Introduction to Computation for Literary Criticism

Randa El Khatib
David Wrisley
Welcome to DHSI 2019!

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 DHSl-related information may be found on our website at dhsi.org.

Do check in there first if you need anything that’s not in this coursepak.

To access the DHSI wifi network, simply go into your wireless settings and connect to the “DHSI” network and enter the password “dhsi2019”.

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....
Regional Map of Greater Victoria

Average Frequency
- **Regional Route**: 15–60 minute service with limited stops
- **Frequent Route**: 15 minute or better service, 7am-7pm, Mon-Fri
- **Local Route**: 20–120 minute service

Legend
- Direction of Travel
- Route Name
- Transit Exchange
- Park & Ride Lot (no overnight parking)
- Major Stop

Legend
- Region
- Major Stop (no overnight parking)
- Park & Ride Lot

Regional Route
- Average Frequency
- 15–60 minute service with limited stops

Frequent Route
- Average Frequency
- 15 minute or better service
- 7am-7pm, Mon-Fri

Local Route
- Average Frequency
- 20–120 minute service
DHSI Wi-Fi

Network name: DHSI
Passkey: dhsi2019
The 2019 schedule is just taking shape nicely! A very few things to confirm, add, etc, still but this is the place to be to find out what is happening when / where ...

Psst: Some Suggested Outings

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

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

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

Suggested Outing 2, Butchart Gardens (self-organised)

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

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

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

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

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

And more!

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

9:00 to 4:00

Early Class Meeting: 4. [Foundations] DH For Department Chairs and Deans (David Strong Building C124, Classroom)

Further details are available from instructors in mid May to those registered in the class. Registration materials will be available in the classroom.

3:00 to 5:00

DHSI Registration (MacLaurin Building, Room A100)

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

Monday, 3 June 2019

Your hosts for the week are Alyssa Arbuckle, Ray Siemens, and Jannaya Friggstad Jensen.

7:45 to 8:15

Last-minute Registration (MacLaurin Building, Room A100)
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30 to 10:00</td>
<td>Welcome, Orientation, and Instructor Overview (MacLaurin A144)</td>
</tr>
<tr>
<td></td>
<td>- Welcome to the Territory</td>
</tr>
<tr>
<td></td>
<td>- Welcome to DHSI: Ray Siemens, Alyssa Arbuckle</td>
</tr>
<tr>
<td></td>
<td>- Welcome from UVic: Jonathan Bengtson (University Librarian), Alexandra D'Arcy (Associate Dean Research, Humanities)</td>
</tr>
<tr>
<td>10:15 to Noon</td>
<td>Classes in Session (click for details and locations)</td>
</tr>
<tr>
<td></td>
<td>- 1. [Foundations] Digitisation Fundamentals and their Application (Clearihue A103, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 2. [Foundations] Introduction to Computation for Literary Criticism (Clearihue A102, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 4. [Foundations] DH For Department Chairs and Deans (David Strong Building C124, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 5. [Foundations] Developing a Digital Project (With Omeka) (Clearihue A031, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 8. [Foundations] Fundamentals of Programming/Coding for Human(s)ists (Clearihue A108, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 9. Out-of-the-Box Text Analysis for the Digital Humanities (Human and Social Development A160, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 10. Sound and Digital Humanities (Cornett A120, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 11. Critical Pedagogy and Digital Praxis in the Humanities (Clearihue D132, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 12. Digital Humanities for Japanese Culture: Resources and Methods (McPherson Library A003, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 15. Retro Machines &amp; Media (McPherson Library 128, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 16. Geographical Information Systems in the Digital Humanities (Clearihue A105, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 17. Introduction to IIIF: Sharing, Consuming, and Annotating the World’s Images (Cornett A121, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 18. Web APIs with Python (Human and Social Development A170, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 19. Ethical Data Visualization: Taming Treacherous Data (Cornett A128, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 20. Linked Open Data and the Semantic Web (Cornett A132, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 21. Palpability and Wearable Computing (McPherson Library A025, Classroom)</td>
</tr>
<tr>
<td></td>
<td>- 22. The Frontend: Modern JavaScript &amp; CSS Development (Clearihue A030, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 23. Modelling. Virtual. Realities. A Practical Introduction to Virtual (and Augmented) Reality (Human and Social Development A150, Lab)</td>
</tr>
<tr>
<td></td>
<td>- 25. Information Security for Digital Researchers (David Strong Building C114, Classroom)</td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch break / Unconference Coordination Session (MacLaurin A144)</td>
</tr>
<tr>
<td></td>
<td>(Grab a sandwich and come on down!)</td>
</tr>
<tr>
<td></td>
<td>Discussion topics, scheduling, and room assignments from among all DHSI rooms will be handled at this meeting.</td>
</tr>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
<tr>
<td>4:10 to 5:00</td>
<td>Institute Lecture: Jacqueline Wernimont (Dartmouth C): “Sex and Numbers: Pleasure, Reproduction, and Digital Biopower”</td>
</tr>
<tr>
<td></td>
<td>Chair: Anne Cong-Huyen (U Michigan) (MacLaurin A144)</td>
</tr>
<tr>
<td>5:00 to 6:00</td>
<td>Opening Reception (University Club)</td>
</tr>
</tbody>
</table>

Tuesday, 4 June 2019

9:00 to Noon   Classes in Session

12:15 to 1:15  Lunch break / Unconference

"Mystery" Lunches

1:30 to 4:00   Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 1 (MacLaurin A144)
Wednesday, 5 June 2019

9:00 to Noon

Lunch break / Unconference

"Mystery" Lunches

Presentation: An Introduction to Scholarly Publishing with Manifold (MacLaurin A144)
Lunch included for those who register here

This presentation introduces Manifold Scholarship, a Mellon-funded digital publishing platform developed by the CUNY Graduate Center, The University of Minnesota Press, and Cast Iron Coding. Manifold allows you to create beautiful, dynamic open access projects that can include text, images, video, embedded resources, and social annotation. We will provide an overview of Manifold and demonstrate how faculty, students and staff in the digital humanities can use Manifold to publish open access scholarly works, conduct and participate in peer review, and create custom edited versions of public domain course texts and OER.

1:30 to 4:00

Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 2 (MacLaurin A144)
Chair: Kim O'Donnell (Simon Fraser U)

Catherine Ryu (Michigan State U), "Tone Perfect: Developing a Multimodal Audio Database for Mandarin Chinese as an Open Source"
Kenzie Burchell (U Toronto Scarborough), "Making Responsible Reporting Practices Visible: Comparing newswire coverage of humanitarian crises in Syria"
Jessica Linzel (Brock U), "The Shopkeeper Aristocracy: Mapping Trade Networks in Colonial Niagara"
Kirsten Painter (U Washington), "From Bogatyr to Bread: Digitization & Online Exhibition of Rare Russian Children's Books at the U Washington"
John Barber (Washington State U), "A Mighty Span"

4:15 to 5:15

"Half Way There!" [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! [A great opportunity for an interest group meet-up ....]

Thursday, 6 June 2019

9:00 to Noon

Classes in Session

Lunch break / Unconference

"Mystery" Lunches

[Instructor lunch meeting]

1:30 to 4:00

Classes in Session

DHSI Conference and Colloquium Lightning Talk Session 3 (MacLaurin A144)
Chair: Kim O'Donnell (Simon Fraser U)

Colleen Kolba (U South Florida), "What Comics can Teach our Students about Multimodal Literacy"
Trish Baer (ETCL; U Victoria), "Preserving Digital Legacies: Archived Websites and Digital Discoverability"
Suchismita Dutta (U Miami), "The Importance of Archival Transcription for Genre Building"
Jeffrey Lawler (California State U, Long Beach), "Twining our way through the Past: Video Game Authoring as History Pedagogy"
Friday, 7 June 2019 [DHSI; ADHO Pedagogy SIG Conference Opening]

9:00 to Noon
Classes in Session

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

1:30 to 1:50
Remarks, A Week in Review (MacLaurin A144)

2:00 to 3:00
Joint Institute Lecture (DHSI and ADHO Pedagogy SIG Conference):
Matt Gold (CUNY Graduate Center and Association for Computers and the Humanities): “Thinking Through DH: Proposals for Digital Humanities Pedagogy”
Chair: Diane Jakacki (Bucknell U)
(MacLaurin A144)

Abstract: How do we teach digital humanities, and how should DH be taught? What, indeed, should we teach when we teach DH? This talk will present a proposal for grounding digital humanities pedagogical practice in the research interests of our students and the epistemological foundations of our methods rather than through an approach grounded more central in data and methods.

3:30 to 5:00
Joint Reception: DHSI and ADHO Pedagogy SIG Conference (University Club)
E-Poetry Event (Chris Tanasescu)
Watch this space for details, including how to participate!
DHSI Conference and Colloquium Poster/Demo Session
Pia Russel (U Victoria); Emily Stremel (U Victoria), “British Columbia’s Historical Textbooks Digital Library”
Cody Hennesy (U Minnesota); Rachael Samberg (U California, Berkeley); Stacy Reardon (U California, Berkeley), “Finding the Haystack: Literacies for Accessing and Using Text as Data”
Paula Johanson (ETCL; Independent Scholar), “Proving Seahorses and Juan de Fuca’s Travels in The Curve of Time”
Tara Baillargeon (Marquette U); Elizabeth Wawrzyniak (Marquette U), “FellowsHub: J. R. R. Tolkien Fanzine Portal”
Graham Jensen (U Victoria), “Canadian Modernist Magazines Project”
Caterina Agostini (Rutgers U), “Art at the Time of Syphilis: A First-Person Medical Narrative in Benvenuto Cellini’s Vita”
Lauren Elle DeGaine (ETCL; U Victoria), “Women at the Front: A Digital Exhibit of Victorian Frontispiece Illustrations”
Adam Griggs (Mercer U); Kathryn Wright (Mercer U); Christian Pham (Mercer U); Gail Morton (Mercer U); Stephanie Miranda (Mercer U), “Digitizing Middle Georgia’s History of Slavery”

Saturday, 8 June 2019 [Conference, Colloquium, and Workshop Sessions]

8:00 to 9:00
Conference / Workshop Registration (MacLaurin A100)

The day's events are included with your DHSI registration. If you're not registered in DHSI, you're very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We'll have a nametag waiting for you!

Coffee, Tea, &c?
Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00
DHSI Conference and Colloquium Sessions
ADHO Pedagogy SIG Conference Sessions
Right2Left Workshop Sessions

9:00 to 4:00
All Day DHSI Workshop Session (click for workshop details and free registration for DHSI participants)

9:00 to 9:10
Informal Greetings, Room Set-up (Lobby, outside Hickman 105)

Session 1
DHSI Colloquium and Conference (Hickman 105)
Digital Humanities & Literature, Chair: Kim O'Donnell (Simon Fraser U)
- Youngmin Kim (Dongguk U), “Transdiscursivity in the Convergence of Digital Humanities and World Literature”
- Caroline Winter (U Victoria), “Digitizing Adam Smith's Literary Library”
- Kaitlyn Fralick (U Victoria); Kailey Fukushima (U Victoria); Sarah Karlson (U Victoria), “Victorian Poetry
9:10 to 10:30

**ADHO Pedagogy SIG Conference (Hickman 110)**

Chair: Katherine Faull (Bucknell U)

- Aaron Tucker and Nada Savicevic (Ryerson U), "Write Here, Right Now: An Open Source eTextbook for the Flipped Classroom"
- Heather McAlpine (U Fraser Valley), "Digital Meters: Using Text Encoding to Teach Literature in the Undergraduate Classroom"
- Tiina H. Airaksinen (U Helsinki), "Digital Humanities in Cultural Studies: Creating a MOOC course for University Students and A-Level Students"

**Right2Left Workshop (Hickman 116)**


10:30 to 10:40

**Break**

10:40 to Noon

**Session 2**

DHIS Colloquium and Conference (Hickman 105)

Digital Humanities & Society, Chair: Eleanor Reed (Hastings C)

- Joel Zapata (Southern Methodist U), "Uncovering the Southern Plains' Mexican American Civil Rights Movement"
- Ayo Oseisanwo (U Ibadan), "Online Newspaper Construction of Agitation for the Sovereign State of Biafra in Nigeria"
- Joseph Jones (U British Columbia), "Testbed for an Approach to Distant Reading: Fictions That Represent Vietnam War Resisters in Canada"
- Brendan Mackie (U California, Berkeley), "Visualizing Long-Term Cultural Change: An Example From The Birth of Civil Society"

**ADHO Pedagogy SIG Conference (Hickman 110)**

Chair: Laura Estill (St Francis Xavier U)

- Jane Jackson (Chinese U of Hong Kong), "Interrogating digital spaces for intercultural meaning-making"
- Ryan Ikeda (UC Berkeley), "Disrupting Digital Literacy: Situating Electronic Literature Among Public Education Initiatives"
- Christopher Church, Katherine Hepworth (U Nevada, Reno), "We’re STEAMed! A call for balancing technical instruction and disciplinary content in the digital humanities"
- Chelsea Milbourne (Cal Poly, San Luis Obispo), "Finding the Right Fit between Technology and Class Content: Reflections on Including Web Development in a Digital Storytelling Course"

**Right2Left Workshop (Hickman 116)**

- Edward "Eddie" Surman (Claremont Graduate U), "Qualitative Digital Text Analysis and #Right2Left Languages: A Demonstration of Atlas.ti using the Hebrew Bible"

Noon to 1:10

**Lunch** (We recommend Mystic Market on weekends!)

1:10 to 2:30

**Session 3**

DHIS Colloquium and Conference (Hickman 105)

Digital Humanities & Community, Chair: Claire Carlin (U Victoria)

- Pia Russel (U Victoria); Emily Stremel (U Victoria), "Mentorship and disability: Supporting disabled employees in digital humanities"
- Amy Lueck (Santa Clara U), "Virtually Emplacing Indigenous Memory"
- Md. Shehabul Alam (National U Bangladesh), "Integrating Library Service with Union Information and Service Center: A Joint Initiative towards Digital Bangladesh"
- Veronica Gomez (Instituto de Humanidades y Ciencias Sociales (HuCoSo) - UNL-CONICET), "Latin American E-literature and Location: The Nation Revisited in Electronic Literature Organization (ELO)"

**ADHO Pedagogy SIG Conference (Hickman 110)**

Chair: Chris Tănăsescu (UC Louvain)

- Laura Estill (St Francis Xavier U), "One Assignment, Three Ways: Assessing DH Projects in a Literature Course"
- Felix Bayode Oke, Stella N. Kpolugbo (Anchor U Lagos), "The Multimodal Technique as a Pedagogical Tool in Pelu Awofeso’s White Lagos: A Definitive and Visual Guide to the Eyo Festival"
- Shu Wan (U Iowa), "A digital "historical gaze" of Chinese students in Iowa, 1911-1930"
- Francesca Giannetti (Rutgers U, New Brunswick), "So near while apart: Correspondence Editions as Critical Library Pedagogy and Digital Humanities Methodology"

**Right2Left Workshop (Hickman 116)**

- Najla Jarkas (American U Beirut) and David Joseph Wrisley (NYU Abu Dhabi), "RTL Software Localization and Digital Humanities: the Case Study of Translating Voyant Tools into Arabic"
Session 4

**DHSI Colloquium and Conference (Hickman 105)**
- Ashleigh Cassermer-Stanfield (U Chicago), “Sonifying Hamlet and Reading the Room”

**ADHO Pedagogy SIG Conference (Hickman 110)**
Chair: Aaron Tucker (Ryerson U)
- Youngmin Kim (Dongguk U), “Teaching Digital Humanities and World Literature in Class”
- Alice Fleerackers, Juan Pablo Alperin, Esteban Morales, Remi Kalir (Simon Fraser U, U Colorado Denver), “Online annotations in the classroom: How, why, and what do students learn from annotating course material?”
- Andie Silva (York C and Graduate Center, CUNY), “Keeping it Local: Undergraduate DH as Feminist Practice”

**Right2Left Workshop (Hickman 116)**
- Joanna Byszuk (Institute of Polish Language, Polish Academy of Sciences, Warsaw/Computational Stylistics Group) and Alexey Khismatulin (Institute of Oriental Manuscripts, Russian Academy of Sciences, Saint Petersburg), “Attribution of Authorship for Medieval Persian Quasidas with Styometry”

---

**Sunday, 9 June 2019 [Workshop Sessions]**

**DHSI Registration (MacLaurin Building, Room A100)**

8:00 to 5:00

The day’s events are included with your DHSI registration. If you’re not registered in DHSI, you’re very welcome to join us by registering here as a Conference / Colloquium / Workshop participant. We’ll have a nametag waiting for you!

**Coffee, Tea, &c?**

Looking for some morning coffee or tea, or a small nibble? Options and hours of operation for weekend campus catering are available here. Mystic Market usually opens around 10.00.

9:00 to 4:00

**All Day Workshop Sessions** (click for workshop details and free registration for DHSI participants)
- 55. Introduction to Machine Learning in the Digital Humanities [8-9 June; All day, each day] (David Strong Building C124, Classroom)
- 56. Pedagogy of the Digitally Oppressed: Anti-Colonial DH Methods and Praxis [9 June; All Day] (Hickman 115, Classroom)
- 57. Natural Language Processing and Network Coding Apps for Text & Textual Corpus Analysis in the Humanities [9 June; All Day] (David Strong Building C114, Classroom)

9:00 to Noon

**AM Workshop Sessions** (click for workshop details and free registration for DHSI participants)
- 59. 3D Visualization for the Humanities [9 June; AM] (Cornett A229, Classroom)
- 60. It’s All Relational: AbTeC’s Indigenous Video Game Workshops as Storytelling Praxis [9 June; AM] (Cornett A121, Classroom)
- 61. Spatial DH: De-Colonizing Cultural Territories Online [9 June; AM] (Clearihue D130, Classroom)
- 63. Creating a CV for Digital Humanities Makers [9 June; AM] (David Strong Building C108, Classroom)

Lunch (We recommend Mystic Market on weekends!)

9:00 to Noon

**PM Workshop Sessions** (click for workshop details and free registration for DHSI participants)
- 65. Indigenous Futures in the Classroom and Beyond [9 June; PM] (Cornett A121, Classroom)
- 66. DHSI Knits: History of Textiles and Technology [9 June; PM] (Fine Arts 109, Classroom)
- 68. Linked Open Datafication for Humanities Scholars [9 June; PM] (McPherson Library A003, Classroom)
- 69. Stylo - WYSIWYM Text Editor for Humanities Scholars [9 June; PM] (McPherson Library A025, Classroom)

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, 10 June 2019
Your hosts for the week are Ray Siemens and Jannaya Friggstad Jensen.

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

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

10:15 to Noon  
Classes in Session (click for details and locations)
- 30. [Foundations] Databases for Digital Humanists ([McPherson Library 210, Classroom](#))
- 32. [Foundations] Databases for Digital Humanists ([McPherson Library 210, Classroom](#))
- 33. Digital Storytelling ([Cornett A120, Classroom](#))
- 34. Text Mapping as Modelling ([Cleanihue D131, Classroom](#))
- 35. Stylometry with R: Computer-Assisted Analysis of Literary Texts ([Cleanihue A102, Lab](#))
- 36. Open Access and Open Social Scholarship ([Cleanihue D130, Classroom](#))
- 37. Digital Games as Tools for Scholarly Research, Communication and Pedagogy ([Cornett A229, Classroom](#))
- 38. Queer Digital Humanities ([David Strong Building C114, Classroom](#))
- 40. Introduction to Electronic Literature in DH: Research and Practice ([Cornett A128, Classroom](#))
- 41. Surveillance and the Critical Digital Humanities ([David Strong Building C108, Classroom](#))
- 42. Text Analysis with Python and the Natural Language ToolKit ([Cleanihue A103, Lab](#))
- 43. Developing Humanities Multimedia ([Human and Social Development A150, Lab](#))
- 44. Digital Humanities Pedagogy: Integration in the Curriculum ([Cornett A121, Classroom](#))
- 45. Accessibility & Digital Environments ([Priestly Law Library 265, Classroom](#))
- 46. Agile Project Management ([Cornett A132, Classroom/Lab](#))
- 47. XPath for Processing XML and Managing Projects ([Cleanihue A105, Lab](#))
- 48. Endings: How to End (and Archive) your Digital Project ([Priestly Law Library 192, Classroom](#))
- 49. Text Processing - Techniques & Traditions ([McPherson Library A025, Classroom](#))
- 50. Introduction to Humanities Data Analysis & Visualization in R ([Human and Social Development A160, Lab](#))
- 51. Introduction to Network Analysis in the Digital Humanities ([Cleanihue D132, Classroom](#))

12:15 to 1:15  
Lunch break / Unconference Coordination Session ([MacLaurin A144](#))
(Grab a sandwich and come on down!)

“Mystery” Lunches

1:30 to 4:00  
Classes in Session

4:10 to 5:00  
Institute Lecture: Angel David Nieves (San Diego State U): “3D Mapping and Forensic Traces of Testimony: Documenting Apartheid-Era Crimes Through the Digital Humanities”
Chair: Constante Crompton (U Ottawa)
([MacLaurin A144](#))

Abstract: In 1989 the killing of a queer, 14-year-old youth in Winnie Mandela’s house named Stompie Seipei (an event that few in South Africa are willing to recall, let alone discuss, in any detail) -- is perhaps one of the most glaring examples where the queer and activist community was suppressed or erased from anti-apartheid/liberation histories. Digital humanities may actually help both reconstruct and recover a history that is still very early in the telling, despite what is commonly believed about the liberation struggle and the contributions of queer activists in the dismantling of apartheid. Perhaps it could explain why a youth such as Seipei was killed -- or at the very least, provide a more complex and messy narrative that permits one to know more how the history of queer anti-apartheid activists was suppressed. This talk outlines a methodology for “messy thinking and writing” in the digital humanities that -- through a queer and feminist intersectional framework -- permits a more complex layering of oral histories and 3D historical reconstructions.

5:00 to 6:00  
Reception ([University Club](#))

Tuesday, 11 June 2019

9:00 to Noon  
Classes in Session

12:15 to 1:15  
Lunch break / Unconference

“Mystery” Lunches
<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30 to 4:00</td>
<td>Classes in Session</td>
</tr>
</tbody>
</table>
| 4:15 to 5:15| DHSI Conference and Colloquium Lightning Talk Session 4 (MacLaurin A144) Chair: Lindsey Seatter (U Victoria) -
|             | • Ashley Caranto Morford (U Toronto); Kush Patel (U Michigan); Arun Jacob (McMaster U), “#OurDHIs anti-colonial: Questions and challenges in dismantling colonial influences in digital humanities pedagogy” -
|             | • Julia King (U Bergen), “Developing Network Visualizations of Syon Abbey's Books, 1415-1539” -
|             | • Luis Meneses (ETCL; U Victoria), “Identifying Changes in the Political Environment in Ecuador” -
|             | • Alicia Brown (Texas Christian U), “Digital Cartography of the Ancient World” -
|             | • Laura Horak (Carleton U), “Building the Transgender Media Portal” -
|             | • Andrew Boyles Peterson (Michigan State U), “Last Mile Tracking: Implications of Rental Scooter Surveillance” -
| 6:00 to 8:00| DHSI Newcomer's Gathering (Grad House Restaurant, Graduate Student Centre) -
|             | Come down, buy meal and a beverage, and make some new friends!        |
| 9:00 to Noon| Classes in Session                                                   |
| 12:15 to 1:15| Lunch break / Unconference                                           |
|             | "Mystery" Lunches                                                    |
| 12:15 to 1:15| Presentation: An Introduction Jupyter Notebooks for Researchers (MacLaurin A144) -
|             | This presentation introduces Jupyter Notebooks for researchers, via a partnership between Compute Canada and the Pacific Institute for the Mathematical Sciences (PIMS) including a large number of Canadian institutions. Read more here. Presenting is James Colliander, PIMS Director and team. |
| 1:30 to 4:00| Classes in Session                                                   |
| 4:15 to 5:15| DHSI Conference and Colloquium Lightning Talk Session 5 (MacLaurin A144) Chair: Lindsey Seatter (U Victoria) -
|             | • Ashleigh Cassemere-Stanfield (U Chicago), “Critical Editions for Digital Analysis and Research Project (CEDAR): Shakespeare Digital Variorum” -
|             | • Calin Murgu (New College of Florida), “Putting local metadata to strategic use: A Dashboard for visualizing 60 years of theses metadata” -
|             | • Jason Lajoie (U Waterloo), “Queer Critical Making and the Logic of Control” -
|             | • John Barber (Washington State U), “Zambezi River Bridge” -
|             | • Kent Emerson (U Wisconsin-Madison), “Digital Mappa and the George Moses Horton Project” -
| 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! [A great opportunity for an interest group meet-up ....] |
| 9:00 to Noon| Classes in Session                                                   |
| 12:15 to 1:15| Lunch break / Unconference                                           |
|             | "Mystery" Lunches                                                    |
| 12:15 to 1:15| [Instructor lunch meeting]                                            |
| 1:30 to 4:00| Classes in Session                                                   |
| 4:10 to 5:00| Institute Lecture: Karina van Dalen-Oskam (Huygens Institute and U Amsterdam; Alliance of Digital Humanities Organizations): "The Riddle of Literary Quality: Some Answers" Chair: Aaron Mauro (Penn State, Behrend C) (MacLaurin A144) -
|             | Abstract: What is literature, and can you measure it? That is the key question of the project The Riddle of Literary Quality. "The Riddle" is a research project of the Huygens Institute for the History of the Netherlands (Amsterdam) in collaboration with the Fryeke Akademy (Leeuwarden) and the Institute for Logic, Language and Computation (University of Amsterdam). The Riddle combines computational analysis of writing style with the results of a large online survey of readers, completed by almost 14,000 participants. In my talk, I will go into
some of the main results of the project.

## Friday, 14 June 2019

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00 to Noon</td>
<td>Classes in Session</td>
<td></td>
</tr>
<tr>
<td>12:15 to 1:15</td>
<td>Lunch Reception / Course E-Exhibits  (MacLaurin A100)</td>
<td></td>
</tr>
<tr>
<td>1:30 to 2:00</td>
<td>Closing, DHSI in Review  (MacLaurin A144)</td>
<td></td>
</tr>
</tbody>
</table>

Contact info:  
institut@uvic.ca  P: 250-472-5401  F: 250-472-5681
Introduction to Computation for Literary Criticism
Digital Humanities Summer Institute
Summer 2019

The syllabus for this course is available on GitHub: https://github.com/djwrisley/dhsi2019. Bookmark this link so that you can find it easily, and so that you can consult updates to the syllabus, readings and project lists. During this course it will be a pinned repository. After the summer, it may be unpinned.

Course description:

This course demystifies, and offers a survey of, the computational tools and techniques being used for literary criticism. Aimed at novice and DH-curious scholars and practitioners, participants gain familiarity with fundamental concepts and methods so that they can better appreciate the potential of computer-assisted critical techniques. Classes are divided between discussions of key theoretical considerations and practical instruction in a selection of tools. Participants are exposed to macro-analytical techniques like most frequent word analysis, collocation, stylometry, topic modeling, digital mapping, and network analysis, gaining experience with environments like Voyant, R, Carto, Palladio, and Gephi. The course also details best practices relating to the preparation and management of digital corpora. Having completed this course, participants will have a better understanding of how computational methods can be used to produce quantitative data for use in the support of literary criticism. More advanced expertise can subsequently be developed at any one of a number of DHSI offerings dedicated to particular methods.

This Course Pack:

In this course pack, will you find scheduling details, as well as some suggested readings (and links to readings). Take a look at these readings in advance of the gathering, as they will be help frame our discussions. Instructions for installing some tools for your laptop will be send in advance of the course, in case you choose to use your own machine. We will curate a list of digital projects to examine during the course on the above-mentioned GitHub page.

Instructor Bios:

Randa El Khatib (U Victoria; returning; @randaelka) is the Assistant Director (Open Knowledge Initiatives) of the Electronic Textual Cultures Labs (ETCL) and a doctoral candidate in English at the University of Victoria. Randa’s research focuses on the representation of space in plays and epic poetry of the English Renaissance. At the same time, she focuses on software prototyping in the digital humanities and across the disciplines, with a particular focus on geospatial methodologies. She’s also interested in community mapping in its many manifestations.
David Joseph Wrisley (NYU Abu Dhabi; @DJWrisley) is associate professor of Digital Humanities at NYU Abu Dhabi. His work in the pre-modern period uses mapping, networks and visualization to model textual mobility and variance, as well as knowledge transfer and intercultural contact. He is interested in the social creation of open corpora and tools for under-resourced medieval languages and building communities of practice for digital medieval studies. His research also addresses the emergence of digital humanities in non-Western, non-English cultures, in particular the contemporary MENA setting, with a particular focus on spatial humanities.

Useful details:

1. We will be in a computer lab that will have all the necessary software installed, so you don’t need to bring a laptop (unless you want to).
2. While many of the tools/methods that we examine can be applied to a range of languages, they are more challenging to use in non-English or non-European languages. It is not impossible to adapt them, but the process can take time and know-how.
3. Bring some data with you! The tools we will work with use either structured data (like Excel spreadsheets or Google sheets) or plain text (.txt format, rather than Word docs or pdfs). If you have lists of persons who are relevant to your research, or lists of places related to your research or teaching interests, bring them along. Look in Project Gutenberg (https://www.gutenberg.org/), the Internet Archive (archive.org) or HathiTrust (https://www.hathitrust.org/, if you have access) before coming to DHSI to see if you can find books of interest to you that you would like to analyze more deeply using some computational methods. Make a list of them and bring it to the workshop! Choosing texts about topics that you know something about will enrich your experience, since familiarity with a “dataset” allows you to draw more relevant conclusions via computer-assisted interpretation.
4. While it is helpful to have data that you know well, if you don’t have any particular data, that is ok too. We will have sets of texts and small structured datasets ready to use for the various assignments (in many cases these have been created by our students who have built them with explicit open licenses for you to reuse and build upon.)
5. The schedule provides an outline of topics to be covered. As we progress through the materials, we may deviate from this depending on how quickly or slowly we are moving as a group. During practical sessions, feel free to return to a previous method if it is more pertinent to your interests.
6. Trial and error. The world of computational tools for humanities research is gaining momentum, but it is still a relatively small world dedicated to the values of open source. Many of the tools we will be using are created by people who devoted a lot of their
time—for no real reward—to the production of tools for our benefit. Compared to the seamless commercial products we use in the office, there will be times when the tools are confusing, give ambiguous errors, do not agree with your system’s unique configuration, or simply do not work. This is the nature of the beast.

We are aiming for an inclusive learning environment and to make our workshops as accessible as possible. In order to make specific accommodations, please let us know specific concerns you may have. Please feel free to reach out khatib@uvic.ca or djw12@nyu.edu.

**Schedule & Readings**

**Day 1:**

10.15 - 12.00

Participants will be invited to introduce themselves to their new cohort, as well as a give a brief outline of their motivations for taking the course, and how it might apply to their future research/practice. Participants will explore the affordances and limitations of computer-assisted approaches to literary criticism, so that they fully appreciate the benefits and drawbacks of the techniques they will be exposed to throughout the week. A number of best practices for the acquisition and preparation of digital corpora will be outlined. In particular, strategies for designing, gathering, cleaning, and managing suitable datasets will be discussed.

1.30 - 4.00

A selection of features from Voyant, a powerful and intuitive tool developed by Stéfan Sinclair & Geoffrey Rockwell, will be demonstrated. Any of the features utilised will be explained in simple terms, so that participants understand how it is that Voyant is producing its results.

**Readings**

Day 2:

9.00 - 12.00
Participants will explore the concept of a "network" and how it might be analyzed using computer-assisted techniques. The theoretical principles behind popular network analysis tools like Palladio will be explained, before participants are shown how to structure their data and analyze it using such software.

1:30 - 4:00
Participants have a hands on session with either text analysis or networks applied to the literary corpora of their choice. If any participants are interested in more advanced applications we can turn to R packages, NodeGoat, Gephi, Cytoscape, etc.

Readings

Day 3:
Participants will be introduced to the field of geospatial humanities, its key concepts and methodologies. They will learn how to extract, geocode, and visualize data in distinguishable ways, as well as to identify what platforms and tools best suit their mapping needs.


Day 4:
Participants will be introduced to the field of Digital Literary Studies commonly referred to as "stylometry," which entails the analysis and classification of literary texts based on authorial signatures.

Readings

Day 5:

9.00 - 12.00
The final morning is used to reinforce a number of key topics from throughout the week, particularly the interpretation of results produced through computational means. Participants will look at examples of DH research and analyze the validity of their scope and methodology, and the extent to which data-driven techniques supported convincing interpretations and robust critical arguments.
If the question, “What do we want the university of the digital future to be?” seems like an impossibly ambitious frame of reference for the question of what we want literary studies in the digital future to be, let’s pause to locate the rhetoric of that question more specifically. Posed by Alan Liu in a keynote lecture for the Texas Institute for Literary and Textual Studies (TILTS) in March 2011, the question was at the top of a branching tree of increasingly precise questions: “What is the ecology of academic, governmental, philanthropic, for-profit, corporate, and other stakeholders in providing digital education?” (31:17); “Will the students of digital education be economic individuals or citizens?” (32:45); “What will be the labor model of the university of the digital future?” (33:34); “What will be the intellectual property model of the university of the digital future?” (34:13). This lecture—delivered in the shadow of an acute economic crisis in higher education—proposed that we remember and reexamine how, as a condition of possibility for humanistic study, the university needs to exist as an economic and ideological set of commitments, and it further proposed that we take seriously the profound changes that might be necessary in those commitments if the university is to continue to exist in some form in a future that is crucially conditioned both by the digital and by the new economy of scarcity. Literary and textual studies, in other words—and digital literary studies as a matter of course—would take place within an institutional and intellectual context that is undergoing radical change, under conditions of extreme competitive pressure. As Liu noted, quoting from Forbes magazine, “The Internet is about to do to America’s universities and colleges what it’s done to media and entertainment—profoundly upend them. And improve them” (25:40).

During that same period, the newly founded 4Humanities site asserted a close linkage between the crisis in the humanities at large and the intervention that the
specifically digital humanities might make in that crisis. The site’s mission statement cites the decline in public and private support for the humanities and warns that the resulting cuts in funding place at risk all that the humanities contribute to our understanding of “the wisdom of the past, awareness of other cultures in the present, and imagination of innovative and fair futures.” It proposes that the digital humanities may prove an important ally to help the traditional humanities “communicate with, and adapt to, contemporary society” (“Mission”).

This alliance marks a crucial shift. The digital humanities has been animated, in its deepest intellectual roots, by the language of change. The “revolution” envisioned by the early theorists of hypertext and electronic modes of authorship suggested a radical restructuring of textuality, authorship, and readership and a significant democratization of the institutions of publication, with potentially dramatic political consequences (see, e.g., Lanham). But in more recent years the language of revolution has not retained its critical bite or its pervasiveness. In its place we can see a softer kind of incremental progressive vision in, for instance, the grant-proposal rhetoric that promises to “broaden access,” improve educational outcomes, and create “new ways of thinking” without upsetting fundamental institutional structures. The language of “innovation” situates these changes in a specifically technological frame of reference, emphasizing the production of “newness” as an ipso facto value. In the shift from revolution to innovation we can see a subtle concomitant shift in the implied framework of responsibility created by the vision of change. In calling for a revolution—in describing what such a revolution could entail—theorists and advocates like Jay David Bolter stipulate a set of responsibilities that are pervasively social: the responsibility of individual authors and readers to act as citizens of a newly democratized “writing space,” the responsibility of publishers and institutions to accommodate such spaces through changes in academic policy, the responsibility of professional communities to take seriously the consequences of such changes for the power and authority structures in such communities. The language of innovation, by contrast, implies a much more narrowly economic set of relations in which responsibility has been replaced by opportunity as a model of interaction. Innovation posits a competitive marketplace of ideas, in which newer (better) tools and ideas will be adopted because of the improvement they offer, following considerations of cost and benefit.

However, the alliance proposed by 4Humanities—and the questions posed in Liu’s TILTS keynote—urges us to recognize that the digital humanities must engage with change not (or not only) as a progressive narrative of steady, technology-driven improvement, but as a more agonistic and uncertain intervention in, or adaptation to, a process of change that is being driven by institutional and economic forces. Turning our attention in this way to changes in human practice—considered less as efficiencies...
or opportunities arising from innovation than as adaptations to complex reconfigurations of the sociotechnical landscape—shows us a set of responsibilities and consequences that are not fully represented in the facile parlance of innovation, or even in the more consequential language of revolution (now somewhat denatured through repeated use). Several such shifts in obligation have been identified in the essays in this volume. For individuals, important responsibilities arise from the changes to scholarly editing that Susan Schreibman describes in her contribution. Increased transparency of process requires readers to take seriously the significance of editorial processes in mediating the texts they consume, and the relocation and dispersal of editorial authority requires readers to play an active and informed role in that mediation. Similarly, the development of editions that perform a set of textual possibilities (as Schreibman describes Clement’s *Versioning Machine* edition of *In Transition: Selected Poems by the Baroness Elsa von Freytag-Loringhoven*) rather than produce a stable editorial product calls for a reader whose use of textual resources can accommodate this play as well. If the traditional edition serves as a grounding so that literary study can proceed without having to first establish its own textual basis, the performative edition asks the literary scholar to work within a much more complex textual universe, as well as to forego certain kinds of interpretive closure.

These examples from digital editing are paralleled by others from textual analysis and data modeling; the analytic power we gain from algorithmic processes on the one hand, and from carefully modeled data on the other, arises from the ways in which those algorithms and models represent (more or less successfully) a strong understanding of the text under examination. The introduction of this layer of formally expressed knowledge into the scholarly ecology creates a burden of responsibility to understand how that layer works and what it is saying, or at least to take its existence seriously. William Kretzschmar’s essay in this collection offers a case in point, illustrating the levels of decision making and complex mathematics undergirding the analysis that in turn permits assessments of things like the randomness or clustering of distribution of some specific geographically distributed feature such as linguistic idioms. As he notes,

The making of any model is a deliberate act of the maker, in part a reflection of the maker’s theoretical foundations and assumptions about what is represented. The more explicitly these ideas are formulated and made known, the more usable the model will be for others besides the maker—and vice versa.

