The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations

Benjamin M. Miller
Graduate Center, City University of New York
THE MAKING OF KNOWLEDGE-MAKERS IN COMPOSITION:
A DISTANT READING OF DISSERTATIONS

by

BENJAMIN MILLER

A dissertation submitted to the Graduate Faculty in English
in partial fulfillment of the requirements for the degree of Doctor of Philosophy,
The City University of New York

2015
This manuscript has been read and accepted for the Graduate Faculty in English to satisfy the dissertation requirement for the degree of Doctor of Philosophy.

Dr. Sondra Perl

______________________________
Date Chair of Examining Committee

Dr. Mario DiGangi

______________________________
Date Executive Officer

Dr. Sondra Perl

Dr. Mark McBeth

Dr. George Otte

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK
ABSTRACT

THE MAKING OF KNOWLEDGE-MAKERS IN COMPOSITION:
A DISTANT READING OF DISSERTATIONS
by BENJAMIN MILLER

Advisor: Sondra Perl

Combining qualitative coding with original algorithmic and quantitative analyses, this project aggregates and visualizes metadata from 2,711 recent doctoral dissertations in Composition/Rhetoric, completed between 2001 and 2010 (inclusive), in order to establish an empirical baseline of what new and established scholars in Composition/Rhetoric agree upon as acceptable research in the field. I find that both subject matter and methodologies largely collocate within a small number of clusters, but not without cross-over among these clusters, and I call for increased dialogue among schools focusing on these different methods and subjects.

Chapter 1, “Disciplinary Anxiety and the Composition of Composition,” reviews the history of Composition/Rhetoric’s search for a shared research paradigm, including its potential rejection of that goal. Following Derek Mueller (2009), I argue for “distant reading” (Moretti), through metadata visualization, as a means of keeping abreast of research trends that would be unmanageable through direct reading alone.

Chapter 2, “From Dissertations to Data: My Exhibits and My Methods,” explains how I obtained, selected, and prepared the 2,711 documents that go into my subsequent analysis.

Chapter 3, “Mapping the Methods of Composition/Rhetoric Dissertations: A ‘Landscape Plotted and Pieced,’ ” takes up the question of whether the field has divided along methodological lines, as Stephen North (1987) predicted. After identifying methods used in
dissertations based on their abstracts, I describe correlations between dissertation methods and the graduate schools where they are most frequently employed. Most dissertations used more than one method. I demonstrate that, while aggregable and empirical methods have not disappeared, few schools focus on them; dialectical and text-hermeneutic methods are far more common across the board.

Chapter 4, “Tapping the Topics: What We Study When We Study Writing in Writing Studies,” turns from methods to content. Drawing on a computer-generated topic model of the full text of 1,754 dissertations, I provide evidence both for high-level clustering of topics and for large numbers of dissertations that cut across these clusters. The most common dissertation topics in this sample address the teaching of writing and, in a largely separate cluster, theories of meaning-making.

In Chapter 5, “Toward a View From Everywhere: ‘Disciplined Interdisciplinarity’ and Distant Reading,” I reflect on the benefits and limitations of the methods I have used, and suggest directions for future study.

Although it is generally clear to doctoral students preparing to begin dissertation work that they have a number of methods to choose from, and a number of ways to construct and usefully constrain their subject matter, Composition/Rhetoric as a field has not generally speaking kept good track of trends across institutions, with the result that individual dissertation-writers do not know whether a particular method or subject they are considering is common or quirky, cutting-edge or passé. By offering a recent, zoomed-out view beyond the vantage point of any one program, these analyses provide a shared map of where Composition/Rhetoric doctoral research has been, so that researchers, thesis committees, and curriculum-planners can make more informed local decisions about where their research should go next.
ACKNOWLEDGEMENTS

There’s something tricky about the genre of the dissertation acknowledgements, because I’m tempted to thank everyone who made my graduate education possible, and I can’t: there are too many of you to name. “I am a part of all that I have met,” as Tennyson would have it, and I am grateful to all my classmates, my students, my teachers, my friends.

I am particularly grateful to my dissertation committee members, who have given me a wonderful combination of encouragement to up my game and encouragement to let it be enough, already. George Otte’s insights into and questions about the shapes emerging from my figures were essential in forming my argument, especially in the chapter on topic modeling. MarkMcBeth’s advice on managing time, workflow, and play has been sustaining me for years, and I know will continue to do so. My advisor, Sondra Perl, has profoundly shaped my trajectory as a scholar and a teacher; from my first pre-registration semester in 2007 through many courses and collaborations until the present day, she continues to embody for me an ethic of care: care for other human beings and care for the quality of the work, achieved through careful attending to both. It has been an honor and a privilege, and I hope I can live up to your example. Thank you.

My peers and colleagues, far and wide, have been enormously supportive: people talk about competitive grad school environments, and I think of all your generosity and good spirit, and I’m glad I don’t know what those other people are talking about. Micki Kaufman, my data visualization guru, provided essential guidance and code as I was preparing documents for topic modeling. I sometimes wish I could channel Amanda Licastro’s inexhaustible focus and drive, and I’m grateful for all our shared projects on which I almost can, a little. Thanks for all our many, many brainstorming sessions and our writing-alone-together meetups, of which I’m sorry
there couldn’t be more. Jill Belli and Dominique Zino, you inspire me with your successes, and showed me multiple ways to get this thing done. And all members past and present of the Graduate Center Composition/Rhetoric Community (however you abbreviate it): thank you for making CUNY feel like home.

Outside of CUNY, I am grateful for the path blazed by scholars of Computers and Writing such as Derek Mueller, Collin Brooke, Doug Eyman, and Cheryl Ball, who not only made it possible for me to justify this ambitious digital project, but were also kind enough to bat around ideas with me at conferences and after speaking gigs. Extra credit goes as well to Mary Kate Hurley, who more than once cut through my writer’s block with a mind sharpened on ancient knights and wandering poets; I should call with my drafts more often and earlier.

I hear the music playing, but oh, so many more people to thank! Duncan Faherty for supportive and inquisitive feedback on my initial dissertation prospectus. Talia Schaffer and all the students in the Fall 2014 dissertation workshop for taking the time to respond to something so far out of your usual norms; thanks, especially, to Andrew Statum, whose on-the-spot summary of what is at stake in my research was spot-on, and became like a motivational poster hovering above the screen as I wrote. You nailed it, Andrew; thank you.

Thanks, Joseph Ugoretz and all the Instructional Technology Fellows, who not only made it possible for me to continue past my first five years, but who made it my job to level up on ever more digital tools. On a related note, a tip of the hat to my teachers in the Instructional Technology and Pedagogy certificate program, including Joe as well as Steve Brier, Matt Gold, and Kimon Keramidas, for giving me the confidence to keep exploring new tools until I find the right one for the job. I probably would not have made the jump to R otherwise.
Thank you to the editors and reviewers of the NCTE journal *College Composition and Communication*, including Kathleen Blake Yancey, who first published an earlier version of chapter 3 in the September, 2014 issue (66.1).

Thanks to my parents, Marc Miller and Judith Miller-Greenberg, my step-parents, Martin Greenberg and Rebekah Alpert Miller, and my parents-in-law, Michele and Allen Morrison, for listening when I tried to explain what I meant by “Composition / Rhetoric”; it may have felt sometimes like I would never finish my PhD, but you had the grace never to say so and to convince me of the opposite.

And my eternal gratitude to my wife, Sara Morrison, who puts up with my late nights, who spent extra time solo-parenting so that I could finish chapters and make presentations, and whose influence on my attitudes toward disciplinarity and research is immeasurable. You’ve made it possible for me to major in meta and somehow neither of us has gone mad. I love you.
# Table of Contents

## 1: Disciplinary Anxiety and the Composition of Composition

- Dissertations as Disciplinary Descriptors ................................................................. 4
- Anxiety of (Outward-facing) Influence........................................................................ 6
- Reports of the Death of Paradigm Hope have been Greatly Exaggerated ............... 8
- What Have the Knowledge-Makers Made, and How Have They Made It? .......... 11

## 2: From Dissertations to Data:

- The Origins and Extent of This Dissertation’s Exhibits............................................. 14
- Casting a Wide Net, Then Filtering ............................................................................ 17

## 3: Mapping the Methods of Composition/Rhetoric Dissertations:

- A ‘Landscape Plotted and Pieced’............................................................................... 22

## 4: Tapping the Topics:

- What we study when we write in writing studies....................................................... 40
- Methods ....................................................................................................................... 44
- Distant Reading Requires Close Reading ................................................................. 46
- Weights and Measures: What are Grad Students Writing About? ....................... 49
- Most Dissertations Address Multiple Topics .......................................................... 53
- Topics and Individual Dissertations: Overview and Examples ............................. 56
- Similarity Clustering of Topics .................................................................................. 64
List of Tables

Table 3-1. Method tags................................................................. 25-26
Table 3-2. Most dissertations use multiple methods. ......................... 30
Table 4-1. Top 10 topics by overall contribution to the corpus. ............ 50-52
Table 4-2. Upper outlier topics for percentage contributed as the top-rank
topics within a dissertation...................................................... 55-56
Table 4-3. Top 5 topics in Anne Whitney’s dissertation....................... 57-58
Table 4-4. Top 5 topics in Nevin Leder’s dissertation.......................... 60-61
Table 4-5. Top 5 topics in Derek Mueller’s dissertation......................... 62-63
Table 4-6. Details of three topics for comparison and contrast............... 64-65
Table 4-7. Similarity between topics discovered through hierarchical
clustering .................................................................................. 67
Table 4-8. Major divisions among topics in the model as determined by
hierarchical clustering................................................................ 69-70
Table 5-1. Reach of Topic Clusters Across Dissertations..................... 82
List of Figures

Figure 2-1. Summary of data and metadata fields used in the dissertation........ 15

Figure 2-2. Dissertation data is U.S.-focused, but not U.S.-exclusive .............. 19

Figure 2-3. Most doctoral programs in the U.S. now have some comp/rhet
dissertations. ................................................................. 20

Figure 3-1. Methods are not evenly distributed across schools, and few
schools span the full range of methods................................. 31

Figure 3-2. Dialectical methods are the most common, followed by
phenomenological.............................................................. 33

Figure 3-3. Consortium schools are significantly less likely to produce
Experimental / Quasi-Experimental studies, and
significantly more likely to produce Philosophical /
Theoretical or Historical / Archival studies, than non-
Consortium schools. ......................................................... 35

Figure 3-4. The top four methods by frequency occur throughout
Consortium schools, but at only a subset of non-Consolidium
schools. ........................................................................ 37

Figure 3-5. Some method clusters of the largest non-Consolidium schools
look like clusters of Consortium schools.............................. 39

Figure 4-1a/b: Three example dissertations in relation to two extracted
topics............................................................................... 42

Figure 4-2: Hierarchical clustering of topics by similarity, as a means of
selecting the number of topics................................................. 46
Figure 4-3: Topic browser in R, preview mode .............................................. 47

Figure 4-4: Topic browser in R, detail mode ........................................... 48

Figure 4-5: Contributions of topics to the text, aggregated across all
dissertations by rank within individual dissertations ................. 54

Figure 4-6. Similarity clusters of topics within the model ....................... 66

Figure 4-7. Major divisions among topics in the model as determined by
hierarchical clustering ................................................................. 68

Figure 4-8a/b. Screenshots of interactive figure for finding topics that tend
to occur in the same dissertations ............................................. 75-76

Figure 4-9. Co-occurring topics overlaid with major divisions discovered
by hierarchical clustering .......................................................... 78
Chapter 1: Disciplinary Anxiety and the Composition of Composition

From the start, then, this field has been marked by its multimodality and use of starting points from a variety of disciplines, all marshalled to investigate a unique and pressing set of problems.

But what are the criteria by which a field may be judged a functioning discipline? The question is an important and tough one to answer.

—Janice Lauer, “Composition Studies: Dappled Discipline,” p. 22

To declare oneself a "compositionist," except in certain circles, is to risk a blank stare.

For all that the field has existed for at least half a century,¹ our terms are not that well-known. One possible reason that people outside the field don’t know what we do is that we don’t know what we do, or even necessarily who “we” are. Even within the world of Composition/Rhetoric,² that is, the methods we use and the very subject matter we engage with are not always agreed upon.

1 Knowing where to begin the count is a matter of some debate, but whether we date from the founding of the Conference on College Composition and Communication (CCCC) in 1949, the Braddock Report in 1963 (cf. North, Making 17), or the Dartmouth Conference in 1966 (cf. Harris 4-5), 50 years seems a safe minimum.

2 Throughout this dissertation I will shift among several terms for Rhetoric, Composition, and Writing Studies; this is by design. In addition to the sonic variety gained from avoiding straight repetition – and I will often need to refer to the field as a whole, so there would otherwise be quite a lot of repetition – Brad Lucas suggests that the fluidity of names for the field is metonymic to the fluid and hybrid identities claimed by its members, perhaps for pragmatic reasons. As he writes,

[... T]o adhere to one label [...] is as easily justified as any other, and the ambiguity surrounding any one term for the field reflects its unanswered problems. Quite often, “composition and rhetoric” and “rhetoric and composition” are equally acceptable terms in the field, and such tolerance reflects the field’s characteristic preference for identity instability. (1–2)
Writing scholars have struggled in professional publications to articulate a disciplinary core since at least the mid-1980s, when two major studies of Composition’s collective efforts appeared in consecutive years, reaching opposite conclusions about the field’s trajectory: George Hillocks’s *Research on Written Composition: New Directions for Teaching* (1986) and Stephen North’s *The Making of Knowledge in Composition: Portrait of an Emerging Field* (1987).

Explicitly calling his work a "meta-analysis," Hillocks (with the help of a team of graduate students) aimed to aggregate the findings of empirical studies of writing process and writing pedagogy, to gain predictive power through increased sample size. Even in the absence of a grand unified model of how writing works and how we know it, he insisted, "systematic and thorough reviews of research can help us to identify variables which might prove significant" (97) – and while "such variables can never be completely controlled, [...] the more teachers involved, the more reliable will be the generalizations emerging from the research" (99). At the core of Hillocks's study, then, was the assumption that the research being done in Composition could be compiled and aggregated, with homogeneity of findings across several contexts the measure of a given conclusion's strength. And, given the findings, he was hopeful: “We have a body of knowledge about the composing process which suggests something about teaching and which raises very interesting questions for further research,” he declared in his introduction (xvi). “The climate for improving the teaching of writing has never been better. In short, although many problems remain, we have reason for optimism” (xvi–xvii). Note that, for Hillocks, the field’s central concern is clear, and it is twofold: gaining “knowledge about the composing process” and “improving the teaching of writing.”
North was less sanguine, on both the clarity of the goals and the prospects of achieving them. In *The Making of Knowledge in Composition*, he called into question both the aggregability of research in the field and the centrality of teaching in that research. Motivated by a student’s failure on his doctoral oral exams to produce a synthetic view of Composition’s knowledge-base (iv), North drew on his own experience and reading to survey the “modes of inquiry” by which knowledge is produced in the field (1), and thus “to provide that image of the whole” for himself (5). Working in this way, he located eight such modes, clustered into three major "methodological communities":

- **Practitioners**, concerned with what works in classrooms on a day-by-day basis, sharing ideas mostly through story-telling (what North calls “lore” [23]);
- **Scholars** (Historians, Philosophers, and Critics), working dialectically, primarily from texts, drawing on humanistic traditions; and
- **Researchers** (Experimentalists, Clinicians, Formalists, Ethnographers), working primarily from empirical observation, drawing on social-scientific traditions.

Each community, North claimed, held to an epistemology that was fundamentally at odds with those of the other two. Rather than working together toward a composite understanding of how writing “works,” then, North saw these groups as merely talking past each other, at best, and at worst, competing unproductively for status (321 ff).

He concludes on a note of dire prophecy:

If composition is working its way toward becoming a discipline in any usual sense of that word, it is taking the long way around.

It might not be too much to claim, in fact, that for all the rhetoric about unity in pursuit of one or another goal, Composition as a knowledge-making society is gradually pulling itself apart. Not branching out or expanding, [...] but fragmenting: gathering into communities or clusters of communities among which relations are becoming increasingly tenuous. [...] It is not difficult to envision what will happen if, as is most likely, these forces continue to operate unopposed in Composition. Quite simply, the field, however flimsily coherent now, will lose any autonomous identity altogether. (364-5)
Almost thirty years later, it seems clear that this has not come to pass: with over 70 doctoral programs identifying with Rhetoric and Composition (“Members”), dozens of long-running academic journals, and yearly attendance at CCCC in the thousands, Composition seems alive and well.

How has this happened? Has Composition/Rhetoric overcome the methodological conflicts North identified by settling on one dominant mode of knowledge-making? Have we instead somehow attained an “inter-methodological peace” (Making 369) based on the mutual understanding North hoped his book would help achieve? Or have we simply fragmented without noticing it, retreating into adjacent but separate rooms at shared conferences, maintaining several conversations that never meet?

To seek some answers, this dissertation will study a central knowledge-making genre in our field: namely, PhD dissertations.

**Dissertations as Disciplinary Descriptors**

As a measure of “disciplinary identity,” dissertations have much to recommend them: as Todd Taylor argues (citing Joseph Moxley), dissertation authorship affords a more democratic view of the field’s membership than articles or books: whereas “it is estimated that about 10 percent of the professionals in any field are responsible for publishing about 90 percent of the journal articles and book titles” (Taylor 143), nearly everyone pursuing a career in the field writes a dissertation.³ Moreover, a journal article is a momentary intervention in a particular

---

³ Major exceptions include those who pursue “alt-ac” careers such as publishing, or those who complete MA programs and teach at the secondary level; in either case, doctorates are not prerequisites for employment. Yet it
argument, whereas a dissertation – given its role in academic hiring, especially at research-focused institutions – is a statement of how one wants to be seen, as what kind of scholar.

Some might object that dissertation research, because of constraints on graduate students, is fundamentally different from the “real” work of the field. I would argue, however, that these constraints make dissertations even more relevant. By definition, dissertations are written by committed scholars who have sought out training in the discipline and sustained effort over a length of time (now averaging over 5 years). Conference presentations, though perhaps the more common form of disciplinary contribution – many people will present multiple times per year – do not require the same sustained engagement. For good or for ill, dissertations serve a gatekeeping function: before it can pass, a dissertation must be approved by a team of established scholars who recognize its work as being relevant to – and advancing the knowledge of – “the field,” as locally construed.

For this reason, although the dissertation-writing population does not directly include the ongoing research agendas of many well-established disciplinary figures, those figures and agendas are in many cases represented in dissertations through advisors and influences: as Marilyn Vogler Urion has argued, building on an idea of Marilyn Cooper, “when advisors ‘teach’ dissertations, they/we (shifting pronouns becomes difficult) are teaching a world view” (Vogler Urion 10). Thus, aggregating these local judgments across a large body of committees and schools should point to overlaps and disjunctions in how the field constructs itself: a multi-authored map of the discipline’s dappled surface.

seems safe to say that for most of the field’s members, a position at college or university is Plan A, and for these positions a doctorate is standard.
In sum, dissertations are key works produced by those entering Composition/Rhetoric’s knowledge-making community, reflecting their bona fides and training in performing the work of the discipline. It remains to be seen, however, how consistently the nature of that work is agreed upon.

**Anxiety of (Outward-facing) Influence**

For anxiety about the field’s viability, and its status as a *discipline* worthy of the word, continues. Richard Haswell, in a much-cited article with the provocative title “NCTE/CCCC’s War on Scholarship,” presented evidence that “for the past two decades, the two organizations have substantially withdrawn their sponsorship of one kind of scholarship,” scholarship which he called “RAD: replicable, aggregable, and data supported” (198). Throwing a gauntlet to the field, he writes,

> What happens when a professional organization is at war with its own scholarship? What happens when the flagstaff organizations of a disciplinary field stop publishing systematically produced knowledge? The answers to these questions are not known because nothing like these events has happened in the history of academic disciplines. (220)

In other words, Haswell claims, “systematically produced knowledge” is part and parcel of disciplinarity in the academy⁴: without it, Composition is not a discipline, no matter how many graduate students or tenured professors.

Similarly, Kurt Spellmeyer argued in 2003 that “comp, in spite of its expressions of contentment, is still not much of a discipline” (84). To become one would, for Spellmeyer, require two things: first, “an adequate *systemic* understanding of how [its] knowledge fit within a

---

⁴ This was not always the case; as Maureen Daly Goggin notes, in a classical college prior to the 19th century, “the goal […] was not to create knowledge; that was not within the province of students or faculty. Rather the goal for faculty was to instill knowledge, moral values, and piety, and the goal for students was to demonstrate that they had attained these ends. In short, it served to construct a particular way of thinking and behaving” (4–5).
larger constellation of knowledges, some rising in value and influence, some declining, some moving to the center, and some moving to the periphery” (85, italics in original); the second, dependent on the first and a sign of its success, is that “the work we do [would] ever travel[] outside of the field” (84). Without being able to articulate to the outside world the nature of what Comp/Rhet’s researchers, scholars, and practitioners know and do, the field renders itself irrelevant, if not invisible, to the rest of academia.

Spellmeyer and Haswell are far from the first to shed ink on the question of Comp/Rhet’s disciplinarity, and they won’t be the last. We now have so many articles and collections debating what Composition is – what Jessica Yood has called a "metadisciplinary turn" in the field, and Russell Durst critiques as "an inordinate amount of time defining the field, cataloging it, classifying it, and critiquing it" (qtd in North, "Death" 196) – that in fact a new backlash has emerged, a call to get over the question of what Composition is. Why, this argument goes, should we care whether Composition is a discipline? Aren't we beyond the need for some shared paradigm? Doesn't post-modernism teach us that everything is radically fragmented anyway?

For example, Stephen North – in a dramatic turnabout from his earlier book – has urged Composition researchers to give up the search for "some (imagined) cumulative disciplinary effort," which he refers to as the "founding Myth of Paradigm Hope" (“Death” 195): a myth that compositionists invoke, he claims, so as to summon or create an illusory collective body. Instead, he calls for a proliferation of place-based studies of writing in practice, predicting with apparent enthusiasm, or at least relief, that “we will have more research more accessible more quickly, but it will also be both far less transportable and – though the term may seem unpleasant – far more disposable” (205).
Along those lines, Thomas Kent directly contradicts the findings and assumptions of Hillocks’ meta-analysis: under the heading “Writing Cannot Be Taught,” Kent argues that “if writing cannot be reduced to a process or system because of its open-ended and contingent nature” – a post-modernist premise he has spent the previous several pages defending – “then nothing exists to teach as a body-of-knowledge” (149).

ECHOING NORTH, David Smit calls on the profession "to capitalize on the fact that it is now localized, historicized, and contingent, both theoretically and pedagogically" (230) by openly declaring that we don't – and can't – know anything cumulative or transferable about writing. Metaphorically speaking, says Smit, "there is no such thing as 'tree-ness'; there are only particular trees" (230).

Tempting though these isolationist positions might be, it remains the case that an oak is more like a pine than a porcupine. That is, despite infinite local variation, too close attention to local details can mask larger patterns and trends – and ignorance of those patterns, to extend Spellmeyer’s argument above, could have serious local consequences if it leaves us no way to argue for the value of our work.

**Reports of the Death of Paradigm Hope have been Greatly Exaggerated**

I began this introduction by suggesting that even those who identify with Comp/Rhet don’t know, necessarily, what it means to study Comp/Rhet. The chapters that follow are my attempt to educate myself and others by investigating what a broad swath of scholars identifying with the field have done recently.

Chapter 2, “From Dissertations to Data: My Exhibits and My Methods,” explains how I obtained, selected, and prepared the 2,711 documents that go into my subsequent analysis.
Chapter 3, “Mapping the Methods of Composition/Rhetoric Dissertations:
A ‘Landscape Plotted and Pieced,’ ” takes up the question of whether the field has divided, as North predicted, along methodological lines, by describing correlations between dissertation methods and the graduate schools where they are most frequently employed.

Chapter 4, “Tapping the Topics: What We Study When We Study Writing in Writing Studies,” turns from methods to content. Drawing on a computer-generated model of the full text of 1,754 dissertations, I provide evidence both for high-level clustering of topics and for large numbers of dissertations that cut across these clusters.

In Chapter 5, “Toward a View From Everywhere: ‘Disciplined Interdisciplinarity’ and Distant Reading,” I reflect on the benefits and limitations of the methods I have used, and suggest directions for future study.

And I should clarify at the outset that I do believe further study is necessary. After all, North’s attempted absolution of the field’s “paradigm guilt” hasn’t taken hold in all quarters. Even writing in the same edited collection as Smit, Kristine Hansen prominently positions the quest for disciplinarity in her title, “Are We There Yet? The Making of a Discipline in Composition”; the fact that her answer remains that “we haven’t arrived yet” (237) doesn’t undermine the element of hope in the word “yet,” or in her concluding call to “conduct more and better research to build a stronger body of knowledge” (260). But we also need to build an index to that body of knowledge, lest it sit inert.

Recently developed digital tools have made such an index increasingly feasible to construct, and my efforts to take the measure of dissertation work extends prior studies, primarily on journals and journal articles. One key precedent is Maureen Daly Goggin's 2000 study

*Authoring a Discipline: Scholarly Journals and the Post-World War II Emergence of Rhetoric*
and Composition, which traces the publication, circulation, editorial stewardship, and university affiliations of the authors in nine major comp/rhet journals, over the years 1950-1997. More recently, Derek Mueller's 2009 dissertation Clouds, Graphs, and Maps: Distant Reading and Disciplinary Imagination examined keywords and citations in 20 years' worth of articles (1987-2006) published in College Composition and Communication (CCC). Mueller has since published articles stemming from that line of research in CCC itself (“Grasping”) and in Kairos (“Views”). Mueller's work is pioneering in the field in its conscious effort to bring recently-introduced techniques from the digital humanities (such as data-mining, tag clouds, and GIS mapping) to bear on the history and present status of scholarship in composition/rhetoric.

My project, like Mueller's, is greatly influenced by the work of Franco Moretti, whose Graphs, Maps, Trees has electrified the community of literary historians, especially with regard to the study of genre. Both Moretti and Mueller argue convincingly in favor of what Moretti calls distant reading, the practice of compiling information about large sets of texts into a series of abstract visualizations – the graphs, maps, trees, and clouds of their titles. Because they enable us to see all the data at a glance, such models can often reveal or suggest systemic patterns that are not easily discernable at more fine-grained levels of detail. Visual models therefore function much like abstracts appended to articles (or dissertations): they simplify, in order to amplify, and give us some indication of what to look for if and when we read on (Mueller, “Grasping” 197–198).

This process can thus contribute to what Mueller calls a "network sense" of the field: “an epistemological capacity for discerning those patterns entangled with a broad set of forces (an actor-network) beyond the text, involving matters of semantic associations, historical orientations, locations, and relationships” (“Clouds” 66) Ordinarily, such associations,
orientations, and so on are formed primarily through local lenses, augmented by personal reading and teaching histories. Distant reading, by compressing texts into metadata about the texts, allows us to “read” far more than we could otherwise, and thus to form new associations, to become aware of new relationships: in short, metadata enables new metacognition. In this way, Mueller writes, “[D]istant reading affords us a new methodology that, by promoting network sense, makes it possible for us to come at the internal problematic of rhetoric and composition differently than has been done before” (“Clouds” 66).

While analysis of metadata does not hold out the promise of perfectly defining the present state of Composition/Rhetoric or of predicting its future, it does offer a widely integrated view as opposed to a purely anecdotal one. In other words, the patterns we abstract from distant reading may enable us to better contextualize the local findings of more traditional reading: they can corroborate – and sometimes challenge – what we have learned to expect through more direct, personal experience. Thus, even if our answers aren't true for all time, they are at least demonstrable, updateable, and comparable to similar studies.

What Have the Knowledge-Makers Made, and How Have They Made It?

That I titled my dissertation as a riff on North’s Making of Knowledge in Composition should signal that I see an affinity between his project in that book and my own here: both are attempts to understand the range of approaches available in Composition/Rhetoric, and especially to explore the existence of cross-talk and mutual understanding among those approaches. At the same time, I think there are essential differences in how North and I go about that task.
In this context, it is useful to consider Lance Massey’s delineation of the contrast between social scientific and humanist approaches to writing – a recurrent trope in histories of Composition – which he achieves by scaling out his examination to a point at which even complex texts can be seen as internally coherent. Writing, from this perspective, is humanist "if the source of its coherence is constructed as the subjectivity of its author rather than the objectivity of its referent" (78, emphasis added). Or again, as he writes later, humanism represents "a commitment to the proposition that interpretive critique is an interesting, valid, and important way to make knowledge in composition" (86). In this light, North's book can be seen as a humanist work at heart, despite its grounding in social scientific theory and its use of the language of participant-observation, because North cites as the advantage of such approaches the unity derived from "the product of a single consciousness" (North Making 5, qtd in Massey 81).

Similarly, Massey defines social scientific writing as "writing in which observation of the external (social) world (irrespective of how much or how little such work acknowledges the fundamentally interpretive nature of such research) is encoded as the primary mode of inquiry" (78, emphasis added). Much of the empirical research published recently in comp/rhet journals would adopt this attitude, because it is no longer defensible to declare that one has found (e.g.) the writing process. Massey sums up by noting that the differences are rhetorical, involving different standards of evidence mobilized by the framework of the text:

And the difference between those standards is precisely the difference between a rational-critical and an empirical-descriptive discourse: one seeks to comment on and, perhaps, change phenomena like the social institutions around and through which we structure our lives; the other seeks to discover what those phenomena are and how they work. (82)

My goal in researching the methodological and topical communities of composition/rhetoric dissertations is, in that sense, opposite that of North in writing MKC. Whereas he set out to
critique and encourage certain community formations, my goal is just what Massey demonstrates North did not do: to empirically "identify which such communities – if any – are present in composition at all" (82).

This dissertation is, therefore, more than another empty invocation of “paradigm hope” – which, despite Smit and North and Kent, has never really disappeared, and if anything seems to be experiencing a recent surge. (Yancey, Robertson, and Taczkak’s Writing Across Contexts, a RAD study of a first-year writing curriculum aimed at teaching for transfer, seems to be pitched in direct opposition to the idea that context is king.) What North most criticizes in the “invocation” of paradigm hope is the Mosaic voice decrying Composition research as bad science in need of reform (“Death” 195) – a voice he identifies first in Braddock, Lloyd-Jones, and Schoer, but which is just as surely visible in his own Making of Knowledge in Composition.

I am issuing no jeremiads. Rather than bemoan something missing or problematic, in the pages that follow I aim to document what has been present in the recent past.
Chapter 2:
From Dissertations to Data: The Origins and Extent of This Dissertation’s Exhibits

Indeed, the RAD methodology is there to deal with research imperfections, which exist in every piece of research ever done. If any scholar questions the inferences [the author of a RAD study] draws from her findings, she has described her system so it can be replicated and her conclusions tested.

—Richard Haswell, “NCTE/CCCC’s Recent War on Scholarship,” p. 203

Although there have been a number of case studies of and reflections by individual dissertation writers, there has not to my knowledge been a large-scale investigation into composition/rhetoric dissertations since Todd Taylor’s “A Methodology of Our Own,” published in 2003 but using data from September 2001. At that time, he found 630 dissertations in *Dissertations Abstracts International* with the subject heading “rhetoric and composition” (Taylor 143). The same search today in ProQuest Dissertations and Theses (PQDT) produces roughly 6,000 results.

PQDT has been the official dissertation repository for the Library of Congress since 1999, as well as a contracted publisher for the National Library of Canada (Palchak); widely available at research libraries worldwide, the database contains metadata (and, in many cases, full text) for over two million dissertations, adding more than 70,000 each year (ibid). For the purposes of the present study, I limited the search to the years 2001-2010 inclusive, selecting full-text-available doctoral-level theses with the subject terms “Rhetoric OR Composition NOT Music.” The time period chosen represents a period late enough to begin after online submission
became common, yet early enough to have allowed two-year embargoed\(^5\) dissertations to become available.

In response to this request, ProQuest provided me with a range of metadata, including abstracts, for 3,013 dissertations (see figure 2-1\(^6\)); they later sent DVD-ROM discs containing full text for 2,949 dissertations. Both numbers are smaller than an online search of PQDT, which for the same time period currently yields 4,122 dissertations meeting my criteria above. In addition, the subject term that ProQuest used to fulfill the request, the term that all the dissertations in the file had in common, was one I had not seen online: “Language, Rhetoric and Composition.” Multiple subject terms (drawn from PQDT’s fixed vocabulary) were merged into a single field, as were keywords (drawn from an open vocabulary of terms provided by authors), with some keywords in all caps and some not; some of the all-caps keywords duplicated some of the lowercase keywords for the same dissertation.

\(^5\) Authors may choose to restrict access to their dissertations for varying amounts of time, often depending on contracts between graduate schools and ProQuest. Two years is common. Reasons for and access to embargoing have been the subject of much recent discussion, especially in the wake of the American Historical Association’s advice that “online dissertations that are free and immediately accessible make possible a form of distribution that publishers consider too widespread to make revised publication in book form viable.” See more at American Historical Association; the WPA-L thread beginning with Wright; and Hawkins, Kimball, & Ives.

\(^6\) Higher resolution versions of all figures are available online at \url{http://majoringinmeta.net/dissertations}. 

