanced stylistic similarities to Hildegard’s former epistolary oeuvre, nor to that of Guibert. These results argue in favor of what Pennebaker (2011) has called the Synergy Hypothesis: when two authors are involved in the same texts, the end result need not resemble the writing style of one of the two individually; the result might rather resemble that of a ‘new’, third author. The evidence offered in this particular case study is valuable in this light, but at the same time still too scant to come to a final verdict on this fascinating topic.

Finally, with respect to our initial research question, we hope to have convincingly disputed the authorship of two texts allegedly attributed to Hildegard: the *Visio de Sancto Martino* and the *Visio ad Guibertum missa*. We argued that these visions are stylistically speaking completely in line with the writing style of Guibert de Gembloux, Hildegard’s last secretary. These results offer
quantitative support to suspicions voiced in earlier, traditional philological research: if Guibert is not to be considered their original author altogether, it is clear that he reworked these texts so profoundly that hardly anything of Hildegard’s writing style is still discernible in them. In fact, it is noteworthy that our analyses could not offer any stylistic evidence at all that Hildegard once authored (even a preliminary or simply oral version of) these texts, although this remains of course an interesting historical possibility.

Acknowledgements

We thank the Corpus Christianorum Library & Knowledge Centre of Brepols (Turnhout) and in particular Luc Joqué for generously putting at our disposal the corpora analyzed in this article. Marco Passarotti (Università Cattolica del Sacro Cuore, Milan) generously provided us with the IT-TB, while Helma Dik (University of Chicago) provided the word list from the Perseus Project (Tufts University). We are moreover very grateful for the valuable feedback from Albert Derolez, Wim Verbaal, Antoon Bronselaer, and Guy De Tré. In addition, we thank the anonymous reviewers of the Digital Humanities Conference 2013 for their helpful comments on this research project, as well as the anonymous reviewers of this journal, in particular, for their extensive feedback on the normalization procedures described. Mike Kestemont developed the stylometric methodology for this article. Sara Moens brought in her domain expertise concerning Guibert of Gembloux and medieval epistolography. Jeroen Deploige, who took the initiative for this collaborative research, contributed from his involvement with Hildegard scholarship. All three authors contributed equally to the end result.

Funding

This work was supported by the Research Foundation – Flanders, of which both Sara Moens and Mike Kestemont are fellows, and by the Flemish Hercules Foundation, which finances the project ‘Sources from the Medieval Low Countries (SMLC)’, directed by Jeroen Deploige.

References


M. Kestemont et al.


Collaborative authorship in the twelfth century


Notes

1. Among the letters written with the help of Volmar, we count those in MS Wien, Österreichische Nationalbibliothek, 963 (theol. 348), which offers a copy of a collection compiled by Volmar before 1173 (Van Acker, 1991, p. xxvi), and the limited number of letters that can be found distributed over MS Stuttgart, Württembergische Landesbibliothek, Cod. theol. phil. 4° 253; MS Wien, Österreichische Nationalbibliothek, 881; MS Berlin, Staatsbibliothek Preussischer Kulturbesitz, Cod. theol. lat. fol. 699; MS London, British Library, Cod. Add. 17292; MS Paris, Bibliothèque Nationale, Nouv. Acquis. Lat. 760; MS Trier, Stadtbibliothek, Cod. 771/1350 and MS Kynzvart, Cod. 40. Among the letters compiled and edited under Guibert’s supervision, we count those in the Riesenkodex Wiesbaden, Landesbibliothek, 2 (dating from 1177-1179/1180), that are not also found in MS Wien, Österreichische Nationalbibliothek, 963 (theol. 348) (Van Acker, 1991, p. xxvii), as well as those copied in MS Berlin, Staatsbibliothek Preussischer Kulturbesitz, Cod. lat. 4° 674, which bear traces of Guibert’s editorial assistance (Klaes, 2001, p. xvi). Among the letters contained in the latter group, compiled under Guibert’s supervision, we obviously encounter all Hildegard’s letters addressed to Guibert and the ones that have been written in the years in which he stayed in Rupertsberg.

2. MSS Brussels, Royal Library, 5397–5407 and 5527–5534 (both originating from Gembloux, early thirteenth century) and MS Brussels, Royal Library, 1510–1519 (originating from Sint-Maartensdal near Louvain, fifteenth century).

3. See www.brepolis.net. The critical editions of the works of both Hildegard of Bingen and Guibert of Gembloux are published in several volumes in Brepols’s own Corpus Christianorum series. For the works of Bernardus, the Brepols Library of Latin Texts relies on Leclercq et al. (1957–1977).

compiled shortly after Bernard’s death, as well as letters transmitted elsewhere. Guibert’s letters were edited by Derolez (1988–1989) on the basis of MS Brussels, Royal Library, 5527–5534.  


6. We supplemented this list with three words—plerumque, utrumque, and quicumque—yet did not allow any of these items into the restrictive set of function words we list below. We did not consider other, much less frequent clitics (e.g. –ne (‘if’) or –ve/ue (‘or’)), because it is difficult to automatically detect these using a simple rule-based approach and to distinguish them from e.g. the –ne in deuotione or the –ue in serue.  

7. We have described our approach in a generic way for future reference. It should be noted, however, that there still remains a small number of possible spelling variants in medieval Latin that are hard to deal with but that were not relevant for the present research because we worked with critical editions that have already normalized orthography to a large extent. One can think here of the interchangeability of –mqu– and –nqu– in some words and the problem of single/double consonants (as e.g. in litera and litera). A lesser frequent, yet still important, orthographical variant that we leave unaddressed is (–)exs– versus (–)ex–, because it is difficult to automatically detect it using a rule-based approach. Nevertheless, this variant hardly affects any of the function words to which we have restricted our analyses.  

8. In these training data too, we have substituted all vs for us and all aes/oes for es.  

9. Note that licet, which strictly speaking derives from the impersonal verb licere, is considered a function word because it is primarily used as a subordinating concessive conjunction.  

10. Other errors in the lemmatization displayed in Table 3 are ‘hildergars’, ‘us’, and ‘ta’.  

11. Note that from this point onwards, we will express the size of textual samples in terms of the number of consecutive lemmatized words they contain (a number which, after tokenization, need not be identical to the original number of surface forms in the original texts).  

12. For the sake of conceptual clarity we shall keep Pennebaker’s original terminology, although it should be stressed that our present use of the term ‘Synergy Hypothesis’ is completely unrelated to the concept of ‘Synergetic Linguistics’ in the field of quantitative linguistics (Köhler, 2005).