The modeling process described here is not simply an articulation of method but also an act of engagement with users of the data, who are expected to work with the model in ways that echo the kinds of performativity noted above. The reciprocity of
that engagement—the ways in which the user’s grasp of the model through reuse of the data constitutes both an acknowledgment of the model’s value and a potential critique of it—puts the modeler and the user into a much more active relationship, one that is also potentially more fraught and risky, as Kretzschmar’s “and vice versa” suggests. In a similar way, Charles Cooney, Glenn Roe, and Mark Olsen’s discussion of text mining and textbases makes clear that the text-mining techniques and algorithms described all incorporate expert decision making into their design and application; these are not systems that can operate blindly. If the quality and appropriateness of these systems has a bearing on the scholarly outcomes they yield, then it follows that as consumers of that scholarship we need to be able and willing to understand what is at stake.

This push is happening in the same academic culture that is seeking to find ways of building and engaging with larger aggregations of data: the “what do you do with a million books?” challenge that has motivated tremendous rhetorical and economic resources in the years since Gregory Crane’s article of that title. As above, the discussions of methodology that have accompanied the “big data” turn in digital humanities have stressed the comparative insignificance of any individual item in the research landscape; in his introduction to Graphs, Maps, Trees, Franco Moretti observes of literary study that “a field this large cannot be understood by stitching together separate bits of knowledge about individual cases, because it isn’t a sum of individual cases: it’s a collective system, that should be grasped as such, as a whole” (4). The individual text recedes here to take its place as a point in one of his book’s eponymous visualizations, through which we can see “a field . . . a collective system” that by implication carries more cultural and intellectual weight. Clearly in Moretti the idea of subordination does not carry the same apparent moral significance as it does in the case of human students. Nonetheless, with the earlier example in mind, our reading of Moretti should be alert to the many potential valences of the relation between individuals and systems: for instance, a “collective” logic through which the meaning of individual cases is most effectively realized or a “system” logic of industrial management in which individual distinctiveness and locality is set aside because it cannot be thought.

If the individual in this ecology is not the text but the scholar, we see an interesting shift of perspective. To return for a moment to the classroom at scale of distance education: when situated in that virtual classroom space, the individual student seems from the perspective of the educational vector (whose intellectual hub is traditionally imagined as the teacher) to be infinitesimally small. But if we reverse the perspective, as if reversing the telescope end for end, the student who was situated in the vast periphery of a subordinatively large number of fellow students is now seen as the hub of an enablingly large number of educational vectors: the educational abundance of the Web. Similarly, Moretti’s subordination of the individual text to the
“collective system” situates the scholar—the “we” who “grasp” that system—in what is by implication a much more empowering geometry: not at one end of a single line of close reading that attaches us to a single text but at the center of an information nebula. But for that geometry to be empowering, our transactions with that nebula must successfully convert its informational scale into something we can think, and the tools and modeling tactics we use to accomplish that conversion must be understood as tools for transacting a vast disproportion, for effecting a gigantic informational change of state. Digital literary study must thus consider, as a central problem, the empowerments and disempowerments contingent on its use of tools, not because they are tools, but rather because of the questions they raise about how we are situated in relation to our objects and methods of study. The human scholar of literary studies must be present in the inquiry at its end points—as the initiator of questions and consumer of answers—and also inside the process, inside the tools, as they mediate between us and the field we are seeking to grasp.

The essays in this volume register this disproportion, this leverage, and engage with it in complex ways. In exploring tools and techniques that offer insight over large bodies of data—GIS, visualization, text mining—they chart in detail the intellectual gains to be had and suggest that an important strand of the research agenda for digital literary studies is to flesh out more fully the potential of such tools through detailed application. But their concluding reflections offer a repeating dialectical figure in which the individual mind is positioned in relation to these systems. In this volume, David Hoover observes, “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.” Schreibman explicitly names a “dialectic between what we might consider the more traditional and intuitive bases of literary interpretation” and the “disambiguating premise of stylometrics, attribution studies, and other statistical methodologies common to computational and algorithmic processing” from which emerge “new forms of analysis, meaning, and insights.” Cooney, Roe, and Olsen also speak of a “dialectic between technological progress and critical inquiry.” Other essays frame the relation more cautiously, opening up an explicit space of value to be occupied by the human scholar. As Stéfan Sinclair, Stan Ruecker, and Milena Radzikowska put it,

It is important for the scholar to know enough about the visualization tools to understand that the interpretive work is being guided and biased by the data and software. Failing that, we need to have methodologies that are sufficiently well tested and understood for scholars to be able to use the tools with confidence. The question remains whether humanistic inquiry lends itself well to well-trodden methodologies when originality and idiosyncrasy are the norm.
And Tanya Clement concludes her essay by invoking the “messiness and doubt” that make “the complexities of concepts like race, gender, class, and culture most immediately relevant.” In this language of “dialectic” and “ambiguity,” I suggest, we may see a second important strand of the research agenda, centering on the nature of the “literary” as another way of asking what we mean by “the human.”

The science of statistics arose as a way of managing scale by distancing the perceiver from the complexity of individual objects (for instance, individual human beings), creating surrogates through which we can know the world as information to be managed. These surrogates are data points, uniform formal packages whose properties are determined by the needs of the knower rather than the exigencies of the object of scrutiny. There is a brutality in this process that exists quite apart from any specific political motivations; the logic of surrogacy turns us away from the individual human narrative through which we could experience pathos, causal explanation, the quiddity of human life. It puts us in possession of a structural understanding, but it thereby also relocates our minds into that structural space so that the appropriate objects of knowledge are trends, relations, aggregations, species, systems.

When does the individual matter? Literary study has in the past made a practice of dealing with individual narratives: the genesis of a novel, the motivations of an author in relation to a specific work, the semantic ecology of a poem. It has also sought to generalize from these narratives to produce an understanding of history, of genre, of regionality. Digital literary study likewise seeks to move from explanatory narratives that apply to individuals to explanatory narratives that apply to populations. This is a move that seeks to relocate the scene of our knowledge, so that by knowing about more things (for instance, more novels) and by grounding our statements in more data, we thereby in virtue of that data appear to know things about culture as a whole. We have, in tools like text mining and visualization, an unparalleled capacity to gain a very large-scale view of culture—of popular fiction, of linguistic variation, and so forth—as a set of observable phenomena that emerge from the aggregation of data points about individual instances. But what has not been clearly articulated is the role—if any—that an individual instance plays in validating or shaping those larger conclusions.

I offer two examples by way of illustration. The first is the tolerance for error in large digital humanities data sets. It is a commonplace of digitization practice that very high rates of accuracy (in transcription, in optical character recognition [OCR]) are unachievable (because of cost) and, in large-scale digitization efforts, unnecessary. Because the statistical techniques underlying text mining and text analysis operate on aggregations, individual textual errors typically have a statistically insignificant impact on the outcome of the analysis; similarly, when searching a large text corpus for
instances of a given word, the user outcome will be only infinitesimally affected by the omission of a single search hit resulting from a transcription or OCR error. However, interesting cases are sometimes cited in which a pattern of error (such as the mistranscription of long s as f) is substantial enough to skew the results. The threshold of significance for error in the aggregate is thus one crucial point; Geoffrey Nunberg’s complaints about the quality of the metadata in Google Books constituted a claim about the basic usability of that data because of the pervasiveness of error. But another crucial point has to do with edge cases and what we can learn from large-scale collections: while the phenomena that are statistically predominant (and comparatively unaffected by small error rates) are naturally of great interest, we may also want to be able to study rarer phenomena in such large collections, and these will be much more subject to interference or obscuration by even small errors. Methods of analysis that look at infrequent rather than frequent words will also need to take the impact of individual errors into account.

¶

17 More interesting than error, though, is the question of how we understand the significance of individual instances in constituting the overall phenomenon being studied. If all words in a text are not equally relevant carriers of meaning—for instance, words occurring within paratextual features such as running heads, indexes, and advertisements or within nonauthorial features such as editorial notes and quoted materials might clearly be excluded from certain kinds of textual analysis—then as our analysis becomes more fine-grained we may find that we need a clearer account of which textual instances are actually contributing to the truth-value of our large-scale observations.

¶

18 In the place of the qualitative divide between the individually irrelevant instance and the aggregation (in which the instance is treated almost like a mathematical infinitesimal), we will need a geometry in which the instance and the aggregation both appear. If we think of the statistical aggregation—the large corpus, the visualization, the database—as a coordinate plane populated by data points, each of which carries its tiny payload of information (metadata, word frequency, demographics, and so forth), then what I am concerned with here is reintroducing a z-axis: in effect, a vector of connection between each data point and its source in the world, whether that is a text or an artwork or a human being or a linguistic transaction. A recent example of traversal of this z-axis appears in an article by Daniel Cohen and Fred Gibbs entitled “A Conversation with Data: Prospecting Victorian Words and Ideas,” in which the authors propose an approach that moves between “distant” analysis and a closer look at individual instances to consider things like the structural location and collocation of specific words. Although the examples they give are still more weighted toward the “distant” side of things, the explicitness of their point that large-scale analysis is complementary to detailed exegesis is important. Put this way, the point may seem
banal, but the fact that it seemed to them necessary and useful to argue for it suggests that, at least in the current cultural moment, there is a presumption against the need for the “close” view, the z-axis.

§ 19 There may be good methodological reasons for insisting on this traversal; for one thing, as we’ve seen, the accuracy of the connection between the plane of the aggregation and the individual sources for its data can’t be taken for granted (and indeed there is a great deal more to be said and learned about methods of ascertaining that accuracy and about whether and how it matters). For another, if our interest in literary study still concerns individual works of literature (however we may define that term), at some point we need to turn our interpretive attention back to these, following whatever insights our look at the broader field may have yielded. But there are also reasons I would like to characterize as moral: in training ourselves and our students to understand the relation between individual cases and broader trends, we train ourselves to treat that same structure responsibly in other arenas—in other words, to understand that trends and populations are consequential narratives of mastery to which we may not be entitled. An earnest contribution to the 4Humanities site on 23 August 2012 makes this point especially vivid: writing about measures taken by the Australian government to prevent asylum seekers from crossing the Indian Ocean, Debjani Ganguly makes clear that part of the problem is precisely the way these refugees are perceptible to and manageable by the observing world: as inputs to a bureaucratic process (the “interminable wait in a queue in Malaysia to get legally processed”), as “a statistical abstraction,” as a mass too large to be absorbed by the countries where they seek asylum. Ganguly proposes that the role of the humanities is to “advance where policy retreats” and, most potently (and pedagogically), to counter these abstractions with a minutely realized view of the human source: “We preserve records of previous habitations. We understand and assemble pasts, presents and futures anew. We imagine spaces and topographies before and after they acquire materiality. We discover facts, affects, metaphors and images.”

§ 20 Can digital literary studies provide a model for how humanists might work at scale without losing sight of why our study matters? To complement this question, here is another: can digital literary studies provide a compelling and distinctive account of interpretation—of how literary scholars might work digitally “close up” in ways that take advantage of the detailed modeling that digital representation affords? How can digital methods help us reexamine the representativeness and the distinctiveness of the individual text in the context of the vast cultural landscape these methods help us grasp?

Notes
1. I here give the minutes and seconds where each quotation can be located in the online video.

2. The site describes itself as “a platform and resource for advocacy of the humanities, drawing on the technologies, new-media expertise, and ideas of the international digital humanities community.” It further notes that “[t]he humanities are in trouble today, and digital methods have an important role to play in effectively showing the public why the humanities need to be part of any vision of a future society.”

Works Cited


DOI: 10.1632/lsda.2013.9
This paper is about data in the humanities. Most of my colleagues in literary and cultural studies would not necessarily speak of their objects of study as “data.” If you ask them what it is they are studying, they would rather speak of books, paintings and movies; of drama and crime fiction, of still lives and action painting; of German expressionist movies and romantic comedy. They would mention Denis Diderot or Toni Morrison, Chardin or Jackson Pollock, Fritz Lang or Diane Keaton. Maybe they would talk about what they are studying as texts, images, and sounds. But rarely would they consider their objects of study to be “data.” However, in the humanities just as in other areas of research, we are increasingly dealing with “data.” With digitization efforts in the private and public sectors going on around the world, more and more data relevant to our fields of study exists, and, if the data has been licensed appropriately, it is available for research. The digital humanities aim to raise to the challenge and realize the potential of this data for humanistic inquiry. As Christine Borgman has shown in her book on *Scholarship in the Digital Age*, this is as much a theoretical, methodological and social issue as it is a technical issue.

Indeed, the existence of all this data raises a host of questions, some of which I would like to address here. For example:
What is the relation between the data we have and our objects of study? – Does data replace books, paintings and movies? In what way can data be said to be representations of them?

What difference does it make to analyze the digital representation or version of a novel or a painting instead of the printed book, the manuscript, or the original painting?

What types of data are there in the humanities, and what difference does it make? – I will argue that one can distinguish two types of data, “big” data and “smart” data. What, then, does it mean to deal with big data, or smart data, in the humanities?

What new ways of dealing with data do we need to adopt in the humanities? – How is big data and smart data being dealt with in the process of scholarly knowledge generation, that is when data is being created, enriched, analyzed and interpreted?

1. What is data (in the humanities)?

As a starting point, it is useful to define what we mean by “data” generally and in the context of research in the humanities. First of all, let’s remember how data is generally defined. Information scientist Luciano Floridi defines data at its most basic level as the absence of uniformity, whether in the real world or in some symbolic system.[4] Only once such data have some recognizable structure and are given some meaning, can they be considered information. Floridi’s very general definition of data also shows why data can be represented in many different formats and on many different supports. Digital data is special in that it is discrete rather than continuous, and is usually represented, at its most fundamental level, in the form of a binary notation involving just two symbols, 0 and 1. On a higher level, digital data are usually represented and processed in data structures that can be linear (for example arrays and matrices, like lists and tables in a data sheet), hierarchical (with a tree-like structure in which items have parent-child or sibling relations with each other, as in an XML file) or multi-relational (with each data item being a node in an interconnected network of nodes, as in graph-based databases).[5]

Some additional distinctions are important. For instance, there is structured and unstructured data as well as semi-structured data. Structured data is typically held in
a database in which all key/value pairs have identifiers and clear relations and which follow an explicit data model. Plain text is a typical example of unstructured data, in which the boundaries of individual items, the relations between items, and the meaning of items, are mostly implicit. Data held in XML files is an example of semi-structured data, which can be more or less strictly constrained by the absence or presence of a more or less precise schema. Another important distinction is between data and metadata. Here, the term “data” refers to the part of a file or dataset which contains the actual representation of an object of inquiry, while the term “metadata” refers to data about that data: metadata explicitly describes selected aspects of a dataset, such as the time of its creation, or the way it was collected, or what entity external to the dataset it is supposed to represent. Independently of its type, any dataset relevant to research represents specific aspects of the object of scrutiny, be it in the natural sciences, the social sciences, or the humanities. Data is not the object of study itself, but “stands in” for it in some way. Also, data is always a partial representation of the object of study. In some cases, however, it is our only window into the object of study. Still, this “disadvantage” of partial representation is small compared to the fact that digital data can be transformed, analyzed, and acted upon computationally.

Data in the humanities is a bit special: one could in fact argue that text in a book or a manuscript, or the visual elements making up a painting, are data already. First, however, this is analog, non-discrete data, which cannot be analyzed or transformed computationally; and second, language, texts, paintings, and music are semiotic systems that have dimensions beyond the physically measurable, dimensions which depend on semantics and pragmatics, that is on meaning in context. For this latter reason particularly, speaking of “data” in the humanities is problematic and has been challenged. Criticism has come from mainstream scholars who see “data” and quantitative methods of analyzing them with suspicion, because the apparent empiricism of data-driven research in the humanities seems at odds with principles of humanistic inquiry, such as context-dependent interpretation and the inevitable “situated-ness” of the researchers and their aims.

Some practitioners of digital humanities, notably Joanna Drucker, have argued that the term “data” is actually inadequate. And indeed, the term’s etymology seems problematic in the context of the humanities: it comes from the Latin datum, which means “that which is given.” This means it carries with it the meaning of an observer-independent fact which cannot be challenged in itself. Johanna Drucker prefers to speak of “capta” instead of data, literally “that which has been captured or gathered”, underlining the idea that even the very act of capturing data in the first
place is oriented by certain goals, done with specific instruments, and driven by a specific attention to a small part of what could have been captured given different goals and instruments. In other words, capturing data is not passively accepting what is given, but actively constructing what one is interested in.\footnote{6}

Similarly, Digital Archivist Trevor Owens has argued that data is not a given, but is always manufactured and created. Moreover, he shows, we can approach data from different perspectives and treat it as an artifact (something actively and purposefully created by people), as text (subject to interpretation, for example by scholars), and as computer-processable information (to be analysed with quantitative methods). According to Owens, this means that data is not a given and not some unquestionable evidence; rather, it is “a multifaceted object which can be mobilized as evidence in support of an argument.”\footnote{7}

Even without using a new term, we can now redefine what we mean by data in the humanities. Data in the humanities could be considered a digital, selectively constructed, machine-actionable abstraction representing some aspects of a given object of humanistic inquiry. Whether we are historians using texts or other cultural artifacts as windows into another time or another culture, or whether we are literary scholars using knowledge of other times and cultures in order to construct the meaning of texts, digital data add another layer of mediation into the equation. Data (as well as the tools with which we manipulate them) add complexity to the relation between researchers and their objects of study.

Basically, I would like to argue that there are two core types of data in the humanities: big data and smart data. These two types of data can be described in two dimensions: the first dimension describes how structured, clean, and explicit the data is; the second dimension describes how voluminous and how varied the data is. I suggest to view big data, in a first approximation, as relatively unstructured, messy and implicit, relatively large in volume, and varied in form. Conversely, I suggest to view smart data to be semi-structured or structured, clean and explicit, as well as relatively small in volume and of limited heterogeneity. Although you could say that these are really just differences of degree, there are more fundamental differences between them when it comes to looking at how each of them are created or captured, modeled, enriched, and analyzed.

2. Smart data (in the humanities)
When we move from books to digitized versions of the text contained in books, we are not necessarily dealing with big or smart data right away. It may very well be small and simple, not to say “messy” data in the beginning. This is probably the least useful type of data. So what do I mean by “smart data?”

First of all, I should mention that “smart data” is not an established or well-defined term. It is not very widespread and does not have a stable meaning. Smart data is data that is structured or semi-structured; it is explicit and enriched, because in addition to the raw data, it contains markup, annotations and metadata. And smart data is “clean”, in the sense that imperfections of the process of capture or creation have been reduced as much as possible, within the limits of the specific aspect of the original object being represented. This also means that smart data tends to be “small” in volume, because its creation involves human agency and demands time. The process of modeling the data is essential to small/smart data; its abstract structure can be defined with elaborate schemas or as predefined database structures.

A prototypical example of smart data are scholarly digital editions produced using the Guidelines[8] of the Text Encoding Initiative. Technically, TEI documents are usually considered semi-structured; usually, they follow a data model expressed in a schema, but such schemas allow for considerable flexibility. In addition to a very clean transcription of the text, digital editions using TEI can make a lot of information explicit: first of all, TEI files contain not just the full text, but also metadata associated with the text (in the teiHeader section); also, the data is structured and explicit: there is markup making the structure of the text explicit, identifying parts, chapters, headings, paragraphs, as well as page and line breaks, for example. Finally, many more types of information can be specified: for example person names in a novel or play, place names in a letters or documents, and many more things; and links to other parts of the documents and to external documents. Making all of these things explicit allows to visualize them in specific ways and to index, count and analyze them computationally.

But let’s move on to another example of “smart data.” This data comes from a study of literary description in the eighteenth century novel which I conducted some years ago.[9] The aim was to identify all descriptive passages in a collection of thirty-two novels published between 1760 and 1800 and to find out how, from the standpoint of literary stylistics, descriptive writing “functioned” at that time. For this, a bibliographic reference management system was used as the front end to a database of descriptive passages which I collected and tagged for dozens of features I
considered relevant for the study. For example, all 1,500 pieces of descriptive writing were tagged for the various textual strategies of integrating or legitimizing them in relation to the narrative context. This allowed me to discover previously unnoticed recurring configurations, patterns of usage, and trends over time. For example, although eighteenth-century novels do not “frame” descriptions as clearly and as symmetrically as some nineteenth-century novels do, most of the features of such framing are present. However, they are used in an asymmetrical way that tends to purposefully create a smooth transition between narration and description rather than a sharp framing contrast. In addition, correlations between such integrating strategies and different narrative perspectives were found, correlations which in turn help relate descriptive technique to long-term trends in French narrative fiction.

Using a database made it possible to deal with the 1,500 examples and their many tags. Also, having all excerpts and their tags at my fingertips changed the way I interacted with the data, as opposed to manual annotation and note-taking. Building the database itself was an ongoing process of explicit iterative modeling via an evolving set of tags and their relations, which involved adding more and more descriptions, adding tags to them, revising the tagging system and hence modifying
the tags, etc. Recurring patterns and correlations could then be discovered; also, most importantly perhaps, outliers could not just be ignored and the resulting models attempt to cover not just a selection of examples judged to be representative but the full actual practice observed in the corpus.

Despite these significant transformations of the way we work with texts when they are available as “data,” I believe the move from print culture to smart digital data is actually a rather small step compared to the steps required by big data, a subject which will be addressed below. In the digital medium, we can also read texts, look at images, make annotations, and write down ideas and syntheses.

Now, this is all very well and good: smart data as we find it in scholarly digital editions, in annotated linguistic corpora and in carefully curated image collections is immensely useful. However, there is an issue with smart, clean data: it does not scale well. Although various aspects of creating smart data such as carefully encoded TEI documents can be at least partially automated, ultimately smart data depends on manual work by real people. Classifying descriptions in their context according to formal, semantic and narratologic categories is not something computers can do just yet. This means that it is very time-consuming to create large volumes of smart data.

Of course, there are ways to deal with this, and Machine Learning will no doubt be one of the keys to these challenges. But what if we actually don’t really need smart data? What if having a lot of relatively unstructured, relatively messy data is just as useful and much easier to create? This kind of data is called “big data,” so let’s have a closer look at this alternative model of data.

3. Big data (in the humanities)

Big data is the buzz-word of the decade. Everyone wants big data and big data technologies; big data experts are telling large corporations they won’t keep their competitive edge without big data. Areas as diverse as online marketing, stock exchange trading, health care, and political campaigns are driven by big data. The European Commission and the German Ministry of Education and Research hold “big data” conferences and fund big data research.

So, what does it mean for data to be “big”? Jonathan Ward and Adam Barker, the authors of a survey paper on definitions of big data, point out that because the term
has been created and defined by industry, media and academia alike, there is a lack of common understanding regarding its definition. Their attempt to define such a common basis results in the following definition: “Big data is a term describing the storage and analysis of large and or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.”[10] Definitions of big data can indeed vary widely depending on the perspective adopted. In a recent best-selling book about the Big Data Revolution, the authors propose a non-technical, outcome-oriented definition of big data: “Big data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value.”[11] Another high-profile albeit more technical definition of big data, by Doug Laney, points to three key qualities of such data, the three V’s: volume, velocity and variety.[12]

Although the three V’s seem to provide a more precise definition of big data, they also show that big data is in fact a relative term and a moving target, depending on context and available technologies. The idea that big data is defined by its (large) volume is seemingly the most obvious of the three V’s. However, when does a large volume of data really become “big data”? You may consider data you want to analyze to be big when it exceeds the memory of your computer, forcing you to move processing to a grid computing system. However, technologies enabling this are becoming more widespread: A solution like “Hadoop”, that allows the distributed but closely coordinated processing of huge volumes of data on hundreds or thousands of machines in a grid, is cutting-edge and fancy now, but will be mainstream very soon.

Second, the idea of “velocity” of data really means two things: first, that data is constantly being generated by sensors (in the natural sciences, or by public surveillance cameras) or as a by-product of people’s activities in a digital environment (in economics or the social sciences), creating a constant influx of new data. Second, this flow of data is being analyzed in real-time and has to be very quick and responsive. In turn, this allows to react immediately to the data. This aspect of big data is probably the least relevant to data in the humanities, at least today.

Finally, the idea of “variety” of big data means that heterogeneous sources and formats of data are being used together, taking advantage of the links and overlap between such heterogeneous datasets to allow all kinds of inferences. What the idea of “variety” also implies is a variety of ways these datasets are structured, or a relative lack of structure in the datasets. The challenges here lie particularly in the fact that all these various datasets cannot be integrated into one unified dataset. The
heterogeneity is probably the biggest challenge of data in the humanities, which may come from a variety of sources, in a variety of formats, and need to be combined flexibly in order to take the greatest possible advantage from them. Similarly to grid computing, however, solutions like “NoSQL”-databases or graph-based databases that avoid some of the limitations that more traditional SQL-databases have when it comes to heterogeneous or unstructured data will soon be part of our normal data analysis toolbox and using them will not be an indicator of “big data” applications anymore.

In addition to this, big data in the humanities is not the same as big data in the natural sciences or in economics. In most cases, velocity does not play a key role in big humanities data right now. Also, the large “volume” is less usefully defined in the humanities by a shift from databases to distributed computing. Variety of formats, complexity or lack of structure does come into play, however. In fact, the distinctive mark of big data in the humanities seems to be a methodological shift rather than a primarily technological one. And it is a huge methodological shift. Paradoxically, the shift from small smart data to big data is much more radical, I would argue, than the shift from print to smart digital data was. Indeed, moving from smart data to big data implies a shift from “close reading” to “distant reading” (in the words of Franco Moretti) or to “macroanalysis” (to use Matthew Jockers’ term). In this paradigm, instead of reading a few selected texts, we analyze an entire collection of relevant textual data.

The first consequence of the macroanalytic paradigm in the humanities, where hundreds or even thousands of texts are analyzed at a time, is that instead of operating on the level of literary forms and conventions, of semantics and context, we operate with quantitative measures of low-level features, on the basis of statistics and probabilities. The second consequence is that instead of so-called “representative” texts or paintings, we can now study the entire set of texts or images relevant to a specific research question. Trends in literature can be observed across the entire literary production of a given time and given genre. Questions of representativeness, of canonization, of literary quality play a much smaller, or at least a different, role in this context.
If this sounds too good to be true, that is because it is. Despite massive digitization efforts by private and public actors, we are still far from the complete record of human culture and creativity, even if we are looking just at text. For the British nineteenth century novel, for example, the calculations go as follows: an estimated 20-30,000 novels were published in the nineteenth century; of these, only around 6,000 are estimated to be still extant in the holdings of libraries and private collections worldwide. Of these 6,000 novels only about half, that is 3,500 novels, have been digitized in full text mode and with sufficient metadata; this is the number of novels contained in the “Stanford Literary Lab Corpus” which Matthew Jockers used in some of the studies described in his book *Macroanalysis*. That corresponds to less than twenty percent of the total production.\(^{14}\) This would certainly be a good sample size, if it were a random sample, but of course it is not. Rather, it is an opportunistic sample. So, the 3,500 novels seem like a small amount and not a number that actually resolves the sampling, representativeness, and canonization issues. Still, such a change of scale is a huge improvement over the mainstream canon which probably does not include more than one hundred novels for the nineteenth century, and may be visualized as a tiny spot at the bottom of the graph.
Such practical limitations in the digital materials available for research mean that examples for “really big” data in the humanities are still relatively rare. Even Google Books is not quite there yet. Google has scanned more than 30 million books, as of April 2013, and continues to scan more. Compared to the estimated 130 million books ever published, this is a large part of the written human record. But of course, this is neither exhaustive nor are books the only medium of print publication there is, so newspapers, magazines and journals would need to be added to this. What really counts, however, from my point of view, is less the volume than the methods used for analysis. And these can be successfully applied to smaller sets of data as well, and imply precisely the methodological paradigm shift I mentioned, from close to distant reading.

I would like to give just one example from my own work, dealing with French drama from the seventeenth and eighteenth century and involving a maximum of 580 individual plays. That’s not big data in the technical sense of the three V’s, but it requires a radical shift from close reading methods to quantitative, statistical analysis. The basic question I have been addressing for the last year or so is how traditional categories of literary history, such as literary genres, forms and periods, relate to classifications made on the basis of the actual linguistic material. What comes out of this type of analysis, which can be pushed further in a variety of ways, is that there are indeed correlations between linguistic features on the one hand, and large categories from literary history on the other hand; but also, that these are not simple and stable correlations, but highly complex and varying ones.

In one case, for example, I analysed a collection of French plays from the seventeenth century written by Thomas and Pierre Corneille using Principal Component Analysis, a technique which discovers correlations in multidimensional data and summarizes such correlations into so-called principal components. The following graph shows how tragedies and comedies by these two authors cluster when plotting them according to the first two principal components.
Some interesting trends become visible: for example, it is remarkable how closely these two components seem to be related one to authorship and one to genre. Most plays in the left half are by Pierre Corneille, with some exceptions especially for the tragedies in the lower half. Most plays on the right side are by Thomas Corneille, again with some exceptions especially in the lower half. So the first component (horizontal axis) seems to be correlated with authorship. The second component (vertical axis), on the contrary, seems to be correlated with genre. Most plays in the upper half of the graph are comedies, and most plays in the lower half of the graph...
are tragedies. Only a few tragedies by Pierre Corneille and even less by Thomas Corneille appear in the upper half of the graph. Also, the dispersion of the data points (or plays) seems to be greater across PC1 for comedies than for tragedies. The tragedies are somewhat lumped together and authorship distinctions are actually hard to make just on the basis of PC1, so much overlap is there! This is not the case for the comedies in the upper half, where overlap seems to be much weaker. French tragedy as a genre, at least in the 1660s, seems to be more stylistically homogeneous than comedy, that is to be a particularly strongly conventionalized genre, a finding which is well-supported by mainstream scholarship.

Conclusion: towards smarter big data or bigger smart data

For most of this paper, I have been opposing big data and smart data. Indeed, big data tends to involve large volumes of raw, plain, somewhat messy text, whereas smart data tends to involve smaller volumes of carefully encoded, very clean text. Big data needs to be analyzed with methods from statistics, such as cluster analysis or principal component analysis and many more, whereas smart data can be analyzed with specific tools allowing to take advantage of structural, linguistic and contextual markup. Big data requires visualization to even start understanding its possible structure, whereas smart data makes its structures explicit. In big data applications, outliers, errors and ambiguities are said to matter little because they get smoothed over by the sheer quantity of information that is good enough, whereas smart data makes exceptions and ambiguities explicit and effectively reduces possible ambiguities.

That said, I believe the most interesting challenge for the next years when it comes to dealing with data in the humanities will be to actually transgress this opposition of smart and big data. What we need is bigger smart data or smarter big data, and to create and use it, we need to make use of new methods. So, how can we enrich big data sufficiently to make more intelligent queries possible? How can we speed up the process of creating smart data so that we can produce larger volumes of it?

Basically, there are two possible ways to do this: one is automatic annotation, the other is crowdsourcing. Automation refers to various heuristics of discovering implicit units, structures, patterns and relations, and of making them explicit in the data. Crowdsourcing, on the other hand, relies on breaking down a large task into such small units that each of these little tasks can be performed in a distributed way.
by a large number of volunteers. Various strategies have been developed for breaking up the tasks, for creating incentive structures to motivate volunteers (like “gamification” or “win-win”-constellations), and to reintegrate the added information into the project.

In fact, automation and crowdsourcing will have to work hand in hand. This is what happens with OCR: better and better optical character recognition systems are still no match to manual double-keying or transcription by experts, especially when it comes to print before 1800 or to handwriting. But state-of-the art OCR combined with algorithms to detect potential areas of error and cleverly crowdsourced and distributed error-correction mechanisms such as the ones implemented by “Captcha” go a long way to producing large amounts of more reliable full text. Similarly, automatic linguistic annotation even of basic linguistic features for well-researched languages is still too faulty to be trusted blindly, at least in a “clean smart data” perspective. We will have to find ways of detecting potentially faulty linguistic annotation, then finding and motivating users to check such annotations, and writing the corrections back into larger and larger collections of clean, structured and well-annotated text.
The story of smart and big data.

To summarize the story of data in the humanities which I have been trying to tell, one could consider that this story has several steps: The first step leads from the study of creative works in the form of books, paintings and movies to their study based on digital representations of these works; this is what digitization at its most basic level as brought about; the first of two possible steps from there leads to smart data, that is to data that has been carefully curated, structured, annotated in a way to make explicit a lot of information that is implicit in the “raw” and messy digitized artifacts. This happens, prototypically, in scholarly digital editions of text or music scores. The second of the two possible steps from “raw digital data” leads to big data, simply by accumulating more and more data and letting the algorithms sort it all out, instead of cleaning it up by hand. The last step in this story is to reconcile, or rather to combine, the smart and the big data approaches.

For my own research in computational genre stylistics, having collections of texts at my disposal that are both larger and smarter than what we have now will be crucial. Collections need to be large, because as soon as you focus on more specific cases, such as a specific sub-genre from a specific period, even a relatively large collection of texts will only yield a small number of samples. And when the number of samples gets too low, statistical approaches loose their robustness and reliability. And similarly, more nuanced and interesting analyses of large text collections depend on having a large array of metadata and annotations regarding each text, including things like the proportion of verse and prose in a play, or of description and narration in a novel. Such information needs to be available so that correlations between stylometric findings concerning a text and relevant attributes of the text in question, can be discovered.

In other words, we need smart big data because it can not only adequately represent a sufficient number of relevant features of humanistic objects of inquiry to enable the level of precision and nuance scholars in the humanities need, but it can also provide us with a sufficient amount of data to enable quantitative methods of inquiry that help us transgress the limitations inherent in methods based on close reading strategies. To put it in a nutshell: only smart big data enables intelligent quantitative methods.

1. [1] This contribution is a revised version of a talk I gave at the European Summer University “Culture & Technology” organized by Elisabeth Burr at the University of Leipzig in Germany, on July 26, 2013. The Summer University brings together graduate students and researchers from a wide range of disciplines in the humanities and in computer science from many different countries around the world. The talk was one of several plenary talks intended to introduce the audience to various topics of more general concern, in addition to the more specific workshop tracks. The talk was the result of a cooperation between the European Summer University and the German branch of the DARIAH initiative in which I am a research associate and where I am mostly concerned with understanding digital methods and with supporting and training mainstream humanities scholars to use such methods. A slightly revised version of the talk was documented on my blog, The Dragonfly’s Gaze, in early August 2013. The research reported here has been supported by DARIAH-DE with funding provided by the German Federal Ministry of Education and Research (BMBF) under the identifier 01UG1110A-M.

2. [2] There are many such sources of digital data for research in the humanities: we have large text archives such as Google Books, Hathi Trust, the Internet Archive, or Gallica; we have scholarly digital text archives such as TextGrid’s Digital Library of German-language literature from 1500 to 1900 or the Théâtre Classique collection of French Drama, and many others; and we have hundreds of elaborate digital editions of literary and historical texts, such as the Van Gogh Letters, the Walt Whitman archives, or Rousseau-Online. We have image databases like “Flickr Commons,” the distributed image archive Prometheus, Getty, and others. And we have sites like Youtube, Open Culture and the Moving Image Archive.


9. [9] This study was done as part of my doctoral dissertation; its results have been published as Christof Schöch: *La Description double dans le roman français des Lumières 1760-1800*. Paris: Classiques Garnier, 2011.


19.7.2013, [http://dh2013.unl.edu/abstracts/ab-270.html](http://dh2013.unl.edu/abstracts/ab-270.html).

About Christof Schöch

Christof Schöch is a researcher at the Chair for Computational Philology, University of Würzburg, Germany. His interests in research and teaching are French Literature (Age of Enlightenment, contemporary novel) and Digital Humanities (quantitative text analysis, digital scholarly publishing). He writes about his research on a blog called *"The Dragonfly’s Gaze."* You can find out more about him on his [personal website](http://www.christof-schoech.de).
1 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").

2 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:

4 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. 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. 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. 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
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, appendices, 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.

17 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.

18 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.

19 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.

¶ 20 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.8

¶ 21 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.

¶ 22 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.

Figure 3. PCA analysis of the same texts with words plotted on the same two dimensions as the 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. 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%.

33 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.

34 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.

§ 36
### Table 1. Lists of the one hundred most distinctive gender marker words in thirteen female and thirteen male poets.

<table>
<thead>
<tr>
<th>One Hundred Distinctive Words in the Works of Thirteen Poets*</th>
<th>One Hundred Distinctive Words in the Works of Thirteen Poets†</th>
</tr>
</thead>
<tbody>
<tr>
<td>mother’s</td>
<td>sign</td>
</tr>
<tr>
<td>father’s</td>
<td>delicate</td>
</tr>
<tr>
<td>skin</td>
<td>fist</td>
</tr>
<tr>
<td>table</td>
<td>hotel</td>
</tr>
<tr>
<td>lovers</td>
<td>heels</td>
</tr>
<tr>
<td>eggs</td>
<td>sea</td>
</tr>
<tr>
<td>mother</td>
<td>door</td>
</tr>
<tr>
<td>hot</td>
<td>itself</td>
</tr>
<tr>
<td>breath</td>
<td>wet</td>
</tr>
<tr>
<td>next</td>
<td>gold</td>
</tr>
<tr>
<td>grow</td>
<td>stones</td>
</tr>
<tr>
<td>grew</td>
<td>gave</td>
</tr>
<tr>
<td>secret</td>
<td>wait</td>
</tr>
<tr>
<td>five</td>
<td>almost</td>
</tr>
<tr>
<td>I’d</td>
<td>mirror</td>
</tr>
<tr>
<td>watched</td>
<td>later</td>
</tr>
<tr>
<td>chair</td>
<td>afternoon</td>
</tr>
<tr>
<td>lines</td>
<td>kept</td>
</tr>
<tr>
<td>shining</td>
<td>shoulder</td>
</tr>
<tr>
<td>watches</td>
<td>thighs</td>
</tr>
<tr>
<td>ear</td>
<td>somewhere</td>
</tr>
<tr>
<td>rare</td>
<td>touched</td>
</tr>
<tr>
<td>wore</td>
<td>trying</td>
</tr>
<tr>
<td>sealed</td>
<td>cream</td>
</tr>
<tr>
<td>wool</td>
<td>angels</td>
</tr>
<tr>
<td>spin</td>
<td>orange</td>
</tr>
<tr>
<td>going</td>
<td>blew</td>
</tr>
<tr>
<td>lay</td>
<td>spent</td>
</tr>
<tr>
<td>sat</td>
<td>remembered</td>
</tr>
<tr>
<td>fruit</td>
<td>bowl</td>
</tr>
<tr>
<td>instead</td>
<td>carefully</td>
</tr>
<tr>
<td>leaving</td>
<td>fair</td>
</tr>
</tbody>
</table>

*This list features words by female poets. †This list features words by 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).