---

Figure 2-1. Summary of data and metadata fields used in the dissertation. In blue, metadata fields provided by ProQuest. In red, what I added: method terms, based on reading of abstracts (see chapter 3), and topics derived from a model of the full-text (see chapter 4).
I wrote to ProQuest to ask about these differences, and to ask for data on departments, which were visible online but not included in the spreadsheet. I got back an answer I didn’t fully understand: that the classification system they used internally was different from the one used on the web-based front-end; that the items they could distribute were only those that they were licensed to sell; and that they had given me all the data they had available to give. Though I was frustrated by the mismatch in expectations, I did not press the issue, figuring that anyone else who made the same request that I did would get the same result from the company. (For verification, the PQDT Dissertation/Thesis Numbers – given as “Pub.number” in the spreadsheet – are listed for all included and excluded dissertations in Appendix B.) In retrospect, perhaps it was a mistake not to push for additional clarity: When did the terms change? Was there a plan to change them again? Was there a consistent, mappable process by which the terms were converted from one system to another? What was the difference between the all-caps keywords and the rest? How do authors select subject terms, and does the answer change at different schools or different times?

Instead, to the best of my ability I inspected the spreadsheet on my own and made some inferences: e.g. the all-caps keywords seemed to be automatically generated from the titles of the dissertations, rather than supplied separately and intentionally. A colon was used in one field to separate multiple keywords, and a pipe symbol in another. Certain unusual characters in titles and abstracts seemed to correspond consistently to quotation marks, non-English letters, and dashes. Having made these inspections privately, I used the Google Refine tool (now OpenRefine) to “transform” the data into something I could work with more easily: I stripped the all-caps keywords, split the fields into multiple rows within records, and used cluster analyses to merge apparent misspellings of keywords. After saving the new file with a different name, so
that I could recover the original data should new information come to light, I saved the scripts I used to make these transformations, so they would be replicable. (See Appendix A for an index of these scripts, which are available online at http://github.com/benmiller314.)

**Casting a Wide Net, Then Filtering**

Taylor had warned that “dissertation authors can and do select the rhetoric-and-composition subject heading somewhat by accident as opposed to trying to locate themselves consciously within the field” (143); his solution was to “eliminat[e …] any dissertation that did not emerge from a PhD program that was included in *Rhetoric Review*’s most recent listing of graduate programs in rhetoric and composition” (ibid). Because one of my guiding questions has been to determine the disciplinary scope of Rhetoric and Composition / Writing Studies, I instead opted to investigate all the search results. Taking a broadly pragmatic definition of the field’s interests as improved understanding of how written language is produced, circulated, and taught, I read through all of these abstracts and identified 2,711 of them (90%) as recognizably work in Composition/Rhetoric. An additional 73 (2.5%) I marked as “false positives,” or work clearly outside the field, and another 227 (7.5%) I marked as “maybes”; these latter 300 are excluded from the dataset in all the analyses that follow.

Encoding of exclusion / inclusion, as well as the encoding of methods described in Chapter 3, took place in Microsoft Excel, working from the output of OpenRefine so that I could view each record (i.e. dissertation) in multiple rows, for cleaner keywords and better autocomplete. Initially, I had the records sorted by Pub.number; however, I soon realized that as a result I was moving through the data chronologically and in clusters by school, meaning that I had a skewed sample from which to gather preliminary summaries. After 100 dissertations, during which time I expanded, revised, and grew confident in my methods coding schema, I
went back to Refine, collapsed the multi-row records into single-row records, and
sorted by both year (to allow me to stratify the sample for obtaining preliminary results, tagging
50 dissertations per year at a time) and alphabetically by last name (to pseudo-randomize the
data within each year). The process of tagging all 3,014 abstracts, including re-reading and re-
classifying the initial 100 dissertations as they came up again in the new order, took place over a
period of roughly one full year.

I base my characterization of the 2,711 included dissertations as “in Composition /
Rhetoric” on my own reading of the abstracts, titles, and author-selected content keywords, from
which I can attest that their questions are questions about writing, language, and literacy.
Regrettably, time and resources did not allow me to conduct tests of inter-rater reliability.
Further analysis for confirmation would be both possible and interesting; for example, analysis
of keywords in context (KWIC) and/or citations (cf. Lang and Baehr; Mueller, “Grasping”;
Lucas and Loewe) could reveal more clearly the scholarly conversations from which these
studies derive their exigence. Until such studies are complete, and in the absence of counter-
evidence, I argue that we take seriously the authors’ decision to select the subject term
“Language, Rhetoric and Composition” as indicative of the possibility for fruitful conversation
and collaboration. Though the data I received from ProQuest did not, unfortunately, include
departmental affiliations, it seems clear – regardless of their nominal homes in English or
History, Linguistics, Education, or Psychology – that research about writing is occurring in
dynamic ways in many more places than prior surveys have reached.
Figure 2-2 shows the locations of all 2,711 dissertations in the dataset,\(^7\) completed at 268 schools in all (see Appendix C). As the map shows, the vast majority of these were completed in the U.S. and Canada, though this may be the result of PQDT’s primary institutional sources. Further studies will be needed to incorporate additional datasets from around the world, but the data analysis programs that I have developed, described in greater detail below, are readily adaptable to new data streams, making this project both replicable and aggregable.

What is perhaps most interesting about the schools mapped in figure 2-2 is that there are far more of them than we usually think of as having graduate programs in Composition. The Consortium of Doctoral Programs in Rhetoric and Composition comprises only 76 of these 268

---

\(^7\) Most data analysis and data figures in this dissertation, including Figures 2-2 and 2-3, were produced using the R Language and Environment for Statistical Computing, version 3.1.0 (2014-04-10), nicknamed “Spring Dance,” using the R.app GUI 1.64 on the x86_64-apple-darwin10.8.0 (64-bit) platform. The programs (R scripts) that I used to generate these figures and analyses are reproduced in Appendix F. In some cases, Adobe Illustrator and/or GIMP (the GNU Image Manipulation Program) were used to improve legibility of axis labels and legends or to highlight key features of a graph.
schools (28%). Though these programs do produce 66% of the dissertations (1,776 of 2,711), that still leaves a full third of the doctoral-level research in writing studies taking place outside of the conversations the Consortium was designed to foster.

As another way of thinking about those numbers, consider figure 2-3, which superimposes the map of dissertation-granting institutions in the dataset with a map of doctoral institutions\(^8\) in the United States. Upward-facing triangles represent the latter, while circles representing the former are placed in the foreground with grayscale intensity indicating dissertation count. Strikingly, there are no visible doctoral institutions \textit{without} at least one comp/rhet dissertation.

\textbf{Figure 2-3. Most doctoral programs in the U.S. now have some comp/rhet dissertations.} Superimposed map of doctoral programs listed in the Carnegie classification database, schools where comp/rhet dissertations in the database were completed, and schools in the Consortium of Doctoral Programs in Rhetoric and Composition.

---

\(^8\) The list of doctoral institutions was derived from the Carnegie Classifications 2010 Graduate Instructional Program Classification, counting schools with the following designations: S-Doc/Ed: Single doctoral (education); S-Doc/Other: Single doctoral (other field); CompDoc/MedVet: Comprehensive doctoral with medical/veterinary; CompDoc/NMedVet: Comprehensive doctoral (no medical/veterinary); Doc/HSS: Doctoral, humanities/social sciences dominant; Doc/STEM: Doctoral, STEM dominant; or Doc/Prof: Doctoral, professional dominant.
True, one or two dissertations in ten years is hardly a concentration; even so, the data testifies to a more successful diffusion of interest in the legitimacy of writing and rhetoric as a subject of graduate study than has been documented before.

A third layer in figure 2-3 marks the locations of schools in the Consortium, using a downward-facing triangle in the background to produce a star-like effect. Many of the schools with the highest dissertation counts are, as expected, among those in the Consortium; surprisingly, though, some of the Consortium institutions (starred) have seen fewer dissertations than some of those outside it (unstarred), especially in the west of the country (see Appendix C). We might therefore consider whether the Consortium now needs to be expanded, and/or future surveys sent beyond the usual locations.

Having established an initial dataset through the methods described above, in the next two chapters of this dissertation I present findings from my analysis of that data to examine the methodology (Chapter 3) and topics studied (Chapter 4) in Comp/Rhet dissertations.
Chapter 3:
Mapping the Methods of Composition/Rhetoric Dissertations:
A ‘Landscape Plotted and Pieced’

All things counter, original, spare, strange;
Whatever is fickle, freckled (who knows how?)
—G.M. Hopkins, “Pied Beauty”

Determining who “we” are is no easy matter, but what
“we” do may be one means of getting closer to that end.
— Brad Lucas, Histories of Research in Composition and Rhetoric

Concluding their write-up of the most recent Rhetoric Review survey of doctoral programs in Rhetoric and Composition, Brown et al. call for further research into graduate student identity and training. Noting the “many impediments to gathering accurate data in a timely fashion” through surveys, they nevertheless “strongly encourage everyone to engage directly with data” (339) when and where it can be found. Doing so from as broadly cumulative a perspective as possible, they argue, “will allow for our disciplinary identity to emerge” (ibid). In this chapter, I take up their call by analyzing dissertation abstracts to address two questions:

• What is the methodological landscape of doctoral research in Composition and Rhetoric? That is, what methods do graduate students turn to in constructing their identities as composition/rhetoric researchers, and in what proportions?
• How do doctoral programs cover this territory? That is, do schools tend to produce graduates specializing in the same one or two methods, or to span the range of possibilities?

A sense of that range can be seen in Kristine Hansen’s recent analysis of textbooks commonly used in introductory graduate research methods courses. Across five books chosen for providing “an overview of a range of methods with enough detail about each that students could use the descriptions to plan and conduct their own research” (245) – North’s The Making of
Knowledge in Composition, Lauer and Asher’s Composition Research, Hayes et. Al’s Reading Empirical Research Studies, Kirsch and Sullivan’s Methods and Methodology in Composition Research, and MacNealy’s Strategies for Empirical Research in Writing – Hansen identified no fewer than twelve methods:

- Practitioner / Teacher research;
- Historical;
- Philosophical / Theoretical;
- Critical;
- Experimental;
- Clinical / Case Study;
- Formalist / Cognitive Studies;
- Ethnographic;
- Survey;
- Interview / Focus Group;
- Discourse or Text Analysis; and
- Meta-Analysis.

Several chapters in The Dissertation and the Discipline: Reinventing Composition Studies (Welch et al.) highlight another kind of practitioner study not included here, namely the use of creative writing, including poetry and fiction, as an act of academic investigation (see especially Moore and Woods; Cook and Fike).

But discussions of possibility do not in themselves tell us what methods students take up for extended projects. As Hansen writes, echoing Brown et al, “In the absence of more reliable data, we don’t know the present state of the field. Even if all graduate programs required a course entitled Research Methods, we wouldn’t know what was taught in those courses or whether they are required or elective without asking more detailed questions” (248, emphasis added). Hansen is talking here about the shortcomings of existing survey instruments, but response rates are also
a source of concern, as is the question of where to send the surveys. Moreover, as
Rebecca Rickly points out, even if we knew perfectly what was supposed to be in those courses,
the experience of students taking it can vary widely depending on which faculty member teaches
it (235).

For these reasons, the dissertations themselves are a particularly promising source of
data. By using subject terms selected by the authors when submitting to the ProQuest
Dissertations & Theses (PQDT) database, I was able to cast a wider net than the usual set of
schools included in the Rhetoric Review surveys, and to obtain a significantly higher rate of
return. As discussed in Chapter 2, reading the abstracts of all non-embargoed dissertations with
the subject term “Language, Rhetoric and Composition” produced a list of 2,711 doctoral
dissertations written at 268 schools.

Methods

To tag the abstracts for their methods, I used a coding schema derived from Hansen (see table 3-1),
with a few important modifications added during the initial round of reading. First, I renamed
some of her tags to maximize clarity: e.g. her “Critical” became “Critical / Hermeneutical” to
avoid confusion with cultural-critical studies; “Formalist / Cognitive Studies” became “Model-
Building” both to avoid confusion with formalist pedagogies and to distinguish the cognitive
subject matter from the approach used to research it; and “Meta-Analysis” became “Meta-
Analytical / Discipliniographic” to link these studies to the work of Maureen Daly Goggin and

9 In addition to the final tags presented here, I also coded for “Cultural-Critical Studies,” following Fulkerson
(“Composition at the Turn”), but later decided that this was less of a method, per se, than a Critical / Hermeneutical
strategy (parallel to, say, Freudian or Feminist criticism). Because Critical / Hermeneutical was separately tagged
(see discussion of “multimodal” dissertations below), the Cultural-Critical tag has been ignored in all of the
analyses below, including method counts.
Derek Mueller, who extends Goggin’s term for scholarly activity that “writes the field” (xviii) to encompass the study of such activity (Mueller, “Clouds” 18-19). Second, I added a new category of “Rhetorical-Analytical” to distinguish between two kinds of work with texts I had

Table 3-1. Method Tags.
The following 14 tags, adapted from Hansen (246), were used to describe the methods and methodologies used in the dataset of 2,711 dissertations. Note that while I have attempted to make tags mutually distinguishable, any given dissertation may engage in multiple methodologies and so receive more than one tag. They are presented here in groups loosely derived from Michael Carter’s “meta-genres” in order to highlight similarities and contrasts.

I. RESEARCH FROM SOURCES

Critical / Hermeneutical (CRIT): Qualitative interpretation of texts' content, meaning, and significance, as in literary criticism: asks, “what can we see in the text if we view it through the lens of _____?” or “what does _____ argue?” Texts are treated as crafted cultural artifacts, so claims about them are subject to disagreements among interpreters. In its “critical” aspect, often involved in curation of value, arguing that some set of texts is worthy of scholarly attention. Similar to Rhetorical Analytical in its subjective analysis of textual features; distinct from Rhetorical Analytical in its emphasis on content – the unique what of the text – as opposed to structure (the repeatable how).

Historical / Archival (HIST): Generally speaking, asks “what happened, and why?” and seeks answers via artifacts (including texts). When paired with other terms, may also indicate explicit “situating” of particular phenomena within historical and contemporaneous cultural contexts. Biographies of historical figures are included here, rather than under Clinical, because textual or second-hand evidence tends to dominate in such studies.

Interview / Focus Group (INTV): Studying some external phenomenon through the reactions and “knowledge about” of many individuals or groups. Distinguished from Clinical / Case Study in that interviews are instrumental (“third person”): the people interviewed are not what is being studied. Likely to have questions set in advance, rather than emerging from open-ended conversation, and as such includes questionnaires distributed directly to participants (as opposed to being widely broadcast, as in Survey).

Model-Building (MODL): What North (Making) called “Formalist” and Hansen called “Formalist / Cognitive Studies”: abstract modeling that looks to capture algorithmically or symbolically the relations among parts of a system, with an understanding of the system's dynamics as a primary goal. For example, actor-network theory would be one rubric (or lens) for formalist analysis; Flower & Hayes' 1979 cognitive model would represent another, drawing on computer science for its rubric. Grounded Theory approaches will generally be tagged Model-Building, as will dissertations that explicitly propose new methodologies. This new name was chosen to distinguish this approach from formalist pedagogies and assessments; see Fulkerson, “Composition Theory in the Eighties.”

Philosophical / Theoretical (PHIL): Inductive or deductive argument based primarily on reason, rather than empirical evidence. Proceeds dialectically from prior arguments. May include claims about what should happen, such as proposed curricula that have not yet been tried. Re-definitions of terms and their significance will generally be classified as Philosophical / Theoretical.

Rhetorical-Analytical (RHET): Attempts to determine extractable writerly “moves” or authorial intent (e.g. with regard to effects on readers) through close or contextual reading of texts. Similar to Critical / Hermeneutical in its subjective analysis of textual features; distinct from Critical / Hermeneutical in its focus on “meta” elements such as motivation, structure, and effect, rather than identifying elements or value in textual content. Genre analysis will generally be tagged Rhetorical Analytical.
observed in the data. Third, I added a category for creative writing, “Poetic / Fictive / Craft-Based,” which bears on the oft-raised question of voice and alternate academic discourses.

Table 3-1 continued.

II. EMPIRICAL INQUIRY

(a) AGGREGABLE

**Discourse / Text Analytical (DISC):** Systematic, often quantitative coding and analysis of formal features in a “text,” broadly construed. Distinct from Critical / Hermeneutical in that whole texts are treated as data archives, so claims are aggregable and findings potentially replicable.

**Experimental / Quasi-Experimental (EXPT):** Hypothesis-driven empirical studies conducted under controlled conditions (or as close as the researchers can get). Whether quantitative or qualitative, the expectation is that the results would be replicable and aggregable.

**Meta-Analytical / Discipliniographic (META):** An analysis that generates and/or analyzes meta-data about disciplinary formation, especially within comp/rhet. In practice, this often takes the form of synchronic analyses of other comp/rhet research materials (e.g. articles, books, conference talks), as a way of capturing the overall state of disciplinary knowledge or identity. May include explicit aggregation of prior research findings (as per Hillocks), or merely aggregation of research or teaching epiphenomena such as authorship (cf. Goggin), conference attendance, curricular requirements, etc. Compare to historiography as opposed to history.

**Survey (SURV):** Research via (widely distributed) quantitative or qualitative questionnaires that do not involve direct interaction between the researcher and those filling out the survey (thus distinct from Interview / Focus Group). Includes quantitative analysis of survey results, as well as data-mining that does not fall under Discourse / Text Analytical or Meta-Analytical / Discipliniographic.

(b) PHENOMENOLOGICAL

**Clinical / Case Study (CLIN):** Rich portraits of individuals to learn about those individuals' behavior or motivations. Distinguished from Ethnographic by emphasis on individuals, as opposed to systems, even though both take context into account. May involve interviews as well as observations, but distinguished from Interview in that the interviews will favor “first person” reflection over “third person” knowledge.

**Ethnographic (ETHN):** Direct (embedded) observations of a community's systems of interaction. Distinguished from Clinical/Case Study by emphasis on community and system vs. individual portraits, and as such includes studies of online / classroom / workplace communities, even when these are referred to as “case studies.” Note that this does not rule out examination of textual evidence, especially transcripts or field notes, but does suggest that such texts will be treated as secondary evidence for context and recall about the studied system, rather than as the primary locus of investigation.

III. PERFORMANCE

**Poetic / Fictive / Craft-Based (POET):** Original poetry, fiction, or creative nonfiction writing (including memoir and autoethnography) composed by the author, perhaps as a way of exploring the process of such composition; see Johnson.

**Practitioner / Teacher Research (PRAC):** Narrative or anecdotal descriptions of “what worked” in a classroom, writing center, writing program, etc, or in the author's personal experiences of writing or performance. Distinguished from Ethnographic classroom studies in its orientation toward future action and enactment vs. understanding of a (possibly unique) system.
Finally, rather than treat “multimodal” as a separate category, I allowed each dissertation to have multiple method tags. Todd Taylor, in his 2003 study of dissertations, claimed that “because these abstracts rarely declare a methodology per se, putting them into appropriate categories is difficult. If there is a pattern among the methodologies in these dissertations, it is that they defy placement in clear methodological categories” (143). To support his claim, he demonstrates that one abstract could arguably fit into seven of the eight methodologies in North’s *The Making of Knowledge in Composition*. Rather than lament this fact, I would argue we should celebrate it. Cultural anthropologist Michael Wesch has suggested that mutually exclusive categories are a holdover from file folders and shelves used for sorting and storing physical objects (books, pages, card-based catalog entries), and that in digital environments, information can and should be “stored” in multiple “places” at once (*Information R/evolution*). In this way, we can avoid the problem of artificially or arbitrarily deciding which method in a hybrid project is “primary,” and instead code for all methods observed. This will also allow for future analyses to examine the correlations among specific methods within individual dissertations.

It is worth noting that in assigning these tags, I paid particular attention to the dissertations’ “exhibits” – Joseph Bizup’s term for “materials a writer offers for explication, analysis, or interpretation” (75). Not only was it instructive to determine what was offered up for examination (full documents, individual sentences or phrases, student behaviors, archival photographs, etc), but methodological affiliations were also revealed by the questions asked of those exhibits, as well as how the exhibits were obtained. One consequence of my focus on exhibits as opposed to other sources is that the presence of other texts in a literature review
would not be sufficient to merit a Critical/Hermeneutical tag, because those texts act instead as Background (“materials whose claims a writer accepts as fact” [Bizup 75]) or as Arguments “whose claims a writer affirms, disputes, refines, or extends in some way” (ibid). Given that virtually all dissertations engage both Background and Argument sources, these would not have been sufficient to distinguish among the methods employed. Similarly, mere mention of pedagogical applications, without the presence of hands-on evidence in teaching situations, would not be tagged as a Practitioner study, but rather as contributing a Philosophical / Theoretical claim with teaching as the content\textsuperscript{10}. Where Method sources “from which a writer derives a governing concept or a manner of working” (Bizup 76) were explicitly mentioned, they did guide my tagging, but, as Bizup notes, such sources often go uncited, slipping instead into prose style or oblique reference.

To illustrate how this non-exclusive tagging works, consider the abstract of Adam Lawrence’s dissertation, “Does it matter what presidents say? The influence of presidential rhetoric on the public agenda, 1946–2003”:

Although scholars have long recognized the president's pre-eminent status as an agenda-setter, there is surprisingly little evidence available to suggest that presidents can and do influence the public agenda. While a modest literature reveals presidential speeches as important determinants of the public agenda, the assumption that rhetoric matters, commonly made by students of the presidency, has been largely unaccompanied by the support of empirical evidence. As a result, the question of whether presidential rhetoric constitutes an important ingredient of agenda setting success remains very much open to debate.

Based on an extensive content analysis of State of the Union Addresses from 1946 to 2003, this dissertation considers in three separate studies the influence of presidential rhetoric as a tool for setting the public agenda. The first considers the influence of several presidential rhetoric variables resulting from the content

\textsuperscript{10} The common impression that a final chapter on teaching must be added to satisfy the “pedagogical imperative” at many schools (Kopelson) may explain the high frequency of Philosophical / Theoretical dissertations identified by my approach, and an investigation into the chapter-by-chapter proportion of this method in particular could prove quite interesting. Such an investigation is, however, beyond the scope of the present study.
analysis on aggregate-level evaluations of the salience of 1,113 issues discussed by 11 presidents from 1946 to 2003. The second study estimates the influence of several moderators of the relationship between presidential rhetoric on the public agenda, based on the individual-level assessments of issue salience expressed by respondents following State of the Union Addresses given by Presidents Ronald Reagan, George H. W. Bush, Bill Clinton, and George W. Bush. Finally, based on an experimental analysis in which 340 subjects were shown edited videos of a presidential speech, the third study examines the influence of the three specific forms of presidential rhetoric used by President George W. Bush in his discussion of the issue of the economy.

The findings demonstrate that (1) presidents respond to environmental conditions fashioning their State of the Union rhetoric, (2) presidents use their rhetoric to move issues onto the public agenda and, by claiming credit, presidents also move issues off the public agenda, (3) presidential rhetoric not only influences the public agenda directly, among those who watch the speech, but also indirectly by affecting media coverage after the speech, and (4) the influence of presidential rhetoric is more pronounced among those who support the president, who share similar political predispositions as the president, and who are politically sophisticated.

The “extensive content analysis” marks this dissertation as Discourse / Text Analytical: the entire body of speeches is treated as an aggregable corpus of words and phrases, coded according to a schema of “presidential rhetoric variables” and analyzed statistically. But Lawrence’s study is also Historical / Archival, using the texts of respondents to triangulate that statistical work with more humanist readings of their “individual-level assessments of issue salience.” Lawrence himself names the Experimental component. Finally, I assigned a Model-Building tag to acknowledge the way the final paragraph lays out interacting components of a system that is presumed to be stable: the work seems intended not merely to describe this one corpus, but to make predictions about how all presidents “use their rhetoric.”

We should not be surprised to see multiple methods in use: after all, as Lynn Z. Bloom points out, “composition studies researchers generally do not choose North's labels (say formalists or clinicians) and most would not restrict themselves to such a categorization system” (Bloom 38–39). Taylor, similarly, celebrates that “the dissertations in [his] study display a wide
array of methodologies for gathering evidence, both within and amongst themselves”

(144). And indeed, multiple methods were more the rule than the exception for dissertations in this study as well. As shown in table 3-2, the great majority of dissertations – over 75% – engage in two or more methodologies.

Table 3-2. Most dissertations use multiple methods.
Frequency with which a given number of method tags was assigned to a dissertation in the dataset. The average dissertation used between 2 and 3 methods. Percentages are given in parentheses.

<table>
<thead>
<tr>
<th>Method Count</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissertation Count (out of 2,711)</td>
<td>601 (22.2%)</td>
<td>1,018 (37.6%)</td>
<td>832 (30.7%)</td>
<td>195 (7.2%)</td>
<td>63 (2.3%)</td>
<td>0 (0%)</td>
<td>2 (0.1%)</td>
</tr>
</tbody>
</table>

Where I disagree with Bloom and Taylor, however, is in concluding that the categories themselves are not useful once we acknowledge this variety. What's important, instead, is to recognize that a researcher – indeed, even a single research project – can wear more than one label at a time. Treating the system as a set of non-exclusive tags, rather than folders into which researchers and their work must be uniquely sorted, we can more aptly represent the work being done and still gain useful perspectives and contrasts.

Updating Taylor’s comment, then, I would say that most (but not all) of these abstracts defy unique placement in such categories. This contrasts with Hansen’s findings that only 10 of 184 articles (~5%) she examined in CCC and RTE were “multimodal.” To some extent, it makes sense that articles edited for publication would be more focused, whereas dissertations spanning multiple chapters can be more expansive. I also wonder, though, whether some approaches were uncounted because they played a secondary, but perhaps still significant, role.

This raises the question: are some methods consistently primary or secondary? To get a better sense of the methodological focus of graduate programs, I aggregated the dissertations by school and examined the frequency with which each method was employed. Figure 3-1
represents this analysis as a heat plot, meaning that numeric values are translated to colors so as to reveal patterns at a glance. Each row of this plot represents the normalized method tag distribution at one school: darker shades indicate greater frequency of the tags represented in the columns, such that white means the method was not used in any dissertations at that school and black means it was used every time. Thus, a consistent neutral shade across a given row would correspond to an even-handed approach to methodology, with any of the tags equally likely; by contrast, a row with some dark and some light boxes would indicate that the school had a methodological focus on the darker columns.

![Figure 3-1. Methods are not evenly distributed across schools, and few schools span the full range of methods.](image)

Across the board, it is clear that most schools exhibit the latter, more divided distribution. This in itself gives us an initial answer to the question of how graduate programs address the challenge of such a methodologically diverse field: by focusing on a subset of possibilities. At
even the largest programs, that is, some methods seem more viable for students to pursue than others.

The order of rows and columns in figure 3-1 has been calculated so as to group together the most similar schools and co-occurring methods; these hierarchical clusters are indicated by the dendrograms located on the left and top of the figure. Reading from the outside of the figure toward the center, the dendrograms divide in half to indicate separation of dissimilar groups. The first major division of schools at the left, which also corresponds to the first major division of methods at the top, divides the schools roughly in half: the upper half of the figure contains a single large bloc of schools that emphasize various combinations of Historical / Archival, Philosophical / Theoretical, Rhetorical-Analytical, and Critical/Hermeneutical methods, while schools in the lower half by and large de-emphasize those four methods. Even without the benefit of the dendrograms, it would be easy to see that the lower half divides into several more distinct clusters, variously emphasizing Clinical / Case Study, Discourse / Text Analytical, Survey, Practitioner / Teacher-Research, and Experimental / Quasi-Experimental methods, with little overlap except between the last two.11 One band across the middle does seem to be more methodologically diverse, but even here (as elsewhere) there remain prominent gaps in the Meta-Analytical / Discipliniographic and Poetic / Fictive / Craft-Based methods.

It is worth noting that the four methods favored in the topmost cluster of figure 3-1 correspond to North’s humanistic “Scholar” community, producing knowledge through dialectical argumentation; and most of the smaller clusters in the lower half of the figure correspond to “Researcher” approaches. Might it be that North’s predictions of a split within the

11 To be fair, some of the sharp clustering may be an illusion caused by a small sample: at some of these schools, the number of dissertations is small and so individual studies may loom large in a normalized heat plot. But even if that's so, the illusion may well hold at those schools, as the local model of research in composition / rhetoric.
field have come true, with a large portion of the Composition community following the pattern of literary studies’ “dissolution,” “for the most part eschew[ing] any other than Scholarly methods, remaining fundamentally a hermeneutical enterprise, supplemented by historical and philosophical inquiry” (366-7), and other portions fragmenting outward into other fields? If we could determine the departments associated with these dissertations – e.g. if the first cluster were consistently associated with English, and the latter clusters with Linguistics, Education, and so on – this would be further evidence in support of that hypothesis. Again, however, that analysis will have to wait for further data.

The methods reported in figure 3-1 are normalized within each school, giving a measure of methodological focus. Does the output from these schools also reflect a high output of dissertations using the methods most prominent in each cluster? As figures 3-2 and 3-3 show, the answer varies somewhat depending on the subset of schools we’re considering.

**Figure 3-2. Dialectical methods are the most common, followed by phenomenological.**
Bar graph of method tag frequency, ordered by descending total across the full dataset (T): overall percentages of dissertations with each tag are given at the right (TP). Each bar is divided to show the breakdown for schools in and outside of the Consortium of Doctoral Programs in Rhetoric and Composition. Because many dissertations are tagged with more than one method, the sum of each column will exceed the total count for that category.

The same four Scholar methods are used most frequently across the full dataset, and in fact one
or more of them is used in 1,667 of the 2,711 dissertations (61%).¹² This agrees with Hansen’s findings in analyzing CCC articles, in which Critical (25%), Philosophical / Theoretical (19%), and Historical (12%) approaches were the methods most frequently used. By contrast, those that Hansen found most prominent in RTE articles — Experimental (31%), Clinical / Case Study (14%), and Discourse or Text Analysis (13%) — are much lower-ranked as choices for dissertation work, with Experimental / Quasi-Experimental methods employed only 8% of the time.

However, these rankings do not tell the whole story: the ratios of Consortium to non-Consortium dissertations using each method vary significantly, as is particularly apparent for Experimental / Quasi-Experimental, the only methodology with which non-Consortium schools produced more dissertations than the Consortium did. Figure 3-3 summarizes the differences across these two subsets of schools, with connecting lines added for ease of comparison.

In addition to confirming that Experimental studies are far more common outside of the Consortium than within it, this analysis offers some surprises. For instance, Meta-Analyses, though still rare, are significantly more common at Consortium schools than non-Consortium schools. Could this be a function of Consortium meetings and listserv conversations inspiring more interest in the field as a field? While the top four methods remain the same, Philosophical / Theoretical and Historical / Archival methods are actually significantly less common outside the Consortium schools, meaning that these methods predominate less over other methods at non-Consortium schools when taken as a whole: the methods within this group are more evenly

¹² Note that this number is less than the sum of 744 + 707 + 677 + 516 shown in figure 3-2 because some dissertations are counted multiple times to account for each method tag.
distributed. Does this indicate that the “Scholar” cluster discussed in the context of figure 3-1 is mostly comprised of Consortium schools, while the methodologically diverse band is mostly non- Consortium?

This would fit with Richard Haswell’s claim that research that is RAD (replicable, aggregable, and data-supported) has been less featured in official NCTE venues, whether in journals or conference submission groups. He clarifies,

It is not that data-infused studies into ‘the lives of those we are teaching’ (Scholes, 1998, p. 81) have died out. As we have seen, they are flourishing but just not under NCTE/CCCC aegis. That labor is turned over to the work hands— to unlicensed apprentices in masters’ theses or dissertations, to ERIC freelancers who are not peer reviewed, to novices in ‘Research Net Forums’ ancillary to the main CCCC convention, or to laborers in the surrounding disciplines presumably at lower altitudes—in discourse and communication studies, technical communication, second-language writing, social sciences, professional schools, and schools of education. (217)

Could it be these other departments and schools that are producing the dissertations in the non-

<table>
<thead>
<tr>
<th>Relative Ranks of Assigned Method Tags</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Consortium Schools</td>
<td>Non-consortium Schools</td>
</tr>
<tr>
<td><strong>Phil (31)</strong></td>
<td>Crit (25)</td>
</tr>
<tr>
<td><strong>Hist (27)</strong></td>
<td>Phil (21) **</td>
</tr>
<tr>
<td>Crit (27)</td>
<td>Hist (20) **</td>
</tr>
<tr>
<td>Rhet (19)</td>
<td>Rhet (18)</td>
</tr>
<tr>
<td>* Ethn (17)</td>
<td>Disc (18) *</td>
</tr>
<tr>
<td>Clin (16)</td>
<td>Clin (17)</td>
</tr>
<tr>
<td>* Disc (13)</td>
<td>Prac (16) *</td>
</tr>
<tr>
<td>* Prac (11)</td>
<td>Expt (13) **</td>
</tr>
<tr>
<td>Modl (10)</td>
<td>Ethn (13) *</td>
</tr>
<tr>
<td>Intv (8)</td>
<td>Modl (10)</td>
</tr>
<tr>
<td>Surv (7)</td>
<td>Surv (8)</td>
</tr>
<tr>
<td><strong>Meta (6)</strong></td>
<td>Intv (7)</td>
</tr>
<tr>
<td><strong>Expt (5)</strong></td>
<td>Poet (5)</td>
</tr>
<tr>
<td>Poet (3)</td>
<td>Meta (2) **</td>
</tr>
</tbody>
</table>

Figure 3-3. Consortium schools are significantly less likely to produce Experimental / Quasi-Experimental studies, and significantly more likely to produce Philosophical / Theoretical or Historical / Archival studies, than non-Consortium schools. Method tags are arranged within in descending order of frequency for dissertations produced at schools in the Consortium of Doctoral Programs of Rhetoric and Composition (N=1800) and dissertations produced elsewhere (N=911), with percentages given in parentheses. Significance computed using Bonferroni-corrected two-tailed Fisher’s Exact Test of Independence.
Consortium subset of data? If so, does that mean that the same pressures acting on NCTE / CCCC are acting as well on the Consortium as a group, or that one influences the other? If the latter, could a change in the Consortium membership exert some influence on the direction of RAD scholarship within the field? I have not yet conducted analyses for temporal changes in dissertation methods, but it would be interesting to see whether there has been a shift in methodological focus or output\textsuperscript{13} since Haswell’s 2005 publication.

In figure 3-4, dividing the heat map into Consortium and non-Consortium allows us to notice some differences. For instance, the Scholar approaches do seem to run strong throughout the Consortium schools: it is not as simple as a few large schools inflating the counts. Thus, while there is still a fair amount of methodological diversity – few if any schools are limited to one method, and most include a fair number of dissertations using methods outside of that cluster, especially the “Researcher” methods of Ethnographic and Clinical / Case Study – some common ground does seem to be suggested by my findings so far, and that ground is to be found more in the text-based and qualitative humanities than the quantifiable or aggregable social sciences.

On the non-Consortium side, the original divide seems to have persisted: a cluster of Scholar-focused schools, here taking up about the top third of the figure, plus several smaller clusters focused on single methods, and finally a broad swath of schools that seem not to form a coherent cluster. Because this pattern could suggest sparseness in the data – too few dissertations at some schools making overlap difficult to attain – consider figure 3-5, which

\textsuperscript{13} It is not possible to determine from abstracts alone whether any of these studies would fit Haswell’s definition of RAD research: as he points out (202), even empirical or quasi-experimental studies can be presented in such a way as to obscure the conditions that would make replication possible or aggregation feasible.
Method Tag Averages by School
Consortium of Doctoral Programs in Rhetoric and Composition

Method Tag Averages by school
Non-consortium Schools

Figure 3-4. The top four methods by frequency occur throughout Consortium schools, but at only a subset of non-Consortium schools. Heat map of dissertation methods (columns) aggregated and normalized by school (rows), including only Consortium schools at left (N=1,800 dissertations, 74 schools) and only Non-consortium schools at right (N=911 dissertations, 194 schools). Darker shades indicate greater likeliness that a dissertation at school Y uses method X (White = 0%, Black = 100%). Dendrograms at top and left indicate similarity clustering.

shows only those non-Consortium schools which averaged at least one composition/rhetoric dissertation per year for the last five years of data.

This procedure reduced the number of non-Consortium schools from 194 to a mere 28, but reduced the number of dissertations from 911 only to 450, making this a particularly
interesting set of schools to examine: why are they not in the Consortium? Is it important that in this subset of schools, Practitioner / Teacher Research is more associated with Experimental and Discourse Analytical approaches than with Model Building, Ethnographic, and Clinical / Case Study approaches, as at the Consortium schools? Or that in many more cases here than in the consortium a Philosophical / Theoretical focus is paired with Ethnography, rather than with Historical / Archival work?

I do not mean to overstate the inferences we can draw from these correlations or contrasts, especially given that these are aggregate counts and do not reflect methodological groupings of individual writers or dissertations: the data here is suggestive, not conclusive, of how methods are taken up at these campuses. Additionally, using abstracts to determine the presence of multiple methodologies makes it difficult to distinguish between a consistent blend of approaches and a series of chapters that each use a single method. Future studies may help to identify the locations of passages in each document associated with given methods\textsuperscript{14}; however, such an analysis was beyond the scope of the present project, which aimed first to determine which methods were present at all.

\textsuperscript{14} One promising approach is the approach to topic modeling used in the Networked Corpus project (Binder and Jennings). Like other topic modeling projects, Networked Corpus uses algorithms to discover clusters of words that tend to co-occur in documents, which then allows human interpreters to associate these clusters with subject matter (thus, “topics”), and then to represent (“model”) each text as if it were produced by all of those clusters in varying proportions (Blei); see Chapter 4 for topic models of Consortium dissertations as the unit of analysis. Binder and Jennings go further by using novel visualizations to locate passages \textit{within} texts that contribute maximally to the topics assigned to that text, and link them to other such passages, making it easier to confirm and refine the model. Pairing their approach with a set of topics derived from single-method dissertations could enable us to see whether multi-method dissertations synthesize these methods or alternate among them.
Our understanding of “the field” is, by and large, local. At our home campuses, we see a range of methods and know the field to be diverse, and so we meet colleagues from elsewhere and agree that yes, this is a dappled discipline. But as this chapter has shown, the variation itself varies from place to place. Understanding the range of methodological options currently in use can help us appreciate both the common ground we share and the paths that are (in Hopkins’s words) “counter, original, spare, strange.”
Chapter 4:  
Tapping the Topics:  
What We Study When We Write in Writing Studies

But what, you will ask, of the content? Methods are all well and good, but only as good as what they help you to accomplish. What are people looking at through all of these lenses? And what are they saying about what they see?

Getting answers to these questions turns out to be a bit harder than one might expect. Although every dissertation in ProQuest is labeled with subject terms – drawn from a limited, or fixed, vocabulary from which authors choose at most three – these subject terms are fairly broad, including such terms as "English literature" or "Higher education," and as such are more suited to gathering a dataset to study than to determining interior contours within that dataset. Nor are the author-supplied "open vocabulary" keywords much more helpful, albeit for the opposite reason: of 5,999 keywords attached to 2,711 dissertations, on average each appeared no more than twice, and the median frequency with which they occurred was only once. The top three keywords, occurring respectively 504, 385, and 292 times, are simply "rhetoric," "writing," and "composition": names of the field whose content we're trying to unpack.

We need another approach. Rather than directly reading the roughly 660,000 pages in the dataset, a project which would undoubtedly have some impact on time-to-degree, we can look again for an algorithmic approach. The simplest form of text mining would be to ask the computer to count the number of times each word occurs, and see what words occur most frequently. (This is the process that generates the tag clouds that were so popular in the late 2000s, with higher word frequencies represented visually by larger font sizes and/or more saturated coloring.) However, basic text mining leads to problems for interpretation in that
simple counts of word frequency cannot distinguish multiple meanings of the same
grapheme: "program" could refer to a department, a curriculum, a piece of computer code, or a
booklet distributed at the theater, each with very different implications for the focus of the texts
that contain these words. Ordinarily, the computer cannot account for the semantic differences
between these various identical *tokens* – individual appearances of the same sequence of
characters – and the underlying word *types* of which the tokens are merely instances.

*Topic modeling* is a family of approaches that attempts to get around this semantic
difficulty. Several excellent explanations of the underlying mathematics are widely available,\(^\text{15}\) so I will here limit myself to a representative overview aimed at humanists. Based on the work of
Blei et al, and introduced to Composition by Clancy Ratliff and Jonathan Goodwin’s analysis of
journal articles (Ratliff), a topic model identifies clusters of words which tend to co-occur within
documents, and in what proportions those clusters combine to form both individual documents
and a large corpus. Though by no means infallible, this approach has a history of success in
identifying semantic themes and subjects at a level of scale larger than any individual researcher
could read (cf. Goldstone and Underwood; Mimno, “Computational Historiography”; Blei).

Suppose, for example, that you had the following three dissertations:

\(^{15}\) See especially Scott Weingart's 2012 roundup of such introductions (Weingart, "Topic Modeling: A Guided
Tour"), which includes links to Matthew Jockers' “The LDA Buffet is Now Open” and Ted Underwood’s “Topic
Modeling Made Just Simple Enough,” as well as several rather less-simple articles by Edwin Chen, David Blei,
David Mimno (“Computational Historiography”), and others.
Through an iterative process of sampling, the computer determines that A and B share one set of words, while B and C share a different set of words:

In this case, the bin of words at the left seems to contain a list of words – a topic – related to classroom writing practices, while the bin at the right seems to be about power relations among social classes. Note that the word “class” appears in both of these topic bins, but with different definitions implied based on the associated words.

Figure 4-1a: Three example dissertations.

Figure 4-1b: Three example dissertations in relation to two extracted topics.
Note, as well, that dissertation B contains words from both topics. This is a significant feature of topic modeling: it assumes that any given text is composed of multiple interacting and potentially overlapping component parts; the goal is to infer these parts, and their proportions, from an observed sample of the text. As Andrew Goldstone and Ted Underwood put it,

The aim of topic modeling is to identify the thematic or rhetorical patterns that inform a collection of documents: for instance, the articles in a group of scholarly journals. These patterns we refer to as topics. If each article were about a single topic, we would only need to sort the articles into categories. But in reality, any article participates in multiple thematic and rhetorical patterns. (Goldstone and Underwood 4, italics in original, boldface added)

What is true for articles is, of course, especially true for dissertations, which not only vary thematically and rhetorically within each chapter, but which can and often do include several different approaches or subjects from one chapter to another. This variation stems both from triangulation, a desire to examine one's subject from more than one angle, and from a writerly desire not to be too one-note: tensions among multiple threads in prose, as in poetry, produce useful energy that drives the writing forward.

Before the chapter is out, I will share my findings from a topic model of Comp/Rhet dissertations, including both individual and composite topics that top the list of concerns in these field-generative texts. But first, I want to clarify what it is I’m looking at and drawing conclusions from.

16 “Text” here is broadly construed: these approaches could equally well describe traditional writing or images or waveforms. Ben Schmidt has used them to identify common routes taken by 19th century whaling ships (Schmidt).

17 The algorithm first "tokenizes" the text, counting the instances of the same words, or tokens, after stripping connective tissue and overly common "stop words" such as "a, an, and, of," and so on that would overwhelm and obscure the more interesting content. Every unique word is assumed to be part of every topic, but each topic associates a different probability with any given token; words with a high probability are used to identify the topic, while many words in a given topic will have a probability at or very close to zero.
Methods

The corpus for this chapter consists of 1,754 full-text dissertations produced at schools with programs in the Consortium of Doctoral Programs of Rhetoric and Composition, a subset of the data discussed in chapter 3. Before the documents could be analyzed, they had to be pre-processed to extract plain-text files from .pdf files provided by ProQuest/UMI; to resolve conflicts in character encoding by shifting from Latin-1 text to Unicode standard; and to organize the cleaned text for reading by the MALLET\(^\text{18}\) topic-modeling software. These pre-processing steps were achieved through a series of Unix shell scripts, provided in Appendix G as `ben_clean_and_consolidate.sh`.

The topic model was generated by MALLET, but parameters were set using R, and are described in the file `r2mallet_with_foreach.R` (see Appendix F). The system was set to run 250 iterations, with alpha optimization every 20 trials after a burn-in of 50 trials; these options follow David Mimno’s defaults in his mallet library for R (Mimno, *Mallet: A Wrapper*). I also used the default set of stopwords, i.e. connective words such as “an” and “um” that appear so often they would overwhelm the content if they were not disregarded.

One challenge of the topic modeling approach I used, Latent Dirichlet Allocation (LDA), is that the algorithm requires the number of topics to be specified in advance.\(^\text{19}\) I ran test iterations with 5, 10, 20, 30, 40, 50, 100, 150, 200, and 250 topics, and looked for a trend in the Log-Likelihood per Token (LL/token) measurements – one way of testing accuracy – that were

\(^{18}\) MALLET stands for “MAchine Learning for Language Toolkit” (McCallum); more information, including source code, is available from [http://mallet.cs.umass.edu](http://mallet.cs.umass.edu).

\(^{19}\) Some methods do exist for validating the selection of this number, often involving iterating over many options and testing the accuracy with which the resulting models can predict the content of held out data (see Mimno, “The Details”; Griffiths and Steyvers)). I regret to say that much of the technique exceeded my ability at this time, though I do plan to return to the problem with more expert collaborators in future iterations of this project.
output automatically by MALLET. The changes in LL/token were, however, minimal once the number of topics exceeded 40 or so. Wallach, Mimno, and McCallum demonstrate that, given certain starting parameters, \(^{20}\) “the risk of using too many topics is lower than the risk of using too few, and that practitioners should be comfortable using larger values of T,” that is, larger numbers of topics (Wallach, Mimno, and McCallum 7). Initially, then, I selected 150 topics.

Underwood and Goldstone suggest in a blog post that the outcome of changing the number of topics may actually reflect the degree of simplification you’re applying to the corpus, rather than something fundamental about the texts themselves:

> [I]f you change the number of topics, you can get results that look substantially different. On the other hand, to say that two models “look substantially different” isn’t to say that they’re incompatible. A jigsaw puzzle cut into 100 pieces looks different from one with 150 pieces. If you examine them piece by piece, no two pieces are the same--but once you put them together you’re looking at the same picture. (Underwood and Goldstone, qtd in Fredheim)

Pairing this insight with Ben Schmidt’s insight (in comments on Underwood, “Visualizing Topic Models”) that hierarchical clustering can preserve more information about how topics relate at various values of T, Rolf Fredheim argues persuasively that hierarchical clusters, visualized as dendrograms, give an overview of the topic assignments that acknowledges the possibility of slicing the corpus in different ways (Fredheim).

To make my dissertation topic model more amenable to inspection while still respecting the complexity of the data, I followed Fredheim’s procedure to visualize a tree structure of similarity among 150 topics (see figure 4-2, left). Close inspection of that tree led to a new cut, with 55 clusters (see figure 4-2, right), that would preserve the essential hierarchy while

---

\(^{20}\) At issue was whether the alpha and beta hyper-parameters for the Dirichlet distribution are allowed to vary. Wallach et al conclude that the optimal conditions are a varying, or asymmetric, alpha parameter and a symmetric, or non-varying, beta parameter. I used these settings in configuring MALLET.
consolidating some smaller topics into hopefully-coherent groups. It is the resulting 55-topic model that I draw on in presenting my findings throughout this chapter.

**Distant Reading Requires Close Reading**

The co-occurring sets of words discovered by MALLET’s algorithms correspond, generally speaking, to topics, but the computer cannot automatically put names on them: it doesn't know what a "teacher" is, other than a string of letters. One key interpretive task, therefore, is to assign labels to the topics that emerge from the algorithm. Rather than rely on the top words alone, I built a tool in R to browse the topics (see Appendix F, 'top docs per topic.R'), a screenshot of which appears in figure 4-3. For each topic, the browser shows the top words (i.e. the words most likely to be assigned to this topic whenever they appear in the corpus) and the top five texts for the topic by weight (i.e. by percentage of words in a single document that are

![Cluster Dendrogram: consorts, 150 topics](image_url)

**Figure 4-2: Hierarchical clustering of topics by similarity, as a means of selecting the number of topics.** The tree formed by 150 topics (left) was cut to produce similarly-sized groups for closer inspection, and the resulting groups are shown by the colored boxes (right). Numbers correspond to topics but are assigned arbitrarily by the algorithm.
assigned to the topic). As a preview, the browser also shows, for each of these top five texts, a unique publication number, the full title, and a grid of assigned method tags (here, 0 means the method is not assigned to this text, and 1 means that it is).

This consolidated preview allows us to begin noticing any particularly striking patterns, and to form a first impression of what content the topic might contain. For example, the topic shown in the figure (Topic 8), which is the second largest topic by one measure (i.e., the percentage of total words aggregated across all documents that are assigned to each topic), shows in the preview above a consistent affinity for dissertations that use Philosophical/Theoretical methods.

To dig deeper into this first intuitive impression, I then examined more details for each of the top five dissertations within the chosen topic (see figure 4-4 on the next page). After restating the title and method tags, the browser shows the top five topics for that individual dissertation, along with their respective topic weights (the percentage of the dissertation's words accounted for by that topic) and the "keys," or top-ranked words, associated with each. In the dissertation shown...
Figure 4-4: Topic browser in R, showing a detailed view of the top document for Topic 32. The browser was used to assign labels to topics.

| Title |

1. In five classrooms: A descriptive study of before writing teaching practices in encouraging college writers to write.

<table>
<thead>
<tr>
<th>Hist Inv</th>
<th>Meta Mod</th>
<th>Phil Poet Proc Proc Surv Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| $doc$tops |

top1 1 wgt1 top2 1 wgt2 top3 1 wgt3 top4 1 wgt4 top5 1 wgt5 1

<table>
<thead>
<tr>
<th>topic</th>
<th>weight</th>
<th>alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32.0</td>
<td>0.3675</td>
</tr>
<tr>
<td>2</td>
<td>1.01481532</td>
<td>0.44727</td>
</tr>
<tr>
<td>3</td>
<td>0.09470423</td>
<td>0.41965</td>
</tr>
<tr>
<td>4</td>
<td>0.0592771</td>
<td>0.33959</td>
</tr>
<tr>
<td>5</td>
<td>0.03637115</td>
<td>0.42419</td>
</tr>
</tbody>
</table>

| top_words |

1. students writing student class teacher classroom teachers paper instructor research study instructors semester college assignment classes write teaching learning.

2. don participants people study time interview research things experience work didn make lot interviews experiences questions feel kind.

3. community research learning project process service work group members development study professional knowledge organization practice team information communities activities.

4. writing composition writers write english essay creative process written language reading work personal essays instruction basic style college.

5. text analysis texts information discourse readers chapter study reader rhetorical content audience context specific reading data based features types.

| Abstract |

1. College writers!Before writing teaching practices!Writing!Freshman composition.

ABSTRACT

1. In this one-semester descriptive study, I investigated how five ENG 101 instructors approached the teaching of writing and what their students thought about the teaching approaches they encountered. I attended each class meeting in all five classroom cultures and observed the before writing teaching practices (BWTP) of each instructor. I examined how instructors used BWTP to help students write, how that preparation helped the students initiate and comply with requests to write, and how students perceived and valued encouragement in their classroom interactions with instructors. I collected data from participants by using the ethnographic techniques of participant observation and interviewing. I also used field notes, informal discussions, focus group sessions, a case study of several BWTPs, and student writings to write a thick description and interpretative-explanatory account of what participants did and how they understood what they were doing in their ENG 101 classroom cultures. Whenever possible, I used participants' words to describe their college writing experiences. While classroom formats and specific BWTP varied from classroom to classroom culture, I found that each instructor's BWTP provided students with opportunities to talk before writing at each class meeting. Both instructors and students said they felt their classroom interactions helped students start new pieces of writing or add to in-progress pieces of writing. Students also said that instructors listening to them helped them to write and that their instructor's interest in students' work encouraged them to write. In each classroom culture, students reported that they felt their instructor's teaching methods were most helpful and encouraging when instructors gave students the freedom to make choices about their writing, trusted students to determine how they would learn, and discussed students' writing choices with them individually. 

Press <enter> for next doc, D for more details, or S to skip to the next topic.
in figure 4-4, for example, Topic 32 ("students, writing, student, class, teacher, classroom") accounts for 53% of the text; Topic 1 ("don't")[21], participants, people, study, time, interview") accounts for an additional 10%; and so on. Finally, this detailed view shows the keywords and abstract. Viewing several of these abstracts in light of the topic keys allowed me to see how those words seemed to be functioning in each topic, and thus to create shorthand labels. Topic 8, for instance, shown in figure 4-3, became “(Critical) Pedagogical Theory,” while Topic 32, shown in figure 4-4, became "Students in the Classroom. " For a full list of topic keys and labels, see Appendix E.

Weights and Measures: What are Grad Students Writing About?

At first blush, no one topic dominates the field. The top 10 topics, presented in table 4-1, each represent only 3-6% of the words in the corpus. (This overall contribution for each topic is provided in the final column of each row. The center column of each row gives the top 19 words and titles of the top 3 dissertations for that topic, along with the percentage of those dissertations accounted for by the topic.)

This top 10 list showcases, in a nutshell, the breadth of subjects that “count” in Composition/Rhetoric. It includes both philosophical theorizing and concrete storytelling, both direct examination of written words and more indirect explorations of writing contexts. Some high-ranking topics confirm our expectations: for a field that Joseph Harris has famously called A Teaching Subject to write frequently on students and pedagogy is not surprising. Nor, given the heavy weight of Philosophical/Theoretical methods that we saw in Chapter 3, is it strange to see language, theory, discourse, and identity appearing in this top list.

21 The tokenizer used in generating this topic model mistakenly treated apostrophes as word breaks, and split words such as “don’t” and “I’ve” into “don” + “t” and “I” + “ve.” This error, discovered late in the process of analyzing the results, will be fixed in future iterations of this project.
<table>
<thead>
<tr>
<th>Rank</th>
<th>Assigned Label</th>
<th>Top Words and Titles</th>
<th>% of Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Students in the Classroom</td>
<td>In five classrooms: A descriptive study of “before writing teaching practices” in encouraging college writers to write (52.8%) A case study of selected ESL students' experiences with writing portfolios in college composition courses (52.2%) Non-traditional students: Age as a factor in the composition classroom (45.3%)</td>
<td>5.4%</td>
</tr>
<tr>
<td>2</td>
<td>(Critical) Pedagogical Theory</td>
<td>Critical contentions: Feminism(s) and critical pedagogy in composition studies (47.2%) Resistance, ontology, and affect (41.2%) The subjects of critical pedagogy and composition: The Asian-American teacher-intellectual and affect (39.1%)</td>
<td>4.8%</td>
</tr>
<tr>
<td>3</td>
<td>Philosophy of Language</td>
<td>Figuration of the folk: The nature and use of a universal linguistic category (65.3%) Minimal foundationalism in literacy studies (59.8%) A pragmatics of power using Juergen Habermas' theory of communicative action (56.3%)</td>
<td>4.3%</td>
</tr>
<tr>
<td>4</td>
<td>Identity Construction</td>
<td>Be-coming subjects: Reclaiming a politics of location as radical political rhetoric (44.1%) Third-space sites, subjectivities and discourses: Reimagining the representational potentials of (b)orderlands' rhetorics (40.3%) Who cares? Rendering care readable in the 21st century feminist writing classroom (33.6%)</td>
<td>4.2%</td>
</tr>
<tr>
<td>Rank</td>
<td>Assigned Label</td>
<td>Top Words and Titles</td>
<td>% of Corpus</td>
</tr>
<tr>
<td>------</td>
<td>----------------</td>
<td>----------------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| 5    | Story and Narrative | life back story time day man people mother don love home good young father stories family long left person  
*The Faithful: Healing through narrative* (73.1%)  
*Writing a young adult novel: An autobiographical account of one non-writer's journey* (73.0%)  
“Life On a Grape” with an introduction on theories of the novel (57.4%) | 4.1% |
| 6    | Process Reflections | don participants people study time interview research things experience work didn make lot interviews experiences questions feel ve kind  
*Telling developments: Narrative interviews with writers as “acts of meaning”* (28.9%)  
*Compelled to connect: A phenomenological study of the experience of writing* (27.9%)  
*Improving the skills of remedial-writing students with strategies for revising* (27.1%) | 3.8% |
| 7    | Community Engagement and Collaboration | community research learning project process service work group members development study professional knowledge organization practice team information communities activities  
*Interdisciplinary group process as an indeterminate zone for collaboration and technical communication: A case study of proposal writing for an immune building and test bed* (50.1%)  
*Online writing labs as sites for community engagement* (48.3%)  
*“Democratizing” clinical research? Efficiency and inclusiveness in an electronic primary care research network* (44.7%) | 3.7% |
| 8    | Capitalism, Marxism, and Activism | public political social economic movement rhetoric society politics power cultural labor university state democracy change action democratic rhetorical class  
*Making change: The role of rhetoric in the politicization of consumption* (46.3%)  
*Inside the teaching machine: The United States public research university, surplus value, and the political economy of globalization* (45.8%)  
*Entering the fray: The slogan's place in Bolshevik organizational communication* (39.7%) | 3.6% |
Somewhat more surprising is the presence in the top five of the topic I’ve labeled “Story and Narrative,” given that the Poetic/Fictive method was the most rarely used; only 61 of 1,800 Consortium-school dissertations were tagged for that method. One possible explanation is that those 61 dissertations, which include a number of novels, might be so dominated by words associated with storytelling that they loom unexpectedly large in the topic model. Another is that narrative (or, at least, anecdotal evidence) may simply be common across many methodologies – especially Historical / Archival work, which was the second most common method among Consortium dissertations. It is likely that both factors are at play.

But perhaps the most surprising appearance in the top 10 is the topic I’ve labeled “Comprehension and Usability,” which deals with concrete applications of syntax and document design. Given the concerns about widespread loss of interest in cumulative or transferable knowledge-making as discussed in Chapter 1, I had not expected to see such a practical, hands-
on topic such as this one ranked so prominently. The possibility of such a surprise reinforces my claim that full-text analytics like topic-modeling are important to include alongside considerations only of metadata, even detailed metadata such as abstracts. Further investigations will be needed to determine whether the appearance of “Comprehension and Usability” as a prominent topic is linked to other patterns or trends: Do these dissertations tend to emerge from programs in Technical and Professional Communication? Has this topic been rising or falling in prominence over time?

**Most Dissertations Address Multiple Topics**

As table 4-1 also shows, there is a good deal of variety with regard to how focused a given dissertation may be on a single topic. Though every dissertation shown above has the associated topic as its most common – i.e. that topic accounts for more of the dissertation’s words than any other single topic – the percentage of the text for this top topic ranges from as much as 73.1% (Story and Narrative, in *The Faithful: Healing through narrative*) to as low as 27.1% (Process Reflections, in *Improving the skills of remedial-writing students with strategies for revising*).

The box plot in figure 4-5 describes the distribution of topic weights within individual dissertations. As is standard for this kind of figure, the thick line in the center of each box represents the median value, the top and bottom of the box represent the upper and lower quartiles, respectively, and the “whiskers” show 1.5 times the inter-quartile distance to identify outliers.\(^{22}\) Thus, e.g., the top-ranked topic represents between 20% and 32% of the text in half of

\[^{22}\text{The interquartile range (IQR) is the difference in the values at the upper and lower quartiles. In the case of the top-ranked topic, e.g., the IQR is 0.32315 – 0.1605 = 0.16265, so the upper whisker shows the maximum observed value below the “fence” of (0.16265 * 1.5) + 0.32315 = 0.567125. In our case, that maximum non-outlier value is 0.51230, and any top-ranked topic representing more than 56.7% of the dissertation is considered an outlier.}\]
the analyzed dissertations, and the second-ranked topic represents between 12% and 18%. The third-ranked topic, for these middle two quartiles of dissertations, represents only 8% to 12%. The circles in this figure represent outliers. Some 53 dissertations out of the 1,754 included here, for example, have a top-ranked topic that accounts for more than 51% of the text, ranging as high
as 88.4% (Quantifying Written Products, in Using social influence messages to examine the effects of matching and adjective laddering on attitudes)\textsuperscript{23}. The three topics that account for the most high-focus dissertations are topics 28, 39, and 41:

\textbf{Table 4-2.} Upper outlier topics for percentage contributed as the top-ranked topic within a dissertation.

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words and Titles</th>
<th>Outlier Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>Political Rhetoric, Mostly of the US</td>
<td>war president american public states united bush national military america speech political nation government people policy press carter york</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>\textit{Part of something larger than ourselves: George H.W. Bush and the rhetoric of the first United States war in the Persian Gulf (76.4%) Foreign policy rhetoric for the post-Cold War world: Bill Clinton and America's foreign policy vocabulary (65.5%) The Reagan rhetoric: History and memory in 1980s America (64.5%)}</td>
<td>10</td>
</tr>
<tr>
<td>39</td>
<td>Writing Process: Formal and Cognitive Studies</td>
<td>writing language english students study writers feedback esl process learning comments reading revision research grammar knowledge sentence learners write</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>\textit{Syswrite: Theory-based writing analysis (74%) How L2 legal writers use strategies for scholarly writing: A mixed methods study (61.6%) Effects of grammar emphasis on the revising processes of ESL students: Two case studies (61.5%)}</td>
<td>10</td>
</tr>
<tr>
<td>41</td>
<td>Quantifying Written Products</td>
<td>test study table results scores group research assessment significant score data participants scale social differences total groups message number</td>
<td>9</td>
</tr>
</tbody>
</table>

\textit{Using social influence messages to examine the effects of matching and adjective laddering on attitudes (88.4%) Measures of writing skills as predictors of high}

\textsuperscript{23} Though there are some outliers at the bottom, these appear to consist solely of dissertations for which the “real” top-ranked topic was non-content-bearing: a marker of a foreign language, for example, or faulty Optical Character Recognition (OCR) that results in garbled text. Where these took up 90-99\% of the dissertation, the remaining content-bearing topics appear to represent an extremely low proportion of the text. In future iterations of this project, such dissertations will be screened for early on and removed from the dataset before building the model.
It is tempting to speculate that something like quantitative methods or linguistic/discourse-analytical approaches are a common element responsible for this level of topical focus, but I see no evidence for this pattern; while 17 of the top 20 dissertations in Topic 41 ("Quantifying Written Products") use Experimental/Quasi-Experimental methods, Topic 28 ("Political Rhetoric") coheres around a very different set of methods – Historical/Archival, 17 of 20, and Rhetorical Analytical, 13 of 20 – and the top 20 dissertations of Topic 31 ("Writing Process") have no clearly dominant set of methods.

In any event, despite the existence of such outliers, their rarity (only about 3% of the dissertations have a top-ranked topic that accounts for more than half of the text) suggests that – as with methods – it is by far more typical for a doctoral dissertation to incorporate multiple topics than to focus exclusively on one.

**Topics and Individual Dissertations: Overview and Examples**

To get a better handle on what these topic pairings can look like, in the next section I consider a few examples of individual dissertations with a range of topic proportions. As discussed above, median values for topic weights in a dissertation are 25% for the top-ranked, 15% for the second-ranked, and 10% for the third-ranked. Anne Whitney’s *The transformative power of writing: Teachers writing in a National Writing Project Summer Institute* (UC Santa Barbara, 2006) is one such dissertation. Her full abstract and top five topics are given below.

This study examines the relationship between teachers' writing experiences and "transformative" professional development. The notion that writing might possess transformative power spans academic disciplines and popular culture, as
seen, for instance, in the scholarship on writing-to-learn, research on writing's physiological and psychological benefits, or in the many self-help books advocating writing as a tool for overcoming life problems. Meanwhile, over the more than thirty years of professional development institutes conducted by sites of the National Writing Project, many participants have claimed, both in their own publications and in research studies, that their experiences in such institutes "changed my life" or were "transformative." This study asks two central questions: first, if transformations are occurring, what are those transformations like; what transforms, exactly, and how? Are these processes akin to those described in Mezirow's (1991) theory of transformative learning? Second, given that writing has often been thought to foster transformation and given that NWP Summer Institutes are writing-intensive environments (in which participants spend much of their time engaged in writing of their own in addition to talking and thinking about writing and its teaching), what role, if any, might writing itself play in these transformations?

Seven K--12 teachers discussed their writing and their learning experiences in two interviews during one NWP Summer Institute, and their activities were observed through participant observation. Writing samples were collected, as were application essays and reflective writing. These data were analyzed as individual cases and in parallel, and the resulting pattern is presented toward a model of teacher transformation in a writing-intensive setting: phases included triggers, accepting the invitation to write, self-examination, reframing, resolving to reorient, trying new roles, building confidence and competence through new roles and relationships, and living in the new frame. Writing played a particularly vital role in self-examination, trying new roles, and building confidence and competence. Writing groups functioned as "audience workshops" in which both written compositions and the compositions of self-presentation were worked out. The study also suggests that self-monitoring was heightened through the writing group and in turn contributed to participants' transformative learning.

Table 4-3. Top 5 topics in Anne Whitney's dissertation.

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words</th>
<th>Proportion of Dissertation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Process Reflections</td>
<td>don participants people study time interview research things experience work didn't make lot interviews experiences questions feel ve kind</td>
<td>24.4%</td>
</tr>
<tr>
<td>35</td>
<td>Community, Engagement, and Collaboration</td>
<td>community research learning project process service work group members development study professional knowledge organization practice team information communities activities</td>
<td>14.8%</td>
</tr>
<tr>
<td>25</td>
<td>History of Composition</td>
<td>writing composition writers writer write english essay creative process written language reading work personal essays instruction basic style college</td>
<td>9.7%</td>
</tr>
<tr>
<td></td>
<td>Students in the Classroom</td>
<td>students writing student class teacher classroom teachers paper instructor research study instructors semester college assignment classes write teaching learning</td>
<td>6.9%</td>
</tr>
<tr>
<td>---</td>
<td>----------------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>32</td>
<td>Writing Process: formal and cognitive studies</td>
<td>writing language english students study writers feedback esl process learning comments reading revision research grammar knowledge sentence learners write</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

That “Process Reflections” takes the top slot in Whitney’s dissertation makes sense given the central role of the case studies of seven teachers, who reflect on their own experiences not only in interviews (which would be directly quoted as well as referenced in later analysis) but also in writing composed as part of the NWP Summer Institute. “Community, Engagement, and Collaboration” here seems to correspond at least in part to the participant observation of the Summer Institute; in this connection it is perhaps worth noting that this topic, Topic 35, is associated with ethnographic methods relatively often (11 of the top 20 dissertations for the topic). This topic may also reflect the model-building component of Whitney’s conclusion, in which she builds “a model of teacher transformation in a writing-intensive setting,” i.e. the collaborative dynamics of a social system (7 of the top 20 dissertations for this topic were tagged for model-building).

“Students in the Classroom,” though not the next topic in order, also seems clearly relevant to analyzing the actions of these NWP participants, who are teachers outside of this context and learners inside of it. Similarly, “Writing Process: formal and cognitive studies,” which reflects a tradition of model-building going back at least to Flower and Hayes, makes sense paired with the other major components we see here. That these latter topics account for less than 7% each may indicate that they contribute more by inflecting the major topics, rather than as stand-out content in their own right.
What to make of the presence of Topic 25, which I have labeled “History of Composition”? Given that Whitney’s dissertation is not primarily historical, why does it come in as the third-ranked topic? At only around 10% – which is, again, a median value for third-ranked topics across the dataset – I believe this could reflect background information that the author included as a framework for her study: the motivating context of “more than thirty years of professional development institutes” referenced in her abstract. To confirm this, we could develop a finer-grained model, breaking down each document by page or even paragraph, which would enable us to “zoom in” to particular passages where a topic concentrates in the text. Binder and Jennings have demonstrated in their Networked Corpus project that such models can provide insights both into the meaning of a topic and how topics interact. However, such a model was beyond the scope of the present project, and will have to wait for future iterations.