<table>
<thead>
<tr>
<th>Cluster</th>
<th>500 Women’s Markers</th>
<th>500 Men’s Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>mother’s, father’s, mother, father, children, ancestral, aunt, baby, birth, child, child’s, cousins, daughters, family, generations, uncles</td>
<td>faith, heaven, hell, prayers, souls, spirit, Christ, gods, myth, paradise, religion, spirits, temple</td>
</tr>
<tr>
<td>Religion</td>
<td>altar, nuns, praying</td>
<td></td>
</tr>
<tr>
<td>Houses/Furniture</td>
<td>table, chair, door, doorway, floor, doors, bathroom, bedroom, carpet, ceiling, cellar, chimney, closet, cupboards, gates, hotel, kitchen, palace, rooms, rug</td>
<td>attic, buildings, ruins, shack, temple, wall</td>
</tr>
<tr>
<td>Song/Dance</td>
<td>danced</td>
<td>song, dancing, sing, dance, song, dancer, music, singer, singing, sings</td>
</tr>
<tr>
<td>Personal Pronouns</td>
<td>he’ll, I’d, mine, ourselves, she’d, she’s, they’d, you’d, you’re, yourself</td>
<td></td>
</tr>
</tbody>
</table>

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

40 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*.

41 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.

42 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.

43 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.).

45 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.

46 7. Lee has a good list of software with commentary.

47 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.

48 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.

49 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


Halliday, M. A. K. “Linguistic Function and Literary Style: An Inquiry into the Language of


Hoover, David L., and Shervin Hess. “An Exercise in Non-ideal Authorship Attribution:


DOI: 10.1632/lsda.2013.3
Visualizing the French Enlightenment Network Using Palladio

Working Paper, 2019


By Melanie Conroy, University of Memphis  mrconroy@memphis.edu

There are numerous digital tools for studying networks that can be of use to humanists. One such tool is Palladio, a digital humanities package developed in the Humanities + Design Lab at Stanford University.¹ Palladio lends itself to qualitative studies because the visualizations that it produces (maps, network diagrams, and tables) are familiar to most humanists, and because it allows for the filtering of data through categories chosen by the user. Digital tools offer an opportunity to pursue such research with lower technical barriers to intervention—including easier communication, tools for sharing data, and collaborating on writing. Visualization tools like Palladio allow academics to produce their own diagrams without necessarily hiring a designer.

I will walk through some examples of diagrams produced in Palladio. I will show how maps can be used to compare the weight, or influence, of cities, as well as travel and communication between cities and other geographical points; then I will show how network graphs can be used in the study of networked people or things. The data are from the Electronic Enlightenment project and the Groupe d’Alembert. I enriched the metadata with Maria Teodora Comsa, Chloe Edmondson, Dan Edelstein, and Claude Willan, according to the methods that we described in our article “The French

¹ Palladio allows for the presentation of multifaceted data, such as network data with date ranges or categories such as network type with location data. Palladio was developed in the Humanities + Design Lab, by Dan Edelstein, Nicole Coleman, Ethan Jewett, Giorgio Caviglia, et al. For more information or try it out, see http://hdlab.stanford.edu/palladio/.
Enlightenment Network."² This article explains how the dataset is structured and why it is structured that way.

I have made the French Enlightenment network dataset available online in multiple formats, with examples of diagrams, at http://bit.ly/VisualizingFEN. While I cannot fully explain how the dataset is structured or how it was created here, it is worth knowing that the data is a list of individuals with basic demographic data (name, birth and death places and years, nationality, gender, etc.); it also places these individuals into other groups, or networks, such as "Letters_Philosophical" for writers on philosophical subjects, and "Arisocracy" for members of the nobility.³ If you are interested in the structure of the data, I encourage you to open the Excel Document the Tab-Separated Values file and take a look at the raw data.

For the sake of this demo, I recommend opening the JSON document—that is, the preformatted version—in Palladio by following these steps:

1) Download the JSON document:


3) Choose “Load an existing project” and then choose the file “French Enlightenment Network.JSON”

4) Select the tab that you would like to see. Choose “map” to view the map of linked birth and death places for the French Enlightenment network. Choose “graph” to see network diagrams.

---


³ The dataset for “The French Enlightenment Network” is available online permanently here: https://searchworks.stanford.edu/view/bc436tm1194. A working folder with all the links contained here can be found at http://bit.ly/VisualizingFEN.
First we will look at the “map” feature in Palladio.

Two Kinds of Maps: Points Versus Point-to-Point

There are two map types: 1) maps with points sized according to defined quantities, which are useful for displaying static geographic information, and 2) point-to-point diagrams, which are more useful for showing movement or change over time. The point-to-point maps are particularly useful for humanities research because they show network connections in a way that is intuitive to humanists by overlaying them on a map. The data for both of these map types can be filtered to show only a subset of the data.

The first type of diagram (Figure 1) consists of points on a map. Producing diagrams of this kind requires LatLong coordinates for all locations. Such maps can easily display quantities associated with a particular place—for example, quantities of people, events, or items. These maps are recommended for population data, data on the production of books, or number of events. The nodes can be sized by the number of records at each LatLong.
Figure 1: Birth Places of Members of the French Enlightenment Network

The second kind of diagram is a point-to-point diagram (see Figure 2).

Figure 2: Linked Birth and Death Places of Members of the French Enlightenment Network with Timespan
Point-to-point maps are ideal for displaying trajectories, such as travel, or communication between population centers, such as the sending of letters. This map type requires LatLong coordinates for two distinct locations. The edges—the connections between points—could represent any number of other network relationships between two locations. The points are sized according to the total count of the items associated with that location.

In Figure 3, the linked birth and death places are displayed on a map alongside a timespan of lifespans. The timespan can be opened or closed with the arrows on the right side. It can also be deleted by clicking on the red garbage can.

Figure 3: Linked Birth and Death Places of Members of the French Enlightenment Network with Timeline by Gender
While maps are not new tools for humanists, digital technology allows for more sophisticated and rapid comparisons between subsets of data. Palladio makes filtering data easy and quick; in fact, producing a different map takes only a few seconds. In this case, the individuals have been coded by gender, nationality, social networks (Elite, Aristocracy, Military, Court), professional networks (Artisan, Finance-banking, Cultural, etc.), and by knowledge networks (Letters, Sciences, etc). These individuals also have birth years, death years, birthplaces, and death places associated with them, where this data is known. Filtering allows us to look at a series of maps and compare them easily. Figure 4 shows the linked birth and death places of aristocrats in the French Enlightenment Network.

![Figure 4: Linked Birth and Death Places of Members of the French Enlightenment Network Filtered (Aristocracy)](image)

*Second we will look briefly at the “graph” feature in Palladio.*

Network Graphs
A network can link people, places, books, or any other entities that are connected to one another. People or groups are commonly nodes in graphs of social networks. Network graphs are useful for seeing connections between people or groups. They are also helpful for understanding how groups are structured. A network graph is a set of points (called nodes), connected by links (called edges).⁴ There are other software packages— notably Gephi, a network analysis computer program designed for the quantitative study of networks. Palladio is particularly good for humanities research but lacks the ability to do complex mathematical analysis.

For this demo, please choose “Graph.” Figure 5 shows the academy affiliations of members of the French Enlightenment Network with the knowledge networks of academy members. Each edge shows a knowledge network that is present in an academy. The nodes are sized according to the number of members of the academy or knowledge network. The edges represent knowledge networks that share members with academy members.

⁴ For an introduction to social network analysis, see Christina Prell, Social Network Analysis: History, Theory, and Methodology, Sage Publications, 2011.
Figure 5: Network of Academy Affiliations and Knowledge Networks

Conclusions

Palladio allows for easy filtering and the construction of maps and network graphs. The French Enlightenment Network dataset contains many variables, such as knowledge networks, social networks, professional networks, academy membership, gender, birth and death places and years. A very large number of diagrams can be generated from this dataset and other tables can be added to any column in the table. One limitation of the dataset is that it only includes birth and death places (for some of the individuals); it doesn’t contain other geographical information. By extending the table in the “data” section of Palladio, further geographic information or other characteristics can be added. When constructing datasets for Palladio, it is important to remember that multiple values can be used in a single cell of a spreadsheet but that one character should function as the delimiter (for example, the comma or the semi-colon). For the sake of easily
processing the data, it’s easier if the same character is used throughout. By adding multiple values in the same cell, datasets can be made more compact.

There are two other major features of Palladio. Palladio can be used to produce tables and galleries. Select the “Table” tab to make lists of items based on a category they are in. For example, you can easily locate all of the French members of the Royal Society or all of the members of the “Letters_Literary” network by producing a table of academies or of knowledge networks. The “Gallery” feature is useful for producing a list of entities who fall into specific categories where there is an image associated with each. The French Enlightenment Network dataset includes Wikipedia Image URLs for portraits of some of the network, which can be used to explore this function.

Melanie Conroy is assistant professor of French at the University of Memphis. She received her doctorate from Stanford University in 2012. Her research explores the intersection of literature, visual studies, and social networks in modern French culture. She is the co-director of the Salons Project, a part of Mapping the Republic of Letters.
Feminist Data Visualization

Catherine D'Ignazio and Lauren F. Klein

Abstract — In this paper, we begin to outline how feminist theory may be productively applied to information visualization research and practice. Other technology- and design-oriented fields such as Science and Technology Studies, Human-Computer Interaction, Digital Humanities, and Geography/GIS have begun to incorporate feminist principles into their research. Feminism is not (just) about women, but rather draws our attention to questions of epistemology — who is included in dominant ways of producing and communicating knowledge and whose perspectives are marginalized. We describe potential applications of feminist theory to influence the information design process as well as to shape the outputs from that process.

Index Terms — Visualization, inclusion, digital humanities, critical perspectives, feminism.

1 INTRODUCTION

When exploring the intersection of data visualization and the digital humanities, one must consider not only how the domain of digital humanities — and of the humanities more generally — can provide new opportunities for the design and application of visualization tools and techniques, but also how theories from the humanities can themselves inform visualization design. Research in the field of data visualization is often framed in terms of how it helps to “reveal” knowledge [15], support narrative storytelling [70], or otherwise facilitate pathways to “insight” [12]. These same keywords are often employed — and challenged — in humanistic theories that explore how knowledge is produced, transmitted, and perceived. Among the earliest and the most enduring of these theoretical schools is what is known as feminist theory. A body of work that owes its emergence to the women’s suffrage movements of the nineteenth century, feminist theory evolved through several “waves” over the course of the twentieth century, and now encompasses a range of ideas about how identity is constructed, how power is assigned, and how knowledge is generated, as well as how a range of intersectional forces [19] such as race, class, and ability, combine to influence the experience of being in the world.

In this paper, we outline a feminist approach to visualization, drawing upon a set of canonical and contemporary theories from the humanities in order to show how visualization research can be adapted to emphasize the situated nature of knowledge and its perception. Our goal is to encourage the development of a range of alternative visualization practices that better emphasize the design decisions associated with data and its visual display. We are particularly interested in exposing the assumptions involved in choices about data type, categorization schema, visual typology, interaction mode, and intended audience; as well as those associated with the qualitative aspects of visualization design and its reception, such as the composition and structure of the design team, the identification and involvement of user communities, the contextual and affective factors that influence both the design and reception of visualizations, and the many forms of labor that contribute to a successful visualization design. By identifying these assumptions and associating them with the core principals of what we term feminist data visualization, we hope to expand the conversation about what visualization for — and with — the humanities could become.

2 RELATED WORK

Feminism is “not (just) a women’s issue,” as Johanna Drucker reminds us, nor does feminist theory help to inform issues of gender alone [25]. As the binary distinction between male and female, as well as the hierarchical relation that posits male above female, have been abstracted to serve as models for a range of structures and systems, feminist theory has been marshaled in order to challenge the validity of a variety of binaristic and hierarchical configurations. By the same token, expansions of feminist theory — crucially, intersectional feminism — have been employed to overturn systems of oppression that cannot be reduced to a single structure or source. We lead with this theoretical lens so as to frame the four related fields of inquiry that have contributed to our formulation of feminist data visualization: feminist science and technology studies, feminist human-computer interaction, feminist digital humanities, and critical cartography & GIS. In the following sections, we summarize the main contributions of each field in more detail.

2.1 Feminist Science and Technology Studies

Science and Technology Studies (STS) is an interdisciplinary field that emerged in the 1960s and 70s. STS examines the social, cultural, and historical aspects of science and technology. Feminist theory and analysis has played a key role in this field, leading to the development of original theoretical frameworks [4, 6, 34, 39] as well as the sustained challenge to dominant epistemological perspectives [37, 47, 56, 80]. One of the key contributions of STS has been to challenge the idea that science and/or technology is objective and neutral by demonstrating how scientific thought is situated in particular cultural, historical, economic, and social systems [77]. Feminist STS, both implicitly and explicitly, looks to the perspectives of those marginalized by current power configurations (including and especially those marginalized because of gender, sexuality, race, and/or ethnicity) as a way of exposing how their perspectives are not included in what is considered “objective” truth [74]. Challenging neutrality, objectivity, and universality does not mean that feminist STS retreats to a position of relativism or solipsism, however. The field rejects neither the scientific process nor quantitative ways of knowing the world. Rather, feminist STS allows us to see how all knowledge is situated, how certain perspectives are excluded from the current knowledge regime, and how multiple true objectivities are possible.

2.2 Feminist Human Computer Interaction

In the field of human-computer interaction (HCI), there is an emerging conversation about how to draw from feminist theory and other critical perspectives for the design of interactive and computational systems. Lucy Suchman’s work has long explored the implications of feminist theory for technology production and use [75, 76]. More recently, Shaowen Bardzell has asserted that feminism can be deployed throughout the design process to produce a “generative contribution” [5]. Feminist HCI design work has included foci on female...
makers and hackerspaces [32], motherhood as a life phase [3], post-partum technologies [22], and talking back to street harassment [23]. While historically, HCI aspired to “universal” usability, the early 2000s saw a proliferation of work that challenged that idea through design practice [24, 60]. Feminist HCI builds on that work and draws on feminist standpoint theory [39] to explicitly valorize marginal perspectives so as to “expose the unexamined assumptions of dominant epistemological paradigms” [5]. In relation to visualization in particular, Peter Hall coined the term critical visualization in 2008 to describe practices that counteract “the technological view” of visualization, a view which emphasizes technique and efficiency while eliding historical, social and rhetorical concerns [35]. Marian Dörk et al. built on this concept to elucidate design principles for working with data [26]. In both feminist HCI and critical information visualization, researchers have introduced design principles that attempt to draw attention to how knowledge resides in specific bodies (disclosure/self-disclosure, embodiment), how power is distributed throughout the design process (empowerment, advocacy, ecology) and how to include more voices and alternative perspectives in the design process, as well as the experience of the resulting artifact (participation, pluralism, plurality).

2.3 Feminist Digital Humanities

Since the field’s inception, digital humanities (DH) has entailed a sustained attention to certain feminist concerns. The Orlando Project [10] and the Women Writers Project [31] are early and enduring examples of how DH has emphasized the recovery of female literary and cultural contributions. In recent years, the fields of DH and STS have begun to converge, resulting in a range of projects that incorporate feminist theoretical perspectives into their digital work. Key work in this area has included wearable representations of Twitter activity related to reproductive health [50] and embodied enactments of historical health data [67], as well as content management platforms [58], social media collectives [9], scholarly networks [30], and educational opportunities [78]. In terms of visualization work, gender – especially as it relates to issues of authorship and style – has long served as a subject of DH research, e.g. [46]. However, the visualizations that accompany such analyses almost always employ standard representational techniques [45]. Recently, Miriam Posner [65] identified the development of new visual strategies for the representation of non-binary gender as one of the most pressing challenges of DH today. Other work that seeks to incorporate embodied and affective modes of perception into new visualization forms, e.g. [49, 79], promises to extend feminist digital humanities visualization work in exciting ways. To date, however, this work has been conducted in isolation from the visualization community. Additional partnerships between DH scholars and visualization researchers, along the lines of interdisciplinary projects to visualize the sonic aspects of poetry [57] or the ambiguities of temporal data [59], constitute a rich site for future inquiry.

2.4 Critical Cartography & GIS

In the late 1980s and early 1990s, geographers challenged conventional academic cartography by linking maps and other visual representations of geographic knowledge explicitly to power using the critical theories of Michel Foucault [17, 18]. Cartographers such as JB Harley challenged the perceived neutrality of the map and introduced notions of ideology and bias [40, 41]. While he did not explicitly draw on feminist theory, Harley argued for the situatedness of maps as historically and culturally contingent documents. During the same period, Denis Wood connected maps explicitly to the rise of the nation-state and showed how maps serve political interests [82]. Other scholars linked Geographic Information Systems (GIS) to an impoverished techno-positivism [64] and militarism [71]. Subsequent scholarship theorized the map more as rhetorical proposition than depiction of “fact” [81]. The declassification of GIS technology and the introduction of locative functions into everyday devices like mobile phones has led to a flourishing of artistic and critical mapping practices [20] like Laura Kurgan’s work that intentionally introduces social and political questions through visualization [51]. Relatedly, there has also been an expansion of participatory design strategies for democratizing geographic information collection and visualization [27] for those who have been excluded from dominant mapping practices, such as indigenous populations [16]. Since 2000, scholars have articulated explicitly feminist approaches to mapping and GIS [28, 42, 62] including nuanced considerations of gender and mapping with new technologies [73]. Mei-Po Kwan used the term feminist visualization to describe how GIS could be used in ways that are compatible with feminist epistemologies and politics [52]. Her design principles include grounding mapping practices in women’s everyday lives and political struggles, as well as incorporating qualitative and narrative components into spatial representations.

3 PRINCIPLES OF FEMINIST DATA VISUALIZATION

In what follows, we introduce six core principles of feminist data visualization. As our intention is to directly impact the design of future visualizations, we follow our explanation of each principle with a set of questions relating to design process and design output. We should note, also, that while our primary focus is on visualization design and the related issues of interaction and display, our feminist approach requires that we expand the design frame so as to account for the range of social forces and material conditions that influence the design process. In other words, a feminist approach to data visualization, while centered on design, insists that data, design, and community of use, are inextricably intertwined.

3.1 Rethink Binaries

Central to feminist theory is the disavowal of binary distinctions—not only between the categories of male and female, but also between nature and culture [37], subject and object [43], reason and emotion [54], and body and world [4], among many others. A feminist approach to data visualization should therefore emphasize representational strategies premised on multiplicity rather than binaries, and acknowledge the limits of any binaristic view [53]. This approach is exemplified by (if not limited to) the representation of gender; typically recorded as binary and discrete variables—e.g. either male or female—gender might be better represented as continuous and multidimensional [29]. Not only a challenge for the visualization phase of research, rethinking the representation of gender, among other binaristic categories, challenges us to inquire how the processes associated with data collection and classification, as well as their visual display, might be made to better account for a range of multiple and fluid categories.

Design Process Questions: Is our data the right type? What categories have we taken for granted? How can we register responses that do not fit into the categories we have provided, even and especially if they are “edge cases” and “outliers”? Design Output Questions: How do we communicate the limits of our categories in the final representation? How can we allow the user to refactor the categories we have presented for view?

3.2 Embrace Pluralism

Feminist theory seeks to challenge claims of objectivity, neutrality and universalism, emphasizing instead how knowledge is always constructed within the context of a specific subject position [8, 38, 39, 54]. In the context of data visualization, a focus on the designer’s own subject position can help to expose the decisions, both implicit and explicit, that contribute to the creation of any particular visual display. Both Bardzell and Dörk et al. have framed this quality around “self-disclosure” [5, 26]. We believe that self-disclosure, and an embrace of pluralism more generally, can do more; it can help to encourage alternatives to the single “view from nowhere” so often favored in visualization design [21]. Ideally, a focus on pluralism would help visualization research move away from its current emphasis on “objective” presentation in favor of designs that facilitate pathways to multiple truths.

Design Process Questions: Whose voices are not represented on the design team but might be important for the conceptualization of the project? Who is being envisioned as the ideal user? How could additional perspectives be accommodated, even those considered marginal? Whose perspectives have been excluded from the
As another example, consider standard practices of data cleaning. As a central premise of feminist theory is that all knowledge is a social good, such as “choice,” “openness,” or “access,” result in disempowerment instead?

3.3 Examine Power and Aspire to Empowerment

Historically, women and other marginalized groups have experienced the negative effects of hierarchical structures of power. Feminist approaches seek to overturn these hierarchies by promoting horizontal systems of knowledge transmission. Such systems insist on a two-way relation between subject and object of knowledge [36, 39]. A feminist approach to data visualization therefore acknowledges the user as a source of knowledge in the design as well as the reception of any visual interface. The creation of knowledge is, after all, always a shared endeavor.

Following from this point is a related principle: that users are bound to the communities that shape them. Aspiring to empowerment, then, may involve designing for and evaluating the success of a visualization at the scale of the community rather than the individual user. This reorientation can help us to acknowledge the communities who provide us with our design challenges, while also ensuring that the outcomes of our design research connect back to the communities that first made them possible. It can also help us to listen to community concerns and co-design visualizations to advance their goals, while building capacity to achieve them within the community.

**Design Process Questions:** How is power distributed across the design team? Whose voice matters more and why? How can end users’ voices be more fully integrated into the design process? Can we build capacity in user communities, or enlarge our internal perspectives, by employing a more participatory design process?

**Design Output Questions:** Can the visualization empower the end user and/or her community, group, or organization? When do values often assumed to be a social good, such as “choice,” “openness,” or “access,” result in disempowerment instead?

3.4 Consider Context

A central premise of feminist theory is that all knowledge is *situated* [36], where “situated” refers to the particular social, cultural, and material context in which knowledge is produced [33]. A feminist approach to data visualization must therefore consider how diverse contexts can influence the production of a visualization, and think through the various ways in which any particular visualization output might be received. In the context of an Enlightenment model of knowledge production, in which additional information leads to increased understanding, a model that allows for the user to “drill down” to more information might be the logical solution for the display of an information system; but this is not the most appropriate choice for more horizontal knowledge frameworks, or those premised on exchange. As another example, consider standard practices of data cleaning. As designers, we often require “clean” data to construct our visualizations. Loukass argues that local context is lost when we homogenize data [55]. An awareness of what we can learn from local context may yield richer and more informative visualizations.

**Design Process Questions:** How can we leverage human-centered design [14] and participatory design [72] methods to learn about and with our end users, including learning more about their culture, history, circumstances, and worldviews? How can we let these insights shape our design practice and change our notions about what constitutes “good” information design?

**Design Output Questions:** What kinds of terminology, symbols, and cultural artifacts have meaning to end users, and how can we incorporate those into our designs? What might we learn if we were to visualize “messy” data [68]? How do we take context into account in the assessment of visualizations?

3.5 Legitimize Embodiment and Affect

Feminist theory recognizes embodied and affective experiences – that is, experiences that derive from sensation and emotion – as ways of knowing on par with more quantitative methods of knowing and experiencing the world [13]. By definition, visualization rests on the production and assimilation of visual knowledge. But even the most efficiency-oriented and task-driven visualizations have embodied and affective impact, if only to communicate their utility, economy, and purposefulness by way of the visual domain. With the rise of popular forms of visualization such as *data journalism*, designers have begun to intentionally leverage affect in order to create an emotional bond with a story or issue [11], or to engage and impress readers with beauty and complexity [61]. These affective dimensions of visualization have been under-explored in traditional visualization research. Acknowledging the importance of embodiment and affect also has implications for how we evaluate visualizations. Not simply about accomplishing a particular task, could we include measures of embodied and affective responses to visualizations as indicators of their effectiveness?

**Design Process Questions:** How can we leverage embodied and affective experience to enhance visualization design and engage users? What kinds of expertise might we need on our design team in order to do that? (e.g. fine art, graphic design, animation, or communication specialists)

**Design Output Questions:** What kinds of embodied and affective experience has meaning to end users? Should we consider tactile, experiential, or social ways of accessing the data visualization? Can we consider visualization outputs in an expanded field, such as data murals [7], data sculptures [1], public walks [2], quilts [48] and installations [63]?

3.6 Make Labor Visible

Information design processes often start with data, but a feminist approach would insist that they begin by working backwards to surface the actors (both individual and institutional) that have labored to generate a particular dataset. Starting with questions of data provenance helps to credit the bodies that make visualization possible – the bodies that collect the data, that digitize them, that clean them, and that maintain them. However, most data provenance research focuses on technical rather than human points of origination and integration [66]. With its emphasis on under-valued forms of labor, a feminist approach to visualization can help to render visible the bodies that shape and care for data at every stage of the process. This relates to the concept of *provenance rhetoric* [44] in which authors of narrative visualizations cite data sources and methods which may help build credibility with the audience. Making labor visible also has implications for fair attribution and credit for the resulting artifact, especially in light of the fact that women and other underrepresented groups have been notoriously excluded from sharing in credit for scientific work [69].

**Design Process Questions:** Can the team work backwards from the given data to document their provenance and talk to their caregivers? Has the team discussed roles, responsibilities, and credit in advance of publication?

**Design Output Questions:** Is it feasible to provide a metadata visualization that shows the provenance of the data and their stakeholders (caregivers) at each step? Have we properly attributed work on the project?

4 CONCLUSION AND NEXT STEPS

In this paper, we have outlined six principles for feminist data visualization: Rethink Binaries, Embrace Pluralism, Examine Power and Aspire to Empowerment, Consider Context, Legitimize Embodiment and Affect.
and Make Labor Visible. These are preliminary and offered for the purposes of beginning a dialogue about how the digital humanities and information visualization communities can productively exchange theories, concepts, and methods. Applying humanistic theories to design processes and artifacts may be new territory for many humanists, just as grappling with questions of subjectivity, power, and oppression may be new territory for many visualization researchers. As data visualization becomes a mainstream technique for making meaning and creating stories about the world [70], questions of inclusion, authorship, framing [44], reception, and social impact will become increasingly important. In this regard, the humanities and specifically feminist theory have much to offer.

ACKNOWLEDGMENTS


Network visualization: mapping Shakespeare’s tragedies

Are Shakespeare’s tragedies all structured in the same way? Are the characters rather isolated, grouped, all connected?

Narration, even fictional, contains a network of interacting characters. Constituting a well defined corpus, the eleven Shakespearean tragedies can easily be compared: We propose here a network visualization in which each character is represented by a node connected with the characters that appear in the same scenes. The result speaks for itself: the longest tragedy (Hamlet) is not the most structurally complex and is less dense than King Lear, Titus Andronicus or Othello. Some plays reveal clearly the groups that shape the drama: Montague and Capulets in Romeo and Juliet, Trojans and Greeks in Troilus and Cressida, the triumvirs parties and Egyptians in Antony and Cleopatra, the Volscians and the Romans in Coriolanus or the conspirators in Julius Caesar.
SHAKESPEAREAN TRAGEDY

TITUS ANDRONICUS
Number of characters: 36
Network density: 50%

ROMEO AND JULIET
Number of characters: 41
Network density: 37%

JULIUS CAESAR
Number of characters: 46
Network density: 34%

HAMLET
Number of characters: 37
Network density: 39%

TROILUS AND CRESSIDA
Number of characters: 35
Network density: 40%

OTHELLO
Number of characters: 24
Network density: 55%

KING LEAR
Number of characters: 33
Network density: 45%

MACBETH
Number of characters: 46
Network density: 25%

TIMON OF ATHENS
Number of characters: 51
Network density: 25%

ANTONY AND CLEOPATRA
Number of characters: 74
Network density: 17%

CORIOLANUS
Number of characters: 63
Network density: 21%

ABOUT
Shakespeare's plays listed chronologically. Two characters are connected if they appear in the same scene. Their size and color intensity are proportional to their weighted degree of centrality. The "network density" measures how close the graph is to complete. A complete graph (100%) has all possible edges.

SHAKESPEAREAN TRAGEDY (The 11 tragedies in chronological order)

ABOUT

Two characters are connected if they appear in the same scene. Their size and color intensity are proportional to their weighted degree. The ‘network density’ measures how close the graph is to complete. A complete graph (100%) has all possible edges between its nodes.
**ROMEO AND JULIET**
Number of characters: 41
37% Network density

**TROILUS AND CRESSIDA**
Number of characters: 35
40% Network density

**MACBETH**
Number of characters: 46
25% Network density

**CORIOLANUS**
Number of characters: 53
21% Network density
JULIUS CAESAR
Number of characters: 46
34% Network density

OTHELLO
Number of characters: 24
55% Network density

TIMON OF ATHENS
Number of characters: 51
25% Network density

HTTP://WWW.MARTINGRANDJEAN.CH/NETWORK-VISUALIZATION-SHAKESPEARE/
Comments

by Michael Hunger on 08/01/2016 23:55

Hi Martin,

Really impressive visualization and comparison.

would it be possible to get your raw network data. This would make an awesome Neo4j GraphGist. (You could also decide to participate in our winter challenge yourself :) http://neo4j.com/blog/neo4j-graphgist-challenge-may-the-graph-be-with-you/ 

by Chris Kapilla on 10/01/2016 00:06

This is awesome -- thanks for all the work you did putting it together!
Do you have any versions where the names of some of the more minor characters are displayed?

• by Martin Grandjean on 10/01/2016 11:22

Something like that :
For graphical reasons, I choose not to display the names on the published version, but I of course have the data behind. For what purpose do you need this? It obviously requires me some time to produce.

by Chris Kapilla on 13/01/2016 00:16

thanks, that's exactly what I was looking for. I don't need it for a particular purpose, I just think it adds a lot of depth to see how the more minor characters come into play.

by agradalahranada on 30/03/2016 18:11

Thanks for the effort it takes to create this amazing graphs. I would like to write an article in Spanish using them and publish it next month for the Bard's death anniversary, the version with all characters names would be required and, of course, you would be mentioned on the credits. Please contact me.

by Dina Terloyeva on 17/04/2016 04:37

Without names these graphs are just circles connected with lines. Even if the graphics are at stake, it would be great to see the names in the circles.

by Martin Grandjean on 19/04/2016 09:19

You're right. In fact, this visualization was just an experimental project, to show the interest of comparing many very similar networks, at a "global" level. Adding all the names would have been visually very heavy/unreadable for this purpose.

by Dina Terloyeva on 17/04/2016 04:41

Amazing work! Thanks for sharing. Have you tired to map his comedies? Is their network density different from the tragedies?

by Martin Grandjean on 19/04/2016 09:23

No comedies yet (as I wrote, this was just a quick experiment), but as this post drove lots of interest, I may produce other Shakespeare graphs soon.

by kpsssport on 10/01/2016 18:26

Hi, just wondering how you chose the degree weight?

by Martin Grandjean on 10/01/2016 18:36
I'm not sure to understand your question. The "weighted degree" measures the number of (weighted) connexions a node has. If character A appear with character B 3x, with character C 1x and with character D 5x, he will have a degree of 3 and a weighted degree of 9.

by kpfssport on 10/01/2016 20:58

Thank you, that was the exact answer I was looking for. I wasn't sure if it was for total number of times the character shared a scene (which it is) or it was per scene shared by each pair of characters.

by Milena Doseva (@millalenna) on 13/01/2016 08:40

- Beautiful, indeed! What data did you use for it? is it connected to http://shakespeare.acropolis.org.uk/ projects ?

by Martin Grandjean on 13/01/2016 19:12

Thanks! No, this small project is not connected to any other major research. Data was manually coded from Shakespeare plays.

by Maddy on 13/01/2016 18:30

- I am studying Othello at school... Can you let me know who the character is on their own in the graph for Othello?

by Martin Grandjean on 13/01/2016 19:18

by Molly Pfaff on 14/01/2016 04:44
Oh my goodness, this is amazing! If you happen have the spare time and don't mind another request, I would love to see a complete version of the Hamlet chart, if only because I'm slightly obsessed with that play. Regardless, thanks for sharing! It's really fascinating.

by Martin Grandjean on 17/01/2016 23:08

Hamlet is a very dense network:

by Molly Pfaff on 23/01/2016 07:46

Thanks so much! :)

by Sara Azul on 14/01/2016 20:18

Are you planning to do that to the comedies as well?
by Martin Grandjean on 15/01/2016 11:22

No, as it is a side project I chose a restricted dataset.

by Chris Pudney on 21/01/2016 06:46

Very nice - I applied a similar technique for that other great work of literature: The Simpsons!


by Martin Grandjean on 21/01/2016 11:21

Great work! the "single episode" clusters are indeed very clear.

by sfasdas on 22/01/2016 14:06

muy buen trabajo, pero seria bueno que publies el dataset para poder profundizar el estudio, saludos!
Dear Mr. Grandjean,
Could you tell me what kind of software u used in order to make these graphs? I find them very interesting and would like to do something similar for my Master's thesis in Dutch literature.
Best wishes,

Tessa

Hi, I'm using Gephi (tutorial here)

Hi,

This is so cool! I was recently an Angus in a production of Macbeth where, just for fun, we hand-drew our own network diagrams of the characters. I love how high Ross's weighted degree is to the network: makes sense, as he is the king of information transmission.

If you again happen to have the time and don't mind another request, I'd really love to see the version of Macbeth with all the nodes labeled.

Hello

The graph you display are computed from the text. But you can apply some algorithms to "extract" information, for example;
- Have you tried to apply some "communities computation algorithms" in these graphs?
- Or to apply the Pagerank algorithm (or any similar algorithms)

Thanks.

Robert.

Right, the purpose here was more to provide a graphical overview than quantitative insights. In terms of communities, I fear that on such simple graphs, the result wouldn't be a revolution for literary scholars ;) The small size of the graphs would also produce very predictable Pagerank results, but it's clear that I'm applying these methods to much bigger datasets, or in cases where precise hypothesis are formulated.

what is the point without the actual data?

Can you share the data? Do you think it would be more precise to do it based on average proximity instead of
count of appearances in scene?
Did you find any interesting on top of density?
Do you run a script for network generation?

by Tanya Qc on 23/09/2016 08:19

○ What did you use to map these beautiful graphs?

by Martin Grandjean on 23/09/2016 10:19

I used Gephi.

by jade on 31/10/2016 09:00

Hi, I love your compilation of networks, and was thinking of analysing the network of Romeo and Juliet in depth -- would it be possible for you to share your data on that particular network? (e.g. labels of nodes/individual edges/individual weightedness/ how you chose to weight them)? Thank you!! :)

by Martin Grandjean on 14/11/2016 12:21

Hi, thank you for your message. Could you please write me an email (see the "contact" tab) so that we can discuss these questions further?
How to Read a Literary Visualisation: Network Effects in the Lake School of Romantic Poetry


Published on 27 Aug 2013 / CC BY 4.0

Abstract

Robert Southey, a member of “the Lake School” of poetry—really, the first “avant garde” in the history of English literature—was as prolific a letter writer as he was of poetry and prose during the Romantic era in England, roughly 1780-1830: “to write to a dear friend,” Southey says, “is to me like escaping from prison.” A massive digital scholarly edition is underway, The Collected Letters of Robert Southey. It is divided into eight Parts: I:1791-1797, II:1798-1803, III:1804-09, IV:1810-15, V:1816-21, VI:1822-27, VII:1828-33, VIII:1833-39. Currently Parts I and II have been completely edited and made available to the public. These letters capture a set of intellectual, amicable, and financial relationships established while Southey lived in and traveled away from his home base of Bristol in “the West Country,” before he actually moved to the Lake District. Because each person and place name in the letters was encoded using TEI P5, we have been able to create a data set that indicates who is mentioned in letters to whom. That data set was fed into a Directed and Undirected Graph to be visualized, which is available to see and manipulate online here (http://dhhub.org/demos/voyeur/).

French abstract

Robert Southey, a member of "the Lake School" of poetry—really, the first "avant garde" in the history of English literature—was as prolific a letter writer as he was of poetry and prose during the Romantic era in England, roughly 1780-1830: "to write to a dear friend," Southey says, "is to me like escaping from prison" (Southey 2009, I.141). A massive digital scholarly edition is underway, The Collected Letters of Robert Southey. It is divided into eight parts: I:1791-1797, II:1798-1803, III:1804-09, IV:1810-15, V:1816-21, VI:1822-27, VII:1828-33, VIII:1833-39. Currently Parts I and II have been completely edited and made available to the public. These letters capture a set of intellectual, amicable, and financial relationships established while Southey lived in and travelled away from his home base of Bristol in "the West Country" before he actually moved to the Lake District. Because each person and place name in the letters was encoded using TEI P5, we have been able to create a data set that indicates who is mentioned in the letters and to whom each letter is addressed. That data set was fed into a Directed and Undirected Graph to be visualised, which is available to view and manipulate online here: http://idhmc.tamu.edu/relate (a quick caveat: if you see nothing in your browser, please try using a different one. It works beautifully in Chrome 22, Safari 5.1.7, Firefox 9.0.1, and presumably anything higher).

Experts are now busy graphing social networks that have been created by people via e-mail or Facebook. Our tool called "Relate," built using the Java Universal Network/Graph Framework (as described below), allows us to graph early social networks created by letters. We used Relate to graph all of Southey's letters. In the dynamic, malleable graph of these letters, uni-directional and bi-directional edges, representing relationships, connect nodes that represent the people to and about whom Southey wrote. Relate visualises the number and kinds of connections that were established among people by Southey's letters. These 877 letters spanning the years 1791-1803 graph a social network of the era. The Relate Graph tells us a lot about Southey's relationships as well as gestures toward what other kinds of data it might be useful to have. But it also prompted me to ask two questions that will be discussed in this essay: first, what are the protocols for reading graphs? The second question has two parts: a) what kinds of new knowledge can be generated by visualising textual data? b) If visualising information helps experts (literary experts in this case) perform deeper research,
then what precisely is the new relationship between experts and tools once tools become, like the Relate dynamic graphing tool, "smart tools?"

In the process, I will argue here that visualising Southey’s letters suggests a new hypothesis about high Romantic literature, knowledge unavailable to the Romantics, from whose self-understanding we have perhaps taken too much instruction. These letters clearly demonstrate something "we"—i.e., disciplinary experts in Romanticism—have always known: that the friendships surrounding Southey, Coleridge, and Wordsworth constitute a "network of ambitious young writers" in the 1790s-1810s (Pratt 2000a, 317). But information visualisation leads us also to hypothesise that high romanticism or the publishing history of Romantic-era writers of the Lake School, may in fact be a "network effect." The network effect has been defined by economists as increasing the value of something through wide adoption of it. Facebook is an example, but so is the telephone or e-mail, two services that are only valuable if other people also have telephones and can get e-mail. Our capacity for seeing Romanticism as a network effect rather than the achievement of specific individuals is augmented not simply by graphs but by the disciplines that have grown up around reading them.