The top 3 topics in Whitney’s dissertation account for about 48% of the dissertation’s content, and as figure 4-5 showed above, it is typical for several topics to combine in this way to reach the 50% mark. Another pattern involves a single topic that by itself accounts for the majority of the text; the opposite is a more “flat” distribution in which no one topic, or even two, emerges as primary.

As an example of a high-focus dissertation, consider Minimal foundationalism in literacy studies, by Nevin Leder (Michigan State University, 2002). His abstract and top 5 topics follow:

Literacy studies – the study of what it is to be literate, how literacy is acquired, and most importantly, how written texts are related to meaning – is currently heavily influenced by antifoundationalist philosophy. According to this perspective, there is nothing “more firm or stable than mere belief or unexamined practice” (Fish, 1989, p. 343). Paradoxically, this position has been taken as axiomatic among numerous literacy theorists, but, by taking this position, these scholars align themselves with classical skepticism, and, therefore, expose themselves to classical refutations of that position, in particular, Kant's argument that human perception is subjective yet informed by a priori intuitions that must be accepted as veridical since denying them entails logical contradiction. With
these arguments Kant established a minimal foundation for both philosophy and science that can be effectively employed in literacy studies.

Most natural scientists reflexively adopt a Kantian position since their work requires a synthesis of rational analysis and empiricism, but social scientists, literary scholars, and some philosophers, particularly since Wittgenstein, have moved increasingly toward a skeptical position in which thought is equated with language and language is seen as merely “contingent” on, rather than reflective of, reality, a position that has led to extreme skepticism in literary interpretation, indeed to the view that texts are nothing more or less than the discourse community takes them to be.

However, literacy studies is also strongly associated with linguistics, which, since Chomsky, has endorsed the very nativist perspective antifoundationalists explicitly reject. Moreover, the generative program in linguistics has sparked a “cognitive revolution,” which is also strongly nativist.

Although routinely portrayed as Enlightenment dogmatists by antifoundationalists, cognitive scientists are acutely aware of the limitations of computational processes and, some, notably Fodor, have concluded that there must also be an “abductive” mental capacity that allows humans to make appropriate decisions quickly in myriad circumstances, but which cannot be modeled by known computational algorithms. Philosophers of language, particularly Davidson, have made similar observations, arguing that computational models of language cannot explain the sorts of ad hoc adjustments interlocutors constantly make in ordinary conversation; these observations are also pertinent to literary interpretation. Although the ultimate source of this free, abductive capacity remains mysterious – and thus susceptible to antifoundationalist claims – models of interpretation that include computational algorithms along the lines of Chomsky, Katz and others, and pragmatic principles along the lines of Grice, offer a much better explanation of how interpretation is possible than antifoundationalism can, and also provide rational methods for choosing among competing interpretations. Because literacy requires mastery of both computational and abductive processes, a rational approach to literacy studies offers one of the best windows on how the mind integrates these processes, and thus simultaneously provides a potential bridge between literary and scientific study.

Table 4-4. Top 5 topics in Nevin Leder’s dissertation

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words</th>
<th>Proportion of Dissertation</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>Philosophy of Language</td>
<td>language theory discourse meaning knowledge system fact point power metaphor question view speech human case social model problem sense</td>
<td>59.8%</td>
</tr>
<tr>
<td>14</td>
<td>Poetics and Semiotics</td>
<td>world experience memory life human art poetry work time process meaning voice language nature metaphors words mind form sense</td>
<td>6.6%</td>
</tr>
<tr>
<td>55</td>
<td>Comprehension and Usability</td>
<td>text analysis texts information discourse readers chapter study reader rhetorical content audience context specific reading data based</td>
<td>6.1%</td>
</tr>
</tbody>
</table>
In Leder’s text we have a clear example of dedication to one topic: in this case, the question of how to understand literacy, i.e. “how interpretation is possible,” and thus “how written texts are related to meaning.” Topic 48, which I’ve labeled Philosophy of Language, makes up almost 60% of the dissertation – more than Whitney’s first four topics combined, and not much beyond her first five topics together. Leder’s second-ranked topic (“Poetics and Semiotics”) and even, to some extent, the third (“Comprehension and Usability”), here seem primarily to reinforce that primary interest in meaning-making and communication. By the fourth-ranked topic we’re looking at contributions of less than 4%.

On the opposite end of the spectrum, consider Derek Mueller’s *Clouds, graphs, and maps: Distant reading and disciplinary imagination* (Syracuse University, 2009), discussed in chapter 1:

24 Leder’s dissertation on the philosophical underpinnings of literacy studies also serves as an interesting boundary case for inclusion in my dataset. Is this rhetoric/composition/writing studies? Why not philosophy? (In fact, the one dissertation more focused on Topic 48 [“Philosophy of Language”] is, indeed, from a Philosophy department: Mark Phelan’s *Figuration of the folk: The nature and use of a universal linguistic category* [UNC Chapel Hill, 2010]. Leder’s degree, as it turns out, is in English; he cites Patricia Bizzell and Janet Emig; his acknowledgments mention sustained work in the Writing Center; and moreover literacy is an explicitly named component of a number of prominent comp/rhet doctoral programs, including at Ohio State, Pitt, and UNC Chapel Hill. These facts reinforce my initial decision, based on his interest in meaning-making through writing, to include the dissertation in my analyses. However, the English program is not the primary home of rhetoric/composition at Michigan State University, but rather the program in Writing, Rhetoric, and American Cultures (“Ph.D. Program Overview”; “WRAC History”). This complicates my understanding of “Consortium schools” as presented in Chapter 3, and reinforces the need to establish department-level data in future iterations of this project. (Departmental affiliations were not included in the metadata provided by ProQuest; see Chapter 2.) It is possible that the presence at Consortium schools of Consortium programs may positively influence the culture of interest in rhetoric and composition in other departments, but at this point it is impossible to say.
Clouds, Graphs, and Maps: Distant Reading and Disciplinary Imagination examines recent efforts by scholars in rhetoric and composition to account for patterns and trends indicative of the discipline's maturation. Many of these "discipliniographic" appraisals resort, on the one hand, to anecdotal, experience-based accounts or, on the other hand, to methods too laborious to reproduce. Within this project, however, I identify and apply new methods that expand our means of apprehending patterns latent in the growing mass of disciplinary materials. Influenced by the work of Franco Moretti, this dissertation theorizes and also carries out variations of a methodology he calls "distant reading," which seeks to mine and aggregate data from large collections of texts to then build experimental models for engaging with non-obvious relationships. After establishing the exigency of this work for the field of rhetoric and composition and after establishing a conceptual groundwork for these methods, this dissertation presents three types of models — tag clouds, graphs, and maps — designed as a means to examine scholarship published in College Composition and Communication from 1987 to 2006. I contend that these models deepen and also complicate existing accounts of the discipline. By shedding light on large-scale patterns, the models also implicitly promote what I describe as a network sense of the field, which is crucial both for introducing newcomers to the shifting terrain of disciplinary knowledge and for sustaining a generalist's wherewithal in the midst of a growing archive of increasingly specialized scholarship. As a consequence of distant reading methods, network sense makes it possible for compositionists both to specialize in their work and also to keep abreast of developments in the field at the periphery of their narrow areas of teaching and research.

Table 4-5. Top 5 topics in Derek Mueller’s dissertation

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words</th>
<th>Proportion of Dissertation</th>
</tr>
</thead>
<tbody>
<tr>
<td>27</td>
<td>visual rhetoric</td>
<td>visual images image figure verbal art space body meaning photographs representations pictures picture representation objects multimodal photograph elements media</td>
<td>13.8%</td>
</tr>
<tr>
<td>43</td>
<td>genre and discipline</td>
<td>research genre genres study knowledge field writing discourse studies academic rhetorical disciplinary activity analysis social practices professional work discipline</td>
<td>11.3%</td>
</tr>
<tr>
<td>55</td>
<td>comprehension and usability</td>
<td>text analysis texts information discourse readers chapter study reader rhetorical content audience context specific reading data based features types</td>
<td>10.4%</td>
</tr>
<tr>
<td>8</td>
<td>(critical) pedagogical theory</td>
<td>students composition teaching pedagogy classroom teachers critical work student teacher theory studies knowledge learning ways education academic pedagogical practice</td>
<td>10.3%</td>
</tr>
</tbody>
</table>
Mueller’s dissertation has no one clearly dominant topic in the model; the top five topics stay within 4 percentage points of each other, and the top-ranked topic accounts for only 13.8% of the words in the dissertation. This balance starts to make sense if we think of these topics as working together more fluidly than in the case of a dissertation like Whitney’s. Whereas she had discrete chapters on participants’ reflections and on the history of the National Writing Project, Mueller is arguing in both his literature review and his original research about how compositionists can come to know about the scope of the discipline (Topic 43); his recommended approach is through visualizing (Topic 27) concrete instances of language (topic 55). That Mueller “theorizes” (Topic 48) the methodology he “also carries out variations of” further integrates and balances the components of his dissertation.

The presence of “(Critical) Pedagogical Theory” here, and in similar balance to the other topics, puzzled me at first: Mueller does not explicitly discuss pedagogy in his abstract. He does talk about “introducing newcomers to the shifting terrain of disciplinary knowledge,” which could bring in pedagogical language. But what is likely more significant is that the objects he examines – including through tag clouds of text-mined keywords – are themselves articles from CCC, meaning that their titles, abstracts, and other metadata may well be included in the model’s rendering of Mueller’s dissertation.
Similarity Clustering of Topics

As we saw earlier with the strong overlap between “Philosophy of Language” and “Poetics and Semiotics,” some of the topics identified by the model are more distinct than others. Consider, for example, the three topics in table 4-6:

Table 4-6. Details of three topics for comparison and contrast.

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words and Titles</th>
<th>Portion of Corpus</th>
<th>Rank</th>
</tr>
</thead>
</table>
| 28           | Political Rhetoric, Mostly of the U.S. | war president american public states united bush national military america speech political nation government people policy press carter york  
Part of something larger than ourselves: George H.W. Bush and the rhetoric of the first United States war in the Persian Gulf (76.4%)  
Foreign policy rhetoric for the post-Cold War world: Bill Clinton and America's foreign policy vocabulary (65.5%)  
The Reagan rhetoric: History and memory in 1980s America (64.5%) | 2.7%              | 16   |
| 36           | Political Discourse       | media news campaign political public blog obama television people issues communication coverage http analysis internet blogs post audience Clinton  
A functional analysis of the 2000 Taiwanese presidential campaign discourse: Advertisements and speeches (36%)  
He said, she said: A functional analysis of differences between male and female political campaign messages (61.9%)  
The impact of interest group and news media framing on public opinion: The rise and fall of the Clinton health care plan (55.9%) | 1.2%              | 31   |
| 45           | Preparation for College   | education school students college teachers educational schools learning higher high student instruction teaching teacher skills colleges educators university state  
Framing first year writing: The conceptual metaphor of journey and the Advanced Placement program (45.2%)  
Self-regulation in college composition: No writer | 1.8%              | 21   |
Topics 28 (which I’ve called "Political Rhetoric, Mostly of the U.S.") and 36 ("Political Discourse") are clearly rather similar to each other, despite having some different top words – *bush* and *carter* in Topic 28 correspond to *clinton* and *obama* in Topic 36, for example. Topic 45 ("Preparation for College"), on the other hand, describes rather a different conversation. To extend this insight and visualize the clusters of similarity among all topics in the model, I constructed figure 4-6, below.

In this figure, the topics are arranged around the circle such that the most similar topics are adjacent to each other. The tree shows clusters of similarity at increasing levels of abstraction, from 48 distinct groups at the outer ring, joined to 21 clusters, then 11, 6, and 4 clusters, and finally 2 large groups that converge at the center. An interactive version is available at [http://majoringinmeta.net/dissertations/figure4-6_consortsk55_radial_clusters.html](http://majoringinmeta.net/dissertations/figure4-6_consortsk55_radial_clusters.html) (or [bit.ly/1CXwbH2](http://bit.ly/1CXwbH2), for convenience), which allows visitors to scroll the mouse over each topic to view details such as the full topic label, the percentage of corpus that topic accounts for, what rank that percentage gives the topic relative to the others, and the words most associated with that topic. Scrolling over the branch points of the tree will display the cumulative percentage accounted for by all nodes included in that branch.

25 Because each topic can be thought of as a vector of probabilities distributed across all the words in the corpus, we can find the distance from one such vector to another using a kind of high-dimensional Pythagorean theorem. The resulting correlation matrix was then sorted to create the hierarchical clusters shown in the tree. The code generating this hierarchical structure can be found in Appendix F, as `frameToD3.R`.

---

**left behind (33.3%)**
*Experiencing composition and literature: Advanced placement and the ends of English (31.4%)***
Topics 28 and 36, which we identified as similar from direct inspection, do indeed turn up next to each other at the bottom left of the figure, as two of four topics that combine in the outermost clustering step. (In other words, if the model had 21 topics instead of 55, we would expect these four topics to combine into one.) Again confirming our expectations, Topic 45, at the middle right of the figure, does not share a common cluster with them – they remain separate even when the full set of topics is divided into just two groups. Instead, Topic 45 is now joined by Topic 6 ("Institutional Context of Writing Instruction"): 

**Figure 4-6.** Screen-capture from interactive figure, available online at ow.ly/J2dVL, for exploring similarity clusters of topics within the model.
As before, the top words seem different enough, but the concepts are clearly related: as students prepare for college, their institutional context shifts. The visualization thus makes it easier to discover and describe related content clusters.

<table>
<thead>
<tr>
<th>Topic Number</th>
<th>Assigned Label</th>
<th>Top Words</th>
<th>Portion of Corpus</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>Institutional Context of Writing Instruction</td>
<td>faculty writing english university program college composition teaching courses year graduate programs academic department time students education professional research</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>University of Louisiana system freshman composition faculty: Instructor working conditions and student learning conditions (76.9%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Writing at the small liberal arts college: Implications for teaching and learning (49.0%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>The state of writing instruction in Southern Baptist colleges and universities (44.6%)</td>
<td>2.5%</td>
<td>17</td>
</tr>
<tr>
<td>45</td>
<td>Preparation for College</td>
<td>education school students college teachers educational schools learning higher high student instruction teaching teacher skills colleges educators university state</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Framing first year writing: The conceptual metaphor of journey and the Advanced Placement program (45.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Self-regulation in college composition: No writer left behind (33.3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Examining composition and literature: Advanced placement and the ends of English (31.4%)</td>
<td>1.8%</td>
<td>21</td>
</tr>
</tbody>
</table>

Therefore, while an initial review of the topic proportions on their own suggests a relatively flat hierarchy of dissertation content, with the largest topic accounting for only 5.39%
of the corpus, it now seems more accurate to describe the model as divided into five or six major divisions, highlighted in figure 4-7 and the accompanying table 4-8, below:

**Figure 4-7.** Major divisions among topics in the model as determined by hierarchical clustering.

Viewed in this way, the two largest individual topics are subsumed into one large content cluster, centered on the teaching of writing (highlighted here in yellow), which accounts for just under a third (31.37%) of the corpus. This should go some way toward alleviating the concerns of some in the field, especially Fredrik deBoer, that the teaching of writing is simply not valued at the dissertation level. Early in a contentious set of threads on the Writing Program Administrator's
<table>
<thead>
<tr>
<th>Region</th>
<th>Assigned Name</th>
<th>Topics included</th>
<th>Total percentage of corpus&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>The Teaching of Writing</td>
<td>Students in the Classroom (5.39%); (Critical) Pedagogical Theory (4.77%); Process Reflections (3.84%); History of Composition (3.07%); Writing Process: formal and cognitive studies (2.89%); Institutional Context of Writing Instruction (2.49%); Quantifying Written Products (2.08%); Preparation for College (1.81%); Literacy and Literacies (1.77%); Multilingualism and World Englishes (1.31%); Writing Center Tutorials (0.98%); Online Learning and Collaboration (0.98%)</td>
<td>31.37%</td>
</tr>
<tr>
<td>Green</td>
<td>Theories of Meaning-Making</td>
<td>Philosophy of Language (4.33%); Poetics and Semiotics (3.2%); 18&lt;sup&gt;th&lt;/sup&gt; and 19&lt;sup&gt;th&lt;/sup&gt; Centuries (3.04%); Close-reading Classical Rhetoric (2.8%); Postmodernist Theories of Meaning and Invention (1.43%); Writing With and About Religious Texts (0.93%); Reading Kenneth Burke (0.8%); Silence plus England&lt;sup&gt;b&lt;/sup&gt; (0.72%)</td>
<td>17.25%</td>
</tr>
<tr>
<td>Purple</td>
<td>Performative Identities, Past and Present</td>
<td>Identity Construction (4.21%); Story and Narrative (4.13%); Women Acting Rhetorically (1.74%); Narrative Theory and Readers as Writers (1.58%); Race and White/Black Power Struggles (1.34%); Performance (0.91%); Native American History and Rhetoric (0.61%); Film Criticism plus Japan&lt;sup&gt;b&lt;/sup&gt; (0.57%)</td>
<td>15.1%</td>
</tr>
<tr>
<td>Red</td>
<td>Audience and Context for Composing (not limited to the written word)</td>
<td>Community Engagement (3.66%); Comprehension and Usability (3.52%); Rhetorical Affordances of the Web (2.35%); Genre and Discipline (1.75%); Technical Communication (1.29%); Visual Rhetoric (1.24%); Games plus Commercial Editing Practices&lt;sup&gt;b&lt;/sup&gt; (0.69%)</td>
<td>14.49%</td>
</tr>
<tr>
<td>Blue</td>
<td>Politics and Power</td>
<td>Capitalism, Marxism, and Activism (3.56%); Political Rhetoric, mostly of the United States (2.7%); Political Discourse (1.23%); Court Decisions and Ramifications (1.09%); International Conflict and Negotiation (0.62%); Civic Discourse plus China and Japan&lt;sup&gt;b&lt;/sup&gt; (0.49%)</td>
<td>9.68%</td>
</tr>
</tbody>
</table>
### Region | Assigned Name | Topics included | Total percentage of corpus
---|---|---|---
White | Other | Workplace and Administrative Histories (3.36%); Health, Medicine, and Disease (0.99%); Science Writing and Environmentalism (0.9%); Museums and Archives\(^b\) (0.54%); Medicine, Disability, the Body, and Identity (0.53%); Military-Industrial Complex\(^b\) (0.44%); Food and Cooking (0.32%) | 7.09%

a. Because non-content-bearing topics are ignored, but the dissertations containing them in high degrees were not removed prior to building the model, the total percentage column will sum to less than 100% – close to 5% of the total corpus was comprised by these non-content-bearing topics.

b. This small topic (< 0.73% of the corpus) seems to combine several even smaller topics, perhaps because a handful of dissertations using them together carried more weight than they would have in a topic represented by a larger sample. These topics may well have been split in a model with a greater number of topics, but possibly at the cost of some coherence in the larger topics.

Email listserv (WPA-L) running to over 85 posts, deBoer writes,

> the message that is sent to doctoral students and young scholars in the field is that the teaching of prose is not valued, that research on teaching prose is not valued. People want careers, and they see what gets published and what gets talked about in conferences, and they take coursework that is about subjects that are very far from traditional prose instruction. The result is a generation of scholars who are producing scholarship that most people outside of the field would not identify as about writing at all. I'm not conservative, I think it's great that some people are writing dissertations on agential realism and Dr. Who and 3D printing. The problem is that the field seems to produce nothing but dissertations on subjects like these, and almost none on prose instruction[]. (DeBoer, emphasis added)

Needless to say, heated email messages are not often known for their high standards of evidence; they are not refereed articles, and deBoer and others may have been simply glib in declaring the presence or absence of certain dissertation topics. Even so, claims like this were repeated and repeatedly grounded only in anecdote. My study, and future distant reading projects like it, provide a means of checking anecdotal impressions against a wider scope, rendering them either falsifiable or defensible.
My data offers strong evidence against deBoer’s final claim above: looking at the topic model of dissertations from 2001-2010, at least, it is simply not accurate to say that there are "almost none on prose instruction." Even with a stricter criterion, e.g. if we left off the next-highest branch within this cluster, which includes more WPA-focused topics such as “Institutional Context for Writing Instruction” (but also theory of writing pedagogy), that would still leave 19.23% on writing instruction *per se*, or nearly one-fifth of the corpus. Considering that in most dissertations the top-ranked topic only accounts for 20-32% of the dissertation (see figure 4-5), 19-31% is on average a fairly high proportion of dissertation content – and the teaching of writing still comprises the highest-ranked cluster at either my initial criterion or the stricter one.

That said, I can imagine a counter-argument that emphasizes the other side of the same statistic: yes, writing instruction is the top-ranked cluster, but nevertheless some 68-80% of what graduate students are writing about in their dissertations is *not* on writing instruction. This may seem surprising, "in a field that sometimes goes by the name of writing studies" (to quote deBoer again), yet he finds himself frequently "having to defend the value and importance of writing pedagogy" even within this field (DeBoer, "Re: Video of Banks' talk?"). While I personally sympathize with the desire to study writing and writing processes, in the sense of how writers generate and revise alphabetic text, such study seems to me to also lead naturally into questions that extend beyond (even as they point back toward) written prose: to what ends do writers engage in these processes, and with what effects?

Several other content clusters focus on such second-order, or indirect, writing questions. In blue above, the cluster I've labeled "Politics and Power" addresses the matter of "toward what end," featuring dissertations such as *Making change: The role of rhetoric in the politicization of*
consumption (Lonni Dee Pearce, University of Arizona, 2003) and Literacies for the long haul: Radical teaching, social movements, and spaces of hope in the age of neoliberal globalization (Kevin T. Mahoney, Miami University at Oxford, 2002). This cluster accounts for just shy of 10% of the corpus.

Another cluster, highlighted in red, includes several topics that examine the audiences addressed by writing and how writers' choices are shaped by their community and context, and includes dissertations such as Online writing labs as sites for community engagement (Jaclyn Michelle Wells, Purdue University, 2010) and Bulleted points and typographic cues: Effects on recall and recall order (Raymond Narveson, University of Minnesota, 2001). This cluster, which includes examination of writing's effects, makes up close to 15% of the total dissertation text.

The second-largest cluster I've identified through the topic model, highlighted in green above and accounting for 17.25% of the corpus, might be thought of as exploring third-order questions about writing: once we've considered the ends and effects of a text, it remains to be determined how we know a text achieves those effects. This cluster, which I've labeled "Theories of Meaning-Making," deals with the philosophical and theoretical underpinnings of writing and language more generally. Typical titles from this cluster include John Dewey on the art of communication (Nathan Crick, University of Pittsburgh, 2005) and Capturing kairos: A theory of rhetorical cunning (Matthew W. Schnackenberg, Washington State University, 2006).

**Connections Between Macro-Clusters**

It is tempting to go even further: The highest-level split in the hierarchy, at the circle’s center, combines “Teaching of Writing” (yellow) and “Audience and Context for Composing” (red) into a joint macro-cluster, distinct from a second macro-cluster containing “Theories of Meaning-Making” (green), “Performative Identities” (purple), “Politics and Power” (blue). It is
possible to see this top-level split as distinguishing, respectively, Composition (centered on practice and generating texts) from Rhetoric (centered on discourse and “the theoretical and historical study of texts,” per Theresa Enos [qtd. in Kopelson 770]). As Kopelson points out (769-70), these two core terms of the field have often seemed at odds, even to the point of disciplinary rupture; she reviews discussions26 of a “rhetoric/composition split” spanning two decades (769), with several possible explanations of the division all pointing in the same direction: “that the seeds of dissolution are indeed being sown” (770). The emergence of Composition and Rhetoric as two largely separate hemispheres, as it were, seems at first to reinforce the idea that rhetoricians and compositionists exist in separate communities of discourse: the hierarchical clustering in figure 4-7 is, after all, based on the similarity of words used across documents. Could it be that compositionists and rhetoricians are speaking entirely different languages?

For good or for ill, the reality is not that simple. Lest we get too worked up about divisions in the field, the data suggest that many dissertation writers work across these levels and content clusters. figure 4-8 recreates the same27 topic clusters as in figure 4-7 around the outer circle; lines connect pairs of topics that occur together in at least four dissertations, where "occur" means that each topic accounts for at least 12% of the dissertation28. The lines in figure 4-8a follow the paths defined in the previous visualization, which show hierarchical similarity of


27 In some clusters, the order of lower branches may have been swapped by the algorithm generating the figure, but the hierarchical relationships remain the same, as do the positions of the six highlighted clusters.

28 The 12% cutoff is based on the lower quartile of the contribution of top-ranked topics within dissertations, as shown in Figure 4-5.
topics, thus allowing us to see how dissimilar topics co-occur within documents. NB: an interactive version of this figure is available at http://majoringinmeta.net/dissertations/figure4-8 Consortsk55 Hierarchical edge bundling.html (or, as before, via the shortcut bit.ly/1KmHa). In that version, hovering on any one topic will both reveal detailed information about that topic and highlight all connections to other topics (as in figure 4-8b).
Figure 4-8a. Screenshot of interactive figure for finding topics that tend to occur in the same dissertations.

Topic labels are arranged around the outside of the circle; blue lines connect topics that contribute more than 12% each to the same dissertation at least 4 times.

Lines are curved so as to follow the hierarchical clustering map between connected topics; thus lines crossing closer to the center indicate connections between more disparate topics.
In this view, it is easy to see that nearly every topic co-occurs with at least one other topic. Our top topic, “Students in the Classroom,” connects with 21 other topics, including all 11 other topics within its top-level branch of the hierarchy (“The Teaching of Writing”); it also reaches beyond that branch to the top four of seven topics in the “Audience and Context for Composing” branch: “Community Engagement and Collaboration,” “Comprehension and Usability,” “Rhetorical Affordances of the Web,” and “Genre and Discipline.” Perhaps most significantly, many lines reach across the center division of the figure: “Students in the Classroom,” to continue the present example, frequently shares dissertation space with “Identity Construction,” “Story and Narrative,” “Race and White/Black Power Struggles,” and “Narrative Theory / Readers as Writers,” as well as the historical/archival-tending topic I’ve labeled
“Women Acting Rhetorically,” all in the cluster on “Performative Identities.” Even the cluster on “Theories of Meaning-Making” – which we might expect to be the most rarified and least tied to everyday pedagogical practices – coexists in a number of dissertations with “Students in the Classroom,” specifically through the topics I’ve labeled “Poetics and Semiotics” and “Philosophy of Language.”

Just three of the topics around the circle lack connections entirely at the 12% level being considered here:

- Topic 54 (“Food and Cooking”), which represents only 0.32% of the corpus, the second-lowest rank overall;
- Topic 46 (“Mostly Museums and Archives”), which similarly represents a tiny fraction of the words, 0.54%; and
- Topic 26 (“Reading Burke”), at 0.9% somewhat larger, but still quite small. Given the small numbers involved, it seems safe to say that these topics may not have emerged as co-occurring with other topics simply because they did not cross the 12% threshold often enough. In the future, were there to be more dissertations giving a great deal of attention to these topics, we could expect them to be less exclusive in their focus – especially, I might add, if the authors were able to examine a figure like this one during the dissertation-planning process. (See Chapter 5 for more thoughts on how such a figure might help.)

29 “Reading Burke” is, intriguingly, also the only topic in the model to center on the hermeneutics of a single author’s work. The significance of this discovery is unclear, however, and I would be interested first to see whether the distinction holds true in repeated iterations of this project before jumping to conclusions.
Although figure 4-8 makes it easy to detect the origins and endpoints of links between topics, without the online interactive features it is harder to trace these connections between specific topics, and harder to see at a glance how frequently overall dissimilar topics occur together. These connections are clearer in figure 4-9, which replaces the curves with straight lines to minimize overlap. Seen this way, the links demonstrate that a large number of topics are discussed in tandem with topics outside their cluster: multiple lines connect the Teaching of

Figure 4-9. Co-occurring topics overlaid with major divisions discovered by hierarchical clustering (see Figure 4-7).
Writing (yellow cluster) to Theories of Meaning-Making (green) and to Performative Identities (purple); Audience and Context (highlighted in red) is especially strongly bound to Teaching of Writing (yellow) and to the various "Other" topics (white). Although it's clear that the two largest groups (red and yellow vs the rest, roughly corresponding to the left and right sides of the figure, what I earlier suggested might correspond to production and reception) do feature more internal topic co-occurrences than external, connections "across the divide" are not uncommon, supporting the idea that there exists a field of Composition and Rhetoric.
Chapter 5: 
Building the View From Everywhere

The foregoing analyses are not intended to establish, for once and for all, the internal and external boundaries of Composition, Rhetoric, and Writing Studies. Such a terrain is in constant flux, as individuals and departments negotiate their ways through overlapping and diverging interests, influenced by national or larger conversations as well as local material conditions, including the time available for research. But to say that these maps of the field are impermanent does not erase their value. Rather than fix the field in place from some unreachable point of objectivity – a "view from above" that aspires to be a "view from nowhere" – they gather together a whole host of subjective judgments of what counts, aggregating these disparate views into what we might call a "view from everywhere at once."

Even this will be a partial view: the questions I've asked so far are limited in scope, and certainly not everything can be seen from abstracts or from topic modeling. Still, getting beyond the perspective of one or two schools can serve as a corrective to strong local tendencies that might otherwise be forced to stand in for the field as a whole. Consider Karen Kopelson's recent survey of "graduate students at two large and long-established doctoral programs in rhetoric and composition" (753), from which she concluded that "current and future scholars are frustrated" by an over-emphasis on classroom applications and pedagogy (757), to the point of "concern for the present and future status and potential knowledge-making contributions of the field as a whole" (ibid). Granting that "80%" of her respondents felt this way, we do not know how many people that 80% accounts for, but it is surely less than 100, given graduation rates at even the
largest schools (see Appendix C) – and it seems significant that her impression based on this sample is the exact opposite of Fredrik deBoer's impression, discussed in Chapter 4, that "the people who oppose a pedagogical focus have already won" ("Re: Video"). Or, as he says earlier in the same post to WPA-L: "I guess this is just the thing, for me: I don't recognize the field that people talk about when they worry that we'll become nothing but a pedagogical discipline" (ibid).

If we remain focused on their small samples, it would seem that deBoer and Kopelson can't both be right, but by zooming out, we can situate their respective pockets of pedagogical and theoretical emphasis in context, and see that they're accurately describing two different corners of a broader landscape. In this way, maps built from distant reading can help people "recognize the field" anew. Equipped with a common understanding of the larger scale, graduate students, thesis committees, and curriculum-planners can make more informed local decisions about where their research should go next.

The View So Far

In Chapter 1, I set out to discover whether the field has stayed together, and how, or whether it has fragmented and fallen apart without really noticing. Four chapters later, I don't think we need to worry too much about the latter. Though there do seem to be, as North put it, "communities or clusters of communities" (Making 364) of both method and subject matter, many connections also exist between these communities, bridging the divide: dissertations engaged in multiple methods, addressing multiple topics. This suggests that one of the core ways in which emerging scholars are trained is to ask more than one question, and answer in more than one way. To put that another way, whereas North worried that “methodological integrity”
was a necessary precursor to disciplinary advancement, most dissertation committees seem rather to agree with Todd Taylor's counterargument: “it may be more the case that the health of today's academic disciplines actually require methodological diversity and interdisciplinarity rather than rigidity and insularity – much like a wide gene pool promotes immunity” (145).

I had further aimed to better articulate to the outside world the nature of what Comp/Rhet’s researchers, scholars, and practitioners know and do. Though the answers remain sufficiently complex that any simple answer is bound to be reductive, some reductiveness is necessary if we are to have a map that is distinct from the territory itself. With that in mind, what have we learned?

While classroom-based research is far from dominant in the field (Practitioner / Teacher research is the 8th-ranked method overall in this dataset), the teaching of writing remains an important touchstone for many dissertation-writers. Not only does the topic cluster on the teaching of writing account for 31.37% of the words in the Consortium school corpus, as discussed in Chapter 4, but of the 1,754 dissertations included in the topic model, over half are concerned enough with the teaching of writing that the sum of contributions from this cluster makes up 20% or more of these dissertations' text. (See table 5-2, which was produced using ‘topic cluster reach.R’, included in Appendix F.)

<table>
<thead>
<tr>
<th>Region</th>
<th>Assigned Name</th>
<th>Dissertations at ≥ 20%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yellow</td>
<td>The Teaching of Writing</td>
<td>970 (55.3%)</td>
</tr>
<tr>
<td>Green</td>
<td>Theories of Meaning-Making</td>
<td>511 (29.1%)</td>
</tr>
<tr>
<td>Purple</td>
<td>Performative Identities, past and present</td>
<td>509 (29.0%)</td>
</tr>
<tr>
<td>Red</td>
<td>Audience and Context for Composing</td>
<td>434 (24.7%)</td>
</tr>
<tr>
<td>Blue</td>
<td>Politics and Power</td>
<td>274 (15.6%)</td>
</tr>
<tr>
<td>White</td>
<td>Other</td>
<td>170 (9.7%)</td>
</tr>
</tbody>
</table>
We also know that Philosophical/Theoretical, Historical/Archival, Critical/Hermeneutical, and Rhetorical Analytical methods are the most common throughout the Consortium schools, where all of my topic modeling data comes from. Given this, we can infer that many of the dissertations studying the teaching of writing are likely to include accounts of historical classrooms, teaching philosophies and applications of theory to the classroom, and close readings of students' or teachers' writing. Further analysis will be able to confirm or improve upon this hypothesis.

At the same time, it is clear from the same statistics that many dissertations do not meet even this low criterion for studying the teaching of writing, but they nevertheless make up some 45% of the successfully approved dissertations at Consortium schools. It is therefore far from a universal imperative to discuss pedagogical applications in order to be fully fledged as a member of Composition and Rhetoric. Rather, what seems to matter more across the board is attention to how meaning is made and conveyed – the conditions that give rise to meaning-making, the processes involved, and what happens then.

If such a definition seems difficult to detangle from that of other disciplines or academic fields, such as Communication, Linguistics, or Psychology, it is perhaps because of Composition's long history of "borrowing" from such fields. It could prove fruitful to conduct related distant readings on dissertations from these fields, as well, to better determine what, if anything, helps give Composition/Rhetoric its own distinct character.30

30 One project that is promising in this regard is the Stanford Dissertation Browser http://nlp.stanford.edu/projects/dissertations/browser.html, which aligns topics to departments, and visually displays similarity among departments based on topic assignments to dissertations (all dissertations from all departments at Stanford, 1993-2008), using a form of topic modeling, called Labeled LDA, related to but distinct from the one I used.
To some, the “findings” I report above might seem obvious, as if we’ve made no advance: we already knew there were disagreements as to the importance of theory vs. pedagogy, and we already knew that there were multiple methods at play in the field. I would respond in two ways. First, having evidence to back up our impressionistic claims is itself an advance, especially because it will allow us to track changes over time. Second, if we wish to communicate the value of our work to stakeholders who are not already invested in the field’s future – whether across the disciplines and administration within the university, or to parents and politicians beyond the university – we need to have data-supported answers to questions about what “counts” as work worthy of a doctoral degree.

**Into the Future**

I would, of course, feel more confident in making the claims above if I could more readily confirm that the people submitting these dissertations, and the faculty members approving them, would identify themselves as doing work in Composition, Rhetoric, and/or Writing Studies. What we do have to go by, as I said in Chapter 2, is the fact that every dissertation in my dataset included the subject tag, "Language, Rhetoric and Composition," paired with my own screening of the abstracts to confirm on an individual basis that the work would not be out of place in a comp/rhet journal or conference. However, spot checking confirms that at least some of these dissertations – including those at Consortium schools – were completed in departments other than the local home of comp/rhet faculty.

To take one example, at the University of Pittsburgh, the dataset includes 31 dissertations. Of these, only 13 were completed in the English department, where the Composition program is housed; 15 were from Communication. (Political Science and Psychology supplied the remainder.) Though faculty from one program have occasionally served
as committee members on dissertations from another, this appears to be more the exception than
the rule. In other words, more than half of the dissertations counted as coming from this one
"Consortium school" are not, in fact, coming from a "Doctoral Program in Rhetoric and
Composition." While I stand by my earlier assessment that these dissertations would be
recognized by compositionists and rhetoricians as doing the work of the field – the
Communications department at Pitt lists specializations in History, Theory and Criticism of
Rhetoric; in Media and Culture; in Public Address and Argument; and in Rhetoric of Science
(University of Pittsburgh University Marketing Communications Webteam) – this discovery
could prove problematic for my characterizations of the programs' methodological output and
focus. Or not: in the particular case of Pitt, there are no statistically significant differences in
either output or focus between the two programs (though the small sample size and many
variables being compared do make such significance a difficult hurdle to achieve).

For this reason, one of my first priorities for continuing the research begun in this
dissertation will be to establish the departments associated with each document in the
dataset, and to re-run some analyses after applying departmental filters. In addition to helping
graduate students and graduates to locate the best-fit departments for their research interests, a
number of questions about the field's composition31 could be addressed by an analysis through
departmental affiliations (some of which I alluded to in earlier chapters):

- Does the predominance of dissertations using "Scholar" methods (Philosophical,
  Historical, Critical, Rhetorical) identified in Chapter 3 come primarily from programs
  associated with English departments? Do independent writing studies programs maintain
  that focus?
- Similarly, do the "Researcher" methods come primarily from departments of Education,
  Linguistics, or Psychology? Or not?

31 This one isn't really a pun, but the reminder of this meaning of the term is intentionally both playful and serious.
I'm even less sorry about this one.
• Is the appearance of “Comprehension and Usability” as a prominent topic associated with programs or tracks in Technical and Professional Communication? If these programs are counted separately (which is a debatable move), what topics then emerge as the top ten?
• How unified are the topic clusters identified in Chapter 4 within Consortium programs alone, leaving out the non-Consortium programs at Consortium schools?
• What are the approximate graduation rates across all programs that house explicitly Composition/Rhetoric faculty and students, and how has that changed over time? How well do such graduation rates track the number of Composition/Rhetoric jobs posted in the same time span? Because such numbers are not directly tracked – and because, when they are, composition/rhetoric graduates of English departments tend to be subsumed into one pot with English literature graduates – it has been difficult for the field to adequately gauge job market pressures on recent graduates, or how those may be changing over time.

Thus far, I have mostly discussed topics and methods as separable entities, but I am also interested in how they relate to one another. We would expect, for example, that some methodological communities function as more or less distinctive discourse communities: think of the abstract language and complex syntax of capital-T Theory (see Nevin Leder’s dissertation in chapter 4, for one example), or of the Greek terms ethos, enthymeme, kairos, and so on used in classical Rhetoric. This leads to the question, are certain topics in the model – which are, after all, merely clusters of co-occurring words – closely tied to certain method tags? A factor analysis of methods as a function of topics, beginning with a sample of single-method dissertations, might be able to identify certain topics (or combinations of topics) as method predictors. If so, they could be used to make relatively fast estimates of the methodological clusters in new data.

---

32 Actual rates should be slightly higher, owing to embargoed dissertations being absent from the dataset and the possibility that ProQuest is not required at 100% of relevant schools.
33 This may be changing, at least for job listings. For 2012-2015, Jim Ridolfo has been archiving job postings in rhetoric and composition (and, for some of those years, technical communication) at http://rhetmap.org.
34 Although the topic modeling was rather time-intensive, it was on the order of months: far less time than the initial tagging of abstracts, which took over a year. Moreover, one result of the time spent on topic modeling is a set of
And I do think extending the dissertation dataset will be important, especially as time goes on and the window of 2001-2010 recedes further into the past. One option for new data would be to go further back in time, for as many dissertations as could be located in digital or (digitize-able) form. As I suggested earlier, a series of field-wide snapshots could help us see shifts, or turns, in the field's collective interests and the ways that those interests are expressed in language. Some of this work has already been done in journal articles (Mueller Clouds), CCCC chair addresses (Mueller "Views"), and job listings (C. Lauer), and it would be interesting to see whether the same trends emerge in doctoral work – and whether doctoral work trails or leads the more traditional indicators of disciplinary currency.

Extending the dataset forward in time (as close to the present as possible) is just as important, and not only for further comparisons like those just described, but also because of the generative nature of diagrams like figure 4-9, which shows topics that co-occur within dissertations. Such a figure presents an opportunity to discover new dissertation questions through combinations that appear or, especially, that do not appear. As an example of how this might work, a student consulting the diagram might note that fewer than four dissertations in the dataset seem to have connected “Writing Center Tutorials” (top right, in the yellow cluster I have labeled "The Teaching of Writing") with any of the topics in the left half of the figure. This suggests that a new study examining writing center practice through the lens of, say, “Identity Construction” (middle left, in the purple cluster of "Performative Identities") or “Writing with and about Religious Texts” (top left, in the green cluster of "Theories of Meaning-Making") functions for preparing and analyzing the data, which I would not have to recreate from scratch. Tagging the methods, by contrast, produced no such replication-friendly product.
could be a welcome intervention. With a more up-to-date set of texts to model, we could anticipate even more useful guidance from an index like this.

**Limits of the Method**

Dissertations, as suggestive as they might be, do not in themselves tell us the full picture of graduate training. Dissertation advisors are not the only source of influence on a student’s approach, and the influence that these students then exert after graduation depends on a number of factors, including when and where they teach. This data has not historically been easy to obtain, and so it is an open question whether graduates of schools in a given cluster of methods or topics go on to work at schools in the same cluster, thus preserving and stabilizing the local admixture of expertise, or whether migrations have occurred across clusters, and in what strength.

One place where we are beginning to gather data on patterns of affiliation between people and institutions is the Writing Studies Tree ([http://writingstudiestree.org](http://writingstudiestree.org)), which is an open-access, crowdsourced database of just such academic genealogies: relationships of mentoring, education, collaboration, and employment. Future studies could look for correlations between dissertation methods or locations and various measures of subsequent influence as recorded in the WST, including job placement:

- are some more likely to work in graduate research departments vs. teaching-intensive undergraduate departments or K-12 schools?
- Do dissertations mentored by faculty whose own dissertations employed a particular range of methods use the same methods, and how much does it matter if that mentoring was as a chair or a non-chair member of the dissertation committee? Does that answer

---

35 The Writing Studies Tree is an open-source and open-access project started at the Graduate Center, CUNY, by Benjamin Miller, Amanda Licastro, and Jill Belli under the guidance of Sondra Perl and Matthew K. Gold, with programming support from Matt Miller and Jeffrey Binder, and funding support from the Graduate Center’s Provost’s Digital Innovation Grants. For a full description, see http://writingstudiestree.org/about.
vary by school?

Because the WST relies on voluntary contributions of data from a large number of people, the picture it paints is only an incomplete approximation of the field’s history; however, by the same token, the WST database continues to grow as more members of the field share what they know, and so comes closer to a representative portrait.

In Conclusion

Christine Farris and Chris Anson recognize that despite the impossibility of a single vantage point from which to unify all our studies, such a position would in any case be less useful than a set of positions from which to triangulate and gain perspective. In their words, “This notion that theory, research, and teaching are all practices providing a location from which to view and critique the others […] offers a way out of battling binaries” (Under Construction 3, emphasis added). Articulating the locations of topic and method, as I have done in the previous chapters, can work to combat a problematic condition that David Smit seems to take as inevitable, that “composition studies has no means for even talking about the differences that divide the profession” (225).

Having “a way out of battling binaries” does not mean simply that we should simply live and let live, free from the task of mapping out the field's contours. On the contrary, it suggests that multiple maps are needed “to view and critique the others” – and, I might add, to celebrate their different contributions – even as we recognize that any one map is insufficient. One major advantage of the distant reading approach that I have taken in this dissertation is that the programs that generated the figures and statistics reported above are replicable functions. As new data becomes available, updating the analyses is a matter of passing that data into the functions.
As more snapshots accumulate, we can begin to see what changes and what stays the same; the more stable the core, the more comfortable we can finally be using the word "discipline."
Appendix A:
Index of GoogleRefine / OpenRefine Scripts Used to Prepare Metadata for Analysis

Dissertation data was supplied by ProQuest as a .xlsx (Microsoft Excel) file, which included fields for Publication Number, School, Author, Advisor, Title, Subjects (closed vocabulary), Degree, Year, Page Count, Abstract, and Keywords (open vocabulary). This data was then prepared for analysis using Google Refine – now called OpenRefine, available at http://openrefine.org/ – as described in Chapter 2.

This appendix serves as an index to JSON files generated in Google Refine through its Undo/Redo > Extract feature. Each file contains instructions which will allow Refine to recreate the steps taken, using the Undo/Redo > Apply feature.

Because of the large amount of whitespace and repetition of JSON files, and the subsequent length of some of these documents, rather than printing them here I have chosen to share them online at https://github.com/benmiller314/Dissertation-Research, in the OpenRefine directory.

One set of files contains the commands used initially to clean the data for coding:

A. clean_titles_and_abstracts.json
B. remove_allcaps_keywords.json
C. split_advisor_from_advisortype.json
D. simple_keyword_merging.json

The commands in file B above were used to remove keywords written in all capital letters, which seemed machine-generated rather than author-supplied. Before this process, there were 20,824 keywords; afterwards, there were 7,974. The commands in file D above were used to locate and merge obvious synonyms among the author-supplied keywords, such as "Kenneth Burke" with "Burke, Kenneth" or "Rhetoric" with "rhetoric." This process further reduced the number of unique keywords to 7,572.

Another set of files can be used to expand and contract records between a single row (for analysis in R) and multiple rows (for ease of reading in Excel). Note that steps 3, 4, and 6 below are now accomplished by `method tag array.R`.

1. compress_and_clean_multirows.json
2. expand_method_tags.json
3. count_methods.json
4. method_words_to_bits.json or method_words_to_bits2.json
5. compress_and_clean_multirows.json
6. exclude_bits.json
7. (export as .csv for use in R)
Appendix B:
ProQuest Dissertation/Thesis Numbers for Included and Excluded Dissertations

ProQuest Dissertations and Theses (PQDT) assigns a unique identifier to each item in its database. Metadata for the items with the following identifiers were provided by ProQuest for use in this research project; researchers wishing to replicate the analyses on the exact same dataset may copy this list and request full text or metadata from PQDT.

Included in the broadest analyses (2,711 items):
3000301, 3000388, 3000749, 3001056, 3001291, 3001292, 3002127, 3002452, 3002516, 3002827, 3002875, 3002940, 3003048, 3003427, 3003732, 3003784, 3004118, 3004152, 3004856, 3004880, 3004928, 3004946, 3005613, 3005620, 3005622, 3006139, 3006147, 3006405, 3006626, 3007265, 3007369, 3007461, 3007720, 3007997, 3008295, 3008469, 3008717, 3008762, 3008778, 3008780, 3378966, 3008896, 3009064, 3009607, 3009697, 3009726, 3009847, 3009855, 3009869, 3009902, 3010100, 3010197, 3010241, 3010265, 3010431, 3010874, 3011064, 3011116, 3011553, 3011736, 3012238, 3012359, 3012360, 3013016, 3149958, 3013429, 3013451, 3013929, 3014374, 3014379, 3014694, 3014703, 3014714, 3014786, 3014815, 3014874, 3014893, 3015010, 3015243, 3015255, 3015734, 3015815, 3016159, 3016164, 3016185, 3016262, 3016698, 3016885, 3017086, 3017089, 3017138, 3017142, 3017397, 3017414, 3017482, 3017487, 3017763, 3019085, 3019098, 3019156, 3019157, 3019158, 3019412, 3019607, 3019745, 3019873, 3020348, 3020532, 3020547, 3020630, 3020942, 3020964, 3021025, 3021474, 3021502, 3021511, 3021635, 3021825, 3022059, 3022147, 3022341, 3022467, 3022736, 3022954, 3022965, 3023005, 3023160, 3023291, 3023367, 3023694, 3023941, 3024078, 3024255, 3024349, 3024485, 3024516, 3008883, 3025246, 3025320, 3025416, 3025536, 3025558, 3144941, 3026213, 3026345, 3026422, 3026428, 3026429, 3026491, 3026525, 3026621, 3026770, 3027030, 3027033, 3027035, 3027056, 3027063, 3027069, 3027535, 3027723, 3027785, 3027904, 3027905, 3028654, 3028763, 3028898, 3029080, 3029597, 3029714, 3029834, 3029847, 3029877, 3029912, 3030087, 3030181, 3030327, 3030341, 3030576, 3030911, 3031323, 3031367, 3031521, 3031770, 3032789, 3033210, 3033462, 3033486, 3033843, 3033863, 3033952, 3033955, 3034147, 3034548, 3035171, 3035397, 3035545, 3035546, 3035547, 3035689, 3035703, 3035715, 3035717, 3036005, 3036049, 3036261, 3036663, 3036739, 3036945, 3036969, 3037095, 3037517, 3037530, 3037535, 3037565, 3037587, 3037649, 3037780, 3037792, 3037859, 3158617, 3038407, 3038422, 3038600, 3038628, 3038744, 3038759, 3038829, 3039131, 3039988, 3040210, 3040300, 3040313, 3040351, 3040369, 3040400, 3040679, 3040680, 3040819, 3040822, 3040865, 3041010, 3041132, 3041314, 3041789, 3042064, 3042241, 3042272, 3042322, 3042343, 3042617, 3042874, 3042944, 3043066, 3043075, 3043218, 3043325, 3043432, 3043703, 3043716, 3043793, 3044015, 3044651, 3044847, 3044857, 3044866, 3045185, 3045905, 3045952, 3046226, 3046309, 3046316, 3047140, 3047166, 3047348, 3047574, 3047596, 3047659, 3047701, 3047756, 3047843, 3047866, 3048038, 3048039, 3048249, 3048486, 3048487, 3048655, 3048657, 3048685, 3049097, 3049136, 3049162, 3049181,
Provided but excluded from all analyses as a “false positive” (300 items):
3000862, 3003970, 3006156, 3006900, 3007762, 3008372, 3369091, 3009325, 3009520, 3010000, 3011382, 3013173, 3016344, 3019449, 3022031, 3022344, 3022369, 3023277, 3023361, 3025227, 3026662, 3030917, 3031634, 3031689, 3031751, 3031845, 3032521, 3033233, 3033959, 3034091, 3035290, 3035558, 3035794, 3036646, 3039055, 3039780,
Appendix C: List of Schools in the Dataset

A list of all 268 schools at which one of the 2,711 composition-and-rhetoric dissertations included in the analysis was completed, with the number of dissertations 2001-2010 given in parentheses.

* = Member of the Consortium of Doctoral Programs in Rhetoric and Composition

<table>
<thead>
<tr>
<th>Rank</th>
<th>School</th>
<th>Rank</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Indiana University of Pennsylvania * (124)</td>
<td>32</td>
<td>Texas Christian University * (27)</td>
</tr>
<tr>
<td>2</td>
<td>Purdue University * (92)</td>
<td>33</td>
<td>New Mexico State University * (26)</td>
</tr>
<tr>
<td>3</td>
<td>University of Arizona * (74)</td>
<td>35</td>
<td>University of Oklahoma Norman * (26)</td>
</tr>
<tr>
<td>4</td>
<td>Pennsylvania State University * (67)</td>
<td>37</td>
<td>University of Nebraska-Lincoln * (25)</td>
</tr>
<tr>
<td>5</td>
<td>University of Minnesota-Twin Cities * (62)</td>
<td>39</td>
<td>University of Rhode Island * (24)</td>
</tr>
<tr>
<td>6</td>
<td>University of Louisville * (61)</td>
<td>43</td>
<td>Regent University (23)</td>
</tr>
<tr>
<td>7</td>
<td>Arizona State University * (46)</td>
<td>44</td>
<td>Texas Christian University * (27)</td>
</tr>
<tr>
<td>8</td>
<td>The University of Texas at Austin * (46)</td>
<td>46</td>
<td>University of Nevada-Reno * (20)</td>
</tr>
<tr>
<td>9</td>
<td>Bowling Green State University * (42)</td>
<td></td>
<td>University of New Hampshire * (20)</td>
</tr>
<tr>
<td>10</td>
<td>Michigan State University * (42)</td>
<td></td>
<td>University of South Florida-Tampa * (35)</td>
</tr>
<tr>
<td>11</td>
<td>University of Illinois at Urbana-Champaign * (41)</td>
<td>16</td>
<td>Wayne State University * (28)</td>
</tr>
<tr>
<td>12</td>
<td>Iowa State University * (36)</td>
<td>32</td>
<td>Texas Christian University * (27)</td>
</tr>
<tr>
<td>13</td>
<td>Texas Woman's University * (36)</td>
<td>33</td>
<td>New Mexico State University * (26)</td>
</tr>
<tr>
<td>14</td>
<td>Michigan Technological University * (35)</td>
<td>35</td>
<td>University of Oklahoma Norman * (26)</td>
</tr>
<tr>
<td>15</td>
<td>University of South Florida-Tampa * (35)</td>
<td></td>
<td>Ohio State University * (25)</td>
</tr>
<tr>
<td>16</td>
<td>University of Washington-Seattle * (32)</td>
<td>37</td>
<td>University of Nebraska-Lincoln * (25)</td>
</tr>
<tr>
<td>17</td>
<td>University of Maryland-College Park * (31)</td>
<td>39</td>
<td>University of Rhode Island * (24)</td>
</tr>
<tr>
<td></td>
<td>University of Massachusetts Amherst * (31)</td>
<td>43</td>
<td>Regent University (23)</td>
</tr>
<tr>
<td></td>
<td>University of North Carolina at Greensboro * (31)</td>
<td>44</td>
<td>Texas Christian University * (27)</td>
</tr>
<tr>
<td></td>
<td>University of Pittsburgh-Pittsburgh * (31)</td>
<td></td>
<td>University of Nevada-Reno * (20)</td>
</tr>
<tr>
<td>21</td>
<td>Illinois State University (30)</td>
<td>46</td>
<td>University of New Hampshire * (20)</td>
</tr>
<tr>
<td></td>
<td>Indiana University-Bloomington * (30)</td>
<td></td>
<td>University of South Carolina-Columbia * (23)</td>
</tr>
<tr>
<td></td>
<td>Texas A &amp; M University * (30)</td>
<td></td>
<td>University of South Carolina-Columbia * (23)</td>
</tr>
<tr>
<td></td>
<td>University of Iowa (30)</td>
<td></td>
<td>University of South Carolina-Columbia * (23)</td>
</tr>
<tr>
<td>25</td>
<td>Georgia State University * (29)</td>
<td>57</td>
<td>University of North Carolina at Chapel Hill * (15)</td>
</tr>
<tr>
<td></td>
<td>Rensselaer Polytechnic Institute * (29)</td>
<td></td>
<td>University of Wisconsin-Milwaukee * (15)</td>
</tr>
<tr>
<td></td>
<td>Temple University (29)</td>
<td></td>
<td>University of Wisconsin-Milwaukee * (15)</td>
</tr>
<tr>
<td></td>
<td>University of Wisconsin-Madison * (29)</td>
<td></td>
<td>University of Wisconsin-Milwaukee * (15)</td>
</tr>
<tr>
<td>29</td>
<td>University of Illinois at Chicago (28)</td>
<td>59</td>
<td>Carnegie Mellon University * (14)</td>
</tr>
<tr>
<td></td>
<td>University of Kansas (28)</td>
<td></td>
<td>Oklahoma State University (14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>University of Hawaii at Manoa (14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>University of Missouri-Columbia * (14)</td>
</tr>
<tr>
<td>Rank</td>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------------------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>University of Southern Mississippi (14)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>Texas Tech University * (13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>University of New Mexico * (13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Auburn University (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catholic University of America (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duquesne University (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kent State University Kent * (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Miami University-Oxford * (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Central Florida * (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Southern California (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Toronto (Canada) (12)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>74</td>
<td>Texas A &amp; M University-Commerce * (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of Alabama * (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Colorado at Boulder (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Denver (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>The University of Texas at Arlington (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-Santa Barbara * (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Memphis (10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Ball State University * (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rutgers University-New Brunswick (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of Texas at El Paso * (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-Los Angeles (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Florida (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Louisiana at Lafayette * (9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>87</td>
<td>Fordham University (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>George Mason University (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>New Orleans Baptist Theological Seminary (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stanford University (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-Davis * (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-San Diego * (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Connecticut * (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Virginia (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>West Virginia University (8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>Capella University (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SUNY College at Buffalo (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Southern Baptist Theological Seminary (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of British Columbia (Canada) (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of Tennessee * (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Chicago (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Pennsylvania (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Walden University (7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>Claremont Graduate University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clemson University * (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Georgetown University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvard University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Idaho State University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lehigh University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Louisiana State University and Agricultural &amp; Mechanical College * (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Northern Arizona University (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Cincinnati * (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Delaware (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>114</td>
<td>Boston University (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Columbia University in the City of New York (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kansas State University (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>North Dakota State University * (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of Texas at Dallas (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Universidad de Valladolid (Spain) (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Universite de Montreal (Canada) (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-Irvine (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of North Texas (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Waterloo (Canada) (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vanderbilt University (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>125</td>
<td>Boston College (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>California Institute of Integral Studies (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Graduate Theological Union (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Loyola University Chicago (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>McGill University (Canada) (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>North Carolina State University at Raleigh * (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Northern Illinois University * (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Saint Louis University (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Simon Fraser University (Canada) (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SUNY at Binghamton (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Alberta (Canada) (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Calgary (Canada) (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of California-Riverside (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Georgia (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Idaho (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of North Carolina at Charlotte (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of North Dakota (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of San Francisco (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>York University (Canada) (4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>144</td>
<td>Brigham Young University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colorado State University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Concordia University (Canada) (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Duke University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>George Washington University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Harvard Divinity School (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Illinois Institute of Technology (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rank</td>
<td>School</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td>Middle Tennessee State University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Old Dominion University * (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Princeton Theological Seminary (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Princeton University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Stony Brook University * (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Alabama at Birmingham (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Colorado Denver (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Kentucky (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Miami (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Mississippi * (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Missouri-Kansas City (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Northern Colorado (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Oregon (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of Ottawa (Canada) (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>University of South Alabama (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Utah State University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Western Michigan University (3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>Alliant International University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Andrews University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Biola University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Brown University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Case Western Reserve University * (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Drew University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Emory University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fielding Graduate University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Florida International University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Georgia Institute of Technology (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hofstra University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Indiana State University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Liberty University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>McMaster University (Canada) (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mid-America Baptist Theological Seminary (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>National-Louis University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Our Lady of the Lake University-San Antonio (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pacifica Graduate Institute (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pepperdine University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Rice University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Saint Joseph's University (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>St. John's University-New York (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The Chinese University of Hong Kong (Hong Kong) (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The University of Auckland (New Zealand) (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Universidad de Castilla La Mancha (Spain) (2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Universidad Pontificia Comillas de Madrid (Spain) (2)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Rank</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>202</td>
<td>University of Akron (2)</td>
</tr>
<tr>
<td></td>
<td>University of Arkansas (2)</td>
</tr>
<tr>
<td></td>
<td>University of Houston (2)</td>
</tr>
<tr>
<td></td>
<td>University of La Verne (2)</td>
</tr>
<tr>
<td></td>
<td>University of Massachusetts-Lowell (2)</td>
</tr>
<tr>
<td></td>
<td>University of Rochester (2)</td>
</tr>
<tr>
<td></td>
<td>Virginia Polytechnic Institute and State University * (2)</td>
</tr>
<tr>
<td></td>
<td>Yale University (2)</td>
</tr>
<tr>
<td></td>
<td>Alabama State University (1)</td>
</tr>
<tr>
<td></td>
<td>Antioch University PhD Program in Leadership and Change (1)</td>
</tr>
<tr>
<td></td>
<td>Aquinas Institute of Theology (1)</td>
</tr>
<tr>
<td></td>
<td>Arcadia University (1)</td>
</tr>
<tr>
<td></td>
<td>Asbury Theological Seminary (1)</td>
</tr>
<tr>
<td></td>
<td>Assemblies of God Theological Seminary (1)</td>
</tr>
<tr>
<td></td>
<td>Azusa Pacific University (1)</td>
</tr>
<tr>
<td></td>
<td>Baylor University (1)</td>
</tr>
<tr>
<td></td>
<td>Brandeis University (1)</td>
</tr>
<tr>
<td></td>
<td>Central Michigan University (1)</td>
</tr>
<tr>
<td></td>
<td>Dalhousie University (Canada) (1)</td>
</tr>
<tr>
<td></td>
<td>East Carolina University * (1)</td>
</tr>
<tr>
<td></td>
<td>Florida Atlantic University (1)</td>
</tr>
<tr>
<td></td>
<td>Fuller Theological Seminary in California (1)</td>
</tr>
<tr>
<td></td>
<td>Gardner-Webb University (1)</td>
</tr>
<tr>
<td></td>
<td>George Fox University (1)</td>
</tr>
<tr>
<td></td>
<td>Georgia Southern University (1)</td>
</tr>
<tr>
<td></td>
<td>Gonzaga University (1)</td>
</tr>
<tr>
<td></td>
<td>Grambling State University (1)</td>
</tr>
<tr>
<td></td>
<td>Johnson &amp; Wales University (1)</td>
</tr>
<tr>
<td></td>
<td>La Sierra University (1)</td>
</tr>
<tr>
<td></td>
<td>Lesley University (1)</td>
</tr>
<tr>
<td></td>
<td>Louisiana Tech University (1)</td>
</tr>
<tr>
<td></td>
<td>Montana State University (1)</td>
</tr>
<tr>
<td></td>
<td>Nihon Fukushi Daigaku (Japan) (1)</td>
</tr>
<tr>
<td></td>
<td>Northcentral University (1)</td>
</tr>
<tr>
<td></td>
<td>Oakland University (1)</td>
</tr>
<tr>
<td></td>
<td>Oregon State University (1)</td>
</tr>
<tr>
<td></td>
<td>Robert Morris University (1)</td>
</tr>
<tr>
<td></td>
<td>Seton Hall University (1)</td>
</tr>
<tr>
<td></td>
<td>Southern Methodist University (1)</td>
</tr>
<tr>
<td></td>
<td>Tennessee State University (1)</td>
</tr>
<tr>
<td></td>
<td>The Chicago School of Professional Psychology at Chicago (1)</td>
</tr>
<tr>
<td></td>
<td>The University of Montana (1)</td>
</tr>
<tr>
<td></td>
<td>The University of Saskatchewan (Canada) (1)</td>
</tr>
<tr>
<td>Rank</td>
<td>School</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>The University of Texas at San Antonio</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>The University of Western Ontario</td>
</tr>
<tr>
<td></td>
<td>(Canada) (1)</td>
</tr>
<tr>
<td></td>
<td>Trinity Evangelical Divinity School (1)</td>
</tr>
<tr>
<td></td>
<td>United Theological Seminary (1)</td>
</tr>
<tr>
<td></td>
<td>Universidad de Cadiz (Spain) (1)</td>
</tr>
<tr>
<td></td>
<td>Universidad de Las Palmas de Gran Canaria</td>
</tr>
<tr>
<td></td>
<td>(Spain) (1)</td>
</tr>
<tr>
<td></td>
<td>Universidad Politecnica de Valencia</td>
</tr>
<tr>
<td></td>
<td>(Spain) (1)</td>
</tr>
<tr>
<td></td>
<td>Universite de Sherbrooke (Canada) (1)</td>
</tr>
<tr>
<td></td>
<td>Universite du Quebec a Montreal (Canada)</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>Universite Laval (Canada) (1)</td>
</tr>
<tr>
<td></td>
<td>University of Alaska Fairbanks (1)</td>
</tr>
<tr>
<td></td>
<td>University of Dallas (1)</td>
</tr>
<tr>
<td></td>
<td>University of Manitoba (Canada) (1)</td>
</tr>
<tr>
<td></td>
<td>University of Maryland-Baltimore County (1)</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>268</td>
<td>University of Massachusetts-Boston (1)</td>
</tr>
<tr>
<td></td>
<td>University of Nebraska at Omaha (1)</td>
</tr>
<tr>
<td></td>
<td>University of Nevada-Las Vegas (1)</td>
</tr>
<tr>
<td></td>
<td>University of New Brunswick (Canada)</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td></td>
<td>University of Northern Iowa (1)</td>
</tr>
<tr>
<td></td>
<td>University of Phoenix-Online (1)</td>
</tr>
<tr>
<td></td>
<td>University of Puerto Rico-Rio Piedras (1)</td>
</tr>
<tr>
<td></td>
<td>University of San Diego (1)</td>
</tr>
<tr>
<td></td>
<td>University of South Dakota (1)</td>
</tr>
<tr>
<td></td>
<td>University of the Incarnate Word (1)</td>
</tr>
<tr>
<td></td>
<td>University of the Pacific (1)</td>
</tr>
<tr>
<td></td>
<td>University of Toledo (1)</td>
</tr>
<tr>
<td></td>
<td>University of West Georgia (1)</td>
</tr>
<tr>
<td></td>
<td>Virginia Commonwealth University (1)</td>
</tr>
<tr>
<td></td>
<td>Westminster Theological Seminary (1)</td>
</tr>
<tr>
<td></td>
<td>Widener University (1)</td>
</tr>
<tr>
<td></td>
<td>Wilmington College (1)</td>
</tr>
<tr>
<td></td>
<td>Wilmington University (1)</td>
</tr>
</tbody>
</table>
Appendix D:

Schools of Interest for each Method Tag

What we see depends in large part on the questions we ask. This appendix presents two alternate views of the ‘top’ institutions for each method in this study. I hope the findings will be of interest to compositionists curious about how their own vantage points are situated within the field’s larger landscape – including, but not limited to, students applying to PhD programs in Composition and Rhetoric; those tracking the job market, whether as applicants or as current faculty; and scholars invested in the ways that research agendas shape and are shaped by graduate education as a map-maker.

At the left, schools for which a dissertation drawn at random, 2001-2010, is most likely to have used the method in question; I call this “methodological focus.” At the right, schools for which the greatest number of dissertations used the method, 2001-2010, regardless of how common it was within the cohort; I call this “methodological output.” These two approaches give variant measures of the “top five” schools which demonstrate the potential dominance of a small number of schools with very high numbers of graduates, and which put the high output of other schools in context. The top 6 schools by total number of dissertations are marked by italics.

\[ ^\text{^} = \text{school is in the top five by number (output), but not by percentage (focus)} \]
\[ \text{v} = \text{school is in the top five by percentage (focus), but not by number (output)} \]
\[ \text{*} = \text{member of the Consortium of Doctoral Programs in Rhetoric and Composition.} \]
\[ P = \text{percentage} \]
\[ D = \text{dissertation count} \]
\[ T = \text{total dissertation count} \]

To be considered for inclusion, each school must have produced at least five composition/rhetoric dissertations in the last five years of the dataset (2006-2010).