For those not yet familiar with the field of Information Visualisation, it "has emerged as a new research discipline" beginning in the 1980s (Keim et. al. 1998, 160). Despite an outpouring of publications in the field between the first and second editions of Chaomei Chen’s Information Visualization (1999, 2006), Ben Shneiderman called it in 2006 a "still emerging academic field" (qtd. in Chen 2006, xi, vii). Colin Ware defines Information Visualisation as "the use of interactive visual representations of abstract data to amplify cognition" (2004, xvii). Before discussing how precisely visual representations might do such a thing, I want to address the bugbear in this statement for literary scholars, viz. "abstract data."

Abstraction in general gives humanists some pause, especially the theoretically inclined who worry about universalising generalisations that hide power plays beneath the surface of abstraction: too often the "Man" starring in Enlightenment discourse is not humanity in general but white, male, Western, and even aristocratic. Normalising, according to Him, as a standard is one way of imposing His form of life upon hapless others who differ. But abstraction itself is a cognitive tool susceptible to being deployed in any kind of politics (Eagleton 1997, 1).
Abstraction via quantification is a process that is key to visualising information about or from texts. Qua "scientific reduction," it has been the focus of the "distant reading" debates surrounding new literary criticism that involves using data-mining techniques to analyse literature; techniques such as those used by the Stanford Literary Lab (http://litlab.stanford.edu/) or by Michael Witmore (http://winedarksea.org/), Robin Valenza (Dr. Valenza’s Visualizing English Poetry Project, supported by Mellon funding to the University of Wisconsin, has not yet been released, as of this writing), and Ted Underwood (http://tedunderwood.wordpress.com/). Whether you side with Franco Moretti or Katie Trumpener, however, one has to admit that reduction is a part of the process of thinking that is impossible to avoid. It is by now a truism that, unless the map is equivalent to the country, the map abstracts data from the country that it represents. But more than that, the country can only be thought—can only enter into anyone’s mind—as a reduction of itself. Although the mind may be wider than the sky, no portion at all of any physical landscape can be put inside a person’s head, forced into a synapse, if that’s where thinking takes place.

In his monograph *Graphs, Maps, and Trees*, where Moretti graphs the world-wide publication history of novels, he points out that his abstractions—dots on the graph—are indeed reductions, but so, he says, are the words "The" and "Novel" in the literary commonplace "The Rise of the Novel." I agree with this—it was a revelation to me—but I also agree with what literary scholar Andrew Stauffer said to me upon seeing one of my graphs: "I know what to do when I see words on page; when I look at this graph, I don’t know what to do, I don’t know how to read it." This candor is incredibly valuable; the mistakes that one can make without knowing how to read data visualisations are tremendous, and I will show some of them while introducing disciplinary instructions for reading visualisations of data. The benefits of rigorous interdisciplinarity, in which the advances made by one discipline are understood and appropriately adapted by another, are valuable both for archiving literary texts and better understanding literary movements.

1. What Are We Seeing?

One is tempted to read, without any training in the field of Information Visualisation, the Google Ngram viewer, showing us the use of the word "presumption" in the graph in Figure 1.
The word is suddenly used very frequently around 1815-20, right around the time that Mary Shelley first published the novel *Frankenstein*. It is obviously Victor who embodies presumption, presuming that a human can take on the powers of a god to give and take life. It is so obvious, in fact, that the first stage rendition of the novel was titled *Presumption*. But a major principle of Information Visualisation is that the first thing we will see when we look at a visualisation is "errors" in our data: "With an appropriate visualisation," Colin Ware says, in one of the most influential textbooks in the field, "errors and artifacts in the data often jump out at you" (2004, 3). All the jokes about the Google Ngram viewer are by now commonplace; we know that it cannot "see" the long ‘s,’ and so, when we search for "prefumption" as well as "presumption," we get a graph that looks like the one in Figure 2.

**Figure 1: Use of the word "presumption" in a sampling of texts published between 1700 and 1900.**

**Figure 2: "Prefumption" in red; "presumption" in blue.**
But rather than calling what we see here "errors" in data, we should see it rather as information about how the data is structured. In this case, we are seeing information; it is not, however, information about the use of the word "presumption." Rather, it is information about the history of typography—and excellent information at that. If one overlays the graphs returned in the Google Ngram viewer for searches of presumption, prefumption, case, cafe, son, fon, curiosity, curiosity, one can see an amazing consistency in pattern (Figure 3).

**Figure 3: Overlay of n-gram searches for prefumption/presumption; cafe/case; son:fon.**

What we see here is that the long-‘s’ was no longer technically necessary from around 1790, and printers began to use type containing short ‘s’ around 1800; however, they still had a lot of older type left over that they wanted to use. Indeed, if you examine the 1798 edition of Wordsworth and Coleridge’s *Lyrical Ballads*, you can see a mixture of long- and short-‘s’ (Figure 4).

**Figure 4: Mixture of long- and short-‘s’ in 1798 text.**
However, by the 1800 edition, the long-‘s’ has been removed (Figure 5).

Figure 5: Long-‘s’ mostly removed in 1800.

But from delay the summer calms were past.
On as we drove, the equinoctial deep
Ran mountain high before the howling blast.

Though publishers changed between the first and second editions, the printer (Nathaniel Biggs of Bristol) and the typeface remained the same (these snippets come from "The Female Vagrant," p. 75 of the 1798 editions [Bristol and London] and p. 73 of the 1800 "second edition" published by Longman. Wordsworth and Coleridge n.d.).

Colin Ware’s stricture that visualisations make data "errors" salient is one that we wish to keep in mind, therefore, as we look at the Southey Letters, we want to think first and foremost about how we might be seeing, not information about Southey’s network of friendships but, something about the data itself—not printing history, in the case of these letters that have been transcribed and coded by hand, but something about the structure of the archive of Southey’s letters. Another moral of the Ngram story is that we do want to follow disciplinary procedures for understanding what we see.

The protocols for reading graphs are laid out in graph/network theory. I will adumbrate these principles while illustrating them in the case of the Southey Letters as loaded into and manipulable within the Relate Tool. This tool was built by an undergraduate (at the time), Jon Jekeli, under the direction of Computer Science Professor Gerald Gannod of Miami University. Jon built it using JUNG (http://jung.sourceforge.net/), the Java Universal Network/Graph Framework. It was Dr. Gannod’s idea to map social networks in the letters—an idea then prominent in computer science research and prompting the building of tools such as Protovis, a JavaScript-based toolkit from Stanford. Travis Brown, Assistant Director of Research and Development at the Maryland Institute of Technology, made the tool viewable on the internet, pre-loaded with data from two Romantic Circles editions: The Letters of Robert Bloomfield and The Letters of Robert Southey. I (Laura Mandell) wrote the XSLT transforms to extract names from these TEI-encoded documents. XSLT, or eXtensible Stylesheet Language Transformations, work with XSLT processors to transform XML files into files of other sorts, HTML for viewing on the web,
text files, database tables, etc. (see http://www.w3schools.com/xsl/default.asp). Travis Brown is working now to make the tool available on the web for use by others using any data set simply by putting it into a specific XML or plain-text format. Mandell will provide XSLTs to use on any TEI-encoded texts, but named-entity extractors could be used as well. This tool is currently available at http://idhmc.tamu.edu/relate, for use on Bloomfield and Southey, and information about how to load data will be available late 2013.

When one first goes to the Relate Tool, one sees, by default, information about Southey's letters to his family. Because it reflects the letters ONLY to family members, the data set initially looks very small. After first arriving at the site, I will first switch to show non-family members, as you can see me doing in Figure 6, using the drop-down menu called Options > Show:

**Figure 6: Default view of Robert Southey's letters; names of family members written to and about in letters written between 1791-1803.**

After I select "Non-Family Only" instead of "Family Only," I see the result pictured in Figure 7.
Figure 7: Non-family members written to and mentioned in letters written by Robert Southey, 1791-1803.

And it is here that I will begin to explain the protocols of reading graphs and demonstrate how to manipulate the views in order to "see" various kinds of information about Southey’s correspondence network.

In the *Semiology of Graphics*, Jacques Bertin tells us that reading a graph begins by identifying the invariant or common ground relating all the elements in the graph (1983, 5). Here, what you see are the people who are mentioned by Southey in letters, either through direct address (the letter is written to them) or through being mentioned in the body of the letter. Not every name mentioned in the letters appears, but only those chosen by the editors to include in the biographies of people associated with the edition. A name like "King George" for instance, would be described in a footnote instead of in this list of brief biographies, which really only contains "information about people with whom Southey is connected" (http://www.rc.umd.edu/editions/southey_letters/people.html). So the invariant or common ground linking all these spheres (people named) and lines (network connections) is this: Southey. It is crucial to remember that Southey himself is not pictured here. Thus, one will not see here how many connections Southey has with his correspondents and the people he mentions in letters. That Southey is connected to them all one or more times is presumed (he is "the invariant"). The strength of Southey’s connection to each person in the network IS visible however, in the number of
times that any name appears in the oeuvre of letters; that is, a name that appears 113 times will have been someone written to and about a total of 113 times. One cannot tell from looking at the graph the breakdown. The person could have been written to 100 times, and mentioned in letters to others thirteen times. Alternately, the person could have been written to five times and about 108 times. In both cases, he or she was on Southey’s mind 113 times during the time-span that these letters were produced. The actual breakdown can be obtained by looking at the list of Southey’s correspondents online.

The variables, the things that one does see figured in the graph, show the connections among the people to and about whom Southey writes, the network of relationships established by and reflected in the letters. A connection is established between two people when one person is mentioned to another. There are two kinds of variables: spheres and edges. Spheres represent people, and they are bigger and a specific colour depending upon how many connections a person has to other people. Each sphere or person is opposed to other spheres or people in the perimeter of the circle (or later, wherever you drag them), and he or she is connected by an edge to one or more of those opposing spheres (people). The spheres vary in both size and colour, which are in fact correlated, except when colour is changed dynamically, by clicking on a sphere—that is, when you are using the tool, you may want to highlight one network, and when you click on a sphere, it and those connected to it change colour. The larger the size of the sphere, the more relationships that person has with other people in the network.

The connections between people (spheres) are represented by straight lines called "edges." The edges are either solid—which means that the connection goes both ways ("bi-directional edges")—or dotted lines, with an arrow on one end ("directed edges"). These directed edges are counter-intuitive to a degree: the arrow points from a person who is addressed in a letter toward the person who is mentioned in it (pointer=person to whom letter is addressed; pointee=person mentioned). I labelled the spheres in the previous image according to number of connections (Figure 8).
Small turquoise spheres have one connection, larger gray ones have four to six connections, etc., up to the largest white sphere, which has eighty-two connections. In Southey’s network during the 1793-1802 time period, there are no people who have twelve or sixteen relationships, just by chance.

As mentioned earlier, this colour scheme changes dynamically: when you click on one of the spheres, it turns red and the name of the person it represents appears (Figure 9).

Another new colour has appeared in the screen capture above: yellow. Yellow marks the spheres representing the people with whom the clicked red sphere is connected—here, there is only one connection. In Figure 9, Elizabeth Smith (red) is connected to the big yellow sphere (we would have to make the mouse hover over it to see who it represents). In Figure 10, one can see the same dynamic
response to clicking on a sphere; in this case, the sphere representing a person has more than one connection.

*Figure 10: Dynamic Response to selecting one sphere.*

Grosvenor Charles Bedford—large and white before clicked, but red in *Figure 10*, after clicking—has eighty two connections, and, if you count the yellow spheres, you would find almost eighty two people to whom he is connected. Some of the eighty two might be family members, in which case they wouldn’t be pictured here because we have selected "non-family only" for viewing in the dynamic graph. Were the image to show both family and non-family, we could count eighty two yellow spheres, Bedford’s eighty two connections.

In both the case of Elizabeth Smith, with her one connection (*Figure 9*), and the case of Grosvenor Bedford, with his eighty two (he’s the largest; *Figure 10*), one cannot see the edges that connect them as the graph is currently configured. There is in the centre of the graph a massive mesh of gray lines, some solid, some dotted. We’ll shortly start manipulating the graph so that you can see the edges and get more information from them.

On the right-hand side of the screen, the lower box has the title "K-Cores." The number of K-Cores is the number of relations that any given person has.
Figure 11: K-Cores.

The K-Core, sphere size, and unclicked colour all represent the number of relationships with other people; Grosvenor has relationships with eighty two others (represented by spheres). Again, the relationship is constructed through Southey’s writing of the letters. A letter either indicates that there is in fact a real relationship between the person to whom it is addressed and the person mentioned in it ("remember me to Duppa," Southey will say to Grosvenor (2009, i.108), or the letter itself will establish a relationship between the person to whom it is addressed and the person mentioned, even if that relationship is only cognitive. So for instance, Southey writes to Horace Bedford that the second edition of his epic Joan of Arc ought "to be dedicated to Mary Wollstonecraft" (2009, i.160), thereby ensuring, presuming the letter was received and read, that Horace has thought about Mary whether in fact he ever physically met her or not. Southey also mentions his epic in connection with Charlotte Cordé (Corday), but the editors of the text have written a footnote explaining who she is rather than encoding her name and linking it to their biographies section. That editorial procedure tends to eliminate mere mentions of unrelated people from the network. Southey writes to Grosvenor Bedford about Plato in Letter 77, for instance, but thankfully Plato is not figured here as a person in his network. Some of the eighty two people have been mentioned to Grosvenor and thereby are connected to him many times. But the K-Core registers not the number of mentions of other people, but rather the number of people,
i.e., the number of people to whom he or she is connected by all the letters in the edition, up through Part II (1803), and the living breathing people who are more or less in or out of this network.

The K-Core boxes are checkable and uncheckable, and we will use this feature to reduce the number of relationships viewed at any time so that we can see the edges. Beginning at K-Core 1, people who have only one connection with one other person and so are very inconsequential members of the network, we uncheck the boxes, moving up in number from only five relationships and up, to nine or more (Figures 12 and 13).

**Figure 12: Unchecking K-Cores.**

![Unchecking K-Cores](image)

**Figure 13: Unchecking K-Cores Continued.**
Again, what we are seeing here is anyone (represented by a sphere) who has nine or more relationships as indicated by the K-Core boxes checked (nothing below nine), the colours (a subset of green, and then all pink, blue, and white), and the spherical sizes. To get a better look at the edges and see who is related to whom and how, we can begin to click on and move the spheres around the workspace (Figure 14).

**Figure 14: Graph with 11+ K-Cores, manually manipulated.**

Among the lower number of connections, blues and pinks (eleven to fifteen), one can see that the edges are most often directed, and I will explain why and what that means momentarily. It must be
emphasised that this is a dynamic graph. Apart from the default positioning of things, spheres and edges can be moved all over to get a better look at the relationships. You will see many screen captures of the graph, and they will almost all differ slightly, freezing one moment in my work with the graph. And other things begin to happen as one moves spheres around on the screen. First, let me show what happens when you click on any one of the spheres in the dynamic graph below (Figure 15):

**Figure 15: Bedford’s Relations.**

When you click on Horace Bedford, all the people who have been mentioned to him and/or to whom he has been mentioned (i.e. all his relationships) turn yellow. We see only ten yellow spheres here, whereas the tag tells us that he has "29 Connections." Again, if all the family and non-family spheres were visible, we would see twenty-nine yellow spheres, but we have selected non-family K-Cores of eleven and over, and among that set, Horace has ten relationships. So Mary Wollstonecraft, to whom he is connected via Southey’s letter saying that he wished he had dedicated Joan of Arc to her, is not one of these yellow spheres because her K-Core is eight:

**Figure 16: Mary Wollstonecraft’s 8 Connections.**
To find a person’s K-Core, one can select all K-Cores, and then scroll through the list of "Families"—the box above K-Core—selecting first "none," and then checking only the surname wanted. The number of K-Cores corresponds to the number of people to whom Wollstonecraft is mentioned plus the number of people who are mentioned to her, if indeed Southey wrote her letters.

Being "written about" and "written to" are two variables that are partly visible in this dynamic graphic via the edges. Figure 17 is a graph of non-family people with K-Cores of ten or above, and I have clicked on an edge between Charles Danvers, Southey’s neighbour in Bristol and longtime friend with whom he stayed while writing his second epic, Madoc, and Charles Biddlecombe, a neighbour in Burton, where he lived temporarily.

**Figure 17: Directed Edge.**

Here, the edge is one way: Danvers has been mentioned in a letter to Biddlecombe, but not vice versa. There is one connection between the two. There is a difference between the meaning of the number of "connections" marked on spheres, which is the number of people with whom that person has connections, and the meaning of the term "connection" as it is used on edges, which is the actual number of mentions. As seen above, Danvers has only been mentioned to Biddlecombe once, and so I surmise that Southey asked Biddlecombe to send something of his to Danvers, if not simply mentioning him in conversation. I checked my hypothesis by going to the Correspondents List, the list of people who
are mentioned in letters, and found a letter to Biddlecombe in which Danvers is mentioned (2011, ii.598); and indeed, Southey is telling Biddlecombe to send the money he makes from the sale of Southey’s furniture to Southey’s mother via Charles Danvers’s address. Now we’ll examine a bi-directional edge (see Figure 18). Some (not many) letters

*Figure 18: Undirected Edge.*

Here I have clicked on the solid line connecting William Taylor, the German translator whom Southey met at Great Yarmouth in 1798, and Humphry Davy, whom Southey knew from living in Bristol where Davy worked under Thomas Beddoes Sr. at the Pneumatic Institute. The solid line means that the edge is bi-directional, that they have both been mentioned in letters to each other, and that there are seventeen connections between them. This means that Taylor was mentioned to Davy X times, and Davy to Taylor Y times, where X + Y = 17, (in this case, as I know from the Correspondents List, Taylor is mentioned to Davy once, as "William Taylor, the all-knowing" [Letter 454, 12 Nov. 1799], and Davy to Taylor sixteen times). However, we cannot see in the edge itself the distribution—we cannot tell from looking at the graph the asymmetry of Taylor’s relationship to Davy that Davy is being discussed with Taylor, and not vice versa. Taylor is Southey’s trusted intellectual advisor, and he discusses with him
Davy’s genius as a chemist and his potential as a poet. Apparently, both Taylor’s judgement of Davy’s early poetry and Davy’s own propensity to put his work at the Pneumatic Institute before all else, scuttled his poetic career.

Now that we know the invariant, the variables, and what we are seeing when we look at any particular symbol—the minimal requirements, according to Bertin, for reading a graph accurately—we can look at and begin actually reading the K-Core 10+ graph, where reading means discovering information that is prominently visible when the relationships are represented in graphical rather than textual form. It is important to be able to see the spheres and edges without clicking on any or all of the spheres (the only way to see the names of the people involved) so I have added numbers and a name key to a screen capture of the graph in order to expose the name labels that the Relate tool will give users as they manipulate it by clicking on the spheres (Figure 19).

**Figure 19:** Name Key added to screenshot of dynamic graph.

Here again the blue spheres are ten to eleven relationships, the pink thirteen to fifteen, and the white seventeen to eighty two. Among the lower number of connections, blues and pinks (K-Cores 11-15), one can see that the edges are most often directed. It may even be the first thing one notices when looking at this graph, that there are a few spheres that have many arrows pointing at them (1-5, 8). This visually
The salient feature is significant as it indicates that these people are being mentioned in letters often, but other members of Southey’s social network are not being mentioned to them. So the question is: why? Except for Mrs. Danvers and Southey’s aunt, Mrs. Tyler, the names of those mentioned are illustrious: head of the Bristol Pneumatic Institute and father of the physician poet, Thomas Beddoes Sr.; William Wordsworth; publisher Thomas Longman; and George Burnett, the only relatively unknown male member. I moved the diagram around a bit to get a better look at Longman (Figure 20).

*Figure 20: Detail of Figure 19* (achieved by moving Thomas Longman’s sphere).

My first inclination in looking at these results was to think that these directed arrows betoken elitism on Southey’s part. He drops the names of people like Longman frequently in letters, and when he writes to such people, says little to them that is personal, and little about his own social network, wanting them to pay attention only to himself. Such a reading of the graph is irresponsible and not supported by the protocols for reading graphs, no more justified than free-associating on what or whom Wordsworth might mean by "Lucy."

A major principle of visualisations is, again, that the first thing one sees when one looks at a visualisation could be "errors" in the data. Upon investigation, I noticed first that many of these people have something in common. Mrs. Tyler and Mrs. Danvers share the fact that members of the network lived with them—Southey and Charles Danvers respectively. In fact, somehow Mrs. Tyler has not been properly designated as "family," revealing either a coding mistake or a problem with the tool. It may not be properly separating family from non-family, for which one would need to submit a bug report. When one looks at the huge number of mentions of Mrs. Danvers by going to the "mentioned" section of The Correspondents List, one finds them indeed in letters to Danvers, and a cursory glance reveals why: Southey almost always ends his letters to his good friend by mentioning her—"Our love to Mrs. Danvers" (*2011*, ii.644). If Mrs. Danvers is always sent wishes through her son Charles, with whom she is
living, then the letter is not so much establishing a relationship between Charles and his own mother as it is between Southey and Charles’s mother. Finally, neither Beddoes, Burnett, nor Longman, the other people who are mentioned to others but not addressed directly, appear on the list of correspondents in Pratt’s edition, which “only names correspondents where one or more letters from Southey to that individual survives” (2009, Correspondents). Southey’s letters to Beddoes, Burnett, and Longman during the period 1791 to 1803 do not survive. We have no record of whom he mentioned to them, only mentions of them in letters to others. From this dynamic graph then, we cannot necessarily conclude that the number of directed edges—arrows—pointing at their spheres, the absence of reciprocity in mentioning others to them, comes from undue deference on Southey’s part, as I had imagined. Others may be mentioned to them in letters that are missing.

Thomas Longman’s publishing house suffered a fire during the mid-nineteenth century, and unfortunately letters were lost, though letterbooks containing Longman’s letters to Robert Southey may indeed survive. It is worth here remembering exactly the invariant in this graph. It shows how people are connected in a network that has been created via letters WRITTEN BY Southey, NOT by any letters written TO him. The Longman Archive Online advertises itself as containing “autograph letters from Longman authors 1799-1900 (authors include . . . Robert Southey . . .).” Despite the implication, it is not true that the archive contains autograph letters from Robert Southey, as I have confirmed via e-mail on Friday 17 August 2012 with Editor Lynda Pratt. She writes,

Some (not many) letters [written by Southey] to Longman do survive for later periods — scattered amongst a number of archives and not always easy to identify. Some of these later letters were published in J.W. Warter’s Selections from the Letters of Robert Southey (1856). To get a sense of how Southey viewed Longman, see Letter 1939: "The people at that house know nothing about books except in the mere detail of trade."

(At the time of final editing, the letter mentioned in the above quotation is now available: http://www.rc.umd.edu/editions/southey_letters/Part_Four/HTML/letterEEd.26.1939.html.)

Noteworthy at this moment in my exposition is the emergence of one very important aspect of our first principle for reading visualisations, one that these three “errors” just discussed concerning this data set makes salient. I put “errors” in quotation marks because what it really means here is incorrectly categorized information. We did get one error and now know one thing that needs to be fixed: —the
name Tyler needs to be designated "family," not "non-family," whether through data correction or fixing the tool. But the other two "errors" are not mistakes as much as information of another kind, information not about Southey’s network but about the data itself. Here follows the first part of Ware’s principle as well as the portion about errors once again:

A visualization commonly reveals things not only about the data itself, but about the way it is collected. With an appropriate visualization, errors and artifacts in the data often jump out at you.

"Errors" are one thing that pops to the fore, but so are "things about" data, "artifacts in the data," and information "about the way it is collected." Visualisations give information about content, but also about the structure of the data—it's medium, genre, and form—and therefore need to be read both ways, attuned to the possibility that at any given moment one could be exposed to information about either or both. Moreover, visualisations can be used to investigate the data collection. We now have a way of looking for which letters of Southey’s might be missing and need to be gathered. For example, imagine a person whom Southey mentions frequently in letters to others but to whom he does not mention the names of other people in his social network (e.g., he writes to Coleridge and Cottle about Longman but, when he writes to Longman, he doesn’t mention Coleridge or Cottle). Such a person will appear, because of the directed edges going toward him, as a sphere surrounded by arrowheads. When one sees such a sphere, there are three possibilities: 1) Southey doesn’t mention other people in his social circle to this person for a potentially interesting reason; 2) he doesn’t write to this person (also interesting); or 3) the letters to that person are missing (interesting as a fact about data more than a fact about the relationship between Southey and the person in question). The latter is the first thing to check. The conventions of closing letters are features of the epistolary genre: we now know one feature—this one, "my best to your mother"—is common sense, but even if such features are noticed, we can now discern a way to mathematically formalise this feature and could then search huge data sets for connections among people while discounting those mentioned in closing. Or indeed, we could look for closings in an unstructured data set of texts using that algorithm in order to find among them texts that are letters or are written in the epistolary genre.
2. Stages of Reading

Bertin lists three stages in the reading process
1. EXTERNAL IDENTIFICATION. In this stage, we need to ask, "What components are involved?"
The components are the invariant plus the "variational concepts." We have identified the invariant
(Southey writing letters). The other components are:
a. People named in the letters, either as addressees or as subjects discussed;
b. Number of relationships that have been forged by the letters between people named in them;
c. Type of connections within any given relationship, whether it is reciprocal or exclusive; and
d. The relative strength of any relationship among Southey’s acquaintance’s relative strength as evinced
by the number of connections: in social networks, "stronger ties represent close friendships and greater
frequency of interaction" (Easley and Kleinberg 2010, 48). Strong ties represent friends, and weak ties
represent acquaintances;
e. And size of the spheres representing the centrality of that relationship to Southey himself, whether in
thought (he mentions the person in a letter to someone) or deed (he actually writes to the person).
2. INTERNAL IDENTIFICATION. "By what variables are the components expressed?"
a. People are expressed as "vertices"—also called "nodes"— in graph theory (they are called "vertices"
because they constitute the point at which edges or lines meet). In the Relate tool, however, each
vertex or node is marked by a sphere.
b. The number of relationships is expressed by the size and colour of the spheres, as well as by the
legend on each sphere when it is clicked ("82 connections" on Grosvenor Bedford means that he has
relationships with 82 people) and by the total of yellow spheres that appear when it is clicked. Those
yellow spheres show, in graph theoretic terms, the red (clicked) sphere’s "network neighbors"
(Newman 2010, 7.2);
c. The type and number of connections within a relationship is expressed by solid or dotted lines,
"edges" in graph theory. The solid lines, called "Bidirectional Edges," indicate that Southey wrote to both
people and mentioned the other to each one, a reciprocal relationship. The dotted lines, called
"Directed Edges," indicate that Southey wrote to the person designated by the sphere where the line
originates and spoke of the person designated by the sphere to which the arrow points. The iconology
here is "counterintuitive" insofar as the arrow points AWAY FROM the person written to by Southey and TOWARD the person mentioned in the letter. Solid lines show reciprocal, strong relationships. Clusters of arrowheads around a sphere show that the person designated by that sphere is more talked about than written to. Crucial to understanding this graph is realising that edges indicate not just connections between people but the flow of information from one about the other through Southey. If an arrow points toward a person, the information about him flows between Southey and the person represented by the sphere from which the arrow shoots out. However, if there is no arrow in a connection between two spheres, then we are indeed seeing an information flow insofar as Southey mentions one to the other and vice versa in letters to him or her. The flow of information between the spheres with bi-directional edges always goes through Southey.

d. Unfortunately, our Relate Tool needs to be tweaked so that the relative strength of edges is made immediately visible through the thickness of the lines, at least for undirected edges. Right now, however, one can at least see the number of connections by clicking on an edge (see Figure 20).

3 PERCEPTION OF PERTINENT CORRESPONDENCES. "This perception is always the result of a QUESTION, conscious or not. What are the questions which one can ask in approaching the information?" The two major questions made possible by this dynamic graph are:

a. Does Southey discuss X and Y with each other?
b. Does Southey mention X to Y (with the arrow direction directly contradicting the syntax)?

One can always ask questions prompted by looking at the graph—e.g., how many of the people whom Southey habitually mentioned to each other actually met each other face-to-face? But that question cannot be answered by looking at the graph itself.

Having worked through our first rule—that the most salient characteristics of any graph represents so-called errors in the data—and thereby having recognised some information as telling us something about the dataset rather than about Southey's relationships, we'll look again at Southey's social network from 1791-1803, those who have ten relationships or more (K-Core 10+; Figure 21). Our ultimate goal, to be achieved by asking lots of specific questions, is to divide conceptually the visible features of this network that indicate, on the one hand, something about how a data set is constituted,
and on the other hand, something about Southey’s relationships with various people, as well as the impact of his network building upon their relationships with each other (Figure 21).

**Figure 21: K-Core 10+.**

By looking at specific details concerning how people are viewed here, and by correlating those facts with what we know about Southey’s biography and/or the data set, can we answer the following question: Do the larger spheres, those to whom Southey is mentioning more people and who are being mentioned more often (when the lines are solid), represent greater intimacy or involvement with Southey? It seems so. All the largest, white spheres (numbers fourteen on) can be said to be part of Southey’s inner circle of close friends. The one exception is Biddlecombe, whose numbers are inflated a bit because Southey is writing to this former neighbour in Burton about selling his things. The pink group, from number six through thirteen, are people with whom Southey is doing intellectual work of some sort or another, with two exceptions: 7) Thomas Lamb, really a father figure for Southey, and 10) Charles Collins, a close school-friend with whom Southey stopped corresponding in 1794. And the blue group, one through five, except for Aunt Tyler and Mrs. Danvers, are mere acquaintances. But it turns out that none of Southey’s letters to them survive, according to the Correspondents List. Whereas the missing letters to Longman can be accounted for by the fire, the missing letters here might in fact be...
accounted for by the level of Southey’s importance to these people, and vice versa. According to Lynda Pratt, Southey had a break with his Aunt Tyler that perhaps prompted both of them to destroy any letters written during a time when their ties were stronger (e-mail message to the author, 17 August, 2012). We can see in this edition that some of Southey’s mentions of Wordsworth are derogatory, and we know that Wordsworth resisted seeing Southey and himself as in one "school" of poetry. These people are more mentioned than written to, and they did not save Southey’s early letters to them because of a relatively low level of intimacy.

Lynda Pratt and I have had a discussion about this argument that I would like to relay. Dr. Pratt points to the element of chance in the survival of letters. In some cases,

> later generations who inherited letters that had been preserved did not care for them and therefore destroyed them, perhaps because they contained information thought to be unsavory, perhaps because they were perceived to have no interest or value. We know, for example, that one of Southey’s granddaughters destroyed family papers. Such a bonfire might account for the lack of Southey’s letters to Caroline Bowles from the period of their courtship (pre-courtship letters have survived – so the gap is an interesting one). One thing that is clear is the element of ‘chance’ in the survival/transmission of MSS – and the Southey letters edition is (like all such editions) a wonderful example of this. A key example of ‘missing’ correspondence is that between Southey and Lovell – only two letters survive (and one of these is a fragment). This makes a key, pre-Coleridge relationship very difficult to reconstruct. (E-mail to the author, 17 August, 2012)

One can see the element of chance factoring into how our current data set has been structured, but also causal elements: Southey’s courtship letters to Caroline Bowles were "unsavory" to a Victorian descendant. Dr. Pratt pointed out to me "the chaotic lifestyle of Coleridge, who discarded letters unread" (e-mail to the author 17 August, 2012), which seems to me to be NOT a shaping of the archive by chance but rather another indication of an argument that I will make below, that Coleridge’s capacity for intersubjective relationships has been injured, and that consequently he serves as a break in network relations, disrupting their ordinary flow. Clearly, though, expert knowledge about Southey, Lovell, Bowles, and Victorian grandchildren is necessary for understanding what we are seeing in this dynamic graph.

Similarly, only literary historians can tell us what kinds of people make up Southey’s social milieu. Among the pink spheres are two doctors, and indeed the presence of three doctors in the K-Core
10+ network—Beddoes (1), King (6), and Davy (9)—demonstrates that some of Southey’s thinking and writing is inflected by these literati of the Bristol Pneumatic Institute. Moreover, two artists make up this group: Mary Barker (11) and Richard Duppa (13), suggesting that, as he was writing and planning publications, Southey was thinking about illustrations of them. That turns as we see in the Betweenness somewhat validated by this set of letters in which Southey was trying to get travel recommendations from Duppa, and so often included his name in letters to friends, mentioning that Duppa was helping Southey himself plan a trip to Italy. Mary becomes, during this period in Southey’s life, a friend as well as an intellectual influence. Nonetheless, Southey’s close social network includes practitioners of medicine and the plastic and visual arts, and he was intensely interested—as his network reveals—in both.

One bit of information stands out if one reads the graph with the question: "What kinds of relationships did Southey have with his publishers?" Joseph Cottle, the Bristol Bookseller, published the first edition of *Lyrical Ballads*, Coleridge’s *Poems of 1796, 1797*, Southey’s *John of Arc, Letters from Portugal*, and the *Annual Anthology of 1799*, among other things. In this graph, one can see substantiated the argument of Cottle’s own Reminiscences; Southey is as close with Cottle as with some of his habitual correspondents, his close friends. Both Southey and Coleridge almost always address Cottle in their letters with the same kind of sobriquet, "My dear Cottle." Although Cottle may not be a typical eighteenth-century bookseller, he plays a role in Coleridge and Southey’s lives well beyond any business relationship.

Of the letters that survive from Southey to Longman, which are not yet published by this edition but will appear in forthcoming parts, and even from Southey to Byron’s publisher John Murray (forthcoming), are addressed "Dear Sir." Southey’s relationship to Longman can also be seen through Coleridge’s printed letters (1956). Despite the fire, we do have two letters from Coleridge to Longman, one of which was copied before being sent. The fact of a letter being copied suggests that its importance is legal, having to do with business. Coleridge addresses him as "Dear Sir," not "My dear Longman" (1956, i.654-55). In Wordsworth’s letters, too, we find "my dear Cottle," whereas Longman is "Mr Longman" and "Dear Sir" (*Wordsworth and Wordsworth* 1967, 306, 307, 309). The content of Coleridge’s letters to Longman differ also from those to Cottle. The latter are playful and often quick invitations sent to Cottle at short notice, asking him to send his servant out for food to cook dinner,
requesting tobacco to aid the writing process, and the like (Coleridge 1956, i.156-57). What book history experts know about Longman is visible in this graph of Southey’s social network, despite data loss—viz., that Longman is the first of the truly modern publishers (Feather 2006, 76) who produced books for business. T. N. Longman took his "bookselling" shop from visible inheritance of a family dynasty to "headquarters of an organization" (Briggs 1974, 9). In other words, the structure of the archive to some extent participates in its content.

3. Expert Knowledge

The structured-graph reading adumbrated above—the reading given so far—offers no new expert knowledge. The graph has helped me become more of an expert on Southey than I might otherwise have become by directing my readings through his letters and the letters of Wordsworth and Coleridge, for the sake of comparison. Reading the graph also required consulting Lynda Pratt, the expert on Southey. I have proven in the first part of the essay that it is not possible to read this graph accurately without knowing the details of Southey’s life well enough to be able to distinguish when we are seeing facts about the dataset, facts about Southey and the people in the network, and facts that are really true about both at the same time. But so far the reading has produced no new knowledge for either literary or information visualisation specialists.

We want it to do more. I will in the remainder of this article prove that it is possible to obtain some knowledge from this dynamic graph of Southey’s Social Network, knowledge that is new and valuable in both literary history and information visualisation.

4. Network Principles

Graph theoretic notions have been boosted by studies of social networks, some of them digital, some of them print. I will now demonstrate how several of these concepts offer new knowledge to literary historians and the scientists who study social networks.
4.1. Cliques

In *Networks: An Introduction*, M. E. J. Newman defines these terms as they are used in analysing graphs. "A clique," he says,

*is a maximal subset of the vertices in an undirected network such that every member of the set is connected by an edge to every other. The word ‘maximal’ here means that there is no other vertex in the network that can be added to the subset while preserving the property that every vertex is connected to every other.* (2010, 7.8.1)

The graph of Southey’s social network reveals several cliques according to that definition of it. In order to find these cliques, however, we need to turn off the "Directed Edges" by unchecking that box at the top right of the Relate tool. A clique can only be found in "an undirected network"—that is, the relationships must be reciprocal. Figures 22 and 23 show two of the small cliques that are possible to see in the graph of Southey’s letters.

*Figure 22: One Clique (arrows added to a screenshot).*

*Figure 23: Another Clique (blue arrows added).*
In Figure 22, I have traced a clique linking Coleridge, Danvers, Wynn, and Cottle; in Figure 23, I have traced Danvers, Wynn, Bedford, and Cottle. Again, a clique is defined as a group of people, all of whom are connected to each other. All the spheres in both Figures 22 and 23 do not together constitute a larger clique because Coleridge is not connected to either Bedford or May. However, another small clique including May can be traced in this cluster represented in Figures 22 and 23. Figure 24 shows a "maximal" clique.

**Figure 24: Maximal Clique.**

Figure 24 displays the maximum set of relationships that can be found in which every member is connected to every other. The clique of Southey’s friends involves a Bristol wine merchant (Danvers), a benefactor (Wynn, who gave Southey an annuity of 160 pounds), a businessman (May), a statistician (Rickman), and a publisher (Cottle).