<table>
<thead>
<tr>
<th>Methodological Focus, Top Five Schools</th>
<th>Methodological Output, Top Five Schools</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ranked by Percentage of Dissertations</td>
<td>Ranked by Number of Dissertations</td>
</tr>
<tr>
<td>Clinical / Case Study</td>
<td>Clinical / Case Study</td>
</tr>
<tr>
<td>P D T</td>
<td>P D T</td>
</tr>
<tr>
<td>v Teachers College at Columbia University</td>
<td>47 9 19</td>
</tr>
<tr>
<td>New Mexico State University *</td>
<td>46 12 26</td>
</tr>
<tr>
<td>University of Massachusetts Amherst *</td>
<td>42 13 31</td>
</tr>
<tr>
<td>Indiana University of Pennsylvania *</td>
<td>40 50 124</td>
</tr>
<tr>
<td>^ Indiana University of Pennsylvania *</td>
<td>40 50 124</td>
</tr>
<tr>
<td>^ University of Louisville *</td>
<td>31 19 61</td>
</tr>
<tr>
<td>Purdue University *</td>
<td>18 17 92</td>
</tr>
<tr>
<td>University of Massachusetts Amherst *</td>
<td>42 13 31</td>
</tr>
<tr>
<td>University of Hawaii at Manoa</td>
<td>36 5 14</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------</td>
</tr>
<tr>
<td><strong>Critical / Hermeneutical</strong></td>
<td>P  D  T</td>
</tr>
<tr>
<td>v The University of Texas at El Paso *</td>
<td>67 6 9</td>
</tr>
<tr>
<td>v Northwestern University</td>
<td>62 10 16</td>
</tr>
<tr>
<td>v Catholic University of America</td>
<td>58 7 12</td>
</tr>
<tr>
<td>Texas Woman's University *</td>
<td>56 20 36</td>
</tr>
<tr>
<td>v University of Utah *</td>
<td>50 8 16</td>
</tr>
<tr>
<td><strong>Discourse or Text Analytical</strong></td>
<td>P  D  T</td>
</tr>
<tr>
<td>v Walden University</td>
<td>57 4 7</td>
</tr>
<tr>
<td>v Howard University</td>
<td>38 6 16</td>
</tr>
<tr>
<td>v University of Colorado at Boulder</td>
<td>36 4 11</td>
</tr>
<tr>
<td>v Carnegie Mellon University *</td>
<td>36 5 14</td>
</tr>
<tr>
<td>University of Massachusetts Amherst *</td>
<td>35 11 31</td>
</tr>
<tr>
<td><strong>Ethnographic</strong></td>
<td>P  D  T</td>
</tr>
<tr>
<td>v University of Illinois at Chicago</td>
<td>39 11 28</td>
</tr>
<tr>
<td>v University of California-Berkeley</td>
<td>38 6 16</td>
</tr>
<tr>
<td>v Kent State University *</td>
<td>33 4 12</td>
</tr>
<tr>
<td>Indiana University of Pennsylvania *</td>
<td>31 39 124</td>
</tr>
<tr>
<td>v New Mexico State University *</td>
<td>31 8 26</td>
</tr>
<tr>
<td><strong>Experimental / Quasi-Experimental</strong></td>
<td>P  D  T</td>
</tr>
<tr>
<td>Fordham University</td>
<td>75 6 8</td>
</tr>
<tr>
<td>v Walden University</td>
<td>57 4 7</td>
</tr>
<tr>
<td>v Capella University</td>
<td>43 3 7</td>
</tr>
<tr>
<td>v University of Connecticut *</td>
<td>38 3 8</td>
</tr>
<tr>
<td>v Auburn University</td>
<td>33 4 12</td>
</tr>
<tr>
<td><strong>Historical / Archival</strong></td>
<td>P  D  T</td>
</tr>
<tr>
<td>v Catholic University of America</td>
<td>67 8 12</td>
</tr>
<tr>
<td>v Northwestern University</td>
<td>56 9 16</td>
</tr>
<tr>
<td>v University of Pittsburgh *</td>
<td>52 16 31</td>
</tr>
<tr>
<td>v University of Utah *</td>
<td>50 8 16</td>
</tr>
<tr>
<td>v Texas Christian University *</td>
<td>48 13 27</td>
</tr>
</tbody>
</table>
### Interview / Focus Group

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>v Walden University</td>
<td>43</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>v Oklahoma State University</td>
<td>29</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>v University of Hawaii at Manoa</td>
<td>29</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>v George Mason University</td>
<td>25</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Indiana University of Pennsylvania *</td>
<td>20</td>
<td>25</td>
<td>124</td>
</tr>
</tbody>
</table>

### Meta-Analytical / Disciplinographic

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>v University of Wisconsin-Milwaukee *</td>
<td>27</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>v SUNY at Albany *</td>
<td>25</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>v University of Connecticut *</td>
<td>25</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>v Texas A &amp; M University-Commerce *</td>
<td>18</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>v University of Central Florida *</td>
<td>17</td>
<td>2</td>
<td>12</td>
</tr>
</tbody>
</table>

### Model-Building

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>v University of Virginia</td>
<td>38</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>v Kent State University *</td>
<td>33</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>v University of Califórnia-Santa Barbara *</td>
<td>30</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>v Syracuse University *</td>
<td>26</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>v University of Nevada-Reno *</td>
<td>25</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

### Philosophical / Theoretical

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>v SUNY at Albany *</td>
<td>69</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>University of North Carolina at Greensboro *</td>
<td>65</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td>v Duquesne University</td>
<td>58</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>v Syracuse University *</td>
<td>57</td>
<td>13</td>
<td>23</td>
</tr>
<tr>
<td>Michigan Technological University *</td>
<td>54</td>
<td>19</td>
<td>35</td>
</tr>
</tbody>
</table>

### Poetic / Fictive / Craft-Based

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union Institute &amp; University</td>
<td>44</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>42</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>SUNY at Albany *</td>
<td>38</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>v Clemson University *</td>
<td>33</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>v University of Denver</td>
<td>27</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>Indiana University of Pennsylvania *</td>
<td>20</td>
<td>25</td>
<td>124</td>
</tr>
<tr>
<td>^ Purdue University *</td>
<td>12</td>
<td>11</td>
<td>92</td>
</tr>
<tr>
<td>^ Arizona State University *</td>
<td>17</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>^ Michigan State University *</td>
<td>17</td>
<td>7</td>
<td>42</td>
</tr>
<tr>
<td>^ University of Louisville *</td>
<td>10</td>
<td>6</td>
<td>61</td>
</tr>
</tbody>
</table>

### Meta-Analytical / Disciplinographic

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>^ Indiana University of Pennsylvania *</td>
<td>5</td>
<td>6</td>
<td>124</td>
</tr>
<tr>
<td>^ University of Louisville *</td>
<td>8</td>
<td>5</td>
<td>61</td>
</tr>
<tr>
<td>Arizona State University *</td>
<td>9</td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>Michigan State University *</td>
<td>10</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>^ Purdue University *</td>
<td>4</td>
<td>4</td>
<td>92</td>
</tr>
</tbody>
</table>

### Model-Building

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>^ Indiana University of Pennsylvania *</td>
<td>10</td>
<td>13</td>
<td>124</td>
</tr>
<tr>
<td>^ Purdue University *</td>
<td>11</td>
<td>10</td>
<td>92</td>
</tr>
<tr>
<td>^ Michigan State University*</td>
<td>19</td>
<td>8</td>
<td>42</td>
</tr>
<tr>
<td>^ Arizona State University*</td>
<td>13</td>
<td>6</td>
<td>46</td>
</tr>
<tr>
<td>^ Rensselaer Polytechnic Institute*</td>
<td>21</td>
<td>6</td>
<td>29</td>
</tr>
</tbody>
</table>

### Philosophical / Theoretical

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>^ University of Arizona *</td>
<td>46</td>
<td>34</td>
<td>74</td>
</tr>
<tr>
<td>^ Purdue University *</td>
<td>34</td>
<td>31</td>
<td>92</td>
</tr>
<tr>
<td>^ Pennsylvania State University *</td>
<td>33</td>
<td>22</td>
<td>67</td>
</tr>
<tr>
<td>University of North Carolina at Greensboro *</td>
<td>65</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td>Michigan Technological University *</td>
<td>54</td>
<td>19</td>
<td>35</td>
</tr>
</tbody>
</table>

### Poetic / Fictive / Craft-Based

<table>
<thead>
<tr>
<th>University</th>
<th>P</th>
<th>D</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union Institute &amp; University</td>
<td>44</td>
<td>8</td>
<td>18</td>
</tr>
<tr>
<td>^ Indiana University of Pennsylvania *</td>
<td>5</td>
<td>6</td>
<td>124</td>
</tr>
<tr>
<td>^ Michigan State University *</td>
<td>17</td>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>SUNY at Albany *</td>
<td>38</td>
<td>6</td>
<td>16</td>
</tr>
<tr>
<td>University of Toronto</td>
<td>42</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>Practitioner / Teacher-Research</td>
<td>P</td>
<td>D</td>
<td>T</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>v Capella University</td>
<td>71</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>v George Mason University</td>
<td>62</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>v Walden University</td>
<td>57</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>v University of Virginia</td>
<td>50</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Illinois State University</td>
<td>47</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>Rhetorical Analytical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v University of Memphis</td>
<td>60</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>v Catholic University of America</td>
<td>50</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Texas Woman's University *</td>
<td>50</td>
<td>18</td>
<td>36</td>
</tr>
<tr>
<td>v The University of Texas at El Paso *</td>
<td>44</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>v Regent University</td>
<td>43</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>Survey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v Texas A &amp; M University-Commerce *</td>
<td>27</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td>v University of Central Florida *</td>
<td>25</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>University of Massachusetts Amherst *</td>
<td>19</td>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>v The University of Alabama *</td>
<td>18</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>v Southern Illinois University Carbondale *</td>
<td>17</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>Rhetorical Analytical</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v University of Pennsylvania *</td>
<td>21</td>
<td>26</td>
<td>124</td>
</tr>
<tr>
<td>Illinois State University</td>
<td>47</td>
<td>14</td>
<td>30</td>
</tr>
<tr>
<td>v University of Arizona *</td>
<td>14</td>
<td>10</td>
<td>74</td>
</tr>
<tr>
<td>v University of Massachusetts Amherst *</td>
<td>29</td>
<td>9</td>
<td>31</td>
</tr>
<tr>
<td>v Georgia State University *</td>
<td>28</td>
<td>8</td>
<td>29</td>
</tr>
<tr>
<td>Survey</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>v Indiana University of Pennsylvania *</td>
<td>10</td>
<td>13</td>
<td>124</td>
</tr>
<tr>
<td>v University of Louisville *</td>
<td>13</td>
<td>8</td>
<td>61</td>
</tr>
<tr>
<td>v Purdue University *</td>
<td>8</td>
<td>7</td>
<td>92</td>
</tr>
<tr>
<td>University of Massachusetts Amherst *</td>
<td>19</td>
<td>6</td>
<td>31</td>
</tr>
<tr>
<td>v Arizona State University *</td>
<td>11</td>
<td>5</td>
<td>46</td>
</tr>
</tbody>
</table>
## Appendix E:
### Complete List of Topic Keys, Assigned Labels, and Ranks

The topics below were generated by the model described in Chapter 4. Numbers associated with the topics are arbitrary; rather than sort by these numbers, therefore, the topics below are arranged in descending order of weight across the entire corpus.

Seven of the 55 topics are labeled with an asterisk (*); these were deemed non-content-bearing and removed from visualizations. An additional six topics are labeled with a double-asterisk (**); these are small topics (< 0.73% of the corpus) that seem to combine several even smaller topics, perhaps because a handful of dissertations using them together carried more weight than they would have in a topic represented by a larger sample. These topics may well have been split in a model with a greater number of topics, but possibly at the cost of some coherence in the larger topics.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Topic Number and Assigned Label</th>
<th>Top Words</th>
<th>% of Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32: Students in the Classroom</td>
<td>students writing student class teacher classroom teachers paper instructor research study instructors semester college assignment classes write teaching learning</td>
<td>5.39%</td>
</tr>
<tr>
<td>2</td>
<td>8: (Critical) Pedagogical Theory</td>
<td>students composition teaching pedagogy classroom teachers critical work student teacher theory studies knowledge learning ways education academic pedagogical practice</td>
<td>4.77%</td>
</tr>
<tr>
<td>3</td>
<td>48: Philosophy of Language</td>
<td>language theory discourse meaning knowledge system fact point power metaphor question view speech human case social model problem sense</td>
<td>4.33%</td>
</tr>
<tr>
<td>4</td>
<td>15: Identity Construction</td>
<td>identity social discourse cultural ways culture power space discourses people practices community identities understanding language personal place construction difference</td>
<td>4.21%</td>
</tr>
<tr>
<td>5</td>
<td>10: Story and Narrative</td>
<td>life back story time day man people mother don love home good young father stories family long left person</td>
<td>4.13%</td>
</tr>
<tr>
<td>6</td>
<td>1: Process Reflections</td>
<td>don participants people study time interview research things experience work didn make lot interviews experiences questions feel ve kind</td>
<td>3.84%</td>
</tr>
<tr>
<td>7</td>
<td>35: Community Engagement and Collaboration</td>
<td>community research learning project process service work group members development study professional knowledge organization practice team information communities activities</td>
<td>3.66%</td>
</tr>
<tr>
<td>Rank</td>
<td>Topic Number and Assigned Label</td>
<td>Top Words</td>
<td>% of Corpus</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------</td>
</tr>
<tr>
<td>8</td>
<td>17: Capitalism, Marxism, and Activism</td>
<td>public political social economic movement rhetoric society politics power cultural labor university state democracy change action democratic rhetorical class</td>
<td>3.56%</td>
</tr>
<tr>
<td>9</td>
<td>55: Comprehension and Usability</td>
<td>text analysis texts information discourse readers chapter study reader rhetorical content audience context specific reading data based features types</td>
<td>3.52%</td>
</tr>
<tr>
<td>10</td>
<td>12: Workplace and Organizational Histories</td>
<td>work time letter working place years letters business people make part job made money workers personal fact long article</td>
<td>3.36%</td>
</tr>
<tr>
<td>11</td>
<td>14: Poetics and Semiotics</td>
<td>world experience memory life human art poetry work time process meaning voice language nature metaphors words mind form sense</td>
<td>3.2%</td>
</tr>
<tr>
<td>12</td>
<td>25: History of Composition</td>
<td>writing composition writers writer write english essay creative process written language reading work personal essays instruction basic style college</td>
<td>3.07%</td>
</tr>
<tr>
<td>13</td>
<td>31: 18th and 19th Centuries</td>
<td>history century american early rhetoric historical nineteenth society life york moral university great education public press literature men america</td>
<td>3.04%</td>
</tr>
<tr>
<td>14</td>
<td>39: Writing Process: Formal and Cognitive Studies</td>
<td>writing language english students study writers feedback esl process learning comments reading revision research grammar knowledge sentence learners write</td>
<td>2.89%</td>
</tr>
<tr>
<td>15</td>
<td>21: Close-Reading Classical Rhetoric</td>
<td>rhetoric rhetorical audience aristotle ethos speech argument classical theory plato communication persuasion invention speaking rhetor arguments good truth philosophy</td>
<td>2.8%</td>
</tr>
<tr>
<td>16</td>
<td>28: Political Rhetoric, Mostly of the US</td>
<td>war president american public states united bush national military america speech political nation government people policy press carter york</td>
<td>2.7%</td>
</tr>
<tr>
<td>17</td>
<td>6: Institutional Context of Writing Instruction</td>
<td>faculty writing english university program college composition teaching courses year graduate programs academic department time students education professional research</td>
<td>2.49%</td>
</tr>
<tr>
<td>18</td>
<td>44: Rhetorical Affordances of the Web</td>
<td>technology web computer online digital technologies computers site information internet media design users page technological access sites software http</td>
<td>2.35%</td>
</tr>
<tr>
<td>19</td>
<td>41: Quantifying Written Products</td>
<td>test study table results scores group research assessment significant score data participants scale social differences total groups message number</td>
<td>2.08%</td>
</tr>
<tr>
<td>20</td>
<td>4: Bad OCR*</td>
<td>en ith ent im em er te ro hat ed om riting ts fo se es le ay stu</td>
<td>1.87%</td>
</tr>
<tr>
<td>Rank</td>
<td>Topic Number and Assigned Label</td>
<td>Top Words</td>
<td>% of Corpus</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>21</td>
<td>45: Preparation For College</td>
<td>education school students college teachers educational schools learning higher high student instruction teaching teacher skills colleges educators university state</td>
<td>1.81%</td>
</tr>
<tr>
<td>22</td>
<td>9: Literacy and Literacies</td>
<td>literacy school reading writing family children community life social practices read stories home study literate literacies research language lives</td>
<td>1.77%</td>
</tr>
<tr>
<td>23</td>
<td>43: Genre and Discipline</td>
<td>research genre genres study knowledge field writing discourse studies academic rhetorical disciplinary activity analysis social practices professional work discipline</td>
<td>1.75%</td>
</tr>
<tr>
<td>24</td>
<td>11: Women Acting Rhetorically</td>
<td>women woman gender feminist men female male sex sexual feminine feminism public work gay rhetorical issues masculine girls domestic</td>
<td>1.74%</td>
</tr>
<tr>
<td>25</td>
<td>23: Narrative Theory plus Fandom plus Readers as Writers</td>
<td>narrative story reading reader readers text literature fiction narratives stories literary book character characters audience read narrator texts author</td>
<td>1.58%</td>
</tr>
<tr>
<td>26</td>
<td>18: Invention plus Postmodernist Theories of Meaning</td>
<td>work subject plagiarism body foucault sense ethics heidegger question place subjectivity postmodern text ethical desire chapter derrida time relation</td>
<td>1.43%</td>
</tr>
<tr>
<td>27</td>
<td>7: Race and White-Black Power Struggles</td>
<td>black white african american race racial south americans people racism blacks rights whites negro slavery whiteness power community civil</td>
<td>1.34%</td>
</tr>
<tr>
<td>28</td>
<td>3: Multilingualism and World Englishes plus</td>
<td>english language cultural american culture linguistic languages al native spanish speakers arabic international university arab world immigrants esl speaking</td>
<td>1.31%</td>
</tr>
<tr>
<td>29</td>
<td>49: Technical Communication Mostly</td>
<td>technical communication design business information company user system users documents participants workplace management engineering research task technology product work</td>
<td>1.29%</td>
</tr>
<tr>
<td>30</td>
<td>27: Visual Rhetoric</td>
<td>visual images image figure verbal art space body meaning photographs representations pictures picture representation objects multimodal photograph elements media</td>
<td>1.24%</td>
</tr>
<tr>
<td>31</td>
<td>36: Political Discourse</td>
<td>media news campaign political public blog obama television people issues communication coverage http analysis internet blogs post audience clinton</td>
<td>1.23%</td>
</tr>
<tr>
<td>32</td>
<td>20: Court Decisions and Ramifications</td>
<td>law legal court state justice case act rights political states supreme decision cases constitution argument trial arguments debate laws</td>
<td>1.09%</td>
</tr>
<tr>
<td>Rank</td>
<td>Topic Number and Assigned Label</td>
<td>Top Words</td>
<td>% of Corpus</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>33</td>
<td>47: Probably ProQuest Boilerplate*</td>
<td>permission copyright reproduction owner prohibited reproduced study practices social fo language oral states press group journal figure activities history</td>
<td>1.03%</td>
</tr>
<tr>
<td>34</td>
<td>38: Health Medicine and Disease</td>
<td>health medical care medicine patient disease patients illness autism depression aids clinical mental people treatment body physicians doctors nursing</td>
<td>0.99%</td>
</tr>
<tr>
<td>35</td>
<td>40: Online Learning and Collaboration</td>
<td>online peer group students discussion face response interaction collaborative participants social communication computer learning peers groups discussions collaboration feedback</td>
<td>0.98%</td>
</tr>
<tr>
<td>36</td>
<td>30: Writing Center Tutorials</td>
<td>writing center tutors tutor student tutoring centers paper session questions tutorial conference writer sessions writers peer work training owl</td>
<td>0.98%</td>
</tr>
<tr>
<td>37</td>
<td>53: Writing With and About Religious Texts</td>
<td>god church religious christian bible faith religion spiritual christ prayer jesus catholic life biblical pastor community congregation sermon sacred</td>
<td>0.93%</td>
</tr>
<tr>
<td>38</td>
<td>16: Performance</td>
<td>performance music humor play culture hip sports hop sport popular audience song football baseball york rap team black press</td>
<td>0.91%</td>
</tr>
<tr>
<td>39</td>
<td>19: Science Writing and Environmentalism</td>
<td>science scientific environmental nature human natural scientists ecological wilderness environment world discourse public population ecology social conservation species management</td>
<td>0.9%</td>
</tr>
<tr>
<td>40</td>
<td>26: Reading Burke</td>
<td>burke terms motives symbolic identification term action act form order language human sense nature scene social arendt principle pp</td>
<td>0.8%</td>
</tr>
<tr>
<td>41</td>
<td>29: Silence plus England**</td>
<td>english england de letter century letters chomsky early sublime hyperbole style renaissance london elizabeth eighteenth texts scrapbooks language scrapbook</td>
<td>0.72%</td>
</tr>
<tr>
<td>42</td>
<td>46: Games plus Commercial Editing Practices**</td>
<td>game review book games author editing authors reviews authorship video editors articles journals editor publishing books published article reviewers</td>
<td>0.69%</td>
</tr>
<tr>
<td>43</td>
<td>52: International Conflict and Negotiation</td>
<td>peace conflict korea korean south colonial people political trade power israeli india lyceum national witch world witches north international</td>
<td>0.62%</td>
</tr>
<tr>
<td>44</td>
<td>37: Native American History and Rhetoric</td>
<td>indian american native indians land cherokee people indigenous white government cultural nation culture sovereignty history americans tribal states rhetorical</td>
<td>0.61%</td>
</tr>
<tr>
<td>Rank</td>
<td>Topic Number and Assigned Label</td>
<td>Top Words</td>
<td>% of Corpus</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------</td>
<td>-----------</td>
<td>-------------</td>
</tr>
<tr>
<td>45</td>
<td>33: Film Criticism plus Japan**</td>
<td>film films japanese documentary media video japan cinema audience popular scene movie cultural camera festival camp man euro foreign</td>
<td>0.57%</td>
</tr>
<tr>
<td>46</td>
<td>34: Museums and Archives, mostly**</td>
<td>university state museum texas writing jewish board kairos editorial archives archive history jews visitors composition holocaust maps exhibit map</td>
<td>0.54%</td>
</tr>
<tr>
<td>47</td>
<td>42: Medicine, Disability, the Body and Identity</td>
<td>disability violence children birth baby deaf child control sanger body mother women disabilities bodies disabled rhetoric welfare mothers people</td>
<td>0.53%</td>
</tr>
<tr>
<td>48</td>
<td>22: Bad OCR and Foreign Characters*</td>
<td>die munoz language ia ms tt la ti oral mopan farm na ce io ni ma jibaro si li</td>
<td>0.52%</td>
</tr>
<tr>
<td>49</td>
<td>2: Bad OCR*</td>
<td>om ae fiom firom fi hom leam die grst tae tke au based aw aid fo hke knowledge wiu</td>
<td>0.5%</td>
</tr>
<tr>
<td>50</td>
<td>54: Civic Discourse plus China and Japan**</td>
<td>chinese china apology discourse mao vico garcia zedong hong humphrey kong cultural western utterance council people taiwanese mr political</td>
<td>0.49%</td>
</tr>
<tr>
<td>51</td>
<td>50: Spanish Language*</td>
<td>de la el en los se del las por mexican es con una lo su para mexico como al</td>
<td>0.44%</td>
</tr>
<tr>
<td>52</td>
<td>51: Military-Industrial Complex**</td>
<td>public risk space nuclear report safety information relevance doe welsh missile image environmental wales topic soviet plant fernald controversy</td>
<td>0.44%</td>
</tr>
<tr>
<td>53</td>
<td>24: Probably ProQuest Boilerplate and Bad OCR*</td>
<td>ission perm ep prohibited reproduction erm copyright ow ner ithout ced ro time eproduced period placement textual criteria nike</td>
<td>0.39%</td>
</tr>
<tr>
<td>54</td>
<td>5: Food and Cooking</td>
<td>food fat cooking quilt janette organic recipe recipes lu guan foods outlaw meat ms eating people quilting weight op</td>
<td>0.32%</td>
</tr>
<tr>
<td>55</td>
<td>13: Italian and Latin Language*</td>
<td>di cicero la de il che lucretius medieval calvino del si isocrates le una roman ovid ad orator da</td>
<td>0.27%</td>
</tr>
</tbody>
</table>
Appendix F:
R Scripts Used for Data Analysis and to Generate Figures

The code provided here was used in generating the figures, tables, and statistics included throughout the dissertation. Individual code files are separated by a full line of pound symbols (#), followed by the filename. The most up-to-date versions of these files will be maintained online at http://github.com/benmiller314/Dissertation-Research.

All files were programmed in the R Language and Environment for Statistical Computing, version 3.1.0 (2014-04-10), nicknamed "Spring Dance," using the R.app GUI 1.64 on the x86_64-apple-darwin10.8.0 (64-bit) platform. Please note that in this environment, the pound symbol (#) sets off the rest of the line as a comment.

The first file included in this appendix, “rerun all analyses.R,” provides an overview of the programs that follow; subsequent files are listed in alphabetical order by filename.

# re-run all analyses.R
#
# This file is a run-down of the programs that could re-create my figures
# from prepared data. For initial data cleaning, I used GoogleRefine (now
# OpenRefine); see files in Appendix A.
#####

#### Preparing the Working Environment ####

# Set working directory to the location of R script files.
setwd("/Users/benmiller314/Dropbox/coursework,
etc/dissertation/data, code, and figures/Dissertation Research")

###
## Global variables called in many functions. remake_figs: If TRUE,
## save new files for figures; if FALSE, display on screen only.
## autorun: If TRUE, call the functions when files are sourced; if
## FALSE, load functions but do not call.
##     remake_figs <- FALSE; autorun <- FALSE;

###
## `dataprep.R`: prepares working environment by loading helper
## functions and setting key variables (such as tagset).
##
## `dataprep 2 - load data.R`: loads in a .csv file of tagged
## spreadsheet data, generates a tag array, and defines various
# subsets. You will be prompted to select the file via file.choose().
# Dependencies: "extract subjects.R", "Factor-Bug fixing.R",
# "thresh.R", "simplifying the schema.R", "check count.R",
# library(data.table)
# NB: These two functions can be called jointly via `source(file="start
# here.R")`
source(file="dataprep.R")
source(file="dataprep 2 - load data.R")


##### Functions for Determining the Scope of the Data ######

# `schools per year.R`: for each year, finds the number of
# institutions and number of dissertations. Optionally plots these
# numbers as a line graph
source(file="schools per year.R")

# `map by school 4 (comp-rhet superimposed on carnegie
# schools).R`: Produces a geographical map of three kinds of data
# points: schools with a Carnegie Classification of doctoral
# institution; schools with programs in the Consortium of Doctoral
# Programs in Rhetoric and Composition; and schools where one of
# the 2,711 dissertations in my dataset were completed.
# Dependencies: package(maps), package(mapdata),
# package(mapproplots), package(maptools), package(scales), map by
source(file="map by school 4 (comp-rhet superimposed on carnegie schools).R")


##### Programs for Analyzing Dissertation Methods ######

# `tags by school.R`: generates heat plots of methods used in
dissertations, aggregated by school. Provides two functions:
# * schoolwise.data(dataset_name, tagset_name): returns a list of
tag means, sums, and counts, each aggregated by school.
# * schoolwise(dataset_name, tagset_name, ...): make one or more
# heatplots from the output of schoolwise.data(). Dependencies:
# library(doBy), library(cluster), library(RColorBrewer)
source(file="tags by school.R")

# `methodcount barplot.R`: produces a bar plot of method-tag
counts per dissertation, for a given method tagset.
source(file="methodcount barplot.R")

##
`subject terms barplot.R`: produces a bar plot of author-provided subject terms counts, by overall frequency. Provides one function:

* `subject.barplot(dataset_name, how.many, ...)`: graphs the top how.many

`keyword barplot.R`: produces a bar plot of author-provided keyword-tag counts, by overall frequency. Median frequency turns out to be 1, making this figure visually not so different from empty axes.

source(file="subject terms barplot.R")
source(file="keyword barplot.R")

##
```
`frequency of method tags.R`: tabulates and plots the number of times a dissertation is tagged with each method. Provides three functions:

* `get_tags(dataset_name, tagset_name)`: returns a named vector of frequencies for each method in the tagset

* `methodfreq_combined(bigset, smallset, diffset)`: plots an overlaid horizontal bar graph of method frequencies; by default the three sets are noexcludes, consorts, and nonconsorts, respectively (but others are possible).

source(file="frequency of method tags.R")
```

##
```
`compare method ranks.R`: creates a side-by-side plot of methods in descending rank order, with lines connecting the same methods to quickly reveal changes in rank across the two sets.

Provides one function:

* `compare_method_ranks(set1, set2, pcts, ...)`

source(file="compare method ranks.R")
```

##
```
`top schools by method.R`: For each method in a given tagset, produces a list of the top X schools by either methodological output (number of dissertations using that method at that school) or methodological focus (percentage of dissertations using that method at that school). Provides one function:

* `toplists(dataset_name, tagset_name, howmany, threshold, ...)`

source(file="top schools by method.R")
```

##
```
`collocation heatmap.R`: If a dissertation is tagged X, how many times is it also tagged Y? Provides one function:

* `sumbytags(dataset_name, tagset_name, doplot, normed, dendro)`: Aggregates methods tags by each method tag, with an option to norm by dividing the sums by the aggregating method's total count. Optionally plots a heatmap of results as an adjacency matrix.

Dependencies: `heatmap_ben.R`

source(file="collocation heatmap.R")
```
##### Functions for Topic Modeling

```r
# generate a (series of) topic model(s)
source(file="r2mallet with foreach.R")
```

```r
## `top docs per topic.R`: browse topics to generate labels. Provides three functions:
## * get.doc.composition(dataset, ntopics): retrieves a pre-existing matrix, output by MALLET, with topic proportions for each document in corpus
## * get.topics4doc(pubnum, dataset_name, ntopics, howmany, showlabels): retrieves top `howmany` topics for a document specified by `pubnum`.
## * top_topic_browser(...): for a specified topic or range of topics, shows the top `howmany` documents and their method tags, with optional detail view showing top topics for each document at a time.
source(file="top docs per topic.R")
```

```r
## retrieve topic information about a dissertation by author name
source(file="get topics for author.R")
```

```r
## find topics that co-occur within documents
source(file="cotopics.R")
```

```r
## get weights of every topic for all documents
source(file="get doctopic grid.R")
```

```r
## find dissertations with high levels of a cluster of topics
source(file="topic cluster reach.R")
```

```r
## `frameToD3.R`: outputs JSON file of topic model data for interactive visualizations. Provides two functions:
## * frameToJSON(dataset_name, ntopics, do.plot, groupVars, dataVars, outfile, bad.topics): given a topic model as generated by 'r2mallet with foreach.R', returns a hierarchical clustering of topics in JSON. For each topic, includes the following metadata: name, size, scaledsize, topwords, topic, rank.
## * cotopic_edges(dataset_name, ntopics, level, min, outfile, bad.topics): given a topic model as generated by 'r2mallet with foreach.R', returns weighted edges between topics and the same hierarchical clustering as above.
source(file="frameToD3.R")
```

```r
## `topics by year.R`: rank topics overall, aggregated per year.
```
Provides two functions:

* `topics.by.year(dataset_name, ntopics, to.plot, do.plot, per.plot)`: charts the rising and falling contributions to the corpus of each topic, or topics specified in `to.plot`, over time.
  Invisibly returns a data frame of these contributions (as `df`) and a list of topics by descending order of total contribution (as `rank.order`).

* `topic.variation(dataset_name, ntopics, to.plot)`: creates a barplot of yearly variation of topics.


source(file="topics by year.R")

```
ivrds of topic proportions.R`: Find out the curve of topic strengths within each document, i.e. how much of the document is the top topic? how much is the second? and so on. Provides one function:

* `topic.proportions(dataset_name, ntopics, bad.topics, use.notch, explore.outliers)`: produces a box plot of contribution (y-axis) sorted by topic rank (x-axis), aggregated over all documents. If `explore.outliers` is true, prints a table of upper outlier values, the topics generating them, and their labels, then starts a browser for dissertations represented in that table. Returns box plot statistics for the top three topics.

Dependencies:

```
ivrds of single topic strength vs rank.R`: are overall top topics high-ranked in few documents, or evenly spread out? Provides one function:

* `strength_v_rank(my.topic, dataset_name, ntopics, bad.topics)`: produces a scatter plot of one selected topic's contributions, with percent of dissertation on the y-axis and rank within dissertation on the x-axis.