Cottle, Danvers, and Wynn are common to all the cliques illustrated here in Figures 22-24. It makes sense that Wynn would be; one can imagine that Southey writes regularly of his activities to his benefactor. Cottle and Danvers, then? I think it is possible to see them as Southey’s closest friends. In his famous letter of 13 or 14 November 1795, when Coleridge "breaks up" with Southey over his withdrawal from their "Pantisocracy" plan, Coleridge defends himself against accusations of spreading rumours about Southey. He says that he answered the inquiries of friends about Southey’s plans only in general terms, but "To Danvers indeed and to Cottle I spoke more particularly—for I knew their prudence, and their love for you" (1956, i.168). These are the people closely connected to Southey, and
so they are those to whom Coleridge turns in a crisis, the crisis of Southey struggling to determine his future career. Coleridge’s letter confirms what the graph shows, that Danvers and Cottle are indeed the people with whom Southey is conversing with most at this moment (1791-1803, as represented by Volume 1 and 2 of the letters), although they are still not necessarily his closest friends. In an e-mail on 17 August 2012 in response to my query, Lynda Pratt writes about these two of Southey’s relationships:

Should we see Danvers and Cottle as Southey’s ‘closest friends’? I wonder if it’s more complex than this. I think it is clear that Danvers is one of Southey’s closest friends – in fact, even something of an older brother, confidante. Cottle is, I think, a slightly different case. He and Southey are close at this time but their relationship is inflected (infected?) by business and a sense of class difference that never enters into Southey’s relationship with Danvers. Yes, Southey is closer to Cottle than to any of his later publishers. He even has a personal relationship with him. However, they are not quite on the same level, and later events make that clear: Cottle’s Recollections and Reminiscences literally writes Cottle himself back into the lives of the Lake School. Cottle’s structuring of his own archive is interesting here. For example, he habitually numbers and renumbers letters sent to him by his poet-friends suggesting that he considers and recon considers the order and significance of those letters. That’s before we get onto his habit of cutting and pasting letters when he incorporated them into his published memoirs which more drastically restructures his archive.

But if Coleridge’s letter and Lynda Pratt’s communication already describe the importance of these relationships to Southey, then what the graph reveals is not news, but only a way of augmenting what we already know. Visualisations should confirm what experts can tell us, or they may be simply wrong. But if we need to check them via expertise, how can they ever show us anything new? Let me push the graph’s meaning as far as I can before we give into that paradox. The person who is written to most often, Grosvenor Bedford, appears in only one of the three cliques pictured above and not in the maximal clique. Coleridge has no reciprocal connection to Bedford; Southey mentions Coleridge in letters to Bedford, but not vice versa. If one takes the common denominators of the most cliques, including the maximal clique, as defining a person’s closest friends, then the graph argues that Coleridge and Bedford are not as close to Southey as Danvers and Cottle. Coleridge breaks up with Southey—Danvers and Cottle never do. Bedford is a close friend but he is not there living in Bristol and is not part of Southey’s daily life in quite the same way as Danvers and Cottle are until 1803, when Southey moves away from Bristol to the Lake District. But the graph argues more: that one’s closest relationships are as much a matter of a person’s imbrication in one’s web of relationships as they are a
matter of individual choice. Pratt’s analysis shows us that Southey’s feelings for Cottle are infected/inflected by business. Thus, the "love" of which Coleridge speaks is less a sentiment and more the power that Cottle and Danvers have to influence Southey as defined by their places in his social network; "love" is a network effect.

4.2. Degree Centrality and Betweenness Centrality

The number of connections is the default setting for what is shown by the spheres in the Relate Tool, but there are two other settings, Degree Centrality and Betweenness Centrality (Figure 25).

*Figure 25: Detail view of Relate’s Drop-down Menu.*

![Relate’s Drop-down Menu](image)

The two are also visible at the bottom of our tool for any individual when their sphere has been clicked (Figure 26).

*Figure 26: Detail view of the table appearing at the bottom of Relate.*

![Table](image)

What do these measures mean? Degree Centrality reveals which "are the most important or central vertices in the network" (Newman 2010, 7.1). In the table in Figure 25, one sees a list of numbers corresponding to a sphere or vertex (here called a "node," but basically, a person). Calculated according to "the number of edges connected to" a person, the number given on Relate at the bottom of the screen for "Degree Centrality" is the percentage of all possible relationships that this person has. There
are 133 vertices or spheres, and in the Relate tool, at the bottom, called "nodes"—133 people in Southey’s non-family social network. Using the drop-down box under "Options," at the top right of the Relate Tool, I click on "Degree Centrality" rather than number of relationships. In Figure 27, we can see what share of those relationships is had by each person in the clique.

**Figure 27: Degree Centrality of Members of the Clique.**

These decimal points are carried out to many places that hamper our view of percentages (and we will fix this view on the tool), but just looking at two place past the decimal gives us a percentage. At the highest number of relations, Bedford garners roughly 62 percent of the relationships to be had in this network.

Betweenness Centrality differs from Degree Centrality; it "measures the extent to which a vertex lies on paths between other vertices" (Newman 2010, 7.7). It measures, in other words, the shortest paths between people. Newman warns us that: "communications do not always take the shortest path. Nonetheless, betweenness centrality may still be an approximate guide to the influence vertices have over the flow of information between others (2010, 7.7)." The higher the number, the greater a person’s Betweenness Centrality is, as we see in the Betweenness Centrality of Southey’s inner circle (Figure 28).
Though Bedford has the highest number of relationships in Southey’s social network, he has lower Betweenness Centrality than Wynn, Cottle, Coleridge, and Danvers. Danvers indeed has the highest. Coleridge, Cottle, and Danvers seem to be the most important clique members in terms of the flow of information, Wynn less so, and Bedford and May, still less. Again, these three are living and working together regularly in Bristol; it makes sense that information would flow through and between them, and so this Social Network re-affirms the hypothesis that high Betweenness indicates information flow.

Notice that Wynn, Danvers, and Cottle hover at around 45 percent Degree Centrality. That may represent the degree of network imbrication necessary for "love." Too much imbrication in a network (Bedford’s 62 percent) as far as letters are concerned may show that the friend is a confessor but not a participant in the network. One cannot talk about everyone to someone who is close to all those other people, only to someone who is less likely to repeat what he or she hears. That Bedford’s Betweenness Centrality is lower than the others indicates that information flows through him less than through others: Southey writes to him about others, but not to others about him. Given the evidence of the Southey social network, one can propose a hypothesis for graph theory: a high number of relationships coupled with a Betweenness Centrality lower than those with lower numbers of relationships signals a person’s role as confessor—someone who listens to discussions about a network without participating in it.
4.3. Clustering Coefficients and Strong Triadic Closure

In another fabulous introduction to network theory, *Networks, Crowds, and Markets*, David Easley and John Kleinberg define clustering coefficients as a measurement of the probability that two randomly selected friends of A are friends of each other (2010, 44). Apparently, the clustering coefficient is found to be very low in teenage girls who have a high suicide rate (2010, 46). There is a natural tendency towards a strong clustering coefficient—natural unless interrupted by depression—as is evinced by the "Strong Triadic Closure Property": "If a node A [sphere, person] has edges to nodes B and C, then the B-C edge is especially likely to form if A’s edges to B and C are both strong ties" (Easley and Kleinberg 2010, 49).

What this means is that the two friends of one person are likely to become friends if the friendships with the first person are strong. The Strong Triadic Closure property predicts that the break between Southey and Coleridge would not last, because, as one can see from the cliques above, Coleridge has strong ties with many of Southey’s friends who have strong ties with each other, as well as strong family ties. Of course, the break didn’t last, and so the graph does have predictive value. It seems that it should also predict that Southey and Wordsworth should become close friends. And yet they never did (Pratt 2006a). At the moment that Coleridge and Wordsworth became close, Coleridge’s ties with Southey had been severely weakened by Coleridge’s anger at him, and by Charles Lloyd’s repetition to Southey of Coleridge’s alleged slurs against Southey (Sisman 2006, 185, 205). Though these ties eventually strengthened again, a graph illustrating the strength of ties through time centred on Coleridge’s letters would reveal successive and alternating strength and weakness. When close to Southey, Wordsworth is only an acquaintance; when close to Wordsworth, Southey has been abandoned: "You have left a large Void in my Heart—I know of no man big enough to fill it," Coleridge writes to Southey in 1795 (1956, i.173). In a later letter in which Coleridge invites Southey to live for a time with him and Wordsworth, he writes, "Wordsworth is a very great man—the only man, to whom at all times & in all modes of excellence I feel myself inferior—the only one, I mean, whom I have yet met with . . . " (Coleridge 1956, i.334, Coleridge’s emphasis. Qtd. in Sisman, who points out that, “in praising his new friend he could not avoid a glancing blow at the old one” [2006, 185]).
Coleridge makes this remark to Southey before Coleridge and Southey are reconciled, but it shows that the Strong Triadic Closure Property can be disrupted by melancholy—in Coleridge’s case, expressed as the necessity of having one idol whom he elevates above all others. He is asking Southey to come worship at the altar of Wordsworth, surely an unattractive proposition even if Southey does not begrudge Wordsworth for having taken over his place in Coleridge’s affections. By this time, Southey seems, to me at least, to be fairly fed up with Coleridge’s demands and so completely willing to pass onto Wordsworth the burdens of being Coleridge’s friend. While one could imagine Southey being willing to befriend two people who are closer to each other than to him, which happens with Triadic closure generally, there is a subtext of their correspondence about working and living together more closely. Coleridge seems (again, to me) to be asking Southey to play a role in what is clearly a transferential game: he wants him to be envious, angry, injured, displaced, while watching someone else be preferred over him in a former idolator’s affections (see Holmes 1998). That is, Coleridge is clearly attempting to arouse Southey’s envy in describing Wordsworth. Whether Southey felt jealous of their relationship or relieved that Coleridge had moved onto another poet to idolise remains an open question. But the fact is that Coleridge’s filial dysfunctions spreads and disrupts the regular operations of an evolving network, Strong Triadic Closure. This fact to me valorises Guinn Batten’s understanding of melancholy as a failure or refusal to engage in economic exchange, even, or especially, in the realm of affect.

4.4. The Strength of Weak Ties Hypothesis

This hypothesis in social network theory calls into question one of literary criticism’s basic tenets. We think it necessary to demonstrate a close relationship between writers in order to argue for significant influence. But social network theory tells us that new information, novelty that could make its way into someone’s writing, comes from "weak ties" or acquaintances (Easley and Kleinberg 2010, 43). For instance, Nicholas Roe establishes only a tentative and circumstantial connection between Wordsworth and George Dyer, Coleridge and Southey’s radical friend in London:

[G]iven Wordsworth’s acquaintance with Dyer in London during Spring 1795, and Coleridge’s intimacy with Wordsworth after 1797, it is not surprising that Dyer’s pamphlet should in some ways foreshadow Wordsworth’s poetry of 1798. (1992, 32)
Roe does not directly claim to find in Wordsworth’s poem “Tintern Abbey” deliberate echoes of Dyer’s pamphlet, *Dissertation on the Theory and Practice of Benevolence*, published by Cottle in 1795, because the connection is tenuous. Wordsworth writes no letters to Dyer at all. Dorothy Wordsworth mentions him once in a letter, when he, along with Basil Montagu, made a short visit to Grasmere: "we have not seen much of Dyer" ([1967], 511). On one of Coleridge’s letters to Dyer, Wordsworth’s London address is jotted down in someone’s hand other than Coleridge’s—Dyer’s, one presumes, or the hand of someone suggesting that Dyer visit Wordsworth while he is in London ([Coleridge 1956], 154; [Wordsworth and Wordsworth 1967], 140 n. 2). It is for this reason that Roe mentions "Wordsworth’s acquaintance with Dyer in London during Spring 1795" ([1992], 32).

Though he can only establish this tenuous connection, Roe quotes a passage from Dyer’s *Benevolence* in which Dyer says, strikingly, "There is a kind of voice that speaks through the universe" (qtd. in [Roe 1992], 31). Roe is obviously showing us an amazing source for the "spirit" that "rolls through all things" in "Tintern Abbey" ([Wordsworth 1798], "Tintern Abbey," ll. 102-104), but he never directly argues that Wordsworth got this line from Dyer’s pamphlet. In network theory, the notion that weak ties can pass the most life-changing bits of information gives us another way to talk about influence. Dyer would be a weak tie that forms a "local bridge" between Wordsworth’s social network and Coleridge’s, as well as his own and Southey’s. Because of this linking position as well as the weakness of Dyer’s tie to Wordsworth, I would argue, one of the most moving lines makes its way into "Tintern Abbey" from Dyer’s *Benevolence*. One might call Dyer’s writing of that pamphlet, in relation to Wordsworth’s poem, a "little, nameless, unremembered, [act] / Of kindness" ([Reiman 1972], 35-36)—or even of benevolence.

### 5. Conclusion: Local Bridges, Giant Components, and Network Effects

Before Southey moved to the Lake District, he was dubbed in the *Edinburgh Review* part of that "new school of poetry" that came to be called "the Lake School" even though he hailed, of course, from Bristol. Southey is called the member of an unnamed "sect" or "school" of poetry in the inaugural issue of the *Edinburgh Review* by Francis Jeffreys. The ER here begins its "unrelenting attack upon the ‘Lake Poets,’" editor Donald Reiman tells us ([1972], 2.a.415), in Jeffrey’s review of Southey’s *Thalaba, the
Destroyer: A Metrical Romance, but the school is not named in this 1802 review. Jeffrey only later designates Southey, Coleridge, and Wordsworth the "Lake School," when reviewing Wordsworth’s Poems of 1807 (2.a.429). In a letter of 1804, Wordsworth balked at the idea of a school of poetry comprised of "Coleridge, Southey, Lamb and myself": "it is scarcely possible that a greater difference should exist between any set of men or Authors, than between these four men with the exception of Coleridge and myself..." (1967, 434). Wordsworth’s poetic tribute to Southey at his death speaks not at all of his poetry, and thus was deeply resented by his second wife Caroline Bowles, who called Wordsworth "‘that other star of the lakes,’” implying of course that Wordsworth wanted no competition in leadership of his school (Pratt 2006a, 221, 233). New work has recently demonstrated Cottle’s leadership in "The Bristol School" or West Country Romanticism (Cronin 1992; Cheshire 1992).

Beginning with Southey’s Joan of Arc, Cottle arguably served as a muse for the epic poems that Southey wrote. Cottle himself wrote epics with "unhappy" results. As Southey puts it in a letter to William Taylor: "From England nothing has reached me but the unhappy Alfred of poor Cottle. I laboured hard & honestly to suppress its birth – & am thrown into a cold sweat by recollecting it" (2011, ii.558), (a search for author "Southey, Robert," and search terms "Cottle" and "Alfred" in NINES [http://www.nines.org/search/browse] will return all the letters by Southey in which he mentions the poem). The account of Joan given by Southey in the Preface to the first edition published by Cottle in December 1795 suggests rapid writing, correcting, and printing, all going on in the same room under Cottle’s auspices; the epic poem was originally written in six weeks, he says, augmented by Coleridge, and completely revised (except for 1000 lines) during its first printing (Southey 2004, 4). A parody akin to Byron’s report of looking forward to "‘an epic from Bob Southey every spring’ (Don Juan III, 97.4)" appeared in the Edinburgh Review of 1808:

A correspondent wrote to us lately an account of a tea-drinking in the west of England, at which there assisted no fewer than six epic poets—a host of Parnassian strength, certainly equal to six-and-thirty bands. ... How unreasonable then is it to complain, that poetry is on the decline among us! (qtd. in Curran 1986, 158)

The tea would have been served at Cottle’s, in Bristol, "in the west," even though by 1808 Cottle is no longer a bookseller, though still involved in provincial printing with Nathaniel Biggs.
This West-Country epic literature always did have an uneasy fit within the Lake School. Dissension in the ranks due to Coleridge’s policy of worshipping only one great mind at a time mirrors what is fundamentally a generic distinction. In a letter written to Cottle shortly after Coleridge met Wordsworth, he repeats approvingly Wordsworth’s criticism of Southey’s writings, objecting to the "ease" and "fluency" with which Southey wrote (Coleridge 1956, i.320, also qtd. in Pratt 1994). Pratt points out that the criticism is articulated by quoting a line from Joan of Arc, an act of irony that "would not have been lost on Cottle" (1994, 336). Coleridge is obviously concerned to distinguish Wordsworth’s lyric condensation from Southey’s expansive epic ambitions.

I have argued that injury to one psyche can block intersubjective exchanges and thus disrupt the smooth operations of network evolution. Whatever the reasons, though, strong triadic closure failed to join Southey and Wordsworth in friendship via their ties with Coleridge and thereby to link the two giant components of Wordsworth’s social network with our graph of relationships constituted by Southey’s letters. "Giant component" is another term found in network theory, and the two giant components mentioned here appear only in an as yet imaginary giant graph of Romantic-era writers’ correspondence networks, a graph that would include Wordsworth, Coleridge, and Southey’s social circles. The separation of the circles of Wordsworth and Southey into two giant components within this imaginary, giant graph—their social networks would be joined only by what are called "local bridges" constituted by Coleridge and a few other common friends—seems to result from, cause, instill, exacerbate, or indicate generic as well as personal differences.

Genre seems to me to be a key, unstated invariant associated with Wordsworth (lyric) and Southey (epic). Wordsworth’s introspective long poem The Prelude, pieces together small lyrical poems about biographically significant moments (the legendary "spots of time" being only one sort of lyric moment, among others). Southey could not write such a sustained introspection, whether in poetry or prose. A history of the growth of his own mind is an account that Southey longs to but never can write (Southey 1969, i.160). On the other side, Wordsworth cannot write the epic Recluse—which would after all be addressed to former sympathisers with revolutionary undertakings such as Pantiscocracy—while Southey’s epic passion continues throughout his writing life, reinvigorated in 1808 while visiting Bristol (Pratt 1994, xvii). There seems to be a generic division of labour between Wordsworth and Southey, and
so the question becomes, how indebted is the production of Lake-School poetry to the evolutions and devolutions of social networks? Should we see the failure of Southey’s introspective and Wordsworth’s epic ones as "personal" problems, as "writer’s block" for Wordsworth that just by chance is accompanied by a weirdly inverse writer’s block on Southey’s part—both of them being otherwise incredibly prolific writers (Leader 1991)? Maybe these ostensibly personal problems are in fact network effects. Literary scholars of the Romantic era have undoubtedly been attuned to "literary circles" throughout the history of literary criticism of the period, most recently visible in the emergence of Romantic Circles in 1993 and the discussion of the "Southey Coleridge Circle" as a "network" in 2000 (Pratt and Denison 2000b). But questioning whether Wordsworth and Southey’s "failings" to write epics and lyrics, respectively, are intersubjective network effects? That question is new.

Notes

[1] Laura Mandell—I—wrote the essay, and so all the errors are my own. The others listed here instigated and made the thinking of this article possible, so that "with" does not really indicate the centrality of their contributions. Furthermore, the editors of this special issue, Susan Brown and Stan Ruecker, pushed me to rethink parts of it in ways that actually changed my argument: it isn’t their argument (I'm responsible, and they may not even agree with it), but the argument is a much deeper one thanks to their careful reading and (to me) groundbreaking questions.

Works Cited


This piece builds on a bunch of my recent blog posts that have mentioned networks. Elijah Meeks already has prepared a good introduction to network visualizations on his own blog, so I cover more of the conceptual issues here, hoping to reach people with little-to-no background in networks or math, and specifically to digital humanists interested in applying network analysis to their own work.

Some Warnings

A network is a fantastic tool in the digital humanist’s toolbox—one of many—and it’s no exaggeration to say pretty much any data can be studied via network analysis. With enough stretching and molding, you too could have a network analysis problem! As with many other science-derived methodologies, it’s fairly easy to extend the metaphor of network analysis into any number of domains.

The danger here is two-fold.

1. **When you’re given your first hammer, everything looks like a nail.** Networks can be used on any project. Networks should be used on far fewer. Networks in the humanities are experiencing quite the awakening, and this is due in part to the until-recently untapped resources of easy tools and available datasets. There is a lot of low-hanging fruit out there on the networks+humanities tree, and they ought to be plucked by those brave and willing enough to do so. However, that does not give us an excuse to apply networks to everything. This series will talk a little bit about when hammers are useful, and when you really should be reaching for a screwdriver.

2. **Methodology appropriation is dangerous.** Even when the people designing a methodology for some specific purpose get it right—and they rarely do—there is often a score of theoretical and philosophical caveats that get lost when the methodology gets translated. In the more frequent case, when those caveats are not known to begin with, “borrowing” the methodology becomes even more dangerous. Ted Underwood blogs a great example of why literary historians ought to skip a major step in Latent Semantic Analysis, because the purpose of the literary historian is so very different from that of the computer scientist who designed the algorithm. This
series will attempt to point out some of the theoretical baggage and necessary assumptions of the various network methods it covers.

**The Basics**

Nothing worth discovering has ever been found in safe waters. Or rather, everything worth discovering in safe waters *has already been discovered*, so it’s time to shove off into the dangerous waters of methodology appropriation, cognizant of the warnings but not crippled by them.

Anyone with a lot of time and a vicious interest in networks should stop reading right now, and instead pick up copies of *Networks, Crowds, and Markets*[^1] and *Networks: An Introduction*[^2]. The first is a non-mathy introduction to most of the concepts of network analysis, and the second is a more in-depth (and formula-laden) exploration of those concepts. They’re phenomenal, essential, and worth every penny.

Those of you with slightly less time, but somehow enough to read my rambling blog (there are apparently a few of you out there), so good of you to join me. We’ll start with the really basic basics, but stay with me, because by part $n$ of this series, we’ll be going over the really cool stuff only ninjas, Gandhi, and The Rolling Stones have worked on.

**Networks**

The word “network” originally meant just that: “a net-like arrangement of threads, wires, etc.” It later came to stand for any complex, interlocking system. **Stuff and relationships.**

![A simple network representation from wikipedia.org](/images/network.png)

Generally, network studies are made under the assumption that neither the stuff nor the relationships are the whole story on their own. If you’re studying something with networks, odds are you’re doing so because you think the objects of your study
are interdependent rather than independent. Representing information as a network implicitly suggests not only that connections matter, but that they are required to understand whatever's going on.

Oh, I should mention that people often use the word “graph” when talking about networks. It's basically the mathy term for a network, and its definition is a bit more formalized and concrete. Think dots connected with lines.

Because networks are studied by lots of different groups, there are lots of different words for pretty much the same concepts. I'll explain some of them below.

The Stuff

Stuff (presumably) exists. Eggplants, true love, the Mary Celeste, tall people, and Terry Pratchett's Thief of Time all fall in that category. Network analysis generally deals with one or a small handful of types of stuff, and then a multitude of examples of that type.

Say the type we're dealing with is a book. While scholars might argue the exact lines of demarcation separating book from non-book, I think we can all agree that most of the stuff on my bookshelf are, in fact, books. They're the stuff. There are different examples of books: a quotation dictionary, a Poe collection, and so forth.

I'll call this assortment of stuff nodes. You'll also hear them called vertices (mostly from the mathematicians and computer scientists), actors (from the sociologists), agents (from the modelers), or points (not really sure where this one comes from).

The type of stuff corresponds to the type of node. The individual examples are the nodes themselves. All of the nodes are books, and each book is a different node.

Nodes can have attributes. Each node, for example, may include the title, the number of pages, and the year of publication.

A list of nodes could look like this:

<table>
<thead>
<tr>
<th>Title</th>
<th># of pages</th>
<th>year of publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphs, Maps, and Trees</td>
<td>119</td>
<td>2005</td>
</tr>
<tr>
<td>How The Other Half Lives</td>
<td>233</td>
<td>1890</td>
</tr>
<tr>
<td>Modern Epic</td>
<td>272</td>
<td>1995</td>
</tr>
<tr>
<td>Mythology</td>
<td>352</td>
<td>1942</td>
</tr>
<tr>
<td>Macroanalysis</td>
<td>unknown</td>
<td>2011</td>
</tr>
</tbody>
</table>
A network of books (nodes) with no relationships (connections)

We can get a bit more complicated and add more node types to the network. Authors, for example. Now we’ve got a network with books and authors (but nothing linking them, yet!). *Franco Moretti* and *Graphs, Maps, and Trees* are both nodes, although they are of different varieties, and not yet connected. We could have a second list of nodes, part of the same network, that might look like this:

<table>
<thead>
<tr>
<th>Author</th>
<th>Birth</th>
<th>Death</th>
</tr>
</thead>
<tbody>
<tr>
<td>Franco Moretti</td>
<td>?</td>
<td>n/a</td>
</tr>
<tr>
<td>Jacob A. Riis</td>
<td>1849</td>
<td>1914</td>
</tr>
<tr>
<td>Edith Hamilton</td>
<td>1867</td>
<td>1963</td>
</tr>
<tr>
<td>Matthew Jockers</td>
<td>?</td>
<td>n/a</td>
</tr>
</tbody>
</table>

We could have a second list of nodes, part of the same network, that might look like this:
A network of books and authors without relationships.

A network with two types of nodes is called 2-mode, bimodal, or bipartite. We can add more, making it multimodal. Publishers, topics, you-name-it. We can even add seemingly unrelated node-types, like academic conferences, or colors of the rainbow. The list goes on. We would have a new list for each new variety of node.

Presumably we could continue adding nodes and node-types until we run out of stuff in the universe. This would be a bad idea, and not just because it would take more time, energy, and hard-drives than could ever possibly exist. As it stands now, network science is ill-equipped to deal with multimodal networks. 2-mode networks are difficult enough to work with, but once you get to three or more varieties of nodes, most algorithms used in network analysis simply do not work. It’s not that they can’t work; it’s just that most algorithms were only created to deal with networks with one variety of node. This is a trap I see many newcomers to network science falling into, especially in the digital humanities. They find themselves with a network dataset of, for example, authors and publishers. Each author is connected with one or several publishers (we’ll get into the connections themselves in the next section), and the up-and-coming network scientist loads the network into their favorite software and visualizes it. Woah! A network! Then, because the software is easy to use, and has a lot of buttons with words that from a non-technical standpoint seem to make a lot of sense, they press those buttons to see what comes out. Then, they change the visual characteristics of the network based on the buttons they’ve pressed. Let’s take a concrete example. Popular network software Gephi comes with a button that measures the centrality of nodes. Centrality is a pretty complicated concept that I’ll get into more detail later, but for now it’s enough to say that it does exactly what it sounds like: it finds how central, or important, each node is in a network. The newcomer to network analysis loads the author-publisher network into Gephi, finds the centrality of every node, and then makes the nodes bigger that have the highest centrality. The issue here is that, although the network loads into Gephi perfectly fine, and although the centrality algorithm runs smoothly, the resulting numbers do not mean what they usually mean. Centrality, as it exists in Gephi, was fine-tuned to be used with single mode networks, whereas the author-publisher network (not to mention the author-book network above) is bimodal. Centrality measures have been made for bimodal networks, but those algorithms are not included with Gephi. Most computer scientists working with networks do so with only one or a few types of nodes. Humanities scholars, on the other hand, are often dealing with the interactions of many types of things, and so the algorithms developed for traditional network studies are insufficient for the networks we often have. There are ways of fitting their algorithms to our networks, or vice-versa, but that requires fairly robust technical knowledge of the task at hand. Besides dealing with the single mode / multimodal issue, humanists also must struggle with fitting square pegs in round holes. Humanistic data are almost by definition uncertain, open to interpretation, flexible, and not easily definable. Node types are by definition concrete;
your object either is or is not a book. Every book-type thing must share certain unchanging characteristics. This reduction of data comes at a price, one that some argue traditionally divided the humanities and social sciences. If humanists care more about the differences than the regularities, more about what makes an object unique rather than what makes it similar, that is the very information they are likely to lose by defining their objects as nodes. This is not to say it cannot be done, or even that it has not! People are clever, and network science is more flexible than some give it credit for. The important thing is either to be aware of what you are losing when you reduce your objects to one or a few types of nodes, or to change the methods of network science to fit your more complex data.

The Relationships

Relationships (presumably) exist. Friendships, similarities, web links, authorships, and wires all fall into this category. Network analysis generally deals with one or a small handful of types of relationships, and then a multitude of examples of that type. Now that we have stuff and relationships, we're equipped to represent everything needed for a simple network. Let's start with a single mode network; that is, a network with only one sort of node: cities. We can create a network of which cities are connected to one another by at least one single stretch of highway, like the one below:

| City          | is connected to |
| ------------------------------- |
| Indianapolis  | Louisville    |
| Louisville  | Cincinnati    |
| Cincinatti  | Indianapolis  |
| Cincinatti  | Lexington    |
| Louisville  | Lexington    |
| Louisville  | Nashville    |

![Diagram of a simple network with cities connected by highways]
The simple network above shows how certain cities are connected to one another via highways. A connection via a highway is the type of relationship. An example of one of the above relationships can be stated “Louisville is connected via a highway to Indianapolis.” These connections are symmetric because a connection from Louisville to Indianapolis also implies a connection in the reverse direction, from Indianapolis to Louisville. More on that shortly. First, let’s go back to the example of books and authors from the last section. Say the type we’re dealing with is an authorship. Books (the stuff) and authors (another kind of stuff) are connected to one another via the authorship relationship, which is formalized in the phrase “X is an author of Y.” The individual relationships themselves are of the form “Franco Moretti is an author of Graphs, Maps, and Trees.” Much like the stuff (nodes), relationships enjoy a multitude of names. I’ll call them edges. You’ll also hear them called arcs, links, ties, and relations. For simplicity sake, although edges are often used to describe only one variety of relationship, I’ll use it for pretty much everything and just add qualifiers when discussing specific types. The type of relationship corresponds to the type of edge. The individual examples are the edges themselves. Individual edges are defined, in part, by the nodes that they connect. A list of edges could look like this:

<table>
<thead>
<tr>
<th>Person</th>
<th>Is an author of</th>
</tr>
</thead>
<tbody>
<tr>
<td>Franco Moretti</td>
<td>Modern Epic</td>
</tr>
<tr>
<td>Franco Moretti</td>
<td>Graphs, Maps, and Trees</td>
</tr>
<tr>
<td>Jacob A. Riis</td>
<td>How The Other Half Lives</td>
</tr>
<tr>
<td>Edith Hamilton</td>
<td>Mythology</td>
</tr>
<tr>
<td>Matthew Jockers</td>
<td>Macroanalysis</td>
</tr>
</tbody>
</table>
Network of books, authors, and relationships between them.

Notice how, in this scheme, edges can only link two different types of nodes. That is, a person can be an author of a book, but a book cannot be an author of a book, nor can a person an author of a person. For a network to be truly bimodal, it must be of this form. Edges can go between types, but not among them. This constraint may seem artificial, and in some sense it is, but for now the short explanation is that it is a constraint required by most algorithms that deal with bimodal networks. As mentioned above, algorithms are developed for specific purposes. Single mode networks are the ones with the most research done on them, but bimodal networks certainly come in a close second. They are networks with two types of nodes, and edges only going between those types. Contrast this against the single mode city-to-city network from before, where edges connected nodes of the same type. Of course, the world humanists care to model is often a good deal more complicated than that, and not only does it have multiple varieties of nodes – it also has multiple varieties of edges. Perhaps, in addition to “X is an author of Y” type relationships, we also want to include “A collaborates with B” type relationships. Because edges, like nodes, can have attributes, an edge list combining both might look like this.

<table>
<thead>
<tr>
<th>Nodel</th>
<th>Node 2</th>
<th>Edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Franco Moretti</td>
<td>Modern Epic</td>
<td>is an</td>
</tr>
<tr>
<td>author of</td>
<td>Franco Moretti</td>
<td>is an</td>
</tr>
<tr>
<td></td>
<td>Graphs, Maps, and Trees</td>
<td>author of</td>
</tr>
<tr>
<td></td>
<td>Jacob A. Riis</td>
<td>How The Other Half Lives</td>
</tr>
<tr>
<td></td>
<td>Edith Hamilton</td>
<td>Mythology</td>
</tr>
</tbody>
</table>
Notice that there are now two types of edges: “is an author of” and “collaborates with.” Not only are they two different types of edges; they act in two fundamentally different ways. “X is an author of Y” is an asymmetric relationship; that is, you cannot switch out Node1 for Node2. You cannot say “Modern Epic is an author of Franco Moretti.” We call this type of relationship a directed edge, and we generally represent that visually using an arrow going from one node to another.

“A collaborates with B,” on the other hand, is a symmetric relationship. We can switch out “Matthew Jockers collaborates with Franco Moretti” with “Franco Moretti collaborates with Matthew Jockers,” and the information represented would be exactly the same. This is called an undirected edge, and is usually represented visually by a simple line connecting two nodes. Notice that this is an edge connecting two nodes of the same type (an author-to-author connection), and recall that true bimodal networks require edges to only go between types. Algorithms meant for bimodal networks no longer apply to the
network above.

Most network algorithms and visualizations break down when combining these two flavors of edges. Some algorithms were designed for directed edges, like Google's PageRank, whereas other algorithms are designed for undirected edges, like many centrality measures. Combining both types is rarely a good idea. Some algorithms will still run when the two are combined, however the results usually make little sense.

Both directed and undirected edges can also be weighted. For example, I can try to make a network of books, with those books that are similar to one another sharing an edge between them. The more similar they are, the heavier the weight of that edge. I can say that every book is similar to every other on a scale from 1 to 100, and compare them by whether they use the same words. Two dictionaries would probably connect to one another with an edge weight of 95 or so, whereas Graphs, Maps, and Trees would probably share an edge of weight 5 with How The Other Half Lives. This is often visually represented by the thickness of the line connecting two nodes, although sometimes it is represented as color or length.

It’s also worth pointing out the difference between explicit and inferred edges. If we’re talking about computers connected on a network via wires, the edges connecting each computer actually exist. We can weight them by wire length, and that length, too, actually exists. Similarly, citation linkages, neighbor relationships, and phone calls are explicit edges.

We can begin to move into interpretation when we begin creating edges between books based on similarity (even when using something like word comparisons). The edges are a layer of interpretation not intrinsic in the objects themselves. The humanist might argue that all edges are intrinsic all the way down, or inferred all the way up, but in either case there is a difference in kind between two computers connected via wires, and two books connected because we feel they share similar topics.

As such, algorithms made to work on one may not work on the other; or perhaps they may, but their interpretative framework must change drastically. A very central computer might be one in which, if removed, the computers will no longer be able to interact with one another; a very central book may be something else entirely.

As with nodes, edges come with many theoretical shortcomings for the humanist. Really, everything is probably related to everything else in its light cone. If we’ve managed to make everything in the world a node, realistically we’d also have some sort of edge between pretty much everything, with a lesser or greater weight. A network of nodes where almost everything is connected to almost everything else is called dense, and dense networks are rarely useful. Most network algorithms (especially ones that detect communities of nodes) work better and faster when the network is sparse, when most
nodes are only connected to a small percentage of other nodes.

Maximally dense networks from sagemath.org

To make our network sparse, we often must artificially cut off which edges to use, especially with humanistic and inferred data. That’s what Shawn Graham showed us how to do when combining topic models with networks. The network was one of authors and topics; which authors wrote about which topics? The data itself connected every author to every topic to a greater or lesser degree, but such a dense network would not be very useful, so Shawn limited the edges to the highest weighted connections between an author and a topic. The resulting network looked like this (PDF), when it otherwise would have looked like a big ball of spaghetti and meatballs.

Unfortunately, given that humanistic data are often uncertain and biased to begin with, every arbitrary act of data-cutting has the potential to add further uncertainty and bias to a point where the network no longer provides meaningful results. The ability to cut away just enough data to make the network manageable, but not enough to lose information, is as much an art as it is a science.

**Hypergraphs & Multigraphs**

Mathematicians and computer scientists have actually formalized more complex varieties of networks, and they call them hypergraphs and multigraphs. Because humanities data are often so rich and complex, it may be more appropriate to represent them using these...
representations. Unfortunately, although ample research has been done on both, most out-of-the-box tools support neither. We have to build them for ourselves.

A hypergraph is one in which more than two nodes can be connected by one edge. A simple example would be an “is a sibling of” relationship, where the edge connected three sisters rather than two. This is a symmetric, undirected edge, but perhaps there can be directed edges as well, of the type “Alex convinced Betty to run away from Carl.” A three-part edge.

A multigraph is one in which multiple edges can connect any two nodes. We can have, for example, a transportation graph between cities. A edge exists for every transportation route. Realistically, many routes can exist between any two cities: some by plane, several different highways, trains, etc.

I imagine both of these representations will be important for humanists going forward, but rather than relying on that computer scientist who keeps hanging out in the history department, we ourselves will have to develop algorithms that accurately capture exactly what it is we are looking for. We have a different set of problems, and though the solutions may be similar, they must be adapted to our needs.

**Side note: RDF Triples**

Digital humanities loves RDF (Resource Description Framework), which is essentially a method of storing and embedding structured data. RDF basically works using something called a *triple*; a subject, a predicate, and an object. “Moretti is an author of *Graphs, Maps, and Trees*” is an example of a triple, where “Moretti” is the subject, “is an author of” is the predicate, and “*Graphs, Maps, and Trees*” is the object. As such, nearly all RDF documents can be represented as a directed network. Whether that representation would actually be useful depends on the situation.

**Side note: Perspectives**

Context is key, especially in the humanities. One thing the last few decades has taught us is that perspectives are essential, and any model of humanity that does not take into account its multifaceted nature is doomed to be forever incomplete. According to Alex, his friends Betty and Carl are best friends. According to Carl, he can’t actually stand Betty. The structure and nature of a network might change depending on the perspective of a particular node, and I know of no model that captures this complexity. If you’re familiar with something that might capture this, or are working on it yourself, please let me know via e-mail.

**Networks, Revisited**

This piece has discussed the simplest units of networks: the stuff and the relationships...
that connect them. Any network analysis approach must subscribe to and live with that duality of objects. Humanists face problems from the outset: data that do not fit neatly into one category or the other, complex situations that ought not be reduced, and methods that were developed with different purposes in mind. However, network analysis remains a viable methodology for answering and raising humanistic questions—we simply must be cautious, and must be willing to get our hands dirty editing the algorithms to suit our needs.

**Part II: Node Degree: An Introduction**

In Part II, I will cover the deceptively simple concept of node degree. I say “deceptive” because, on the one hand, network degree can tell you quite a lot. On the other hand, degree can often lead one astray, especially as networks become larger and more complicated.

A node’s *degree* is, simply, how many edges it is connected to. Generally, this also correlates to how many *neighbors* a node has, where a node’s neighborhood is those other nodes connected directly to it by an edge. In the network below, each node is labeled by its degree.

![Network diagram](https://en.wikipedia.org/wiki/Network)

*Each node in the network is labeled with its degree, from wikipedia.org*

If you take a minute to study the network, something might strike you as odd. The bottom-right node, with degree 5, is connected to only four distinct edges, and really only three other nodes (four, including itself). *Self-loops*, which will be discussed later, are counted twice. A self-loop is any edge which starts and ends at the same node.

*Why* are self-loops counted twice? Well, as a rule of thumb you can say that, since the degree is the number of times the node is connected to an edge, and a self-loop connects to a node twice, that’s the reason. There are some more math-y reasons dealing with matrix representation, another topic for a later date. Suffice it to say that many network algorithms will not work well if self-loops are only counted once.

The odd node out on the bottom left, with degree zero, is called an *isolate*. An isolate is any
node with no edges.

At any rate, the concept is clearly simple enough. Count the number of times a node is connected to an edge, get the degree. If only getting higher education degrees were this easy.

Centrality

Node degree is occasionally called degree centrality. Centrality is generally used to determine how important nodes are in a network, and lots of clever researchers have come up with lots of clever ways to measure it. “Importance” can mean a lot of things. In social networks, centrality can be the amount of influence or power someone has; in the U.S. electrical grid network, centrality might mean which power station should be removed to cause the most damage to the network.

The simplest way of measuring node importance is to just look at its degree. This centrality measurement at once seems deeply intuitive and extremely silly. If we’re looking at the social network of Facebook, with every person a node connected by an edge to their friends, it’s no surprise that the most well-connected person is probably also the most powerful and influential in the social space. On the same token, though, degree centrality is such a coarse-grained measurement that it’s really anybody’s guess what exactly it’s measuring. It could mean someone has a lot of power; it could also mean that someone tried to become friends with absolutely everybody on Facebook. Recall the example of a city-to-city network from Part I of this series: Louisville was the most central city because you have to drive through it to get to the most others.

Degree Centrality Sampling Warnings

Degree works best as a measure of network centrality when you have full knowledge of the network. That is, a social network exists, and instead of getting some glimpse of it and analyzing just that, you have the entire context of the social network: all the friends, all the friends of friends, and so forth.

When you have an ego-network (a network of one person, like a list of all my friends and who among them are friends with one another), clearly the node with the highest centrality is the ego node itself. This knowledge tells you very little about whether that ego is actually central within the larger network, because you sampled the network such that the ego is necessarily the most central. Sampling strategies—how you pick which nodes and edges to collect—can fundamentally affect centrality scores. The city-to-city network from Part I has Louisville as the most central city, however a simple look at a map of the United States would show that, given more data, this would no longer be the case.
A historian of science might generate a correspondence network from early modern letters currently held in Oxford’s library. In fact, this is currently happening, and the resulting resource will be invaluable. Unfortunately, centrality scores generated from nodes in that early modern letter writing network will more accurately reflect the whims of Oxford editors and collectors over the years, rather than the underlying correspondence network itself. Oxford scholars over the years selected certain collections of letters, be they from Great People or sent to or from Oxford, and that choice of what to hold at Oxford libraries will bias centrality scores toward Oxford-based scholars, Great People, and whatever else was selected for.

Similarly, the generation of a social network from a literary work will bias the recurring characters; characters that occur more frequently are simply statistically more likely to appear with more people, and as such will have the highest degrees. It is likely that the degree centrality and frequency of character occurrence are almost exactly correlated.

Of course, if what you’re looking for is *the most central character in the novel* or *the most central figure from Oxford’s perspective*, this measurement might be perfectly sufficient. The important thing is to be aware of the limitations of degree centrality, and the possible biasing effects from selection and sampling. Once those biases are explicit, careful and useful inferences can still be drawn.
Things get a bit more complicated when looking at document similarity networks. If you’ve got a network of books with edges connecting them based on whether they share similar topics or keywords, your degree centrality score will mean something very different. In this case, centrality could mean the most general book. Keep in mind that book length might affect these measurements as well; the longer a book is, the more likely (by chance alone) it will cover more topics. Thus, longer books may also appear to be more central, if one is not careful in generating the network.

**Degree Centrality in Bimodal Networks**

Recall that bimodal networks are ones where there are two different types of nodes (e.g., articles and authors), and edges are relationships that bridge those types (e.g., authorships). In this example, the more articles an author has published, the more central she is. Degree centrality would have nothing to do, in this case, with the number of co-authorships, the position in the social network, etc.

With an even more multimodal network, having many types of nodes, degree centrality becomes even less well defined. As the sorts of things a node can connect to increases, the utility of simply counting the number of connections a node has decreases.

**Micro vs. Macro**

Looking at the degree of an individual node, and comparing it against others in the network, is useful for finding out about the relative position of that node within the network. Looking at the degree of *every node at once* turns out to be exceptionally useful for talking about the network as a whole, and comparing it to others. I’ll leave a thorough discussion of degree distributions for a later post, but it’s worth mentioning them in brief here. The degree distribution shows how many nodes have how many edges.

As it happens, many real world networks exhibit something called “power-law properties” in their degree distributions. What this essentially means is that a small number of nodes have an exceptionally high degree, whereas most nodes have very low degrees. By comparing the degree distributions of two networks, it is possible to say whether they are structurally similar. There’s been some fantastic work comparing the degree distribution of social networks in various plays and novels to find if they are written or structured similarly.

**Extending Degree**

For the entirety of this piece, I have been talking about networks that were unweighted and undirected. Every edge counted just as much as every other, and they were all symmetric (a connection from A to B implies the same connection from B to A). Degree can be extended to both weighted and directed (asymmetric) networks with relative ease.
Combining degree with edge weights is often called strength. The strength of a node is the sum of the weights of its edges. For example, let’s say Steve is part of a weighted social network. The first time he interacts with someone, an edge is created to connect the two with a weight of 1. Every subsequent interaction incrementally increases the weight by 1, so if he’s interacted with Sally four times, Samantha two times, and Salvador six times, the edge weights between them are 4, 2, and 6 respectively.

In the above example, because Steve is connected to three people, his degree is 1+1+1=3. Because he is connected to one of them four times, another twice, and another six times, his weight is 4+2+6=8.

Combining degree with directed edges is also quite simple. Instead of one degree score, every node now has two different degrees: in-degree and out-degree. The in-degree is the number of edges pointing to a node, and the out-degree is the number of edges pointing away from it. If Steve borrowed money from Sally, and lent money to Samantha and Salvador, his in-degree might be 1 and his out-degree 2.

**Powerful Degrees**

The degree of a node is really very simple: more connections, higher degree. However, this simple metric accounts for quite a great deal in network science. Many algorithms that analyze both node-level properties and network-level properties are closely correlated with degree and degree distribution. This is a pareto-like effect; a great deal about a network is driven by the degree of its nodes.

While degree-based results are often intuitive, it is worth pointing out that the prime importance of degree is a direct result of the binary network representation of nodes and edges. Interactions either happen or they don’t, and everything that is is a self-contained node or edge. Thus, how many nodes, how many edges, and which nodes have which edges will be the driving force of any network analysis. This is both a limitation and a strength; basic counts influence so much, yet they are apparently powerful enough to yield intuitive, interesting, and ultimately useful results.


I plan to continue blogging about network analysis, so if you have any requests, please feel free to get in touch with me at scbweing at indiana dot edu.

Scott B. Weingart is an NSF Graduate Research Fellow and PhD student at Indiana University, where he studies Information Science and History of Science. His research focuses on the intersection of historiographic and quantitative methodologies, particularly...
as they can be used to study scholarly communications in the past and present. He also writes a blog called the scottbot irregular, aiming to make computational tools and big data analytics accessible to a wider, humanities-oriented audience. When not researching, Scott fights for open access and the reform of modern scholarly communication.
Introduction: Do Maps Lie?

Béatrice Joyeux-Prunel

École normale supérieure, beatrice.joyeux-prunel@ens.fr

Follow this and additional works at: http://docs.lib.purdue.edu/artlas

Part of the Arts and Humanities Commons

Recommended Citation


This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

This is an Open Access journal. This means that it uses a funding model that does not charge readers or their institutions for access. Readers may freely read, download, copy, distribute, print, search, or link to the full texts of articles. This journal is covered under the CC BY-NC-ND license.
Do Maps Lie?
Editor-in-Chief

Catherine Dossin
Purdue University

Béatrice Joyeux-Prunel
École normale supérieure

Editorial Board

Michela Passini
Centre national pour la recherche scientifique

Daniel Quiles
School of the Art Institute of Chicago

Blaise Wilfert-Portal
École normale supérieure

Editorial Assistant

Michaelene Werth
Purdue University

Cover: Katherine E. Bash,
Physical Polylogue. Photograph.
Santa Caterina, Brazil. 2004.
© Itinerant Laboratory for Perceptual Inquiry.

Editorial Statement

The ARTL@S BULLETIN is a peer-reviewed, transdisciplinary journal devoted to spatial and transnational questions in the history of the arts and literature.

The journal promises to never separate methodology and history, and to support innovative research and new methodologies. Its ambition is twofold: An insistence on the “transnational” as constituted by exchange between local and international or transnational, and an openness to innovation in research methods, particularly the quantitative possibilities offered by digital mapping and data visualization.

By encouraging scholars to continuously shift the scope of their analysis from the national to the transnational, ARTL@S BULLETIN intends to contribute to the collective project of a global history of the arts and literature.

Copyright Statement

ARTL@S BULLETIN (ISSN 2264-2668) is published biannually by the École normale supérieure, 45, rue d’Ulm, 75005 Paris, France and the Centre national pour la recherche scientifique 16, rue Pierre et Marie Curie, 75005 Paris, France. The online version of the ARTL@S Bulletin is hosted by Purdue Scholarly Publishing Services.

Compilation copyright © 2013 ENS-CNRS. Contents copyright © 2013 by the respective authors, artists, and other rights holders. All rights in the ARTL@S BULLETIN and its contents reserved by the École normale supérieure, the Centre national pour la recherche scientifique and their respective owners. Except as permitted by the Copyright Act, including section 107 (the fair use doctrine), or other applicable law, no part of the contents of the ARTL@S BULLETIN may be reproduced without the written permission of the author(s) and/or other rights holders. The opinions expressed in the ARTL@S BULLETIN are those of the authors and not necessarily of the editors or publishers.
Table of Contents

Introduction: Do Maps Lie?
Béatrice Joyeux-Prunel

Spaces of Arts: Wrap-up Comments
David M. Lubin

Mapping Eastern Europe: Cartography and Art History
Katarzyna Murawska-Muthesius

Visualizing Spaces, Flows, Agents, and Networks of the Art Markets in the 18th century: Some Methodological Challenges
Sophie Raux

Les XX in the City: An Artists’ Neighborhood in Brussels
Laurence Brogniez and Tatiana Debroux

Filling the Blank Space of Global Art Peripheries: Measurements of Art Mobility and their Ambivalence in Nairobi, Kenya
Olivier Marcel

ARTL@S at Work
Les « marchands de tableaux » dans le Bottin du commerce : une approche globale du marché de l’art à Paris entre 1815 et 1955
Félicie de Maupeou and Léa Saint-Raymond

Poiesis
Walking as Experimental Poiesis
Katherine E. Bash
To spatialize and globalize art history is the main objective of the ARTL@S Project and, hence, its Bulletin. We want to contribute to the horizontalization of art historical narratives and especially that of modernism, a project that Piotr Piotrowski identifies in the work of many scholars coming from or writing on supposedly “peripheral” places and movements.1 Make maps: you will see and think differently, more broadly, and maybe more fairly.

This “new seeing” through maps, however, implies a very scrupulous methodology, and a mistrustful attitude towards the images our data seem to produce with the help of our computers, hands, and prejudices. Maps lie. Spatial art historians should always be aware of this. Their readers, even more. Maps lie as they have always lied. Throughout history, maps have provided “scientific” justifications to imperialist visions.2 Even in art history: “Kunstgeographie” in the context of 19th and 20th century German speaking countries offers examplies of how cartography justified pangermanist ambitions in Central Europe.3 As the French specialist of geopolitics Yves Lacoste put it: “Geography is first used for making war.”4

Maps can be handled as a kind of objective evidence that is in fact in no way objective: who selected the data? What was forgotten, what was erased? Why those colours, and not these, and what do they underline, or hide? Why this basemap? Everything is significant, hence potentially manipulated, or biased. It is good, and fair, that some art historians revolt against the symbolic violence of maps, as they have against the symbolic violence of numbers in our symposia and conferences.5 A map, much like a number, is not a proof just because it is supposedly scientific. In France, cigarette brands must indicate the risks of smoking with horrible photographs of cancerous lungs, or with the very clear announcement: “fumer tue” (smoking kills):

---

5 On the question of quantitative method in art history, see Béatrice Joyeux-Prunel, éd., L’art et la mesure histoire de l’art et méthodes quantitatives (Paris: Éd. rue d’Ulm, 2018). On the question of maps, the conference Spaces of Arts at Purdue University in September 2012 saw fascinating controversies, especially on the last day, during a round table animated by David Lubin.
In turn, why not oblige art historians to add an official warning to their maps: “Maps LIÉ”? Art historians, who have not been critically trained in manipulating maps, charts, or graphs, would thus have their own version of a “mind the gap” warning:

![Figure 1](image)

Foreign artists at the 1913 Parisian Salon d’Automne 1913, and their addresses in Montparnasse. Source: Mapping or the addresses of non-French artists of the catalogue on the base map of Paris published by E. Andriveau-Goujon, 1885.6

Of course, this proposition is ironical. But at the very least, a standard methodological caveat should be a reflex to signal to our readers our own awareness that “maps lie,” and to spare many unnecessary controversies. We know that maps lie, but we also know that they expose things that a chronological, or descriptive, or even critical narrative (capable of being every bit as deceptive as any map), cannot show.

Still, none of this should prevent a sincere reflexion on the potential and limits, of the cartographic approach to art history. This new issue of the *ARTL@S Bulletin* presents two positions: one which questions the longstanding mendacity of the map in the modern era, and another that makes a case for the map’s bird’s eye view of art history—what Franco Moretti calls “distant reading”—as essential for transnational histories of modernism. The authors in this issue demonstrate a clear understanding of the constructed character of the data and the questionable dimension of their results. In the first approach, maps lie; in the second, it is the canon that lies. Ultimately maps reveal how other narratives may be possible, without necessarily insisting on the “truth” of this new narrative over and above dominant canonical accounts.

As spatial art historians, we see value in a kind of “flattening” of the discipline, especially when encouraging quantitative and serial approaches to historicizing the arts. Count, compare, map, question! The French phrase “mise à plat” I am translating here by “flattening” has a double meaning:

1. A general evaluation of things, putting every artist, every work, and every question on the same level. This, again, is Moretti’s “distant reading,” in which there is no canon, just objects—and millions of them.7 No hierarchy – just a large horizontal landscape to study.

2. The “mise à plat” leads to a flat reading, without taste. Making maps for art history, and for global art history, pushes out (or at least seems to) questions of reception, taste and creation. Is it still art history? Where is the interest in artworks? How do we stop on one special work, if one is obliged to browse around all possible works? The work of the art historian, traditionally passionate and

---


spiritedly engaged, one could say, is now dispassionate. No more fascination. Flatness.

What is the added value of spatialization for art history, beyond the deconstruction of the canon, or the construction of some new digital humanities gimmicks? Perhaps we yield a more critical approach. At the very least, we are reminded of our obligation to be prudent in our conclusions—in addition to writing a serial, transnational art history that goes further than merely adding new pieces to the impossible puzzle of the “Global” art historical project.⁸

To open this special issue entitled “Do Maps Lie?” we asked a colleague, David Lubin, to help us better reflect on the implications of art historical mapping. David has a subtle approach to visual objects and the politically misleading, or downright deceptive, potential of images.⁹ We greatly appreciated the friendly but severe comments he first provided during an international conference we organized at Purdue University in September 2012 and then through lively conversations and email exchanges. We thank him warmly for accepting to rework his remarks for publication, and for the new directions he is charting for how we will continue to map... lies.

---

⁹ See for instance David M Lubin, Picturing a Nation: Art and Social Change in Nineteenth-Century America (New Haven; Yale University Press, 1994).
SPATIAL HUMANITIES:
AN AGENDA FOR PRE-MODERN RESEARCH

David Joseph Wrisley

Digital maps are objects of our time. We move about through cities in a different way today thanks to global positioning systems (GPS). We find digital content that is more relevant to us because our mobile devices share our location with providers. We interact with annotated map layers today, rather than turning the huge pages of printed atlases found in our library’s reference room.

This article provides some basic context and proposes some paths forward for researchers interested in exploring digital mapping in humanities research, particularly that of the pre-modern period. It begins with a simple observation: whereas access to geographic information systems (GIS) was once limited to specialized domains due to both the exorbitant cost of licenses and the technology’s steep learning curve, visualizing map information today has been significantly democratized, with a variety of simple entry points available for professional and public researchers. Whereas other humanists may find this article useful, the article is written with the pre-modern humanities researcher in mind. It will discuss some of the challenges in making basic maps with pre-modern humanities data and propose a list of practical suggestions.

Scholars have grown increasingly interested in exploring maps as a way of thinking about research materials, and the humanities provide us with endless use cases. A whole field of the digital humanities known as the spatial humanities has emerged centered on the question of location, pioneering research across and between traditional disciplines. Most research topics lend themselves to some kind of location-based approach and there are usually several units of analysis in humanities research that can be associated with a place on the earth. Place of publication, place of birth or death, stops on a pilgrimage route, location of shrines, city wall borders—these are the stuff of maps. It goes without saying that the more associations we can make with those locations—that is, the more metadata, the information about information, we can collect—the more complex and dense the maps we can make are. I use the term “digital mapping” because it is in common parlance, although it is perhaps more accurate to use the term map visualization, or visualization of spatial information on a map interface. Lev Manovich has claimed, generally speaking, that the main goal of information visualization is “discovering the structure of a (typically large)
In the pre-modern humanities, by contrast, the size of the dataset will likely range from small to medium, information can be highly uncertain and the data is usually manually collected.

Map visualizations can be found in all kinds of older printed materials: encyclopedias, linguistic atlases, historical atlases, as well as inserted as figures in the prose of textbooks or scholarly monographs. Depending on the age of the document, and no doubt also the budget of the publication, they can be either professional graphic representations or simple hand-drawn figures. Presentation of research findings in the form of a map usually has the rhetorical function of summary, distilling down complex information into what might be called visual argument, and yet this is only one of the uses of visualization. In recent years with the democratization of basic screen-based mapping, the researcher—guided by hypothesis or even just trial and error—is able generate multiple maps with ease and without the costly constraints of paper-bound publication. Visualization has become part and parcel of the research process. We think through our research through seeing data on maps.

We might link this to the rise of what has been called “spatial literacy” or “spatial thinking.” The static representations of canonical atlases are shattered, when the areas of a map are reoriented, when their constituent layers are broken down or recombined, or when the graphical features, colors or backgrounds employed are changed on demand. Current GIS-based software easily accommodates such transformation. Take as an example a static visualization of a key research document many of us deal with, the hand list of manuscripts:

This map has 108 points on it.\textsuperscript{318} It represents a complex, multilingual transmission network, that of al-Mubashshir ibn Fātik’s eleventh-century Arabic wisdom collection \textit{Mukhār al-Hikam} (The Choice Sayings) in five European languages. The glyphs (the dots) on the map indicate known manuscript witnesses and the Spanish incunabula in the tradition.\textsuperscript{319} If you explore this map live you realize that in certain locales, glyphs overlap with each other. It makes Paris with twenty five manuscripts look as important as Valladolid with only one. Using the same data, one way of countering this is to allow the glyphs to aggregate producing weighted clusters:

\textsuperscript{318} The map entitled “Mukhar al-Hikam translations” is available live at \url{https://djw.cartodb.com/viz/bbcedb408-8224-11e4-8dba-0e9d821ea90d/public_map}.

\textsuperscript{319} This hand list expands on manuscript information found in the Archives de littérature du moyen âge (ARLIMA) entry for Guillaume de Tignonville, the translator of the French version: \url{http://www.arlima.net/eh/guillaume_de_tignonville.html} [accessed 13 December 2014].
Neither of these maps attempts to represent time; they show only the spatial distribution of different modern holding institutions across languages. Since some of the more effective maps for humanities are ones that adopt a diachronic, comparative approach, an interesting extension of this map would be to place it, side by side, with archival data about the original provenance of such manuscripts to depict the historical movement of this sort of manuscript. Alternatively, a fascinating map might compare manuscript distribution of other Arabic texts, such as the *Kalila and Dimna* or the *Secret of Secrets*, that were translated in Spain during the same period and witnessed similar, widespread transmission.

One of the problems of traditional GIS is the difficulty with which popular standalone software deals with the temporal element of research data. As map visualization moves into web-based formats, mapping technologies are becoming increasingly dynamic. This means, for example, that map visualizations can change as the information contained in them is updated from a live stream, such as in the case of meteorological maps or traffic flow patterns, although the use of live stream data in the pre-modern humanities is unlikely. More commonly, maps can be “animated” in time, such as the FRENCH OF ITALY PROJECT map designed by the Center for Medieval Studies at Fordham University. It illustrates centers of literary
production of French documents on the Italian peninsula, with glyphs appearing and disappearing as the x-axis timeline advances. Pop-up windows appear with metadata and some manuscript thumbnail images when glyphs are selected on the map.

Mapping possibilities abound in the humanities. Allow me to suggest a few hypothetical examples. Pre-modern chronicles contain a plethora of toponyms and indications of time, and historians might want a way of organizing that data visually. A static map of a chronicle might include each place mentioned throughout the text indicated by a unique glyph, where color is used to indicate the time of the event. Take, for example, this map made from a dataset of place names occurring in medieval French texts from the twelfth to fifteenth centuries curated by my project VISUALIZING MEDIEVAL PLACES:

---

320 This map was created using NEATLINE (http://neatline.org/), a plug-in for the digital exhibit platform, OMEKA (http://omeka.org/). The Fordham map can be explored at http://frenchofoutremer.com/omeka/neatline/show/french-of-italy-timeline [accessed 8 December 2014].

321 The VISUALIZING MEDIEVAL PLACES project is described at the project blog http://visualizingmedievalplaces.wordpress.com.
The map depicts the place names of the Eastern Mediterranean found in some one hundred works. The rough time of composition of the work is represented by one of four colors, standing for each of the four centuries included in the project data. This map visualizes, therefore, how places are mentioned at different moments of historical time. On the other hand, maps visualized in the PELAGIOS PROJECT’s viewer allow us to filter the toponyms found in narrative time. The following map illustrates all the places found one single work, Pomponius Mela’s *De Chorographia* where the user can filter the data by book. The way that a researcher models spatial data naturally affects the way that such maps need to be read.

---

322 The data used in this article can be found on the project blog: [http://pelagios-project.blogspot.ca/2014/06/what-have-romans-ever-mapped-for-us.html](http://pelagios-project.blogspot.ca/2014/06/what-have-romans-ever-mapped-for-us.html) [accessed 7 December 2014].

323 The map visualized in the project’s Recogito viewer can be found at [http://pelagios.org/recogito/map?doc=18](http://pelagios.org/recogito/map?doc=18) [13 December 2014].
Urban historians might choose to map specific architectural patterns or population density by neighborhood. Art historians might choose to map places associated with patronage together with places of collection or artistic production. Scholars of material culture might want to collect inscriptions, visualized according to language, script or location. Manuscript historians might make point-to-point maps representing provenance trajectories. Scholars of religion might map pilgrimage sites, procession routes or shared sites of devotion. Depending on the complexity of the spaces involved, researchers may need to represents place by more than just a dot on a map, opting for polygonal descriptors. This problem takes us, however, beyond the scope of this article.

Furthermore, the kind of map layer we choose for visualizing that data can be chosen by the researcher. Given that the spatial information of Pomponius Mela’s work is largely situated on the Mediterranean coast line and largely in Greece and Italy, the fixed map projection that includes northern Europe in the Recogito viewer seems like a bit of an afterthought. What if we were to map the same data onto a historical map, say, the Prima Asiae Tabula 1478 early printed map of Ptolemy?324

324 The open-access version of this map can be found at http://commons.wikimedia.org/wiki/File:Prima_Asiae_Tabula_%28Asia_Minor_%26_Cyprus%29_-_Geography_%28Ptolemy%29,_Rome,_1478.jpg.
Or what if were to use today’s map layers, but visualize them using a tilted view looking south from the Bosphorus?\textsuperscript{325}

\textsuperscript{325} This map was created in Google Earth using a KML file of the data from the Pelagios website mentioned in notes 8 and 9 above.
The last two maps provide very different ways of viewing the segment of the historical data of Pomponius Mela’s text from Greece and Asia Minor, the former emphasizing a coastal perspective and the latter a sea populated with places. (Incidentally, it was possible to make these maps easily because the PELAGIOS PROJECT published their spatial data openly on their project blog for other researchers to use.)

The places that researchers interested in geography have obtained spatial data have changed over time. The traditional source of such information was the official geographic service of contemporary nation-states, but obviously such data come with limitations for the pre-modern researcher. Increasingly, web-based open, community gazetteers provide both contemporary and historical information. Since places are often palimpsests, meaning that their names having changed across languages and time, it is often possible to reach an ancient or a medieval place through its contemporary toponym. GEONAMES, a very large open gazetteer, has multilingual lists of names for common locations. A search for the common Latin place name Caesarea yields a variety of places that once bore that name: for example, Zaragoza in Spain, Kayseri in Turkey or Cherchell in Algeria.
If, for example, we are interested in the coordinates of Mauretania Caesariensis, a former Roman administrative division, the capital of which was located at the present site of Cherchell in Algeria, we are lucky since GEONAMES has a stable uniform resource identifier (URI) for it: www.geonames.org/8354632/mauretania-caesariensis.html.

Stable URIs are not currently available for all historical locations. Scholars in coming years will need to create them so that linked data can consistently be created for the places of our research. In general, we must
be creative and flexible and accept a certain margin of uncertainty or incompleteness in our spatial data.

The knowledge-sharing community WIKIPEDIA has a large amount of spatial data associated with their articles and is a good place to look for pre-modern locations, although it is important to say that the data is not equal across all languages. The classical Roman world is well documented in gazetteers, as well as archeological sites important for contemporary national identities. An excellent community-built gazetteer named Pleiades for the ancient Greek and Roman world will be expanding to include the Near East, Byzantium and the early medieval world. A detailed historical GIS data is available for China. A gazetteer for Syriac has recently been created. A community sourced gazetteer exists for the ancient world. A few pilots are underway for the historical Arabograph world. Funding has been recently secured for a gazetteer of Byzantine Cyprus. VISUALIZING MEDIEVAL PLACES is collecting pilot data for a multi-dialect gazetteer of medieval French.

When collecting spatial data, a structured table can be as basic a multi-column spreadsheet with an identifier for the research object—here a medieval toponym—spatial coordinates of that place and the source of the data (the URI). It can also be helpful to have a disambiguated contemporary place name associated with it.

<table>
<thead>
<tr>
<th>medieval name</th>
<th>real place</th>
<th>coordinates</th>
<th>source geodata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tarsus</td>
<td>Turkey</td>
<td>39.35</td>
<td><a href="http://www.geonames.org/298795/republic-of-turkey.html">http://www.geonames.org/298795/republic-of-turkey.html</a></td>
</tr>
<tr>
<td>Tyre</td>
<td>Lebanon</td>
<td>33.27333, 35.19389</td>
<td><a href="http://www.geonames.org/267008/tyre.html">http://www.geonames.org/267008/tyre.html</a></td>
</tr>
<tr>
<td>Warwick</td>
<td>England</td>
<td>52.82853, -1.58333</td>
<td><a href="http://www.geonames.org/2654725/warwick.html">http://www.geonames.org/2654725/warwick.html</a></td>
</tr>
<tr>
<td>Akko</td>
<td>Israel</td>
<td>32.927778, 35.081667</td>
<td><a href="http://www.geonames.org/275721/akko.html">http://www.geonames.org/275721/akko.html</a></td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td>51.5, 10.5</td>
<td><a href="http://www.geonames.org/3921044/federal-republic-of-germany.html">http://www.geonames.org/3921044/federal-republic-of-germany.html</a></td>
</tr>
<tr>
<td>Aalst</td>
<td>Belgium</td>
<td>50.93604, 4.0355</td>
<td><a href="http://www.geonames.org/2830448/aalst.html">http://www.geonames.org/2830448/aalst.html</a></td>
</tr>
<tr>
<td>Andalusia</td>
<td>Spain</td>
<td>37.6, 4.5</td>
<td><a href="http://www.geonames.org/2521750/andalusia.html">http://www.geonames.org/2521750/andalusia.html</a></td>
</tr>
<tr>
<td>Artois</td>
<td>France</td>
<td>50.5, 2.5</td>
<td><a href="http://www.geonames.org/5056644/artois.html">http://www.geonames.org/5056644/artois.html</a></td>
</tr>
<tr>
<td>Auxerre</td>
<td>France</td>
<td>47.8, 3.56667</td>
<td><a href="http://www.geonames.org/3035843/auxerre.html">http://www.geonames.org/3035843/auxerre.html</a></td>
</tr>
</tbody>
</table>

Since historical places are not always represented in contemporary gazetteers, the researcher needs to make all kinds of educated guesses. Historical GIS is, after all, always based on human created data! Consider sharing data when you have brought it to an acceptable level of completion, so that a spatial “community of practice” in your specialized research field can develop.

There are many platforms for visualizing spatial data on a map interface, each with its own balance of entry costs, openness and functionality. Geographical Information Systems emerged in the 1960s and their powerful standalone applications that are most commonly used today include ArcGIS (by ESRI) and the open-source QGIS. Web-based visualization engines have such as MapBox or CartoDB have also become available in recent years. Companies such as Google, Yahoo and Microsoft are increasingly integrating map visualization technologies into their internet-related services. Map creation solutions are also emerging rapidly for both collaborative research and the mobile environment. Both the technology and the services provided are evolving rapidly at the time of the publication of this article. Digital humanities institutes and training opportunities are the best ways to learn the ropes of more complex platforms. The individual researcher can begin collecting some sample data and experimenting with simpler platforms such as CartoDB and Google Maps.

In lieu of a conclusion, I offer a list of tips for working with historical spatial data:

1. Consult a variety of gazetteers to collect your spatial data, including WIKIPEDIA. Make sure it is in decimal form and not in degrees, minutes, seconds (e.g. Constantinople 41.01384, 28.94966).
2. Save the canonical URI for the spatial data you capture, as you would a citation of a printed or written source (e.g. for Constantinople http://www.geonames.org/745044/istanbul.html or for Herion http://pleiades.stoa.org/places/521034).
3. Make sure that you create structured data, that is, it should be organized in tabular form. It is as simple as using Excel or Access, or their equivalents.
4. Use a consistent vocabulary for your metadata.
5. Take advantage of pre-digital, spatial information found on historical maps. Historical maps are atlases are scholarship and they can help us. Consult digital atlases that have spatial data such as the Digital Atlas of Roman and Medieval Civilization (DARMC) http://darmc.harvard.edu/icb/icb.do. Learn how to geo-rectify analog maps.
6. Do not hesitate to capture the uncertainty of your data, using a typology appropriate for the kind of data you have (e.g. educated guess vs. occasionally attested toponym vs. commonly attested toponym).
7. Record the indigenous terms of place and time in your sources as you work with them (e.g. Julian vs. Gregorian calendar). Find a way to translate these as adequately as possible to match the formats your computer understands.
8. Make maps as you proceed in your research. Think through maps. Do not think of them only as final products.
9. Share your maps on a web-based map visualization engine like Google Maps or CartoDB. Learn to embed them in your blog or websites. Consider sharing portions of your spatial data in CSV (comma separated value) file so that others can work with it, and build upon it.
10. Consider starting a scholarly blog in which you discuss your pre-modern data. As more researchers use spatial technologies, the more useful infrastructure such as region-specific historical gazetteers become.
Locating Medieval French, or Why We Collect and Visualize the Geographic Information of Texts

By David Joseph Wrisley

This article focuses on geographic information contained in the body of medieval French texts composed over the period of the eleventh to the fifteenth century. By “geographic information” we mean textual references made to different kinds of place names at different scales within sustained prose or poetic narrative—landmarks, settlements, regions, and countries—real and imaginary. Collecting such geographic information across a large corpus of texts and analyzing it with the digital methods that have become available to scholars in recent years allow us to create new contexts in which we can reexamine a variety of questions in literary history. The scholarly context for the article is a digital project launched in 2014, Visualizing Medieval Places,¹ that seeks to arrive at a “complex understanding of cultural space-time” for the large corpus of literature written in medieval French.²

The methodological approach adopted here diverges from classic studies of medieval cartography inasmuch as it is not concerned primarily with analyzing medieval notions of space or their conceptual, visual, and cartographic expression in historical maps.³ The article looks at where, when, and how often locations are mentioned in a literary-historical corpus. We proceed from the basic idea that an evocation of a geographical location within a text is not gratuitous, and is worthy of critical exploration like any other subject. Place is, after all, as much a topical consideration as it is a geographical one. Textual geographies, as they have been called by some, visually collate spatial references found in a work, but they are not simply the depiction of a text’s realism.⁴ Instead, they are indicative of a more layered set of relationships related to geographic space: provenance, patronage, setting, even the overlap between spaces of textual production and more symbolic or imaginary realms. This article does not deal explicitly with the concept of mapping fictional narrative, that is, the spatializing of text from a theoretical, narratological perspec-

¹ Visualizing Medieval Places, digital project, http://medieval.place. At the time of publication of this article, the corpus was made up of spatial data from over 150 works. An updated list of the corpus is kept at the project site. All links in this article were last consulted 18 February 2017.
A collection of geographic information found in a database of places, otherwise known as a gazetteer, however, does not preclude such analysis by others in the future.

This article presents a case study of data derived from medieval French texts, and it argues more broadly for scholars in medieval studies to collect geographic information at small and large scales, as others in classical and modern studies have, and to use that data in order to ask new questions about our sources. Mapping is an important research method that uses geographical location as an index of the above-mentioned relationships in and between texts. It is important to note that by the term “mapping” we do not mean simply the act of compiling different forms of geographical information together in a cartographic medium, but we refer instead to the combined intellectual effort of curating such information in structured form and its iterative, visual representation. If this larger sense of mapping as a way of categorizing, organizing, and studying information about place is to take root in medieval studies, it must be spearheaded by research centers or other communities of practice that can balance the necessary diversity of approaches with some consensus on technical standards, as has been the case for the textual digital humanities.

Visualizing the places mentioned throughout a corpus is not a mode of inquiry new to literary studies, although in the past it generally took the form of static, paper maps and was, therefore, limited in its scope. Resources that scholars use on a daily basis, such as atlases or sometimes critical editions, contain diagrammatic maps that display research data, and these visuals have a strong rhetorical function. Whereas certain strands of literary and cultural studies resist mapmaking that asserts an unproblematic relationship between the space of texts and geographic space, others have embraced the practice of mapping, understood in the larger sense of both curation and representation, as an innovative, even “ludic” or “conjectural,” activity of the *studia humanitatis*. Access to the data behind visualizations and to the variety of computational platforms in which such information can be analyzed and represented was not, until recently, granted to scholars in all disciplines. The situation has changed significantly with the democratization of digital forms of cartography, bringing geographic information systems (GIS) to a wide base of users. This has given rise to an era of what has been called by Andrew Turner and others “neogeography.”

---

1 Two works taking this approach are Barbara Piatti, *Die Geographie der Literatur: Schauplätze, Handlungsräume, Raumphantasien* (Göttingen, 2008); and Marie Laure Ryan, Kenneth Foote, and Maoz Azaryahu, eds., *Narrating Space / Spatializing Narrative: Where Narrative Theory and Geography Meet* (Columbus, 2016).


Speculum 92/1 (October 2017)
The visualizations found in the body of this article might offend traditional cartographic sensibilities since they go beyond what has typically been considered the practice of geography and they do not have the qualities of traditional maps. They include data extracted from medieval sources visualized on contemporary base maps, fashioned according to specific research questions we have about those texts. In the branch of the digital humanities most concerned with modeling and analyzing space, known alternatively as the “spatial humanities” or “geohumanities,” the digital study of place has taken a turn away from a traditional concern of cartography, namely topography, and is increasingly conceived though a ‘topological’ lens: as a set of overlapping reticulations in which the nature and frequency of links among different sites matter more than the physical distances between them.”

In this respect, collecting and visualizing geographic information could be imagined either on a traditional map grid, as is the case in this study, or in more abstract ways, such as network visualizations. Medievalists will be inspired by the research of digital classicists, who have created an open infrastructure for complex digital projects that model different elements of the classical Mediterranean world. Such infrastructure has paved the way for innovative research with a significant spatial dimension. Medieval studies must focus, in our opinion, on similar infrastructure that reflects both the complexity and diversity of our field.

Geovisualization of information is sometimes explained by the need to expand a researcher’s interaction with data beyond a simple “text search” to provide a visual means of “sifting meaningful patterns from such massive datasets.” Collecting spatial data from medieval sources, as opposed to larger digitization projects, such as creating newspaper collections, is of an entirely different scale but stems from the same general research impulse. Large source bases are likely to contain “meaningful patterns,” and the ability to examine them, that is, to gain a bird’s-eye view with the assistance of a computer, is tantamount to having a team of readers, even many teams of readers, at one’s disposal. There is an important distinction, however, between harvesting medieval data and that of digitized newspaper collections. Whereas it is possible that the digitization of materials such as newspapers could be carried out by state-funded agencies or large library consortia, and that researchers might even be alienated from the process of its creation, data in medieval studies are a very specialized kind of information, and those involved in digital projects in medieval studies will, chances are, be quite intimately involved in their creation. For that reason, we agree, following Franco Moretti, that mapping does not play a “wholly peripheral” or “decorative” role in the study of history and culture, that is, it does not simply illustrate the results of research but can rather be part and parcel of the com-

13 Of particular mention are Pleiades, “a community-built gazetteer and graph of ancient places” (https://pleiades.stoa.org/); and the Pelagios Commons (http://commons.pelagios.org/).
14 For example, the ORBIS project is a geospatial network model of the ancient world (http://orbis.stanford.edu/).
plexities of the interpretative process. Groups of medievalists should work to
design the infrastructure for their geographical information.

The Visualizing Medieval Places (VMP) project has been collecting the place
names found in different genres and periods of textual production of what scholars
commonly call medieval French. The spatial and temporal coverage of the project is
admittedly vast. At the time of writing this article, it contains the geographic infor-
mation of some one hundred and fifty texts, with over ten thousand place-name oc-
currences documented. It is difficult to say what percentage of all sustained narra-
tive texts written in medieval French this represents, although our guess is that it is
only a small fraction of them.