Dependencies:

```

carnegie 1 (setup).R
# Read in and parse the Carnegie Classification data.
# Called by `map by school 4`.
####
carnegie.all <- read.csv(file=paste0(dataloc, 
"cc2010_classification_data_datasheet_06.03.2013.csv"))
attach(carnegie.all)

# identify and subset out schools in my dataset
# all schools with at least one doctoral program
cdoc2010 <- carnegie.all[which(carnegie.all$IPGRAD2010 > 11),]
cdoc2005 <- carnegie.all[which(CCIPGRAD > 11),]
cdoc2000 <- carnegie.all[which(CC2000 %in% c(15,16)),]

detach(carnegie.all)
if (!exists("cdoc2010.geo")) {
  action <- readline("Geocoding data is missing for Carnegie
  Classification doctoral schools. To load a
  pre-created file, press L; to geocode now, press G.")

  if (tolower(action) == "l") {
    invisible(readline("Select the geocoding csv file from geocode.R.
    (Filename is like 'geocoding by school, cdoc2010,
    N449.csv'; press <Enter> when ready.")")

    cdoc2010.geo <<- read.csv(file=file.choose())

    # trim the first column, which is just the row number
    # added when the file is saved
    cdoc2010.geo <<- data.frame(cdoc2010.geo[,2:ncol(cdoc2010.geo)])
    head(cdoc2010.geo)
  } else if (tolower(action) == "g") {
    if(!exists("geoCodeAll", mode="function")) {
      source(file="geocode.R")
    }

    # takes about 15 minutes to geocode from scratch
    cdoc2010.geo <<- geoCodeAll("cdoc2010", "NAME")
  } else {
    warning("Selection for geocoding action not understood;
    trying default for cdoc2010.")
    filename <- paste0(dataloc, 
"geocoding by school, cdoc2010, N",
    nrow(cdoc2010),".csv")
    cdoc2010.geo <<- read.csv(filename)
# trim the first column, which is just the row number
# added when the file is saved
cdoc2010.geo <- data.frame(cdoc2010.geo[,2:ncol(cdoc2010.geo)])
head(cdoc2010.geo)
}
} else {
message("Found cdoc2010.geo, using existing data frame.")
}

# cdoc2010.geo <- merge(cdoc2010, all_schools.geo[, c("all_schools", "Lat", 
"Lng")], by.x="NAME", by.y="all_schools")

# inspect the results
if(any(is.na(cdoc2010.geo$Lat))) {
  warning(paste("Still missing", length(which(is.na(cdoc2010.geo$Lat))),' of ",
              nrow(schools.geo),' schools. Try OpenRefine.")
} else {
  message("All Carnegie-indexed doctoral schools geocoded
          and saved as cdoc2010.geo.")
}

# TO DO: determine whether these schools changed classification between 2000
and 2010

# TO DO: related analyses to try
# 1. all comp/rhet dissertations by school classification (try different
levels of drill-down; see 2010classifications_logic.pdf / 
http://classifications.carnegiefoundation.org/methodology/grad_program.php)
# 2. correlation table of school classification vs. aggregate method tags
# `check count.R`
#
# Find methods that couldn't be determined and re-tag appropriately.
# Called by `dataprep 2 - load data.R`.
####

```r
m <- noexcludes$Method.Terms
m1 <- as.character(m)
m2 <- sapply(m1, FUN=function(x) strsplit(x,"|",fixed=TRUE))
# find everything that might need checking
checkcount <- sapply(m2, FUN=function(x) length(grep("~check", x, ignore.case=TRUE)))
# filter out the ones that are maybe worth checking from those that def need checking
checkmaybes <- sapply(m2, FUN=function(x) length(grep("~check \?", x, ignore.case=TRUE)))
sum(checkmaybes) m5 <- noexcludes[which(checkmaybes < checkcount),] nrow(m5)
# save the file that still needs checking
filename <- paste0(dataloc,"noexcludes in need of checking.csv")
write.csv(m5, file=filename)
```

```r
## test
data.frame(noexcludes[1:20, which(names(noexcludes) %in%
# c("Method.Terms"))], as.factor(checkcount[1:20]), row.names=NULL)
# allchecks <- which(checkcount == noexcludes$Method.Count) backup <-
# noexcludes[allchecks,]

levels(noexcludes$Method.Terms) <-
levels(factor(c(levels(noexcludes$Method.Terms),"Other")))
noexcludes[allchecks, which(names(noexcludes) %in% c("Method.Terms"))] <-
"Other"
```

```r
## Make sure that worked!
noexcludes[allchecks, which(names(noexcludes) %in%
# c("Method.Terms","ABSTRACT"))] noexcludes$Method.Terms
# print(paste("Converted questionable method terms in",length(allchecks),"rows. Row indices affected:"))
print(as.numeric(allchecks))
```

```r
## To restore replaced rows:
levels(noexcludes$Method.Terms) <- levels(factor(m))
noexcludes[allchecks,] <- backup noexcludes[2697,]
```
Now add an "Other" column

```r
m3 <- as.character(noexcludes$Method.Terms)  # recalculate with new Others

other.index <- grep("Other", m3, ignore.case=TRUE) noexcludes$Othr <- 0
noexcludes[other.index,]$Othr <- 1
```

And, finally, let's recalculate method counts.

```r
m4 <- sapply(m3, FUN=function(x) unlist(strsplit(x,"|",fixed=TRUE)))
noexcludes$Method.Count <- sapply(m4, FUN=length)
```

remove interim variables to save memory

```r
rm(m, m1, m2, m3, m4, m5, allchecks, backup, other.index)
```

```
#############################################################################
# `collocation heatmap.R`: Given method tags, collocate them and
# construct a heat plot. That is, if a dissertation is tagged X,
# how many times is it also tagged Y?
#
# NB: diagonals in the resulting matrix are for solo tags, i.e.
# the number of times a dissertation tagged X is *only* tagged X.
# The total number of times dissertations are tagged X is returned
# separately.
###

# 1. Calculate tag collocations, total dissertations per tag, and solo
# counts per tag.

sumbytags <- function(dataset_name = "noexcludes",

tagset_name = "tagnames",

doplot = TRUE,
	normed = FALSE,  # should we divide by total dissertations per row?
dendro = FALSE  # should we output dendrograms showing method clusters?
)
{

# get values from variable name; we'll use names later
# for filenames and figure titles
dataset <- get(dataset_name)
tagset <- get(tagset_name)

# make a fresh start...
sum.by.tags <- total.counts <- solo.counts <- c()

# ... then build up
for (i in 1:length(tagset)) {

# select the tag
tag <- tagset[i]

# sum columns where the tag is 0 and where it's 1;
# this produces an array with two rows.
tagsum <- aggregate(dataset[, tagset], list(dataset[, tag]),
    FUN=sum)

# Save the row in which the tag is "on" (i.e. set to 1).
# If no such row exists, fill with zeroes to avoid NA results.
# First column is the on/off status, so leave it out.
if (nrow(tagsum) == 1 & tagsum[, 1] == 0) {
    sum.by.tags <- rbind(sum.by.tags, rep(0, ncol(tagsum)-1))
} else {
    sum.by.tags <- rbind(sum.by.tags, tagsum[which(tagsum[,1] == 1),
        2:ncol(tagsum)])
}

# Name the row we've just added by the tag we're currently
# summarizing.
row.names(sum.by.tags)[i] <- tag

# Now the diagonals will dominate, so find the tag's solo count...
solosum <- sum(dataset[which(dataset$Method.Count==1), tag])
solo.counts <- c(solo.counts, solosum)
names(solo.counts)[i] <- tag

# ... and replace the diagonal with that solo count
# (but save the true count, i.e. the total)
total.counts <- c(total.counts, sum.by.tags[i,i])
names(total.counts)[i] <- tag
sum.by.tags[i,i] <- solosum

} # end for loop

# print(sum.by.tags)
# print(total.counts)

to.return <- list("dataset" = dataset_name,
    "correlations" = as.matrix(sum.by.tags),
    "solo.counts" = solo.counts,
    "total.counts" = total.counts)

if(doplot) {
    if(!exists("heatmap.ben", mode="function")) {
        source(file="heatmap_ben.R")
    }
if(!normed) {      # 2. Basic heatmap
  if(remake_figs) {
    filename <- paste0(imageloc,"Method Tag Co-Occurrence, ",
                       dataset_name, ", N", nrow(dataset), ".pdf")
    pdf(filename)
  }

  heatmap.ben(to.return, diags=TRUE, dendro=dendro)
  title(main="Method Tag Co-Occurrence, 
  sub=paste0(dataset_name, ", N", nrow(dataset))
  mtext("A box in row Y, column X gives the number of 
  dissertations tagged Y that are also tagged X", side=4)

  if(remake_figs) {
    dev.off()
  }
}
else {           # 3. Normed heatmap
  if(remake_figs) {
    filename <- paste0(imageloc, "Method Tag Co-Occurrence 
                       (normed by row), ", dataset_name, ", N", 
                       nrow(dataset), ".pdf")
    pdf(filename)
  }

  heatmap.ben(to.return, rowscale=TRUE, diags=TRUE, 
              dendro=dendro)
  title(main="Method Tag Co-Occurrence 
          (normed by row)", 
          sub=paste0(dataset_name, ", N", nrow(dataset))
  mtext("A box in row Y, column X gives the probability that a 
  dissertation tagged Y is also tagged X", side=4)

  if(remake_figs) { dev.off() }
}
# end of if(do.plot)

return (to.return)
}
# end of wrapper function sumbytags()

# Run it when the file is called
if (autorun) {
  remake_figs
  # sum.by.tags <- sumbytags()
  sumbytags("consorts.plus")
  sumbytags("consorts.plus", normed=T)
  sumbytags("top.nonconsorts")
  sumbytags("consorts", dendro=T, normed=T)
}
# compare method ranks.R
#
# GOAL: Find the difference in method frequency between two sets
# by arranging method tags in two columns, and
# connecting matching methods with lines for ease of comparison.
#
# For set1 and set2, use text strings naming variables, not the
# variables themselves, so we can use them to label the figure.
####

compare_method_ranks <- function(set1="consorts",
                                  set2="nonconsorts",
                                  pcts=TRUE,
                                  colorful=FALSE,
                                  betterlabels=NULL) {

  if(!exists("get_tags", mode="function")) { source(file="get tags.R") }

  b <- get_tags(set1)
  d <- get_tags(set2)

  # Line up tag names
  # set1 first:
  b0 <- b[!names(b) %in% "Othr"]                # Exclude "other" tag
  b1 <- names(b0)[order(b0, decreasing=T)]     # Sort by rank

  # repeat for set2:
  d0 <- d[!names(d) %in% "Othr"]                # Exclude "other" tag
  d1 <- names(d0)[order(d0, decreasing=T)]     # Sort by rank

  # Add percentages or diss counts
  if (pcts) {
    # Add percentages to each tag
    b2 <- paste0(b1, " (", round(100*b0[order(b0, decreasing=T)] / nrow(get(set1)), 0), ",")
    d2 <- paste0(d1, " (", round(100*d0[order(d0, decreasing=T)] / nrow(get(set2)), 0), ",")

    filename <- paste0(imageloc, "Ranks of methods in ", set1, " v ",
                        set2, ", no Othr, pcts.pdf")
  } else {
    # Add diss counts to each tag
    b2 <- paste0(b1, " (", b0[order(b0, decreasing=T)], ")")
  }
}
d2 <- paste0(d1, " (", d0[order(d0, decreasing=T)], ")")

filename <- paste0(imageloc, "Ranks of methods in ", set1, " v ", set2, ", no Othr.pdf")
}

## Test significance of any differences

# Strategy: For each tag, construct a 2x2 contingency matrix with columns
# = \{set1, set2\} and rows = \{this-tag, not-this-tag\}; try to reject the
# null hypothesis that the ratio within each column is the same. Account
# for the fact of multiple comparisons, and thus higher chance of
# randomly low p value somewhere in the set, via Bonferroni correction.
# Return asterisks or blank space to add to the label.

onetag.fisher <- function(tag="Clin", verbose=F) {
  mat <- matrix(nrow=2,
    data=c(b[tag], sum(b[!names(b) %in% tag]),
          d[tag], sum(d[!names(d) %in% tag])),
    dimnames=list(c(tag, paste("Not", tag)),
                   c(set1, set2))
  )

  fish <- fisher.test(mat)
  if(verbose) { print(mat); print(fish) }

  # Bonferroni correction: divide target significance levels
  # by the number of comparisons in the set
  if(fish$p.value < 0.001 / length(b)) {
    message(paste(realtags(tag), "is very significantly different
             (Bonferroni corrected p < 0.001) between", set1, " and", set2))
    return(" ** ")
  } else if(fish$p.value < 0.05 / length(b)) {
    message(paste(realtags(tag), "is significantly different
             (Bonferroni corrected p < 0.05) between", set1, " and", set2))
    return(" * ")
  } else {
    message(paste(realtags(tag), "is not significantly different
             between", set1, " and", set2))
    return(" ")
  }
}

# Add significance labels
sig.b <- sapply(b1, FUN=function(x) onetag.fisher(x, verbose=F))
sig.d <- sapply(d1, FUN=function(x) onetag.fisher(x, verbose=F))
b2 <- paste0(sig.b, b2)  # on left, add labels to the left;
d2 <- paste0(d2, sig.d)  # on right, add labels to the right.

if(remake_figs) { pdf(file=filename) }

# set up a blank plot
plot(x=0:length(b)+1,
y=0:length(b)+1,
axes=FALSE,
type="n",
xlab="",
ylab="")

# arrange set1 in descending rank order on the left, set2 on right
text(labels=b2,
x=rep(5.4, length(b2)),
y=length(b2):1,
pos=2)
text(labels=d2,
x=rep(length(d)-5.4, length(d2)),
y=length(d2):1,
pos=4)

## connect matching methods with lines for ease of comparison

# optionally add color to lines to detangle spaghetti
if(colorful) {
  require(RColorBrewer)
  mycol <- brewer.pal(4, "Dark2")
} else {
  mycol <- c("#000000")
}

tag <- b1[1]
lapply(b1, mycol=mycol, FUN=function(tag, mycol) {
  # locate each tag on the plot
  y.left <- length(b2) - grep(tag, b1) + 1
  y.right <- length(b2) - grep(tag, d1) + 1
  col.index <- (y.left-1) %% length(mycol) + 1

  # draw a line between tag's positions on left and on right
  segments(x0=5.7,
           y0=y.left,
           x1=length(b)-5.7,
           y1=y.right,
           col=mycol[col.index]
})
# extend those lines to point horizontally to the tags, # to remove ambiguity
segments(x0=5.4, y0=y.left, 
   x1=5.7, y1=y.left,  
   col=mycol[col.index])
segments(x0=length(b)-5.4, y0=y.right,  
   x1=length(b)-5.7, y1=y.right, 
   col=mycol[col.index])
}

# label the two columns
if(!is.null(betterlabels)) {
  if(length(betterlabels)==2) {
    text(labels=betterlabels, 
      x=c(4,length(b)-4), 
      y=rep(length(b)+1,2)
  }
  else {
    warning("Incorrect number of betterlabels: must be vector of length 2. Using set names.")
    text(labels=c(set1, set2), 
      x=c(4,length(b)-4), 
      y=rep(length(b)+1,2)
  }
}
else {
  text(labels=c(set1, set2), 
    x=c(4,length(b)-4), 
    y=rep(length(b)+1,2)
}


text(labels=c(paste0("(N=", nrow(get(set1)), ")")), 
  paste0("(N=",nrow(get(set2)),")")), 
  x=c(4,length(b)-4), 
  y=rep(length(b),2), 
  cex=0.8)

# add legend for significance
if(any(grep("\*", sig.b))) {
  mtext("\* Bonferroni corrected p < 0.05 \
        \** Bonferroni corrected p < 0.001", 
       cex=0.8, 
      side=2
    }

if (remake_figs) { dev.off() }
} # end of wrapper function compare_method_ranks

if(autorun) {
    remake_figs=F
    compare_method_ranks("consorts", "nonconsorts",
        betterlabels=c("Consortium", "All Non-Consortium"))
    compare_method_ranks("consorts", "top.nonconsorts",
        betterlabels=c("Consortium", "Top Non-Consortium"))
}

#############################################################################
# "cotopics.R"
#
# GOAL: find topics that co-occur within individual dissertations at a level
# greater than (say) 10 or 5%. Map these into a (non-directed) source-target
# edge table, for use in http://bl.ocks.org/mbostock/7607999 (hierarchical
# edge bundling).
#
# Strategy:
# 1. in each row i of X, find all columns with X[i,j] > level; call that A.
# 2. For all combinations of two elements in A, create a new row in a
# source-target table called "cotopics."
#####

getcotopics <- function(dataset_name="consorts",
ntopics = 55,
    level = .12, # what fraction of the doc (out of 1) must
        # each topic account for?
json = F, # export to JSON?
    min = 3, # how many times must these topics co-occur
        # to be "co-topics"?
bad.topics = c("2", "4", "22", "24", "47") # exclude non-content-bearing topics
    )
{
    require(data.table)

    if (!exists("get.doctopic.grid", mode="function")) {
        source(file="get doctopic grid.R")
    }

    grid <- get.doctopic.grid(dataset_name, ntopics)$outputfile
    head(grid)
    grid <- grid[, !names(grid) %in% bad.topics]
    head(grid)

    # start empty, build up.
    cotopics <- data.frame(row.names=c("source","target"))
for (i in 1:nrow(grid)) {
  # loop through the documents (rows).
  
  # find which topics (columns) make up a big chunk.
  A <- which(grid[i, 2:length(grid)] > level)

  # can't combine just one thing.
  if (length(A) >= 2) {
    # don't forget to get topic names, not col numbers!
    A <- as.integer(names(grid[, 1+A]))

    # find all pairs of those big-chunk topics.
    cotopics <- cbind(cotopics, combn(A,2))
  }
}

# the data.frame gave us a wide array; switch to a long one.
# as a data.table, we can do a fast sort and more besides
# for example, let's find unique source/target pairs, # and count their occurrences! in one line! whee!
cotopics <- data.table(cotopics, key=c("source", "target"))

# to reduce complexity, set a minimum number of co-occurrences
cotopics <- cotopics[which(weight > min), ]

# print and optionally save the result
if(autorun) {
  print(cotopics)
}

if(remake_figs) {
  if(json) {
    require(jsonlite)
    filename <- paste0(imageloc, dataset_name, "k", ntopics,
      ":_edges:", level*100, ".json"
    cat(toJSON(cotopics), file=filename)
  } else {
    filename <- paste0(imageloc, "co-topic edge table, ",
      dataset_name, ", k", ntopics, ", ", level*100,
      "pct_nobads.csv")
    write.csv(cotopics, filename)
  }
}
# and pass it back to the calling environment
return(cotopics)
}

if(autorun) {
  get.cotopics(level=0.2, min=2)
}

#`dataprep.R`
# A file to configure my usual working directories, variables, and functions.
# Follow up by running `dataprep2 - load data.R`; see `run all analyses.R`
# for further steps.

# define some broad parameters, since this file will always be run first
# make a shortcut for retrieving the last entered value
ans <- function() {
  .Last.value
}

# set the working directories, taking into account the GitHub setup
sourceloc <- "/Users/benmiller314/Dropbox/coursework, etc/dissertation/data, code, and figures/Dissertation Research/"
setwd(sourceloc)
imageloc <- paste0(sourceloc, "/Dissertation Research - Figures/")
dataloc <- paste0(sourceloc, "/")
malletloc <- "/Users/benmiller314/mallet-2.0.7"
webloc <- "/Users/benmiller314/Documents/Webdev/datavis_testing"

tagnames <- c("Clin", "Crit", "Cult",
              "Phil", "Poet", "Prac", "Rhet", "Surv", "Othr")

tagnames.long <- c("Clinical / Case Study",
                   "Critical / Hermeneutical",
                   "Cultural-Critical",
                   "Discourse or Text Analytical",
                   "Ethnographic",
                   "Experimental / Quasi-Experimental",
                   "Historical / Archival",
                   "Interview / Focus Group",
                   "Meta-Analytical / Disciplinographic",
                   "Model-Building",...
"Philosophical / Theoretical",
"Poetic / Fictive / Craft-Based",
"Practitioner / Teacher-Research",
"Rhetorical Analytical",
"Survey",
"Other"
)

# provide a function to convert tag column labels to real tag names

realtags <- function(tag, tagset_name="tagnames") {
  tagset <- get(tagset_name)
  index <- grep(tag, tagset, ignore.case=TRUE)
  tagset.long <- get(paste0(tagset_name, ".long"))

  return(tagset.long[index])
}

sumnames <- sapply(tagnames, FUN=function(x) paste0(x,".sum"))
meannames <- sapply(tagnames, FUN=function(x) paste0(x,".mean"))
topnames <- sapply(tagnames, FUN=function(x)
  as.list(tolower(paste0("top.",x))))
topnames <- lapply(topnames, FUN=function(x) substr(x,1,8))

# If remake_figs is true (e.g. if set by 'rerun all analyses.R'),
# new pdf files will be created; otherwise, they'll display on screen only.
if(!exists("remake_figs")) {
  remake_figs <- FALSE
}
if(!exists("autorun")) {
  autorun <- FALSE
}

## prep some useful functions
# source(file="function scratchpad.R")
source(file="extract subjects.R")
source(file="Factor-Bug fixing.R")
source(file="heatmap ben.R")
source(file="heatmap fixedcols.R")
source(file="method tag array.R")
source(file="thresh.R")
source(file="simplifying the schema.R")

#############################################################################
# dataprep 2 - load data.R
#
# A file to read in dissertation metadata from a csv file. Binds key subsets
# of data to variables and encodes method tags for easier analysis.
# NB: To identify schools in the Consortium of Doctoral Programs in Rhetoric
# and Composition, requires a separate csv file listing those schools.
### if (!exists("tagnames")) {
  source(file="/Users/benmiller314/Dropbox/coursework,
  etc/dissertation/data, code, and figures/Dissertation Research/dataprep.R")
}
## now get the data
# The most recent file of dissertation metadata
ignore <- readline("Select the most recent file of dissertation metadata.
(Press <Enter> to continue.)")
bigarray <- read.csv(file=file.choose())
rm(ignore)
# parse the method tags... including for the collapsed schema
bigarray <- parse_tags(bigarray)
bigarray <- short_schema(bigarray)

# filter out false positives
noexcludes <- bigarray[bigarray$Exclude.Level==0,]
justexcludes <- bigarray[bigarray$Exclude.Level>0,]
diss.count <- nrow(noexcludes)
false.positives <- nrow(justexcludes)

message(paste("In this data set, there are",diss.count,"dissertations, not
counting",false.positives,"false positives."))

# refactor levels for noexcludes alone
refactor.index <- which(names(noexcludes) %in% c("Subject", "KEYWORDS",
  "School", "Advisor.type", "Advisor.Name", "Degree",
  "Method.Terms", "pages", "Flag.notes"))
for (i in refactor.index) {
  noexcludes[,i] <- factor(noexcludes[,i])
}

# redefine methods that are all "check" or "check?" as "Other,
# and recalculate "Method.Count"
source(file=paste0(sourceloc,"/check count.R"))

# get tag index columns on their own, for simplicity down the road
# TO DO: See whether we still need this
tagarray <- noexcludes[,tagnames]
row.names(tagarray) <- noexcludes[, "Author"]
data.matrix(tagarray) -> tagarray.m

# tag.totals <- tagtotals(tagarray, skip=0)
# barplot(tag.totals)

consortium <- read.csv(file=paste0(dataloc, "doctoral-consortium-schools-programs, reconciled to carnegie.csv"))
conschools <- factor(consortium$University)
consorts.index <- which(noexcludes$School %in% conschools)
consorts <- noexcludes[consorts.index,]
conschoolsfound <- factor(consorts$School)
consort.count <- nrow(consorts)

# print("Consortium Schools Found:")
# print(levels(conschoolsfound))
# print("Did you remember to reconcile schools?")

# figure out which consortium schools are not showing up
missing_conschoolsf <- setdiff(levels(conschoolsf), levels(conschoolsfound))
non_conschoolsf <- setdiff(levels(noexcludes$School), levels(conschoolsf))
nonconsorts <- noexcludes[(which(noexcludes$School %in% non_conschoolsf)),]

# confirm that nonconsorts gets all the schools not in consorts
setequal(nonconsorts, (noexcludes[-consorts.index,]))

# find top nonconsorts
top.nonconsorts <- thresh("nonconsorts")$thresh.data
consorts.plus <- rbind(consorts, top.nonconsorts)

# re-factor all factor columns in all data subsets
consorts <- refactor.all("consorts")
nonconsorts <- refactor.all("nonconsorts")
top.nonconsorts <- refactor.all("top.nonconsorts")
consorts.plus <- refactor.all("consorts.plus")

# make noexcludes easy to index and search
library(data.table)
noexcludes.dt <- as.data.table(noexcludes)
setkey(noexcludes.dt, Pub.number)

## Export file lists for subsets of data
write(levels(factor(noexcludes$Pub.number)),
     file=paste0(sourceloc, "/Shell scripts and commands/file list noexcludes.txt", sep="\n"))
write(levels(factor(consorts$Pub.number)),
     file=paste0(sourceloc, "/Shell scripts and commands/file list consorts.txt", sep="\n"))
write(levels(factor(nonconsorts$Pub.number)),
     file=paste0(sourceloc, "/Shell scripts and commands/file list nonconsorts.txt", sep="\n"))
file=paste0(sourceloc, "/Shell scripts and commands/file list
nonconsorts list.txt", sep="\n")

# TO DO (maybe): split out multiple advisors

# extract subjects.R
#
# Expand subject terms that had been combined in a single cell of a data csv.
# Called by `data prep2 - load data.R`.
#
extract_subjects <- function (s) {
  if(!is.character(s)) { s <- as.character(s);}  
  s2 <- sapply(s,FUN=function(x) unlist(strsplit(x,|",fixed=TRUE)));  
  output <- c();  
  for (i in 1:length(s2)) {  
    output <- c(output,s2[[i]]);  
  }  
  output <- factor(output)  
  return(output);  
}

# Factor-Bug fixing.R
#
## First function: fix_factor
## Rationale:
## When dealing with text, R likes to pre-determine what counts as a valid
## possibility (because it thinks everything is an experimental observation
## with controlled variables). Trying to add new rows in a text column,
## therefore, sometimes causes problems. fix_factor allows you to add new
## items to (or edit old ones in) your factor-ish vectors.
##
## Parameters:
## f a factor, i.e. a text column, in which you want to add or edit
## some entry
## to.add the entry you wish to add, or the revised value if editing.
## required.
## to.remove the entry you wish to replace, if editing. optional.
##
## Usage:
## some.factor <- fix_factor(some.factor, to.add="some.new.text")
```r
# fix_factor <- function(f, to.add, to.remove = NULL) {
#   ff <- as.character(f)
#   if (!is.null(to.remove)) {
#     ff[which(ff %in% to.remove)] <- to.add
#   } else {
#     ff <- c(ff, to.add)
#   }
#   return(factor(ff))
# }

## Second function: refactor.all
## Rationale: When you subset a data.frame, the factors can have more levels
## than there are rows. We want to fix that.
## Usage:
# consorts <- refactor.all("consorts")
# refactor.all <- function(dataset_name="consorts") {
#   dataset <- get(dataset_name)
#   for (i in 1:ncol(dataset)) {
#     if(is.factor(dataset[,i])) {
#       dataset[,i] <- factor(dataset[,i])
#     }
#   }
#   return(dataset)
# }
```

```
# frameToD3.R
# Outputs JSON file of topic model data for interactive visualizations.
# Provides two functions:
# * frameToJSON(): given a topic model as generated by
#   'r2mallet with foreach.R', returns a hierarchical clustering of
#   topics in JSON. For each topic, includes the following metadata:
#   name, size, scaledsize, topwords, topic, rank.
# * cotopic_edges(): given a topic model as generated by `r2mallet with
#   foreach.R`, returns weighted edges between topics and the same
#   hierarchical clustering as above.
# Forked from Rolf Fredheim at
# https://github.com/benmiller314/frameToD3/blob/master/frameToD3.r as
# discussed in
```
frameToJSON <- function(dataset_name = "consorts",
             ntopics=55,
             do.plot=TRUE,  # Ben: Use this the first time to
              # find good cuts in the dendrogram.
groupVars=NULL,  # Ben: If not provided by the calling
dataVars=NULL,    # environment, these 3 parameters
outfile=NULL,     # will be set to defaults.
bad.topics= c("2", "4", "22", "24", "47",
              # exclude non-content-bearing topics
                "50", "13")
              # (incl. Spanish & Italian languages)
}

#packages we will need:
require(data.table)
require(jsonlite)

# Ben: Get topic weights for every document we have
if(!exists("get.doctopic.grid", mode="function")) {
  source(file="get doctopic grid.R")
}
dt <- as.data.table(get.doctopic.grid(dataset_name, ntopics)$outputfile)

# Ben: Exclude non-content-bearing topics
if(!is.null(bad.topics)) { dt <- dt[, !names(dt) %in% bad.topics, with=F] }

# Set parameter defaults if needed
if(is.null(groupVars)) {
  groupVars <- c("Pub.number")  # Group by ID column
}
if(is.null(dataVars)) {
  dataVars <- colnames(dt)[!colnames(dt) %in% groupVars]
  # any column that's not an ID is a datapoint
}
if(is.null(outfile)) {
  # the desired location of the JSON file produced by the function
  outfile <- paste0(webloc, "/", dataset_name, "k", ntopics,
                  "_clusters_", ntopics-length(bad.topics), ".json")
}

#Rolf: Here you may want to sort by colSums()
#to keep only the most relevant variables.
#Rolf: calculate the correlation matrix
t <- cor(dt[, dataVars, with=F])

#Rolf: calculate the hierarchical cluster structure
# from the correlation scores
hc <- hclust(dist(t), "ward.D2")

# Ben: I'm making this section optional,  
# because it makes the most sense early on and has diminishing returns.
if(do.plot) {
    #Rolf: take a look at your structure:
    # Ben: optionally save clustering figure
    main <- paste0("Cluster Dendrogram, ", dataset_name, ", ",
                   ntopics - length(bad.topics), " topics")

    # Ben: Try various cut levels until you find a set that seems  
    # interesting; Then adjust the memb_ variables below, accordingly.

    # with 5 bad.topics removed
    if(dataset_name=="consorts" && ntopics==55 && length(bad.topics) == 5) {
        if(remake_figs) { pdf(file=paste0(imageloc, main, ",.pdf")) } 
        plot(hc, main=main)
        abline(1.35, 0, col="#99FF99")
        rect.hclust(hc, k=32, border="#99FF99")
        abline(1.55, 0, col="#009900")
        rect.hclust(hc, k=16, border="#009900")
        abline(1.7, 0, col="#FF9999")
        rect.hclust(hc, k=12, border="#FF9999")
        abline(1.85, 0, col="#9999FF")
        rect.hclust(hc, k=7, border="#9999FF")
        abline(1.95, 0, col="#990099")
        rect.hclust(hc, k=6, border="#990099")
        abline(2.33, 0, col="#009999")
        rect.hclust(hc, k=4, border="#009999")
        abline(3.37, 0, col="#999900")
        rect.hclust(hc, k=2, border="#999900")
        if(remake_figs) { dev.off() }
    }

    # with 7 bad.topics removed
    else if(dataset_name=="consorts" && ntopics==55 && length(bad.topics) == 7) {
        if(remake_figs) { pdf(file=paste0(imageloc, main, ",.pdf")) } 
        plot(hc, main=main)
        abline(1.45, 0, col="#99FF99")
        rect.hclust(hc, k=21, border="#99FF99")
    }
}
abline(1.73, 0, col="#009900")
rect.hclust(hc, k=11, border="#009900")
abline(1.955, 0, col="#FF9999")
rect.hclust(hc, k=6, border="#FF9999")
rect.hclust(hc, k=4, border="#009999")
rect.hclust(hc, k=2, border="#999900")
if(remake_figs) { dev.off() }

# TO DO: Find splits for model with 150 topics
else {
  if(remake_figs) { pdf(file=paste0(imageloc, main, ".pdf")) }
  plot(hc, main=main)
  if(remake_figs) { dev.off() }

  # If we're plotting, we probably wanted to locate splits.
  # Exit the function here.
  message("Exiting function.")
  message("Using abline() and rect.hclust(), try various cut levels until you find a set that seems promising.")
  return()
}
# end of if(do.plot)

# Rolf: now we split the data based on membership structure. We will take four levels: (basically this means we will calculate which group each variable belongs in for different levels of the tree structure)

## Ben: so, essentially, we're going to look at plot(hc) and decide what the major branch points are, then cut the tree to find group assignments above/below those splits. NB cutree() also allows us to split the tree at specific heights (on the y axis of that plot), if we don't want to count the groups.