Other humanities gazetteers, by which we mean databases of historical place
names and their geographic coordinates, such as the Pleiades Gazetteer for the
Greco-Roman world or the al-Thurayyā Gazetteer for the Arabograph world, do
not begin with a corpus of texts per se but instead with a discipline-specific canonical
atlas as their starting point for associating places with texts and spatial data.
The geographic information generated from the VMP project has not begun in this
way. It resembles what Janelle Jansted has defined as a literary gazetteer, that is, a
list of names built from a corpus of texts, names that are significant on account of
their evocation in these texts. The VMP project bears a resemblance to the Map of
Early Modern London (MoEML) project inasmuch as it is a “hybrid historical and
literary project.” Jansted continues,

Literary toponyms range from spaces as large as the city to places with relatively small
footprints, such as a conduit. One passage in a literary text might mention an adminis-
trative unit (a ward with clear legal boundaries), a neighborhood (an area identified pri-
marily by perceptions and/or use), a church (with a precise footprint we can know from
the archeological record), a long-gone building with uncertain coordinates, a monument,
and a shop. Spatial references in a text are necessarily place names, or strictly speaking,
“referring strings” of characters that people have chosen to indicate spaces made mean-
ingful by human habitation, practices, travel, and events.

The MoEML aims to describe London as it has been evoked in a large corpus, and
its spatial boundedness resembles the approach taken in many projects adapting
Geographic Information System technology to social and economic history (a field
known as “historical GIS”). For the sake of comparison, let us mention several other
projects in medieval studies that deal with geographic information and adopt either
a corpus approach or bounded-spatial approach. The Mapping Medieval Chester
project provides dynamic, diachronic views of one city, Chester, from 1200–1500,
focusing on archaeology, built space, and literature and their relation to Chester’s po-

---

17 David Joseph Wrisley, “How Many Texts Were Composed in Medieval French?,” Visualizing
Medieval Places (blog), 2 August 2014, http://vmp.djwrisley.com/2014/08/02/how-many-texts-were
-composed-in-medieval-french/.
Enriching and Integrating Gazetteers, ed. Merrick Lex Berman, Ruth Mostern, and Humphrey Sou-
thall (Bloomington, 2016), 129–43.
19 Ibid., 130–31.

Speculum 92/1 (October 2017)
tion as a historical border town. The Mapping Dante project, on the other hand, takes Dante Alighieri’s *Commedia* and links mention of places with occurrences of linguistic phenomena found in the corpus as well as graphical, nonmap representations of these occurrences. At the Austrian Academy of Sciences, there is an ambitious project, Mapping Medieval Conflict, aiming to represent medieval political dynamics across various means of digital representation. This project’s placement in the midst of a digital-humanities group with other researchers in text editing, prosopography, and network mapping has led to a social-research dynamic of interdisciplinary cross-pollination, especially between traditional medieval sources and innovative modes of visualization. Likewise, the Icelandic Saga Map began as a literary corpus visualization project but has expanded into a complex, multidisciplinary project concerning landscape, environmental memory, and historical climate change. These four projects illustrate varieties of spatial humanities research being carried out currently in medieval studies in the United States, the United Kingdom, and Europe, each with different research parameters but all attesting to the appeal of the spatial humanities’ use of the place name as a common denominator for potential interdisciplinary.

* * *

The VMP project does not begin with an atlas or a map representation, nor does it limit itself to representations of one region, but is geo-graphic in scope. It draws upon a representative sample of genres and periods, but also considers zones of literary production and the circulation of texts written in mutually comprehensible written varieties of French. The project includes texts composed in England, France, Italy, and the eastern Mediterranean and will expand to include many more texts. Its data have begun to trace the contours of the known world, or more precisely, the “named world,” according to medieval French. In this respect, the project responds to scholars’ calls to expand the study of medieval French beyond the traditional borders of France and the constraints of national literary studies. Research in the spatial humanities with medieval French materials contributes in two ways to the study of medieval literature in a global context: we examine textual production from a diversity of sites, including understudied ones; and we look at the different worlds that

---


*Speculum* 92/51 (October 2017)
are referred to therein. The VMP project data exists as a “digital object,” a prototype of a future gazetteer, that begins to make an argument about the importance of the geographical embeddedness of medieval French texts, and we believe the model is generalizable across other languages and corpora.25 Another digital project that used and enriched VMP data and enriched VMP data from ten texts of the Franco-Italian corpus illustrates how diverse those geographies can actually be.26

Another main difference between the aforementioned MoEML gazetteer and the research carried out in the VMP project is the principle of collection. The MoEML is built using a collection of rich digital editions of early modern texts in which strings referring to places have been tagged manually in TEI XML, the de facto standard for formal markup and representation of digital texts, a process that other text mapping projects have followed.27 The VMP project, by contrast, relies on a combination of manual and semiautomatic means of extraction in order to find place names in the corpus. Lacking a large open-access initiative for the publication of medieval French texts, we have adopted a pragmatic approach of collecting toponyms in the medium in which they are most easily available to us, namely from print critical editions. Corpora do exist for medieval French (such as the Base de français medieval or the Nouveau Corpus d’Amsterdam); however, they are oriented largely towards linguistic annotation, their licenses are not necessarily open, and they are relatively small, not reflecting the overall diversity of texts found in medieval French.28 Open-culture enthusiasts may be disappointed by our approach; and medievalists will also object that there are theoretical problems with it, considering that editorial principles define the kind of text available to us and that there are known cases where different manuscript witnesses and translations exhibit very different toponymy.29 Hopefully, such obstacles will be surmounted in the future; however, the labor required to create such a corpus of new digital editions and to do such data collection at the manuscript level across the whole corpus is, for now, infeasible. We have proceeded with the process of building the database of place names in the hope that others will follow.

The VMP project harvests its place names either from a corpus of texts that have fallen into the public domain, from the paratexts of critical editions, or from editions have been digitized and processed by Optical Character Recognition (OCR) into searchable electronic text. Issues of copyright prevent us from reusing the majority of those texts openly in the way that the MoEML project has. One of the ad-

---


Speculum 92/S1 (October 2017)
vantages of the latter’s strong relationship between spatial description and large-scale editing is that it facilitates a form of innovative reading whereby it is possible to move from map to text and text to map. Øyvind Eide reminds us, however, that such shifting across media, from text to map, is riddled with a number of conceptual problems, and one medium is never quite a reflection of the other, no more than moving from speech to writing is a smooth passage, or, to use a medievalist’s analogy, moving from manuscript to edited text.30

The VMP project organizes geographic information in a simple database along with other metadata about the works at hand. These include the form of the text, that is, whether it is written in prose or verse, or both; an approximate date range of composition; and an abbreviated name of the work, taken from a canonical online resource, the Dictionnaire étymologique de l’ancien français (DEAF), to ensure that data extracted from a particular edition are linked to it and have traceable editorial principles.31 The places in the database are then associated with contemporary locations, where possible, and with their corresponding geographic coordinates. Incidentally, the last step—disambiguation and association of textual data with place names and latitude-longitude pairs—is the phase of data collection and curation that is the most time-consuming and is the most often subject to debate. A text-centered approach to mapping means that the locatable places can be compared visually on a digital map interface as a way of performing a “spatial reading” of those texts—alone, comparatively, or in the aggregate. In his maps of collocations of topics in large linguistic corpora and their related, collocated places, Noah Bubenhofer calls this process “visual corpus analysis” (visuelle Korpusanalyse).32

Data about texts can be compared and aggregated with those of other texts, allowing for comparative analysis and pattern discovery, and ideally can be used by researchers in tandem with close reading. The process of formalizing place names found in medieval texts, however, is characterized by a degree of uncertainty. Typical GIS functionality, such as pattern detection and statistical spatial analysis, therefore, must be used with care in our case. The VMP project to some extent exchanges precision for scale. Furthermore, not all texts in the medieval French corpus lend themselves to such analysis, since not all texts contain numerous spatial references identifiable as real geographic locations.

The VMP project attempts to take a large view of medieval French literatures after the spatial digital turn, but the gesture of curating place names is not a new one. We have place-name indices for medieval French, but they are concerned largely with the imaginary toponymies of romance and chanson de geste, and so none of them would serve as a canonical starting point for a gazetteer.33 A more useful part

31 Dictionnaire étymologique de l’ancien français (DEAF), http://www.deaf-page.de/french/.
33 There are three such place name indexes. Ernest Langlois, Table des noms propres de toute nature compris dans les chanson de geste imprimées (Paris, 1904); L.-F. Flutre, Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Âge écrits en français ou en provençal et actuellement publiés ou analysés (Poitiers, 1962); André Moisan, Répertoire des noms propres de personnes et de lieux cités dans les chansons de geste françaises et les œuvres étrangères dérivées (Geneva, 1986).
of the scholarly apparatus that favors real geographic names is a standard para-
textual feature of the print critical edition, namely the index. The kinds of indices
found in an edition depend on a number of factors: the nature and complexity of
the textual tradition, the guidelines of a publication series, and even editorial idio-
syncracy. Medievalists are familiar with glossaries of *hapax legomena*, textual var-
iants, and forms of versification found in the indices of the editions they use, but also
with lists of references made to other texts (called an index locorum) as well as of
proper names and place names. The rationale for creating such indices seemingly
lies in being able to know quickly—that is, without reading the text again—that
a given work contains a reference to any such piece of information. These lists serve
as an organizational infrastructure for different kinds of both textual and contextual
research data designed to facilitate critical study of the edited text. The index of place
names was a crucial starting point for the VMP project to begin to collect geographic
information in medieval French.

For example, a scholar might be interested in which mythological figures are men-
tioned by Christine de Pizan in her *Epistre Othea*, a question quickly answered by
glancing at the “Table des noms propres et des personnages anonymes” of the crit-
ical edition by Gabriella Parussa (Geneva, 1999). Likewise, historical figures men-
tioned in a medieval chronicle would be listed in an index nominum of a modern
dition. Scholarly publishing insists on such indices because the critical edition is
a social text; it is meant to communicate salient textual features to a community
of researchers for the collaborative advancement of knowledge.34 When an edition’s
index is incomplete or is less than meticulously constructed, access to the text’s infor-
mation is limited. When an edition has a very complete index, a reader’s ability to
discover topics shared with other texts—say, the mention of a particular prince or of
a castle—is more likely to occur. Whereas indices of place names have been a per-
sistent element of the key publishing series for critical editions of medieval French
texts (including the Société des Anciens Textes Français, Classiques du Moyen
Âge Français, Textes Littéraires Français, and the Anglo-Norman Text Society),
and they have no doubt contributed to many scholars’ understanding of those texts,
oddly no comprehensive monograph-length study of the geographies evoked across
medieval French literature has, to my knowledge, appeared. This article’s focus on
place names found in texts asks some preliminary questions about how location-
based features of texts afford us new avenues for interpretation. Let us return to
the example of Christine de Pizan mentioned above. Following the method suggested
in this article, of extracting place names in a textual corpus and associating them with
geographic coordinates, a map that uses color as a distinguishing feature of texts in a
corpus might look like the visualization of the various geographies found in Christine
de Pizan’s oeuvre found in Fig. 1.35

The existence of printed place-name indices was no doubt what allowed several
classic studies of the image of Paris in medieval literature to be carried out. Alice
Planche contrasted the historical capital’s image across a corpus of Latin and French

---

34 Unfortunately one series publishing recent editions of medieval French texts, namely *Les lettres
gothiques*, has done away with the index.
35 David Joseph Wrisley, “The Literary Geographies of Christine de Pizan,” in MLA Approaches to

*Speculum* 92/1 (October 2017)
Locating Medieval French

Fig. 1. A visualization of the places mentioned in the oeuvre of Christine de Pizan. An interactive version is available at http://djwrisley.com/maps/cdep and the data are available for download at https://zenodo.org/record/35350 with a CC0 1.0 license.

sources up through the mid-thirteenth century, arguing for a study of both the presence and absence of the city in key texts. Her approach combined geographic references to places within the city of Paris with close readings of texts, whereby the place name served as an entry point into what Matthew Wilkens has called the “literary-imaginative geography.” Leo Olschki carried out a similar study, focusing on the representation of specific landmarks of medieval Paris, such as city walls and gates, royal residences, churches, and bridges, as well as nearby faubourgs, within the generic field of the chanson de geste. These studies share two features: an interest in urban geography and, more specifically, an interest in Paris, the city that has been the political center of France for many centuries. A geographical model of the medieval French corpus will have substantial data on its main city, but should not make the mistake of ignoring the many other places mentioned with high frequency as well. Texts are full of places mentioned with varying frequency, places about which we have greater or lesser amounts of information, real and imaginary places, and near and far places, in addition to geographic denominations of varying scale. Of course, the judgment of whether a place is real or imaginary can be fraught with difficulty.


Speculum 92/51 (October 2017)
The VMP project collects all place names that are found within the texts of its target corpus. The linked processes of disambiguation and associating geographic coordinates with place names is non-trivial. In some cases, the place names either have not been identified as a geographic location (for example, the monastery of Valberte from Adenet le Roi’s *Berte aus grans piés*) or cannot be disambiguated by close reading. In such cases, the geographic names are included in the database, but are tagged as unidentifiable, leaving open the possibility of future revision. Judging a place to be “unmappable” within what we know about historical geography is a weaker claim than deciding that a place is imaginary. Interestingly, we can still speak of patterns of recurring unidentified or imaginary places. It is not the main purpose of the project, but the data are collected nonetheless.

Drawing an analogy to another field that developed in the late twentieth century, the academic field of proposography, might be useful. Proposography attempts to document people and to explore relationships among minor historical personages about whom we might know very little, but who, when studied in the aggregate, reveal patterns, or networks, that provide insight into new lines of historical inquiry. The principle of studying the relationships among parts of an interconnected system provides an important cue, we believe, for thinking about geographic references made throughout a corpus. Much early analysis in the digital humanities also focused on pattern searching in a large amount of textual data. Looking for the pattern is what Stephen Ramsay calls “the strongest point of intersection between the computational strictures of text analysis and the open ended landscape of interpretative literary studies.”

An assumption of the VMP project is that we might make the same pattern-based conjectures about places. Spatial patterns in and across texts exist, and they provide a portal into the interpretative work of literary studies.

* * *

In the research done so far with the geographic information of the VMP project an emphasis has been placed on visualization. These visualizations provide a distinct view of a work’s historical, political, or narrative situation. The VMP project makes the double wager that the use of geographic information can, and should, be an important part of medieval literary history; and that different ways of enriching and visualizing that data must be employed to represent the data with conceptual depth and theoretical rigor. The goal of mapping texts cannot be simply to produce a set of static maps, but rather to create a rich database of spatial information associated with literary metadata that can be combined with other interdisciplinary knowledge about the Middle Ages. Practitioners in the spatial humanities remind us that the process of mapping is “topological and relational, revealing the ties that places have with each other and tracing their embeddedness in networks that span scales and range from the local to the global.”

As a way of thinking about texts and as the production of a visual object, both a method and an output, mapping

should ideally be explored in the medium, that is, dynamically. The constraints of publishing this article limit us, unfortunately, to static views. What follows in the rest of the article is a number of screen shots of visualizations generated from the project data. The process of visualizing geographical information is an iterative one based on a contextual knowledge of literary history that allows us to make targeted queries. It is also an invitation to return to close reading and to thinking more deeply about source material, or to collect more data.

The first examples here focus on mapping the locations in individual works written in medieval French. Fig. 2 is a close-up of locations mentioned in the Roman de Rou, Wace’s mid-twelfth century account of the origins of the Normans, offering a clear depiction of the cross-Channel theater of their invasion, with data sparsely scattered across northern France and dense clustering around the city of Caen and in the region of Kent, which is not surprising for the subject at hand.

The visualization in Fig. 2 was created from a place-name index of the Roman de Rou using automatic geoparsing with the web service provided by Google known as Fusion Tables. Such automatic means of mapping a text’s locations is a quick but crude way of visualizing a textual geography. The emphasis in Fig. 2 here is not on producing a clean, authoritative data set, but rather on rapid visualization for the purpose of discovery. Medievalists with some coding experience might learn how to extract toponyms and associate them with normative names and geographic coordinates using any number of online tutorials designed for humanists, thereby bypassing Google’s “black box” approach. Alternatively, medievalists might try tools developed for such semiautomated geoparsing with modern English translations of works that interest them. In a semiautomatic fashion, but not without human decision making, a basic text map can be generated.

Fig. 3 illustrates the geographic information of another single text, that of Joinville’s Vie de saint Louis, visualized as a heat map. According to that map style, the highest density of place names occurs where the map is orange and yellow. The question that arises immediately from the process and product of mapping is how such visuals can be interpreted. One of the elements of this particular visualization of Joinville’s text worthy of remark is how it seems to turn received ideas about a text on their heads. The Vie de saint Louis chronicles Joinville’s voyage accompanying Louis IX on the Seventh Crusade, and yet, while the map does feature a number of southern and eastern Mediterranean locales important for that voyage (Tunis, the Egyptian delta, and the Crusader domains, in increasing order of density), northern France is visually much more important. Out of some 290 place names found in Joinville, only some 70 are outside Europe. This is not because we have traditionally misread this text, but rather, given the number of noble families mentioned

41 “Fusion Tables,” web service, available online at https://support.google.com/fusiontables/answer/2571232.
Fig. 2. A map of places mentioned in Wace’s Roman de Rou created using Google Fusion Tables.
in Joinville’s *Vie de saint Louis*, toponymy and onomastics are collapsed in the process of place-name extraction. Put simply, maps of the place names found in texts are not representative of the text’s narrative setting alone, but also represent the narrator’s sense of belonging to, and the associations he has with, geography.

Fig. 4 uses a map visualization as a way of contrasting different versions of a text, the verse *Roman de la Violette*, dating from the first third of the thirteenth century (in red), and its fifteenth-century rewriting in prose (in blue). Whereas the sheer numbers of place names in both verse and prose are practically the same (68 versus 65), meaningful geographic transformations have taken place in the text’s rewriting. This map suggests visually how late medieval prose rewriting could involve not only the updating of language but also the localization of an older narrative within new regimes of power. It is visualized in the Google Earth interface not using the familiar Mercator projection used for web mapping, but in 3D and from the side. What is interesting about such a perspective is how it offers an alternate, if not clearer, view of those geographic transformations of the *mise en prose*. There are significant new clusters of blue in the regions of the Savoy, the Nivernais and the Nord. Whereas the additional places in the region of Nevers seem to align with the prose text’s resituating the hero as hailing from this region, other clusters raise potential questions about the impact of fifteenth-century court cultures, the houses of Valois Burgundy and the Savoy, on the formation of the literary geographies of the time.

Speculum 92/51 (October 2017)
The digital map has emerged as an important research tool for thinking about and representing location-based humanities data. A decade has passed since Martyn Jessop outlined some of the reasons that digital humanists have been reluctant to embrace GIS.44 At present, we are witnessing a moment of intellectual effervescence, with digital mapping projects—both medieval and nonmedieval—increasingly appearing in academe. Such map visualizations are both worthwhile and challenging; they provide new lenses for seeing texts through the geographic information they contain, as well as new ways of combining, if sometimes only experimentally, different kinds of research data by means of assemblage, layering, or filtering that mapping from a geotagged data set affords us.

Figs. 2, 3, and 4 illustrate ways that the geographic scope of individual texts can be visualized, and these examples maintain a strong resemblance to forms of close reading that have dominated textual studies over the twentieth century. The second set of visualizations presented in this article looks at distant readings of geographical information from the larger corpus of medieval French. These corpus-level projects have been championed in modern literatures, where large corpora are openly available; and both the gazetteers and the computational infrastructure have facilitated such reading. Wilkens has employed computational means to extract large numbers of place names (with degrees of uncertainty that go unmentioned) that occur in the large corpora of novels written over several decades of the American Civil War.45 In addition, he is directing a project that has begun work on developing tools for the spatialization and visualization of texts contain in the HathiTrust Dig-

---

45 Matthew Wilkens, “Geographic Imagination.”

Speculum 92/5 (October 2017)
The research objective in his work bears a resemblance to that of the VMP project: to chart patterns in the mention of place names in texts over time and space, suggesting relationships with either sociopolitical context or the investment of places in temporal slices of cultural memory. There are, however, three significant differences between the two projects. First, as we have mentioned above, open, representative, and computationally tractable corpora for the five centuries of medieval French do not yet exist. Second, the computational techniques that work well with modern English-language corpora are still underdeveloped for the linguistic particularities of medieval French, most importantly, the language’s high degree of orthographic variance. Third, determining the date of composition of medieval works is not possible with precision.

This does not mean that distant readings of place for the Middle Ages are not possible. It will take time before we have the research infrastructure to carry them out, and our conclusions about change over time will remain speculative, given the uncertainty that we have about dating the composition of medieval texts. Fig. 5 illustrates the fifty most common place names found in a large corpus of medieval French texts, spanning the entire five centuries of documented textual production. The orientation along an axis from northwest Europe to the southeastern Mediterranean generally replicates the axes of textual production in medieval French. Outside France, the most common locations mentioned in texts cluster in the British Isles, the Italian peninsula, Carthage, the Greek islands, Constantinople, and the eastern Mediterranean. Whereas the predominance of those general geographic zones will not surprise scholars of medieval French, taking a closer look at other recurrent locales reveals some that are not so familiar: Blois, Pavia, Galicia, Apulia, Gascony, the Red Sea, and Cologne. More research needs to be done on these networks of places across the corpus and how they might be explained.

Fig. 6 illustrates what might be called a distant, genre-specific mapping experiment. From the geographic information extracted from all the texts categorized as chansons de geste, dates of composition were normalized to their midpoint in order to construct a choropleth map of time. Choropleth maps allow for a “color ramp” to indicate increasing numerical values, here used to represent the passage of time. The results are quite stunning, as they illustrate a slow shift in epic geographies from Old French texts on a northeast-southwest France-Iberia axis to the mirror image, Middle French texts on a northwest-southeast France-Italy-eastern Mediterranean axis. This map confirms a general claim of literary history with empirical evidence.

Whereas Figs. 5 and 6 approach the medieval spatial data of texts with no attention paid to textual provenance, the latter can also become the focal point of analysis. As mentioned above, the methodology of the VMP project has been extended to deal with a specific subcorpus of medieval French, namely Franco-Italian, in the Exploring Place in the French of Italy (EPFOI) project. Fig. 7 illustrates a detail of EPFOI data. The project site argues for the importance of a distant approach to place, especially given the corpus’s composition in “a non-native language—that is French instead of Latin or a local Italian dialect.” In Fig. 7, the researcher can identify a spatial “imprint” or “signature” of individual works as a way of thinking about their relation to geography. Note the specific emphasis in Martin da Canal’s

Matthew Wilkens et al., Textual Geographies, digital project, http://txtgeo.net/about.

Speculum 92/51 (October 2017)
Estoire on the Veneto and the Adriatic (red), the pattern of places found in the Piemonte region of Italy in the anonymous *Aquilon de Bavière* (magenta), and the emphasis on the western and central regions of the Italian peninsula in Brunetto Latini’s *Trésor* (dark green). In the composite map of the French of Italy, clusters of points in the British Isles, Asia, the empire, and the Adriatic do not occur in all texts. One can even see texts that bear witness to medieval French’s status as a Mediterranean lingua franca, such as the *Roman de Hector de d’Hercule*, the *Moamin* and the *Ghattrif*, since these texts, although in French, do not contain a single mention of a place in France. The insights provided by the EPFOI project suggest that the work should be also carried out on texts composed and copied for the Anglo-Norman corpus. When moving to other regions, we will have to take into consideration the different generic emphases in those subcorpora.

EPFOI has productively challenged one of the aspects of the original data model of the VMP project. The former carries out its collection of spatial information at the manuscript level. Using editions of French texts based on manuscripts composed or copied in Italy, the EPFOI project found significant orthographic variance in some place names, but also widespread confusion (and perhaps purposeful exchange) of names in different manuscripts in certain genres. Modeling place at a witness level was possible, given the kinds of editions available for the Franco-Italian corpus and the awareness of cultural crossings going on in its texts documented in editions.

Perhaps the most distant visualization of the VMP project is found in the maps of the aggregate of its data. Fig. 8 represents the data with date of composition normalized to the midpoints of the accepted ranges. The benefit of this particular visualiza-

Fig. 5. The top fifty places mentioned in a corpus of 250+ texts composed in medieval French, sized by frequency, and visualized in CARTO.
zation is the implementation of a time slider, allowing for queries about location frequencies across a user-defined time period. It provides a faceted geoview that allows the contours of named places in Europe, the Mediterranean, and the Middle East found in medieval French texts to be explored as a function of time.

One of the surprising findings of the VMP project is how infrequently, and by how few texts, Europe east of the Rhine is mentioned. Another way of exploring this finding about a map’s silences has been to resort to historical data on Europe’s populated areas. The aggregate of the VMP geodata extracted from texts was overlaid onto a historical map layer of major towns c. 1200 (Fig. 9). For the region east of the Rhine, generally speaking, locations mentioned in the VMP data did not correspond to areas with major towns. Data are also available for major towns for the years 800, 1000, and 1450, and similar results were observed for those periods as well. Fig. 10 is a close-up of north central Europe from the same map in Fig. 9, except that the texts mentioning the places have been labeled. We can observe how a few later medieval texts are responsible for the visible points. (Fig. 10). Not only are there very few cities that are mentioned in this contemporary corpus of medieval French, in contrast to the quite dense urbanization of medieval Lower Saxony, but also the toponyms found from imperial lands were largely regional names (Franconia, Thuringia, Bavaria, Bohemia). This observation suggests to us that not only

Speculum 92/51 (October 2017)
should we look more deeply into the image of Germany in medieval French literature to test the hypothesis of a “borderland,” but also more effort can be made to put medieval demographic and environmental data into dialogue with textual geographies. Other regions, such as the Midi of France and the Midlands of Britain, are also not mentioned. It is important to investigate a map’s silences as well as its data patterns.

The third, and final, set of visualizations in this article focuses on locating medieval French vis-à-vis other vernacular and vehicular languages of the Mediterranean. The work of mapping geographical references in corpora should not be limited to French but, as we have been suggesting in this article, should extend to other co-existing languages: medieval Latin, Arabic, and Greek. Fig. 11 illustrates a comparative aggregate view of the VMP data with the spatial distribution of toponyms annotated in the Pelagios project CW2 Latin tradition data set. Such a comparison is quite crude, since the different data sets, the states of their data, and the parameters for generating the heat map are different, but it does suggest, from a distant perspective, general contours of the geographies of medieval French texts as opposed to those of early Latin geographic treatises.

For such comparative views to be valuable and meaningful, we must assemble and interpret them carefully, since the means of data creation in one may not match the nature of another data set. In this particular case, the data from the Latin corpora come explicitly from geographic texts. It is not an equal temporal comparison, since Speculum 92/S1 (October 2017)
Fig. 8. The aggregate of the Visualizing Medieval Places (VMP) data with a time slider allowing for custom temporal views. This interactive visualization is available at http://djwrisley.com/maps/VMP.
the Latin map draws upon sources much earlier than the French. The impetus behind making such a comparative visual corpus analysis might be, however, to test the hypothesis that many place names are used in medieval French on account of those writers’ knowledge of Latin geography. Much more analysis to test such a hypothesis is required than the simple views shown here. The data are coming into existence for such analyses to be attempted, but we are very much at the beginning of the enterprise. At first glance, the differences do suggest, however, the extent to which medieval French texts refer to the larger Roman Empire of late antiquity. Northern France is much better represented in the French case, along with Britain. Conspicuously absent from the French corpus is coverage of Tunisia; most parts of Greece, except Achaia; the Black Sea coastline; and Asia Minor. Such comparisons would be revealing if we were to move forward in time with the other vernaculars of Europe and medieval Latin (and in collaboration with projects such as GeoLat).47 This world of medieval data would require the social engagement of digital medievalists specializing in a wide number of disciplines and spatiotemporal concentrations, a vision of which has recently been articulated in regard to the use of linked open data about the Middle Ages.48

---


Fig. 12 illustrates a textual comparison akin to those found at the beginning of this article’s visualizations, focused on a specific genre but differing in the fact that the corpus is constituted across European vernaculars. It illustrates three collections of Marian devotional poetry, a transnational and translinguistic phenomenon of the high Middle Ages. Represented are the place names found in the oeuvres of Gautier de Coincy and Gonzalo de Berceo as well as in the Cantigas of Alfonso X, shown in rust, lavender and green, respectively. Whereas the cases of Gautier and Gonzalo illustrate quite well how those works localize Marian devotion within relatively restricted regions, the example of Alfonso demonstrates how the textual space of the Cantigas manages to inscribe all Iberia and also make reference to many places in Western Europe. Scale in this map illustrates the contrast between the local geographies of two poets close to local communities of Benedictines and that of the man who almost became emperor. Extracting and visualizing these geographies entailed working in three vernacular languages: Castilian, Galician, and French.

Fig. 13 illustrates another way to contextualize medieval French texts within their larger Mediterranean frame. This visualization is an attempt at mapping comparative geographies of historical writing in the late medieval Mediterranean. Currently it visualizes the place names found in Guillaume de Machaut’s narrative poem the Prise d’Alixandre (yellow), about the so-called crusade of Peter I of Cyprus on the city of Alexandria in 1365, along with that of the almost contemporary Mamluk chronicle recounting the same event, the Kitāb al-Ilmām (Book of Considerations) by al-Nuwairī al-Iskandarānī (green).
Fig. 11. A comparative heat map view of VMP data of places mentioned in medieval French texts (above) and places mentioned in the Pelagios Latin geographic tradition data (below). Pelagios data are available with a CC-BY 4.0 license: http://commons.pelagios.org/2014/06/what-have-the-romans-ever-mapped-for-us-results-from-the-latin-geographic-tradition/ and the base map is available from http://awmc.unc.edu/wordpress/tiles/ with a CC-BY 4.0 license.
This map is unfinished, as many projects in spatial humanities can be. At the time of the writing of this article, the data from the Kitāb al-Ilmām have not been completely georeferenced, and other texts that recount the same event have not yet been added (Philippe de Mézières’ *Vita sancti Petri Thomae* and Leontios Machairas’s *Chronicle*). Already visible in the map are the different civilizational spheres that are evoked by the French and the Arabic accounts, but when we look more closely, especially with the help of the DARMC map layer of Islamic cities, we also notice multiple zones of shared geography in specific “microecologies” around the Mediterranean. Such comparative views have potential in enriching Mediterranean approaches to medieval studies, given their ability to chart the notion of the named world from different nodes of textual production in Euro-Asia.

* * *

In conclusion, this article has attempted both to articulate in theory the various reasons why we map geographical information from texts and to show these in practice with examples at different scales and with diverse cultural horizons in mind. The principal objective of the visualizations contained in this article is argument-based assemblage—for hypotheses testing, pattern discovery, and investigating the chang-
ing relationships across different geographic data sets. These mash-ups, like re-mixed music, bring together varied data sources in the same visual field to produce enriched results that go beyond the original intention of the source. This, in our opinion, is the work of spatial humanities argumentation. After the digital spatial turn, which introduced the possibility of extracting, organizing, and visualizing the geographic information present in a large number of texts of the historic cultural record, the medievalist cannot help but be humbled by the complex space-time problem that the French corpus represents. From the neogeographic representations in this article, which are based on a fraction of the geographic information that could be collected, there are numerous directions in which to expand spatial research on the period. It cannot be the research domain of one person.

Above we evoked the critical edition, suggesting how its paratext, designed for organizing extracted features of textual data, functions as a social infrastructure for academic print culture, guiding generations of readers and scholars through the intricacies of argument and textual criticism. If we did not want to stop with the printed paratext, but rather to integrate all the scribbled notes of successive scholars, and the corrections and elucidations in the books of our university libraries, we could imagine a perpetually expanding paratext. This idea is, of course, possible in digital environments, and in the case of data repositories where scholars share their data openly, with special open licenses encouraging their reuse, it has already been realized. The preliminary data of the VMP project have produced results that would...
not have been possible without digital platforms, and yet additional research infra-
structure is required to take the model of the VMP project and turn it into a social
research space for cocreation and annotation. Future plans for the further-enriched
VMP data include the creation of a digital gazetteer, which would open up many
new avenues for further research in medieval French studies. Perhaps in subsequent
phases the VMP project might be reinvented in the image of Pleiades, the community-
built gazetteer of the ancient world, or the Syriac Gazetteer, or might even be built, as
was the al-Thurayyā Gazetteer, within the open framework of GitHub.49 This work
would extend the project beyond collecting and visualizing to sharing that geographic
information. Classicists and now Middle Eastern scholars have led the way in human-
ities spatial infrastructure research and in the creation of open text bases. The cre-
ative and innovative scholarship in their fields can hopefully inspire Western and Med-
iterranean medievalists to do the same, taking into account the particularities of our
multilingual medieval past.

49 These three gazetteers can be found at https://pleiades.stoa.org/, http://syriaca.org/geo/index.html,
and https://althurayya.github.io/, respectively.
‘Stylo’: a package for stylometric analyses

Maciej Eder
Pedagogical Univ. of Kraków

Jan Rybicki
Jagiellonian University

Mike Kestemont
University of Antwerp

June 19, 2017

Abstract

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

Keywords

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

Contents

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

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

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

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

8 Options unavailable on GUI 27
  8.1 Cluster analysis linkage 27
  8.2 Network analysis support 27
  8.3 Undocumented arguments 28
1 Introduction

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

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

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

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

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

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

2 Installation

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

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

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

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

- `assign.plot.colors`
- `define.plot.area`
- `delete.markup`
- `delete.stop.words`
- `dist.argamon`
- `dist.cosine`
- `dist.delta`
- `dist.eder`
- `dist.simple`
- `draw.polygons`
- `gui.classify`
- `gui.oppose`
- `gui.stylo`
- `load.corpus.and.parse`
- `load.corpus`
- `make.frequency.list`
- `make.ngrams`
In most cases, these lower-level functions provide very basic processing functionality and they are therefore not intended to be invoked by everyday users. Hence, they will not be discussed in this manual. However, if you are interested how they work and how to use them, you can invoke the help pages for these functions: `help(load.corpus)`, `help(make.ngrams)`, etc. Help pages routinely contain some insightful examples as to how to use the code: refer to them if you want to understand what a particular function does. The examples can be copy-pasted into an active R console. (Don’t be afraid of the lines ‘## Not run’ – they prevent R to run some automatic checks on interactive functions; you can use these examples safely).

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

- novels
- galbraith
- lee

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

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

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

4.1 Corpus preparation

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

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

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

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

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

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

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

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

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

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

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

4.3 Options available on GUI

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

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

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

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

### 4.3.1 Input

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

- plain text: plain text files.  
- XML files; this option will remove all tags and TEI headers. 
- XML files of plays; with this option, all tags, TEI headers, and speakers’ names between `<speaker>...` tags are removed. 
- XML contents only: all tags, TEI headers, and chapter/section (sub)titles between `<head>...` tags are removed. 
- the option will attempt to remove HTML headers, menus, links and other tags.
UTF-8: if you use Linux or Mac, this option is immaterial; however, if your operating system is Windows, then you need to set it depending whether your dataset is encoded in Unicode (then check the option), or in ANSI (then leave it unchecked).

4.3.2 Language

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

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

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

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

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

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

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

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

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

4.3.3 Features

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

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

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

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

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

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

### 4.3.4 MFW settings

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

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

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

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

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

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

4.3.5 Culling

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

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

- Maximum: this setting specifies the last (and possibly, only) culling setting in your analysis (cf. the maximum MFW setting). (This parameter setting is espe-
cially important when working with the bootstrap consensus trees in *stylo()*, a procedure which involves running several analyses in a row).

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

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

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

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

### 4.3.6 Statistics

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where:

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

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

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

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

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

• Cosine Distance: a classical measure, introduced to this package in the version 6.3:

\[ \delta(AB) = 1 - \frac{A \cdot B}{\|A\| \cdot \|B\|} = 1 - \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \]

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

4.3.8 Sampling

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

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

- **Onscreen**: check this if you want to display the graph on the screen.
  - `display.on.screen= TRUE|FALSE`
- **PDF**: check this to obtain a PDF file with your graph.
  - `write.pdf.file= TRUE|FALSE`
- **JPG**: check this to obtain your graph in JPEG format.
  - `write.jpg.file= TRUE|FALSE`
- **SVG**: check this to produce a SVG vector file (XML-based scalable image format). Useful if you want to embed your plot in HTML code, or to edit it further with, say, Inkscape.
  - `write.svg.file= TRUE|FALSE`
- **PNG**: check this to obtain your graph in PNG format (probably the safest option if you want publication-ready resolution).
  - `write.png.file= TRUE|FALSE`
4.3.10 Plot area

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

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

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

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

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

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

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

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

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

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

4.3.11 PCA/MDS

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

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

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

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

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

4.3.12 PCA flavour

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

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

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

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

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

4.3.13 Various

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

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

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

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

5 classify()

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

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

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

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

5.1 Corpus preparation

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

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

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

### 5.2 Calling the function

The function is evoked by the command `classify()`.

### 5.3 Options

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

#### 5.3.1 Options inherited from `stylo()`

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

#### 5.3.2 Statistics

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

- **ALL culling**: the culling procedure (cf. 4.3.5) is based on the percentage of samples containing a given word. To compute this ratio, one might want to use the texts from the first set only, or from both sets.
- **ALL wordlists**: the both tables of frequencies are build using the pre-prepared word list of the whole primary set (default). Alternatively, one might want to prepare this list of both sets by activating this option.
- **ALL z-scores**: how the $z$-scores should be calculated. If the variable is set to FALSE, the $z$-scores are relying on the primary set only (this should be better in most cases; after all, this is the classical solution used by Burrows and Hoover). Otherwise, the scaling is based on all the values in the primary and the secondary sets. (This option is applicable to Delta only).

5.3.4 SVM options

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

- **Linear**: linear kernel of SVM; probably the best choice in stylometry, since the number of variables (e.g. MFWs) is many times bigger than the number of classes.
- **Polynomial**: polynomial kernel of SVM.
- **Radial**: radial kernel of SVM.
- **Degree**: parameter needed for kernel of type “polynomial” (default: 3).
- **Coef0**: parameter needed for kernel of type “polynomial” (default: 0).
- **Cost**: cost of constraints violation (default: 1); it is the C-constant of the regularization term in the Lagrange formulation.

5.3.5 k-NN options

- **k value**: the $k$ value in $k$-Nearest Neighbour algorithm, or number of neighbours to be considered. Certainly, the bigger the number the better the performance, but, on the other hand, to set this value to 3, you need to have at least three samples per class in the training set. If you keep just one text per class in the training set, the performance is *ex definitione* suboptimal.
- **l value**: minimum vote for definite decision, otherwise “doubt”. (More precisely, less than $k - l$ dissenting votes are allowed, even if $k$ is increased by ties).

5.3.6 Output: general

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

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

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

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

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

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

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

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

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

6.2 Calling the function

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

6.3 Options

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

6.3.1 Features

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

6.3.2 Sampling

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

6.3.3 Colors

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

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

7.1 Corpus preparation

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

7.2 Calling the function

The function is evoked by the command oppose().

7.3 Options

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

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

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

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

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

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

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

8 Options unavailable on GUI

8.1 Cluster analysis linkage

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

8.2 Network analysis support

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

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

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

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

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

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

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

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

8.3 Undocumented arguments

• features

• frequencies

• training.frequencies

• test.frequencies
• parsed.corpus
• training.corpus
• test.corpus
• cv
• cv.folds
• encoding
• relative.frequencies

9 Advanced topics

9.1 Batch mode

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

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

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

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

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

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

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

<table>
<thead>
<tr>
<th></th>
<th>ABronte</th>
<th>Austen</th>
<th>CBronte</th>
<th>Conrad</th>
<th>Dickens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agnes</td>
<td>4.57</td>
<td>4.24</td>
<td>4.25</td>
<td>4.19</td>
<td>4.47</td>
</tr>
<tr>
<td>Emma</td>
<td>3.11</td>
<td>3.29</td>
<td>3.43</td>
<td>3.14</td>
<td>3.71</td>
</tr>
<tr>
<td>Jane</td>
<td>3.19</td>
<td>3</td>
<td>3.08</td>
<td>2.85</td>
<td>2.81</td>
</tr>
<tr>
<td>Lord</td>
<td>2.6</td>
<td>3</td>
<td>2.63</td>
<td>2.43</td>
<td>2.86</td>
</tr>
<tr>
<td>Bleak</td>
<td>2.17</td>
<td>2.2</td>
<td>2.13</td>
<td>2.42</td>
<td>2.22</td>
</tr>
<tr>
<td></td>
<td>2.24</td>
<td>1.92</td>
<td>1.92</td>
<td>2.21</td>
<td>1.92</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

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

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

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

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

9.2 Custom splitting rule

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

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

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

Some other regular expressions may include:

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

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

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

Then your splitting rule might look as follows:

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

9.3 Splitting rule and batch mode combined

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

All_NNP true_JJ histories_NNS contain_VBP instruction_NN ;_: though_RB ...

, in_IN some_DT ,_, the_DT treasure_NN may_MD be_VB hard_JJ to_TO find_VB ,_, and_CC when_WRB found_VBN ,_, ...

31
Suppose one wants to drop the lemmas and to get tags only: NNP JJ NNS VBP NN RB IN DT DT NN MD VB JJ TO VB CC WRB VBN RB... It is possible to extract the grammatical annotation via the function `parse.pos.tags()`, using the following code:

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

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

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

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

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

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

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

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

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

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

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

### 9.4 Custom similarity measures

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

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

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

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

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

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

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

9.5 Large-scale stylometric tests

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

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

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

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

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

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

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

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

10 Error messages and troubleshooting

[TBD]

References


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


Research


Published: 27 March 2018

Peer Review:
This is a peer-reviewed article in Digital Studies/Le champ numérique, a journal published by the Open Library of Humanities.

Copyright:
© 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Open Access:
Digital Studies/Le champ numérique is a peer-reviewed open access journal.

Digital Preservation:
The Open Library of Humanities and all its journals are digitally preserved in the CLOCKSS scholarly archive service.
RESEARCH

Measuring Joycean Influences on Flann O’Brien

James O’Sullivan¹, Katarzyna Bazarnik², Maciej Eder³ and Jan Rybicki²

¹ University College Cork, IE
² Jagiellonian University, PL
³ Pedagogical University of Kraków, PL
Corresponding author: James O’Sullivan (james.osullivan@ucc.ie)

This paper examines the stylometric similarities between James Joyce and Flann O’Brien, demonstrating which works from the latter’s oeuvre are stylistically the most Joycean. We will outline the results of a series of quantitative enquiries focused specifically on Joyce and O’Brien, before offering a number of literary interpretations. It has long been argued that Brian O’Nolan, operating under the pseudonym of Flann O’Brien, is a disciple of James Joyce. This relationship remains a concern for scholars, and so our purpose here is to contribute some computational evidence to the discussion. We pinpoint those exact moments where O’Brien’s style is quantitatively similar to that of Joyce, using our results to re-engage existing arguments with renewed statistical precision.

Keywords: Stylometry; James Joyce; Flann O’Brien

1 Two Irish Birds

O’Brien’s *At Swim-Two-Birds*, despite considerable critical acclaim, was initially ill-received because of its “Joycean undertones”, with commentators “tend[ing] to condemn the work as inferior imitation” (Hopper 1995, 46). Seán Ó Faoláin remarks that the novel has “a general odour of spilt Joyce all over it”, while the *New Statesman* brands it as “dull” on account of its “long passages in imitation of the Joycean parody” (Hopper 1995, 46). Asbee (1991), while critical of *The Hard Life*, accepts that some comparisons can be drawn between the comic novel and Joyce’s collection of short stories, *Dubliners*. Hopper (1995) argues that “O’Brien is usually lumped in with Joyce” as a result of “their historical and cultural proximity”,
but that this is “an assumption which is unfair to both writers” (14). Stylistically, O’Brien’s novels are littered with parodic tributes to Joyce (O’Grady 1989). Indeed, while O’Brien demonstrated “repeated efforts to escape his influence” (Dotterer 2004, 59), “At Swim had everything in the world to do with James Joyce” (Taaffe 2004, 253). Some critics maintain that the “omnipresence of Joyce … was to be expected” on account of O’Brien’s shared affiliation with University College Dublin (Taaffe 2004, 249). While Joyce may have been a “talismanic figure” at UCD, O’Brien’s Joycean parodies are not always interpreted as positive. Taaffe (2004) suggests that O’Brien’s “attitude towards the elder writer … is equivocal, at the very least” (253); Anspaugh (2004) traces his “profound ambivalence toward Joyce – an almost constant vacillation between admiration and denigration, devotion and denial, love and hate” suggesting that its quality is Oedipal (3), while McMullen (1993) argues that “At Swim-Two-Birds enters into dialogue not with James Joyce alone” (63). As summarized by Dotterer (2004):

Critical comparison with Joyce has been frequent, as have analytical comparisons of their fiction, but less often has an awareness of this link to Joyce been seen as central and persistent in Brian O’Nolan’s formation of his own work. This link with James Joyce was one O’Nolan embraced, at times begrudgingly or unwillingly, but always out of some inner artistic and psychic necessity. (54)

By offering a fresh appraisal based on quantitative methods, this paper identifies the specific points at which O’Brien’s Joycean parodies are most prominent, so that literary interpretations can be focused, with computational precision, on the relevant passages. In using such methods, this paper does not offer any advances in the techniques of this field, but rather, uses computer-assisted criticism to shed new light on a pre-existing concern. This essay is far from the first to suggest that there are stylistic affinities between O’Brien and Joyce—it is, however, the first to use quantitative methods as a means of pinpointing, with statistical precision, where such affinities are most pronounced.
2 Methodology and Results

Stylometry, also referred to as computational stylistics, is usually aimed at analysing authorial style, e.g. unique authorial stylistic features, using statistical methods. Unlike traditional approaches to style, computational stylistics assumes that the usage of bare function words—such as “the”, “of”, “in” etc.—is as equally important as the distribution of meaningful content words when one is seeking to measure fundamental stylistic affinities. Moreover, being very frequent, function words exhibit properties suitable for statistical inference. Stylometry, then, tries to measure stylistic profiles of the most common function word frequencies in order to identify meaningful similarities between (groups of) texts. While conceptually straightforward, stylometry is rather advanced when it comes to its mathematical background; simultaneously comparing several frequencies of function words requires techniques which are referred to as multivariate, or multidimensional, because they involve multidimensional geometry to compute similarities between texts.

A number of multivariate stylometric methods are used in this study. Cluster analysis provides a preliminary insight into the dataset, identifying main groupings. As cluster analysis is very sensitive to the number of features—or most frequent words—analyzed, we measure the 100 most frequent words, expanding this range from 100 to 1000 in intervals of 100 in order to produce a number of virtual dendrograms combined into one consensus plot. The distance measure in each of our tests is derived from Burrows’ Delta (Burrows 2002; Hoover 2004), which performs clustering on the basis of authorial fingerprints derived from the most frequent words in a text. Finally, to identify possible peculiarities in sequential development of the analyzed texts, we use Rolling Delta (Rybczki, Hoover and Kestemont 2014), which forms an authorial signature based on one set of texts, and then applies that fingerprint to another text. Authorial signatures are plotted over the text in question, with stylistic similarity indicated through proximity to the baseline. The aforementioned methods were applied using the R package, “stylo” (Eder, Rybicki and Kestemont 2016). Delta-based approaches to stylistics “builds on the subtle differences in high-frequency phenomena” (Herrmann, van Dalen-Oskam and Schöch
The emergence of the Digital Humanities has given rise to definitions of style which are “broader and more abstract than most of the earlier definitions” (Herrmann, van Dalen-Oskam and Schöch 2015, 45), but has been demonstrated in foundational works (Burrows 2002; Craig 1999, 2004; Jockers 2013); and indeed, in the great many works of literary criticism that have applied such methods since first formulated, this form of evidence is empirically valid, and valuable to interpretation. As discussed by Hermann, van Dalen-Oskam and Schöch (2015), definitions of style have been broadened to include those which might be considered quantitative, and such constructions of style provide evidence beneficial to the critical process. As with any critical technique, computer-assisted or otherwise, there are limitations—the loss of context, for example, should be accounted for across all manner of distant reading. However, our macroanalytical approach is such that we are using the Delta analysis to identify those areas where quantitatively distinct stylistic similarities exist, using such as preliminary evidence for a more nuanced qualitative interpretation.

Initially, we generated a cluster analysis using a selection of English-language Irish modernists. Using the 100 most frequent words, it is interesting that O’Brien’s *At Swim-Two-Birds* clusters with Joyce’s works. Figure 1 shows a dendrogram, which represents stylistic similarity in terms of height. In other words, the less distance that has to be traversed between texts, the more similar their authorial fingerprint.

This prompted further exploration, so a bootstrap consensus tree, a more robust measure of style, was conducted. Figure 2 also represents stylistic clusters, using tree-like branches to show stylistically-similar groupings. Interestingly, O’Brien’s novels continue to cluster with Joyce, with *At Swim Two-Birds* clustering more closely with *Ulysses* and *Finnegans Wake*.

With our cluster analysis and bootstrapping confirming the common belief that O’Brien’s style was strongly influenced by Joyce, we adopted Rolling Delta as a means of pinpointing specific passages of interest within the relevant corpora. Rolling Delta analyses compare the style of one or multiple texts against that of one test text, broken into segments or windows. The closer each line is to the x-axis, the greater the stylistic affinity with that particular segment of the test text. There are a number of places in these texts where O’Brien’s authorial signature is particularly clear. We can identify
Figure 1: A dendrogram representing stylistic similarities between Irish modernist writers.

Figure 2: A further representation of stylistic clusters, using tree-like branches to show stylistically-similar groupings.
these sections as having a distinct crossover between the style of the two authors. O’Brien-like idiom of *At Swim-Two-Birds* emerges, quite strongly, in two sections of *Ulysses* (see Figure 3), and in several sections of *A Portrait of the Artist as a Young Man* (see Figure 4). Interestingly, The Hard Life is stylometrically similar to *Dubliners* throughout, consistently more so than any of Joyce’s own texts (see Figure 5).

**Figure 3:** A Rolling Delta analysis of *Ulysses.*

**Figure 4:** A Rolling Delta analysis of *A Portrait of the Artist as a Young Man.*
3 Literary Interpretations

These results contribute significantly to scholarship surrounding Joyce and O’Brien in that they offer a clear picture of where the styles of both authors are most similar. Throughout this section, we offer literary interpretations based on these findings, giving specific focus to those correlations that we deem most significant. While we analyzed the entire oeuvre of both authors, it is clear that the greatest stylistic similarities exist between *At Swim-Two-Birds* and *Ulysses*. This confirms much of the existing criticism which underlines the overtly Joycean nature of the style adopted by O’Brien in this particular novel. **Figure 6**, taking advantage of the consensus network method as an extended version of the above consensus trees (Eder 2017), simply rearticulates what is represented in the dendrogram (see Figure 1), though limited to a corpus comprised only of works by Joyce and O’Brien. The thicker the line between texts, the greater the stylistic similarity.

Our Rolling Delta analyses also produces significant results in relation to the similarities between *The Hard Life* and *Dubliners*, as well as *At Swim-Two-Birds* and *A Portrait of the Artist as a Young Man*. These are perhaps more easily explained than our findings relating to *Ulysses*, but they are nonetheless worth mentioning as they lend statistical evidence to existing scholarship.

![Figure 5: A Rolling Delta analysis of *Dubliners*.](image)
3.1 Dialects and Protégés

O’Brien’s tendency to present an archetypal Dublin dialect across many of his novels is one explanation for this proximity to Joyce’s style. Clune (1986) argues that it is O’Brien’s Ulster Irish that allows him to reproduce the Dublin dialect so convincingly. He himself claimed that Joyce had the edge on him in this, but there are those who disagree. They argue that only a non-Dubliner could have “caught” his Dubliners so precisely, pinning them down “phrase by phrase” (Clune 1986, 6). Indeed, “Dublin dialogue has a special relish for Brian O’Nolan”, and he praises Joyce for his authorship’s “supernatural skill” (Mays 1974, 246). It is perhaps unsurprising that both writers’ affection for the Dublin dialect results in their styles being so similar. While most of O’Brien’s novels are situated in Dublin, The Hard Life is the closest novel to Dubliners. Published forty-seven years after Joyce’s collection, the proximity of O’Brien’s style to that of Dubliners demonstrates that O’Brien, though not a Dubliner himself, mastered a style long dominated by Joyce. This counters much of the novel’s criticism, which accuses O’Brien of being the overt protégé, too conscious in his attempts at achieving the ideal Joycean parody. Asbee (1991) finds comparisons between The Hard Life and Joyce’s work to be “almost insulting” (91). Our

Figure 6: A further visualisation of the stylistic clusters represented in Figure 1.
analysis illustrates (see Figure 5) that O’Brien’s attempts to capture this style are far from insulting. Instead he has managed to replicate, perhaps not with “supernatural skill”, but certainly with some measurable success, the style encountered in Joyce’s collection.

Other comparisons are often drawn between Stephen Dedalus and O’Brien’s unnamed protagonist from *At Swim-Two-Birds*. Our results present two interesting findings in this respect. In *Ulysses*, the sections most closely aligned with O’Brien’s style are “Oxen of the Sun” and “Eumeaus” (see Figure 3). Incidentally, “Oxen of the Sun” and “Eumeaus” are among the few episodes in which Stephen and Bloom appear together (the other two are “Circe” and “Ithaca”, which, however, clearly stand out stylistically, the former laid out as a drama, and the latter having the catechism-like form of questions and answers). One could put forward the hypothesis that the reason for this similarity is the presence of Stephen, the literary predecessor to O’Brien’s student. We do not consider this a valid explanation: in both episodes, Bloom’s consciousness seems more prominent, while the earlier episodes, where Stephen features more heavily, show little proximity to O’Brien’s style. Furthermore, in “Eumeaus”, Stephen is written with an ironic distance, featuring in passages that could not, over the course of the character’s appearances, be considered as entirely representative of the style to which he is typically attached. We can conclude from this that connections between the young artists in *At Swim-Two-Birds* and *Ulysses* are more symbolic than stylistic.

However, the same cannot be said of *A Portrait of the Artist as a Young Man*, where our results present an interesting correlation. The style of *At Swim-Two-Birds* is very similar to the final section of *A Portrait* (see Figure 4). This passage is dominated by a maturing, morally assured, Stephen Dedalus. Claims that O’Brien’s character is a refiguring of Dedalus are supported, as the style of *At Swim-Two-Birds* is most closely aligned with the section concerning Stephen’s career where he is also a student. Not only does this substantiate considerations of O’Brien’s protagonist as a protégé, it is also a significant finding in terms of how we treat Stephen’s progression from *A Portrait* to *Ulysses*. Our findings suggest that O’Brien’s student has more in common with the Stephen, who is looking to “fly by those nets” (P 231), than with the Stephen
we encounter in Joyce’s longer epic. If we take O’Brien’s protagonist as the test-case, it would seem that his contemporary has written two Stephens, lending evidence to the hypothesis that Joyce consciously constructs his later Dedalus, returned from his travels, as being different from the young artist that we first encounter.

3.2 The Cracked Looking-glass

A close reading of the Joycean passages most similar to *At Swim-Two-Birds* reveals further complexity in intertextual relations. The greatest convergence points between the styles of *At Swim* and *A Portrait of the Artist* occur approximately 64,500 and 74,000 words into Joyce’s novel (see Figure 4). The first of these is when Stephen discusses aesthetics and perceptions of beauty with the dean of studies, but, unexpectedly, their conversation swerves into the linguistic question “whether words are being used according to the literary tradition or according to the tradition of the marketplace” (Joyce 2007, 164). But, as Stephen feels, the difference does not lie only in generic and sociological contexts; it is also deeply entangled in history and politics. – It is in fact the question of dominance and “ownership” of language, which the Joycean hero expresses in such a poignant way:

> The language in which we are speaking is his before it is mine. How different are the words home, Christ, ale, master, on his lips and on mine! I cannot speak or write these words without unrest of spirit. His language, so familiar and so foreign, will always be for me an acquired speech. I have not made or accepted its words. My voice holds them at bay. My soul frets in the shadow of his language. (Joyce 2007, 166)

We could easily imagine O’Brien repeating this confession, admitting that the language he writes is “Joyce’s” before it is his. Perhaps it is “an acquired speech” for him, but he mastered it so well that the languages of the two authors tend to converge and become the closest at the point of Stephen’s reflection. However, O’Brien’s student-narrator does not share Dedalus’ sense of alienation from “his” language—he recognizes the heteroglossic nature of literature, especially the novel (which anticipates Bakhtin’s characterizations of discourse in this genre). Since nothing can
be done to revive the utopian dream of originality, consequently, he decides to make
dialogism in the form of patchwork, pastiche, and parody the fundamental principle
of his writing. As he explains to Brinsley:

The entire corpus of existing literature should be regarded as a limbo from
which discerning authors could draw their characters as required, creating only
when they failed to find a suitable existing puppet. The modern novel should
be largely a work of reference. Most authors spend their time saying what has
been said before – usually said much better. A wealth of references to existing
works would acquaint the reader instantaneously with the nature of each
character, would obviate tiresome explanations and would effectively preclude
mountebanks, upstarts, thimble-riggers and persons of inferior education from
an understanding of contemporary literature. (O’Brien 2001, 25)

He considers literary tradition not only a reservoir of characters, motifs, and topoi,
but also a repository of styles from which writers can and should draw freely. As
our analyses demonstrate, O’Brien not only preaches this through his hero, but also
effectively puts this into practice.

At this point it is worth noting how the author of *At Swim-Two-Birds* exploits
references to his master, appropriating the words Stephen used in the above-
mentioned quote to make them his own. While “Christ” occurs in O’Brien’s novel
only six times, and always in connection with Sweeny, usually in his verses, rather
surprisingly, “God” features among the 100 most frequent words. This high frequency
is probably connected with the exclamation “by God”, used by many characters. This
marks the language of O’Brien’s novel as distinctly Irish (Camden cited in Walshe
2009, 129–30), and even more as “literary spoken Irish” (Walshe 2009, 130). The
frequent use of “by God” may be O’Brien’s (deliberate or unconscious) strategy
for parodying the well-known cliché of Irish speech. But it may also be read as his
“domesticating” response to the word so disquieting to Stephen by translating the
doubly foreign “Christ”—because the name itself comes from Greek and because
Stephen attributes it to the Englishman—into the more familiar name “God” (As
little Stephen reflects, “God’s real name was God” [Joyce 2007, 27]). Moreover, the frequency of the word is close to the words “uncle” and “Trellis”, whose names might be seen as “authority figures” attempting to control the undisciplined characters: the former the student-narrator, the latter a whole group of his literary creations: Furriskey, Pooka MacPhellimey, the Lamonts, Shanahan and Orlick. God is also a figure of ultimate authority, and the supreme creator. However, in O’Brien’s novel, unlike Stephen’s artist, the author does not, emphatically, remain “like the God of the creation […] within or behind or beyond or above his handiwork, invisible, refined out of existence, indifferent, paring his fingernails” (Joyce 2007, 187). On the contrary, – God features overtly at all levels of narration, incarnated as the student, Trellis, Finn, Sweeny, and the implied author responsible for composition of the whole novel. We could perhaps read the exceptionally high frequency of “God” as a metafictional gesture hinting at the writer-creator’s controlling presence in the text. However, the omnipresent authors in O’Brien’s novel are hardly omnipotent, or able to control their creations, thus satirizing Stephen’s ideal.

Poking fun at the writer as “master” is evident in O’Brien’s usage of this word in his novel. It is one of the least often encountered lexemes in At Swim-Two-Birds, appearing in the text only ten times, excluding compounds such as ‘master-printer” and “master-key” from the count. It is used four times by Trellis’ maid in reference to her “master” (O’Brien 2001, 215–6), once when Trellis’ characters use it when speaking about him (O’Brien 2001, 181), and in two more occurrences when they are in plural, referring to great Russian novelists, so most of them are in fact synonymous with “author.” Two further instances might be interpreted as allusions to Joyce. The first one (and the first time the word appears in the novel) occurs in an excerpt from Christian Brother’s Literary Reader describing alcohol as a “terrible and merciless master” (O’Brien 2001, 22); Joyce was known for his drinking problem, so it may (or may not) be associated with him. This may also prefigure the sleeping drugs with which Trellis’ characters intoxicate their master in order to outwit him. However, the other case leaves no doubt as to its Ulysses provenance. When the Pooka visits Trellis to torture him, the half-asleep writer mistakes him for a servant, but McPhellimey corrects him in rather lofty style, concluding with the statement:
“It is not false that a servant is a servant but truth is an odd number and one master is a great mistake. Myself I have two” (O’Brien 2001, 173–4). Puzzled and upset by the unwanted company, Trellis wishes to “re-enter the darkness of his sleep” (O’Brien 2001, 174) but is, nevertheless, attracted by “allogamy” and “arachnoid”, two words unknown to him, and asks the Pooka for explanations. The interest in rare words may be seen as another of Trellis’s Joycean traits.

His words echo Stephen’s witty conversation with Haines in the opening of *Ulysses*, in which Dedalus admits bitterly that he is “a servant of two masters […] an English and an Italian”, “[t]he imperial British state […] and the holy Roman catholic and apostolic church”, and perhaps even “a third” wanting him for odd jobs (Joyce 2008, 1: 638, 643–4, 641 respectively). By twisting Stephen’s confession, the Pooka embraces the fact that he is “authored” by others, forced to speak and be spoken about in “their language”, seemingly not minding this. However, on the narrative level the conversation leads directly to the climactic moment of violence when Trellis’s imaginary characters turn vengefully against him, while the description of tortures inflicted on their master is communicated in the language of elaborate parody smacking of the Oxean, Cyclopic, and Eumaean style of *Ulysses*. Clearly, in O’Brien’s literary universe, it is the “Liber Lord” (Joyce 2012, 250) who plays the role of oppressor. We can only wonder at how different these words are now on Stephen’s lips and on the leaves of *At Swim-Two-Birds*.

“Ale” continues to contribute to this network of allusions. It occurs in Joyce’s novel only once, and in O’Brien’s book twice: first, in the phrase “a proprietary brand of ale” printed in large letters on the student’s mirror in which he tries to get a glimpse of his face when shaving, and secondly, in a conversation about quality of beer, as an ironic example of a miracle in which spirit is turned into water (O’Brien 2001, 11, 46 respectively). The first use reminds us of Stephen’s definition of Irish art as “the cracked looking glass of a servant” (Joyce 2008, 1: 146). Once we notice the presence of the Joycean “ale” in the passage, the seemingly naturalistic description of the mirror becomes symbolic of O’Brien’s position in the literary tradition, supplied “gratis” by the compatriots of “Messrs Watkins, Jameson and Pim.” Tellingly, it is on the washstand with the mirror where the student keeps his books, “generally
recognized as indispensable to all who aspire to an appreciation of the nature of contemporary literature […] ranging from those of Mr. Joyce to the widely read books of Mr. A. Huxley, the eminent English writer” (O’Brien 2001, 11). For O’Brien, this tradition bears additionally an imprint of his gigantic predecessor, looming large over the whole territory of Irish letters, “between the words of which [he] had acquired considerable skill in inserting the reflection of [his] countenance” (O’Brien 2001, 11). He responds to this challenge with humour and irony so typical of him, first proposing his own theory of aesthetics of the novel grounded in non-originality, recycling, pastiche and parody, and then putting it masterly into practice.

3.3 The Style of Parody
As already noted, in Ulysses, our Rolling Delta analysis demonstrates significant similarities between the style of At Swim-Two-Birds and the “Oxen of the Sun” and “Eumaeus” episodes (see Figure 3). Our working hypothesis for explaining this proximity is parody, a dominant feature of these two episodes. Neil Corcoran (1997) has suggested that the basic structure of O’Brien’s novel “is derived from the ‘Cyclops’ episode of Ulysses, whose technique (named ‘gigantism’ by Joyce) involves alterations of more-or-less realistic passages, which reproduce conversations of a Dublin pub with parodies, often very extensive, of various conventions and cliches of Irish writing” (22). While Corcoran’s observation clues us to consider parody as an important source of the stylometric similarities of the two authors, our analyses show that it is not “Cyclops” but these other episodes that are closer to the style of O’Brien’s novel (Figure 3). “Oxen of the Sun” is particularly parodic, as it presents a panorama of historical styles of the English language, starting from Old Anglo-Saxon chronicles and elegies, to modern slang contemporary to the author. Very much a Menippean satire, At Swim-Two-Birds is also intensely parodical, and like “Oxen of the Sun”, draws upon a wide range of sources, from “high” modernist works, through translations of old Celtic legends, to a range of non-literary styles, including correspondence with a horse racing pundit. According to Leighton Pratt, O’Brien parodies as many as thirty-six different literary styles, and “produces forty-two extracts by way of pastiche” (62).1

1 Our thanks to Barbara Szot for bringing our attention to this article.
Thus, stylistic similarities between this episode and O’Brien’s novel may be due to their polyphonic—in the Bakhtinian sense—texture, rather than affinities between styles of particular heroes. In fact, the Russian scholar recognizes such a mosaic of different types and forms of language as constitutive of the novelistic genre *per se*. Interestingly, the high frequency of the word “said”, the ninth item in the list of words in *At Swim-Two-Birds*, may be interpreted as the narratorial signal of languages of the others, laying bare the novel’s heteroglossia. A comparison with word frequency lists for the British National Corpus (Leech, Rayson and Wilson 2001) and the American English (Davies 2014) suggests that it may be a specific feature of O’Brien’s style, at least in this novel. The BNC-based frequency list includes “said” in 51st position (53rd for the written language), while the AE frequency list features it in 19th position, this being close to the 21st position of “said” in the frequency list for the whole *Ulysses*. This also indicates that *At Swim-Two-Birds* contains much more directly reported dialogues than *Ulysses*. Predominance of dialogue may be an additional reason why their styles converge at the end of “Oxen” and in the beginning of “Eumaeus”, where the style becomes distinctly conversational (see Figures 7 and 8).

While “Oxen of the Sun” places its emphasis on literary imitations, in “Eumeaus” attention is turned to the bourgeois. The aforementioned similarity with the ironic Dedalus the student, and the corresponding similarity with the equally ironic Dedalus the bourgeois, underlines O’Brien’s treatment of the bourgeoisie as being utterly Joycean. *At Swim-Two-Birds* appears as a novel-length continuation of the satire started in the second half of *Ulysses*, and indeed in the final section of *A Portrait of the Artist as a Young Man*. Significantly then, our results support ironic readings, rather than autobiographical readings, of this text.

As regards “Eumaeus”, another interpretive possibility is connected with W. B. Murphy, Skin-the-Goat Fitzharris and the anonymous narrator, the storytellers of this episode, weaving fantastic tales in a rambling style, have parallels in *At Swim-Two-Birds*. In both novels we come across very long, complex sentences beginning with participles or participle phrases. It may be sufficient to compare the opening sentences of both texts: “Having placed in my mouth sufficient bread for three minutes’ chewing, I withdrew my powers of sensual perception and retired into
the privacy of my mind, my eyes and face assuming a vacant and preoccupied expression” (O’Brien 2001, 9), and “Preparatory to anything else Mr Bloom brushed off the greater bulk of the shavings and handed Stephen the hat and ashplant and bucked him up generally in orthodox Samaritan fashion which he very badly needed” (Joyce 2008, 16: 1–3). Then, the narratorial voice of “Eumaeus” engages in elaborate, extended, often convoluted and clichéd explanations, full of digressions and aside comments that span over a dozen lines. Similar elaborateness and length can be found in O’Brien’s narrator-student when he gives accounts of his everyday activities, and in parodies of old Irish tales retold by Finn McCool about Sweeney. Such long,
complex sentences are built of enumerations connected with “and”, prepositional phrases containing “in” “with”, “for”, “on” “to” (also used as part of the infinitive) and comparisons “as … as” (“as” also functioning as a conjunction). Both authors have a tendency to use extended periphrastic expression abounding in the preposition “of” and the definite article “the”; for example: “an absence of movement on the part of the cerebral mechanism” (O’Brien 2001, 13) and “the harmless necessary animal of the feline persuasion” (Joyce 2008, 16: 870). All of these lexemes are found among the 30 most frequent words in both texts. Admittedly, they are nearly the same as the list of word frequencies for the British National Corpus, with the exception of “as”, which appears as the 40th item in it (Leech, Rayson and Wilson 2001). So, from the perspective of close reading, it is the way in which both authors use the words, rather than their frequency, that accounts for the similarity of styles in this regard.

While length is not a complete measure of complexity, it does offer some indication of how O’Brien and Joyce structure their sentences, and the flow of discourse. Segmenting the texts into sentence sets, we find that in *At Swim-Two-Birds* and “Eumaeus”, while the average is negligible, both authors oscillate from typical length to more complex extremes (see Figures 9 and 11). Arguably, this is typical of modernist aesthetics and the multi-fractal nature of the narrative modes that one encounters throughout these texts. The oscillation evident in the graphs indicates that colloquial speech interchanges with passages stylized as literary language — of exaggerated and parodied novelistic narration or stylized oral narration.

Regarding “Oxen of the Sun”, the complexity of its sentences declines as the episode draws to conclusion (see Figure 10). This would account for the vernacular style that Joyce adopts at this point. What is noteworthy, however, is that a further Rolling Delta analysis (see Figure 7) reveals that, as “Oxen of the Sun” progresses, it gains proximity to the style of *At Swim-Two-Birds*, first approximately 125,000, and then 155,000, words into the novel. The first proximity point coincides with line 942 in “Oxen”, where we find the switch to Edward Gibbon’s style. It is interesting to note that the historian of the Roman empire also demonstrates a strong tendency towards periphrastic expressions with the possessive “of”, rather than the Saxon genitive, and parallel noun phrases, which Joyce picks up and imitates in this
passage (cf. McKenna and Antonia 1994, 85, who say this may be responsible for a higher frequency of “the” and “of” in Gibbon’s passage in contrast to other authors Joyce emulates in “Oxen.”). As we have already suggested, such parallelisms are also
a prominent feature of O’Brien’s student narrator, which may explain why *At Swim-Two-Birds* gets so close to the style of “Oxen of the Sun” at this point. The 155,000-word threshold coincides with a switch in style, initiated by the paragraph parodying John Henry Cardinal Newman (Joyce 2008, 14: 1344), succeeded by imitations of John Ruskin and Thomas Carlyle, before the episode concludes with the contemporary slang, full of highly colloquial, short, idiomatic sentences and phrases, which Joyce himself describes as “a frightful jumble of pidgin English, nigger English, cockney, Irish, Bowery slang and broken doggerel” (Joyce 1957, 1: 138–9, 13 March 1920; cf. Gifford and Seidman 1989, 441). However, we need to note that the style of *At Swim-Two-Birds* diverges from the slang concluding the episode (see Figure 7). This is understandable because nowhere in O’Brien’s novel could we find its obvious stylistic counterpart. Yet, the fact that the stylometric proximity of these texts increases at a steady rate is testament to the manner in which Joyce chronologically parodies the Western literary canon and popular literature of his day. As the episode continues, the style becomes increasingly closer to that which his contemporaries would seek to mimic. Taking both findings into account (see Figures 7 and 10), and remembering our discussion of dialect, we can conclude that O’Brien’s imitation of Joyce is most
successful when centred on his predecessor’s use of more contemporary, and often vernacular, style.

In “Eumaeus”, the moment of greatest stylometric proximity occurs between 6,500 and 8,500 words (i.e. roughly between 16:570 and 16:760, see Figure 8), when Murphy, “the doughty narrator”, begins to brag about travels, which his listeners clearly doubt. It is the excerpt in which the sailor’s voice is heard most directly, and the style becomes most conversational. Dialogues are interspersed with extensive paragraphs in the narrator’s voice, containing such extremely extended sentences as the one quoted above, while the length of sentences oscillates between very short and extremely long ones (see Figure 11). A similar oscillation is also observable in At-Swim-Two-Birds over longer stretches of text, especially when Finn weaves his tale about Sweeny and is interrupted by Furriskey, Shanahan and Lamont’s conversations (see Figure 9). The moment of the stylistic convergence between “Eumaeus” and At Swim-Two-Birds may again be marked by a subtle intertextual connection. The poet John Casey and his “most trying declamation piece”, “a bit of perfect poetry in its own small way” (Joyce 2008, 16:427–8), whom Bloom recalls when Murphy launches into his tales, reminds us of the poet Jem Casey, described by Shanahan as “Poet of the Pick, […] a labouring man, […] but as sweet a singer in his own way as you’ll find in the bloody trees there of a spring day, and that’s a fact” (O’Brien 2001, 74–5).

3.4 Nature of Similarity: Catechistic Style
The proximity to “Eumaeus” is open to alternative interpretation, in that it may not only be “Eumaeus” that our analyses are detecting, but also the beginning of “Ithaca”, with this penultimate episode beginning after 218,700 words, and our Rolling Delta indicating the similarity at approximately the 215,000 mark (see Figures 3 and 12).

It is evident that in At Swim-Two-Birds, O’Brien parodies the catechistic style of “Ithaca” with short, verbless phrases functioning as questions, and definition-like answers suggestive of the style of the penultimate episode in Ulysses. As evidenced (see Figure 12), it is the beginning of “Ithaca” that is stylometrically the closest to At Swim-Two-Birds, the point of convergence marking another scene in which language is in focus. The Rolling Delta indicates that it is approximately at 7,500
words, or line 750 in the episode – the moment when Bloom and Stephen discuss
linguistic and historical contacts between ancient Hebrew and Irish – enumerate
similarities between them, and quote whatever scraps of Hebrew and Irish texts they
remember. It can be argued that their discussion is concerned with intertextuality
entailed in “their archaeological, genealogical, hagiographical, exegetical, homiletic,
toponomastic, historical and religious literatures comprising the works of rabbis and
culdees, Torah, Talmud (Mishna and Ghemara), Massor, Pentateuch, Book of the Dun
By mentioning such books as the Bible and the Book of Dun Cow, both exemplary
repositories of different genres and styles (As explained in Annotations to ‘Ulysses’,
the Book of the Dun Cow, the oldest transcription of Irish literature, contains as many
as “sixty-five different pieces: romantic tales in prose, an elegy on St. Columcille, a
copy of the Voyage of Maeldun, etc.” [Gifford and Seidman 1989, 578]), this passage
directs the reader’s attention to the fundamentally polyphonic and heterogenous
nature of literature, which both writers explore in their novels. This part of Bloom’s
and Dedalus’ conversation concludes with another intertextual interpolation, an
anti-Semitic ballad “Little Harry Hughes” (Joyce 2008, 17: 724–830), chanted by
Stephen. Incidentally, this may offer an additional explanation as to why O’Brien’s
novel, containing several ballads and verses, approximates Joyce’s text at this point.
Bloom’s and Dedalus’ impressions when they listen to each other’s languages can be extrapolated on Joyce and O’Brien to define their ambivalent relation. While the younger one “heard in a profound ancient male unfamiliar melody the accumulation of the past”, the elder saw in “a quick young male familiar form the predestination of a future” (Joyce 2008, 17: 777–80). In other words, the author of *At Swim-Two-Birds* heard in *Ulysses* the language simultaneously “so familiar and so foreign” and by listening to it learned from its author how to accumulate voices and styles of the present and the past in his own polyphonic novel. The latter recognized in his younger compatriot an apt disciple eager to outwit and beat his master, and appreciated a true comic spirit worthy of his praise. But the sinister tone of the song emerging at a seemingly friendly moment of the conversation prefigures O’Brien’s troubled attitude to his literary forefather. As we could see in *Trellis* even a fictional beating can be a painful case.

**Acknowledgements**

Previous iterations of this research were presented at the *Thirteenth International Conference on English and American Literature and Language*, Krakow, April 13th, 2014, and *Digital Humanities*, Lausanne, July 9th, 2014. The authors would like to thank Dr. Laura Pomeroy for giving this paper a final proofread.

**Competing Interests**

The authors have no competing interests to declare.

**Author Contributions**

- Conceptualization: K.B.; M.E.; J.O.; J.R.
- Methodology: M.E.; J.O.; J.R.
- Software: M.E.; Mike Kestemont; J.R.
- Writing – Original Draft Preparation: K.B.; M.E.; J.O.; J.R.
- Writing – Review & Editing: K.B.; M.E.; J.O.; Laura Pomeroy; J.R.

**References**

**Anspaugh, Kelly.** 2004. “Agonizing with Joyce: At Swim-Two-Birds as Thanatography.”


How to cite this article: O’Sullivan, James, Katarzyna Bazarnik, Maciej Eder and Jan Rybicki. 2018. “Measuring Joycean Influences on Flann O’Brien.” Digital Studies/Le champ numérique 8(1): 6, pp. 1–25, DOI: https://doi.org/10.16995/dscn.288

Submitted: 04 November 2017 Accepted: 25 February 2018 Published: 27 March 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Digital Studies/Le champ numérique is a peer-reviewed open access journal published by Open Library of Humanities.