# Ben: splits for consorts with 55 topics (i.e. including bad.topics)
if(dataset_name="consorts" && ntopics==55 && is.null(bad.topics)) {
  splits <- c(2, 5, 10, 22, 55)

  memb2 <- as.character(cutree(hc, k = 2))
  memb5 <- as.character(cutree(hc, k = 5))
  memb10 <- as.character(cutree(hc, k = 10))
  memb22 <- as.character(cutree(hc, k = 22))
  memb55 <- as.character(cutree(hc, k = 55))
}

# Ben: splits for consorts with 50 topics
# (i.e. 5 bad.topics removed for bad OCR or boilerplate)
if(dataset_name="consorts" && ntopics==55 && length(bad.topics) == 5) {
  splits <- c(2, 4, 6, 7, 12, 16, 32)
memb2 <- as.character(cutree(hc, k = 2))
memb4 <- as.character(cutree(hc, k = 4))
memb6 <- as.character(cutree(hc, k = 6))
memb7 <- as.character(cutree(hc, k = 7))
memb12 <- as.character(cutree(hc, k = 12))
memb16 <- as.character(cutree(hc, k = 16))
memb32 <- as.character(cutree(hc, k = 32))

# Ben: splits for consorts with 48 topics (i.e. 7 bad.topics removed
# for bad OCR, boilerplate, or non-English lang)
if(dataset_name=="consorts" && ntopics==55 && length(bad.topics) == 7) {
  splits <- c(2, 4, 6, 11, 21)
  memb2 <- as.character(cutree(hc, k = 2))
memb4 <- as.character(cutree(hc, k = 4))
memb6 <- as.character(cutree(hc, k = 6))
memb11 <- as.character(cutree(hc, k = 11))
memb21 <- as.character(cutree(hc, k = 21))
}

# TO DO: Add splits for model with 150 topics

# Make note of group names for later;
# same operation for all numbers of bad.topics
membVars <- paste0("memb", splits)

# Ben: get topic labels, which you've composed elsewhere using
# 'top docs per topic.R'
if(!exists("get_topic_labels", mode="function")) {
  source(file="get_topic_labels.R")
}
topic.labels.dt <- get_topic_labels(dataset_name, ntopics)
  # str(topic.labels.dt)

# exclude non-content-bearing topics
if(!is.null(bad.topics)) {
  topic.labels.dt <- topic.labels.dt[res Topic %in% bad.topics]
}

#Rolf: Now put this information into a table, together with the labels and
#the order in which they should appear:
# Ben adds: use gsub to remove spaces (this seems to help the d3
# scrollover); add topic number to aid in merging w/ edge table later
b <- data.table(sapply(membVars, FUN=function(var){
    get(as.character(var))
  }),
       label = gsub(' ', '_', topic.labels.dt[, Label]),
topic = topic.labels.dt[, Topic],
topwords = topic.labels.dt[, Top.Words],
rank = topic.labels.dt[, Rank],
order = hc$order)

#Rolf: We might want to know the size of each node. Let's add that.
# Ben: for a topic model, this will find the total %-point contribution of
# the topic to all docs; that means we could divide by number of docs to
# scale to [0,1], but no need: it's proportional.
b$size <- colSums(dt[,c(dataVars),with=F])
b$scaledsize <- b$size/nrow(dt)

#Rolf: sort the data so it aligns with the structure calculated using
#hclust()
setkey(b,order)

#Rolf: drop the order variable:
b[,order:=NULL]

# Ben: Save this data table to a csv for later inspection; this table will
# also be returned by the function.
if(remake_figs) {
  filename <- paste0(imageloc, "topic clusters - ", dataset_name, ", K", ntopics, "bad topics removed.csv")
  write.csv(b, filename)
} else {
  # print(b)
}

## Hierarchical Clustering of Topics by Similarity
#Rolf: we define a function which will create a nested list in JSON format:
#From here:
#http://stackoverflow.com/questions/12818864/how-to-write-to-json-with-
#children-from-r
# Ben: but see also, now, http://bit.ly/1jXAC5M

makeList <- function(x) {
  if (any(names(x) %in% membVars) & ncol(x)>2) {
    listSplit<-split(x[-1],x[1],drop=T)
    grp <- names(x)[1]
    grpnum <- substr(grp, 5, nchar(grp))
    names(listSplit) <- paste0(names(listSplit), "of", grpnum)
    lapply(names(listSplit), function(y){
      list(name=y,children=makeList(listSplit[[y]]))
    })
  } else {
    lapply(seq(nrow(x[1])), function(y){
      list(name=x[,"label"][y],
           size=x[,"size"][y],
          ...}
```r
scaledsize=x[, "scaledsize"] [y],
topwords=x[, "topwords"] [y],
topic=x[, "topic"] [y],
rank=x[, "rank"] [y])
}
} # end of if-else
} # end of makeList

# Rolf: This will not work on a data.table
b.df <- data.frame(b)
out <- makelist(b.df)
# str(out)
# toJSON(out)

# Have a look at the structure this creates:
if(autorun) { print(head(out)) }

# Rolf: Basically we have made a list of lists containing the information
# from the tree diagram. Finally we put everything into a list, convert this
# to json format and save it as data.json
jsonOut <- toJSON(list(name="1of1", children=out), digits=6, pretty=TRUE)

# Rolf: We use the cat function here, because in some cases you may want to
# add separators, or a prefix and suffix to make the formatting just right
# Ben adds: to avoid overwriting, only save this file if remake_figs is
# TRUE
if(remake_figs) { cat(jsonOut, file=outfile) }

# Ben: Return the data.table for use in edge bundling, below
return(b)

## Hierarchical Edge Bundling between (possibly unrelated) topics. This one's
# all Ben, but trying to reconstruct a figure like Rolf's
#
# Plan: From the combined hierarchical data structure above (named `b`), for
# each topic (row):
# 1) pull out the "name" field that combines location in hierarchy with
#    label information
# 2) loop through the targets, and find the "name" corresponding to
#    that target
# 3) convert to JSON.

cotopic_edges <- function(dataset_name="consorts",
ntopics=55,
level=0.12,    # topic must constitute how much of each doc?
min=3,        # how many times must a pair of topics co-occur?
outfile=NULL,
bad.topics= c("2", "4", "22", "24", "47", "50", "13")
```
# exclude non-content-bearing topics

{}

## set default parameters if needed

# the desired location of the JSON file produced by the function
if(is.null(outfile)) {
    outfile <- paste0(webloc, "/", "edges ", dataset_name, "k", ntopics, " ", ntopics-length(bad.topics), " ", level*100, "pct_min", min, ".nobads.json")
}

# get co-occurring topics, for hierarchical edge bundling
if(!exists("get.cotopics")) { source(paste0(sourceloc, "cotopics.R")) }

cotopics <- get.cotopics(dataset_name, ntopics, level, min)

cotopics_flip <- data.table( source=cotopics$target,
    target=cotopics$source,
    weight=cotopics$weight)

cotopics_both <- rbind(cotopics, cotopics_flip)

# aggregate all edges by source
edges <- cotopics_both[, .SD[, list( 
    "targets"=paste(target, collapse=" "),
    "weights"=paste(weight, collapse=" "))
], by=source]

setkey(edges, source)
# head(edges)

# Bring in the node table
b <- frameToJSON(dataset_name, ntopics, bad.topics=bad.topics, do.plot=F)
setkey(b, topic)
# head(b)

# merge
b <- edges[b, ]
str(b)

# Create a "name" column that collapses the hierarchical structure and # topic label, as per # http://fredheir.github.io/dendroArcs/pages/hierarc/test.JSON This is # what the d3 edge bundling code in packages.js will parse to recreate # the hierarchy

# first re-derive 'membVars' from the names of b that include "memb"
membVars <- names(b)[grep("memb", names(b))]

b$name <- sapply(1:nrow(b), FUN=function(x) {
    paste(b[x, c(membVars, "label"), with=F], collapse=".")
})

# We're going to build our JSON for edge bundling with a name, size, and
# (to take advantage of Mike Bostock's
# http://mbostock.github.io/d3/talk/20111116/packages.js) we'll call the
# edges "imports"

# We start empty...
edge_bund <- data.table(name=rep("NA", max(b$source)),
    topic=0,
    size=0.0,
    imports=list("NA")
)

# ... and then build up
for (i in b$source) {
    edge_bund[i, "topic"] <- i
    edge_bund[i, "name"] <- b[source %in% i, name]
    edge_bund[i, "rank"] <- b[source %in% i, rank]
    edge_bund[i, "size"] <- b[source %in% i, size]
    edge_bund[i, "scaledsize"] <- b[source %in% i, scaledsize] * 100
    edge_bund[i, "topwords"] <- b[source %in% i, topwords]

    # extract targets' topic numbers
    imports <- lapply(strsplit(b[source %in% i, targets], ","),
        FUN=function(x) {
            x <- as.integer(x) # convert from string to numeric
            b[source %in% x, name] # match topic numbers to sources
        })

    # weights correspond by position in array
    weights <- lapply(strsplit(b[source %in% i, weights], ","),
        FUN=function(x) {
            x <- as.integer(x)
        })

    if(!anyNA(imports[[1]])) { # list edges if there are any
        edge_bund[i, "imports"] <- list(imports)
        edge_bund[i, "weights"] <- list(weights)
    } else { # otherwise...
        # ... give it a loop back to itself
        edge_bund[i, "imports"] <<- list(b[source %in% i, name])

        # and call the weight "1"
        edge_bund[i, "weights"] <- list(1)
    }
}
# Now remove any empty rows introduced by cutting bad.topics
edge_bund <- edge_bund[!(name %in% "NA")]

jsonEdge <- toJSON(edge_bund, pretty=TRUE)
cat(jsonEdge, file=outfile)
return(jsonEdge)

if(autorun) {
  remake_figs
  # debug(frameToJSON)
  frameToJSON(do.plot=F)
  frameToJSON(ntopics=150, bad.topics=NULL)

  # 12% determined by `variation of topic proportions.R` to include nearly
  # all primary topics and 3/4 of secondary topics; see `Variation of Topic
  # Proportions, Top 10 Topics per Document.pdf`
cotopic_edges(level=0.12, min=1)
cotopic_edges(level=0.12, min=2)
cotopic_edges(level=0.12, min=3)
cotopic_edges(level=0.12, min=4)
cotopic_edges(level=0.12, min=5)
}

# geocode.R
#
# Following geocoding instructions from
# http://allthingsr.blogspot.com/2012/01/geocode-your-data-using-r-json-
#
# NB: This file will be called by `map by school 1 (setup).R`
####

library("RJSONIO") #Load Library
getGeoCode <- function(gcStr,  # what location should we find lat/long for?
throttle = 2)  # how many seconds to wait between requests{

  # correct for locations that GoogleMaps API can't find the names of
  if (gcStr == "Aquinas Institute of Theology") {
    gcStr <- "3 South Spring Avenue St. Louis, Missouri"
  } else if (gcStr == "Emmanuel College of Victoria University (Canada)") {
    gcStr <- "75 Queens Park Crescent East, Toronto, ON M5S 1K7, Canada"
  } else if (gcStr == "Fuller Theological Seminary, School of Theology") {
    gcStr <- "135 N. Oakland Avenue, Pasadena, CA 91182"
} else if (gcStr == "Seton Hall University, College of Education and Human Services") {
    gcStr <- "400 South Orange Avenue, South Orange, NJ 07079"
}

# Encode URL Parameters
gcStr2 <- gsub(\',\"%20\", gcStr)

# Open Connection
connectStr <-
paste('http://maps.google.com/maps/api/geocode/json?sensor=false&address=', gcStr2, sep=")
con <- url(connectStr)
data.json <- fromJSON(paste(readLines(con), collapse=""))
close(con)

# Flatten the received JSON
data.json <- unlist(data.json)
lLat <- data.json["results.geometry.location.lat"]
lLng <- data.json["results.geometry.location.lng"]

# Progress report
if(is.na(lLat)) {
    print(paste("Unable to match ", gcStr))
} else {
    print(paste("Successfully matched", gcStr))
}

gcodes <- c(lLat, lLng)
names(gcodes) <- c("Lat", "Lng")

Sys.sleep(throttle)
return (gcodes)

} # end of getGeoCode()

# Now apply the function above across a set of data
geoCodeAll <- function(dataset_name = "noexcludes",
                    schoolColName = c("School", "NAME")) {

dataset <- get(dataset_name)
if (schoolColName == "NAME") {
    all_schools <- levels(factor(dataset$NAME))
} else {
    all_schools <- levels(factor(dataset$School))
}

# For each school found, use the function above to create new Lat
# and Long columns.
geoCols <- lapply(all_schools, function (val) { getGeoCode(val) } )

# geoCols gives Lat/long in rows; flip it and add to school names
all_schools.geo <- data.frame(all_schools, t(data.frame(geoCols)) )
head(all_schools.geo)
row.names(all_schools.geo) <- NULL
head(all_schools.geo)

# now to fix the ones that didn't match:
# 1. get city and state data from "carnegie I (setup).R"
all_schools.geo <- merge(all_schools.geo, carnegie.all[,c("NAME", "CITY", "STABBR")], 
by.x="all_schools", by.y="NAME", all.x=T, all.y=F)
head(all_schools.geo)

# 2. combine the city and state for ease of search
all_schools.geo$City.State <- paste0(all_schools.geo$CITY, 
", ", all_schools.geo$STABBR)

# 3. find blank Lat or Lng
blanks.index <- which(is.na(all_schools.geo$Lat))

# 4. search based on city and state
for (i in blanks.index) {
  lookup <- getGeoCode(all_schools.geo[i, "City.State"])

  # add the new Lat/Lng values to the levels,
  # so they won't get rejected
  levels(all_schools.geo$Lat) <- c(levels(all_schools.geo$Lat), 
    lookup["Lat"])
  levels(all_schools.geo$Lng) <- c(levels(all_schools.geo$Lng), 
    lookup["Lng"])

  # add the new values to the appropriate Lat/Lng cells
  all_schools.geo[i, "Lat"] <- lookup["Lat"]
  all_schools.geo[i, "Lng"] <- lookup["Lng"]
}

# check output
print(all_schools.geo)

# convert Lat/Lng factors to numbers
all_schools.geo$Lat <- 
as.numeric(levels(all_schools.geo$Lat)[all_schools.geo$Lat])
all_schools.geo$Lng <- 
as.numeric(levels(all_schools.geo$Lng)[all_schools.geo$Lng])

# save that file!
if(remake_figs) {

filename <- paste0(dataloc, "geocoding by school ", dataset_name, ", N", nrow(dataset), ".csv")
write.csv(all_schools.geo, file=filename)
}
return(all_schools.geo)
}  # end of wrapper function geoCodeAll

# get doctopic grid.R
#
# GOAL: Read in a table of documents with all topic proportions for each.
# Return both this table and total topic proportions across documents.
#
# STRATEGY:
# Edit the reshapeMallet.py script (in TextWrangler) to update filenames,
# then run it here and read in the output.
####
get.doctopic.grid <- function(dataset_name="consorts", ntopics=55, doplot=F) {
  # get packages in case we've just restarted R
require(data.table)

  # Locate the doc/topic grid, or create it if it doesn't yet exist.
filename <- paste0(malletloc, "/", dataset_name, "k", ntopics,
"_doc-all-topics.txt")
scope <- paste("cd", shQuote(sourceloc),
"; cd 'Shell scripts and commands' ; ls ", filename)

if (system(scope))  # runs only if file not found,
  # which returns a non-zero error value
{
  command <- paste("cd", shQuote(sourceloc),
    "; cd 'Shell scripts and commands' python reshapeMallet.py")
go <- readline("Have you updated reshapeMallet.py
to reflect your current dataset/ntopics? (Y/N)\n")
if(tolower(go) != "y") {
  stop("Better fix that, then")
}

  print("Converting topic/weight pairs into doc/topic grid...")
  if(! system(command)) {  # runs only if command exits successfully
    print("Done.")
  } else {
    print("Oh, good, the file exists. Moving on...")
  }
}
# Read in the doc/topic grid
outputfile <- read.delim(filename, header=F)

# switch from θ-indexed to 1-indexed, so the topic numbers in
# topic_keys.dt are the same as row numbers
# NB: this seems to be necessary to avoid searching for column "θ"
head(outputfile)
names(outputfile) <- c("Pub.number", (1:(ncol(outputfile)-1)))
head(outputfile)

outputfile.dt <- as.data.table(outputfile)
head(outputfile.dt)

## Find overall top topics
# Each cell gives the percentage which the topic in that column
# contributes to the dissertation in that row. Summing these percentages
# and sorting gives us a rank based on percentage points.

colsums <- colSums(outputfile.dt)
names(colsums) <- names(outputfile.dt)
head(colsums)
colsums.sort <- colsums[order(colsums, decreasing=TRUE)]
head(colsums.sort)

# Divide the percentage point totals by the number of dissertations
# to get an overall percent contribution

colsums.sort.pct <- round((colsums.sort / nrow(outputfile)), 4) * 100

# Optionally save to file
if(remake_figs) {
    filename <- paste0(imageloc, dataset_name, "k", ntopics,
    "_topic-ranks.csv")
    write.csv(colsums.sort.pct[2:length(colsums.sort.pct)], filename)
}

# Optionally get an overview of the topic sizes, as a scatterplot
if(doplot) {
    plot(x = 2:length(colsums),
         y = colsums.sort[2:length(colsums)],
         xlab = "topic numbers (arbitrary)",
         ylab = "sum of contributions",
         xaxt = "n"
    )

    # barplot(colsums.sort[2:length(colsums)],
    #         xlab="topic numbers (arbitrary)",
    #         ylab="sum of contributions",
    #         xaxt="n",
    #         main="Contributions by Topic"
}
get_topic_labels <- function(dataset_name = "consorts", ntopics = 55) {
    filename <- paste0(image_loc, "topic labeling - ", dataset_name, ", K", ntopics, ".csv")
    topic.labels.dt <- tryCatch(
        data.table(read.csv(filename), key = "Topic"),
        error = function(e) {
            message("File not found; using top words instead.")
            keys <- get.topickeys(dataset_name, ntopics)
            outfile <- paste0(webloc, "/", dataset_name, "k", ntopics, "_clusters_topwords.json")
            return(data.table(Topic = 1:ntopics,
                              Label = keys$top_words))
        },
        finally = { message("done.") }
    )
}
# get topickeys.R
#
# GOAL: Given a dataset and number of topics, read in the top words for each
# topic in that topic model.

get.topickeys <- function(dataset_name="consorts", ntopics=55)
{
  # get packages in case we've just restarted R
  require(data.table)

  filename <- paste0(malletloc, "/", dataset_name, "k", ntopics, "_keys.txt")
  topic_keys.dt <- as.data.table(read.delim(filename, header=F))
  setnames(topic_keys.dt,
    c("V1", "V2", "V3"),
    c("topic", "alpha", "top_words"))

  names(topic_keys.dt)
  head(topic_keys.dt)

  # switch from 0-indexed to 1-indexed so the topic numbers in
  # topic_keys.dt are the same as row numbers
  # NB: this seems to be necessary to avoid searching for column "0"
  topic_keys.dt$topic <- 1:nrow(topic_keys.dt)

  head(topic_keys.dt)
  setkey(topic_keys.dt, topic)

  return(topic_keys.dt)
}

# get topics for author.R
#
# Exploring topics within individual dissertations. Now with author name
# search, for convenience. NB: Author names from ProQuest are in all caps,
# like this: "LASTNAME, FIRSTNAME"

if(!exists("get.topics4doc", mode="function")) {
  source(file="top docs per topic.R")
}

get.topics4author <- function(authorname,
    dataset_name = "consorts",
    ntopics = 55,
    howmany = 10, # How many topics to show?
    showlabels = TRUE) # Use existing labels for
    # topics if TRUE, or just
    # topic numbers if FALSE
{

pubnum <- noexcludes[grep(authorname, noexcludes$Author, ignore.case=T),
           "Pub.number"]
if(length(pubnum) == 1) {
  print(noquote(paste("$author: ",
              noexcludes.dt[as.character(pubnum),
              list(Author)]$Author)))
  print(get.topics4doc(pubnum,
              dataset_name,
              ntopics,
              howmany,
              showlabels))
}
else if (length(pubnum) > 1) {
  message("More than one match;
            please use exact author name from list below.")
  results <- noexcludes.dt[as.character(pubnum), list(Author, Title)]
  print(results)
  a <- as.integer(readline("Search again using row number... "))
  pubnum <- results[a, Pub.number]
  print(noquote(paste("$author: ", results[a, Author])))
  print(get.topics4doc(pubnum,
              dataset_name,
              ntopics,
              howmany,
              showlabels))
}
}
if(autorun) {
  get.topics4author("MUELLER, DEREK")
  get.topics4author("Lucas")
}

# heatmap_ben.R
#
# Construct a map of squares, shaded according to value. Based heavily on R's
# core heatmap function, modified to print the value for each square of the
# grid.
#
heatmap.ben <- function (sum.by.tags,                     # A list output by sumbytags() containing
diags = FALSE,                # Should we outline diagonals?
           # a correlation matrix, solos, & totals
           # A list output by sumbytags() containing
           # a correlation matrix, solos, & totals
           # Should we outline diagonals?
highval = "#818181",    # Darkest color
lowval = "#FAFAFA",    # Lightest color
numCols = 10,          # How many different shades?
rowscale = FALSE,      # Should we norm each row by tag totals?
verbose = TRUE,        # Should we add a subtitle about solo tags?
legend = TRUE,         # Should we output a separate file with a legend
                      # for the color map?
dendro = FALSE,        # Should we output dendrograms showing method
                      # clustering?

{
  # extract the matrix, if need be
  if(!is.matrix(sum.by.tags)) {
    sum.by.tags.s <- sum.by.tags$correlations
    if(is.null(sum.by.tags.s)) {
      sum.by.tags.s <- as.matrix(sum.by.tags)
    }
  } else {
    sum.by.tags.s <- sum.by.tags
  }

  # is it symmetrical?
symm <- all(sum.by.tags.s == t(sum.by.tags.s))

  ## sort the matrix (borrowed from heatmap())
  Rowv <- rowMeans(sum.by.tags.s, na.rm = TRUE)    # Find row means
  hcr <- hclust(dist(sum.by.tags.s))              # Cluster based on
distances
  ddr <- as.dendrogram(hcr)                      # Convert to dendrogram
                      # (which we might use
  later)
  ddr <- reorder(ddr, Rowv)                      # Reorder the dendrogram
  rowInd <- order.dendrogram(ddr)                # Extract the row order
  Colv <- colMeans(sum.by.tags.s, na.rm = TRUE)  # Find column means
  hcc <- hclust(dist( if(symm) {sum.by.tags.s} # Cluster based on
distances
  perspective)
  else {t(sum.by.tags.s))})
  ddc <- as.dendrogram(hcc)                      # Convert to dendrogram
                      # (which we might use
  later)
  ddc <- reorder(ddc, Colv)                      # Reorder the dendrogram
  colInd <- order.dendrogram(ddc)                # Extract the column
  order
  sum.by.tags.s <- sum.by.tags.s[rowInd, colInd]  # Apply row and column
                      # orders from above
# make variables more readable for later
n.col <- ncol(sum.by.tags.s); # print(n.col)
n.row <- nrow(sum.by.tags.s); # print(n.row)

# norm by tag totals
if (rowscale) {
  # get the totals from the sumbytags() list object
totals <- sum.by.tags$total.counts

  # put it in the same order as the rows
totals <- totals[rowInd]

  # divide each row by the total of that row's tag
  sum.by.tags.s <- apply(sum.by.tags.s, 1, FUN=function(x) { x/totals })

  # round to make it prettier
  sum.by.tags.s <- round(sum.by.tags.s, 2)
  sapply(sum.by.tags.s, FUN=function(x) { if(is.na(x)) x <- 0 })
}

# color function
# if we're norming rows, use white for 0 and black for 100%
if(rowscale) {
  max.val <- 1
  min.val <- 0
  highval <- "#000000"
  lowval <- "#FFFFFF"
} else {
  if(any(sum.by.tags.s == 0)) { lowval <- "#FFFFFF" }
  max.val <- max(sum.by.tags.s)
  min.val <- min(sum.by.tags.s)
}

pal <- colorRampPalette(c(lowval, highval))
cols <- pal(numCols)

colorme <- function (val) {
  colIndex <- round(numCols * (val - min.val) / (max.val - min.val))
  colIndex <- max(1,colIndex)
  return(cols[colIndex])
}

if(legend) {
  if(remake_figs) {
    if(rowscale) {
      filename <- paste0(imageloc, "color legend for ",
    } else {
      filename <- paste0(imageloc, "color legend for ",
    }
  }
}
sum.by.tags$dataset,  
" method correlations, normed.pdf")
} else {
  filename <- paste0(imageloc, "color legend for ",  
  sum.by.tags$dataset,  
  " method correlations, raw.pdf")
}

pdf(filename)

xleft <- seq(0, 1, length.out=numCols)
xdiff <- xleft[2]-xleft[1]
plot(x = 0,  
  y = 0,  
  xlim = c(0,1+xdiff),  
  ylim = c(0,1),  
  type = "n",  
  xaxt = "n",  
  yaxt = "n",  
  xlab = "",  
  ylab = "",  
  bty="n"
)
rect(xleft = xleft,  
  xright = xleft + xdiff,  
  ybottom = 1,  
  ytop = 1 - xdiff,  
  col = cols)

text(x = xleft + xdiff / 2,  
  y = 1 - 2 * xdiff,  
  labels = round(seq(min.val, max.val, length.out=numCols), 1))

if(remake_figs) { dev.off() }

} # end of if(legend)

# set up a blank canvas of the right size
plot(0, 0, xlim=c(0.5,0.5+n.col), ylim=c(0.5,0.5+n.row), type="n",  
xaxt="n", yaxt="n", xlab="", ylab="", bty="n")

# map each square
for (i in 1:n.row) {
  for (j in 1:n.col) {
    # print(c('i' = ',',i,' j' = ',',j))
    diagcheck <- NULL # outline the diagonals if need be
    if (diags & i == j) {
      diagcheck <- "black"
symbols(x = j,  
y = 1 + n.row - i,  
squares = 1,  
add = TRUE,  
inches = FALSE,  
fg = diagcheck,  
bg = colorme(sum.by.tags.s[i,j]))

text(x = j,  
y = 1 + n.row - i,  
round(sum.by.tags.s[i,j], 2),  
cex=0.65)

# add axis labels
axis(side = 2,  
at = n.row:1,  
labels = rownames(sum.by.tags.s),  
pos = 0.5,  
las = 2,  
col = "white"
)

axis(side = 1,  
at = 1:n.col,  
labels = colnames(sum.by.tags.s),  
pos = 0.5,  
las = 2,  
col="white"
)

plot(ddc)

# add subtitle indicating scaled / not scaled
if (verbose) {
  if (rowscale) {
    h2 <- paste("Each row normed by dividing over total number of dissertations for that row's tag.\n",  
    "Diagonals represent tags occurring on one-method dissertations."
  } else {
    h2 <- "Diagonals represent tags occurring on one-method dissertations."
  }
  title(sub=h2)
}
if(dendro) {
    if(remake_figs) {
        if(rowscale) {
            filename <- paste0(imageloc, "dendrogram for ",
                              sum.by.tags$dataset,
                              " method column correlations.pdf")
        } else {
            filename <- paste0(imageloc, "dendrogram for ",
                              sum.by.tags$dataset,
                              " method correlations, raw.pdf")
        }
        pdf(filename)
    }
    plot(ddc)
    if(remake_figs) { dev.off() }
}

if(rowscale) {
    if(remake_figs) {
        if(remake_figs) {
            filename <- paste0(imageloc, "dendrogram for ",
                              sum.by.tags$dataset,
                              " method row correlations.pdf")
        }
        pdf(filename)
    }
    plot(ddr)
    if(remake_figs) { dev.off() }
}

} # end of if(dendro)

} # end of wrapper function heatmap.ben()

# heatmap fixedcols.R
#
# Slightly modifies the built-in heatmap function to allow for pre-set column
# order (lines 46-69). Loaded by `dataprep.R`, and optionally called in `tags
# by school.R`.

heatmap.fixedcols <- function(x,
    myColInd,
    Rowv = NULL,
    Colv = if (symm) "Rowv" else NULL,
    distfun = dist,
    hclustfun = hclust,
    reorderfun = function(d, w) reorder(d, w),
    add.expr,
    symm = FALSE,
    revC = identical(Colv, "Rowv"),
    scale = c("row", "column", "none"),
    ...)

na.rm = TRUE,
margins = c(5, 5),
ColSideColors,
RowSideColors,
cexRow = 0.2 + 1/log10(nr),
cexCol = 0.2 + 1/log10(nc),
labRow = NULL,
labCol = NULL,
main = NULL,
xlab = NULL,
ylab = NULL,
keep.dendro = FALSE,
verbose =getOption("verbose"), ...)
{
  scale <- if (symm & missing(scale))
    "none"
  else match.arg(scale)
  if (length(di <- dim(x)) != 2 || !is.numeric(x))
    stop("'x' must be a numeric matrix")
  nr <- di[1L]
  nc <- di[2L]
  if (nr <= 1 || nc <= 1)
    stop("'x' must have at least 2 rows and 2 columns")
  if (!is.numeric(margins) || length(margins) != 2L)
    stop("'margins' must be a numeric vector of length 2")
  doRdend <- !identical(Rowv, NA)
  doCdend <- !identical(Colv, NA)
  if (!doRdend & identical(Colv, "Rowv"))
    doCdend <- FALSE
  if (is.null(Rowv))
    Rowv <- rowMeans(x, na.rm = na.rm)
  if (is.null(Colv))
    Colv <- colMeans(x, na.rm = na.rm)
  if (doRdend) {
    if (inherits(Rowv, "dendrogram"))
      ddr <- Rowv
    else {
      hcr <- hclustfun(distfun(x))
      ddr <- as.dendrogram(hcr)
      if (!is.logical(Rowv) || Rowv)
        ddr <- reorderfun(ddr, Rowv)
    }
    if (nr != length(rowInd <- order.dendrogram(ddr)))
      stop("row dendrogram ordering gave index of wrong length")
  }
  else rowInd <- 1L:nr

  # Ben says: Here's where the original function sets column order
  if (doCdend) {
    if (inherits(Colv, "dendrogram"))
ddc <- Colv
else if (identical(Colv, "Rowv")) {
  if (nr != nc)
    stop("Colv = \"Rowv\" but nrow(x) != ncol(x)\")
  ddc <- ddr
}
else {
  hcc <- hclustfun(distfun(if (symm) x
  else t(x)))
  ddc <- as.dendrogram(hcc)
  if (!is.logical(Colv) || Colv)
    ddc <- reorderfun(ddc, Colv)
}
if (nc != length(colInd <- order.dendrogram(ddc)))
  stop("column dendrogram ordering gave index of wrong length")
else colInd <- 1L:nc

# Ben says: Okay, whatever, I'm over-riding all that
# with my pre-chosen column order
colInd <- myColInd

x <- x[rowInd, colInd]
labRow <- if (is.null(labRow))
  if (is.null(rownames(x)))
    (1L:nr)[rowInd]
  else rownames(x)
else labRow[rowInd]
labCol <- if (is.null(labCol))
  if (is.null(colnames(x)))
    (1L:nc)[colInd]
  else colnames(x)
else labCol[colInd]
if (scale == "row") {
  x <- sweep(x, 1L, rowMeans(x, na.rm = na.rm), check.margin = FALSE)
  sx <- apply(x, 1L, sd, na.rm = na.rm)
  x <- sweep(x, 1L, sx, "/", check.margin = FALSE)
}
else if (scale == "column") {
  x <- sweep(x, 2L, colMeans(x, na.rm = na.rm), check.margin = FALSE)
  sx <- apply(x, 2L, sd, na.rm = na.rm)
  x <- sweep(x, 2L, sx, "/", check.margin = FALSE)
}
lmat <- rbind(c(NA, 3), 2:1)
lwid <- c(if (doRdend) 1 else 0.05, 4)
lhei <- c((if (doCdend) 1 else 0.05) + if (!is.null(main)) 0.2 else 0, 4)
if (!missing(ColSideColors)) {
  if (!is.character(ColSideColors) || length(ColSideColors) !=
nc)
    stop(paste("'ColSideColors' must be a character vector",
               "of length ncol(x)"))
lmat <- rbind(lmat[1,] + 1, c(NA, 1), lmat[2,] + 1)
lhei <- c(lhei[1L], 0.2, lhei[2L])
}
if (!missing(RowSideColors)) {
  if (!is.character(RowSideColors) || length(RowSideColors) != nr)
    stop(paste("'RowSideColors' must be a character vector",
               "of length nrow(x)"))
lmat <- cbind(lmat[, 1] + 1, c(rep(NA, nrow(lmat) - 1),
                           1), lmat[, 2] + 1)
lwid <- c(lwid[1L], 0.2, lwid[2L])
}
lmat[is.na(lmat)] <- 0
if (verbose) {
  cat("layout: widths = ", lwid, ", heights = ", lhei,
       "; lmat=n")
  print(lmat)
}
dev.hold()
on.exit(dev.flush())
op <- par(no.readonly = TRUE)
on.exit(par(op), add = TRUE)
layout(lmat, widths = lwid, heights = lhei, respect = TRUE)
if (!missing(RowSideColors)) {
  par(mar = c(margins[1L], 0, 0, 0.5))
  image(rbind(if (revC)
               nr:1L
             else 1L:nr), col = RowSideColors[rowInd], axes = FALSE)
}
if (!missing(ColSideColors)) {
  par(mar = c(0.5, 0, 0, margins[2L]))
  image(cbind(1L:nc), col = ColSideColors[colInd], axes = FALSE)
}
par(mar = c(margins[1L], 0, 0, margins[2L]))
if (!symm || scale != "none")
x <- t(x)
if (revC) {
  iy <- nr:1
  if (doRdend)
    ddr <- rev(ddr)
  x <- x[, iy]
}
else iy <- 1L:nr
image(1L:nc, 1L:nr, x, xlim = 0.5 + c(0, nc), ylim = 0.5 +
c(0, nr), axes = FALSE, xlab = "", ylab = "", ...)
axis(1, 1L:nc, labels = labCol, las = 2, line = -0.5, tick = 0,
cex.axis = cexCol)
if (!is.null(xlab))
  mtext(xlab, side = 1, line = margins[1L] - 1.25)
axis(4, iy, labels = labRow, las = 2, line = -0.5, tick = 0,
    cex.axis = cexRow)
if (!is.null(ylab))
  mtext(ylab, side = 4, line = margins[2L] - 1.25)
if (!missing(add.expr))
  eval(substitute(add.expr))
par(mar = c(margins[1L], 0, 0, 0))
if (doRdend)
  plot(ddr, horiz = TRUE, axes = FALSE, yaxs = "i", leaflab = "none")
else frame()
par(mar = c(0, 0, if (!is.null(main)) 1 else 0, margins[2L]))
if (doCdend)
  plot(ddc, axes = FALSE, xaxs = "i", leaflab = "none")
else if (!is.null(main))
  frame()
if (!is.null(main)) {
  par(xpd = NA)
  title(main, cex.main = 1.5 * op["cex.main"])
}
invisible(list(rowInd = rowInd, colInd = colInd,
    Rowv = if (keep.dendro && doRdend) ddr,
    Colv = if (keep.dendro && doCdend) ddc)
}

# map by school 1 (setup).R
#
# Goal: create a table with columns for school, lat, lng, sum of each tagname
# and total number of dissertations
#
# Load required packages
require(doBy)

# Begin wrapper function
maptags1 <- function (dataset_name="noexcludes", tagset_name="tagnames") {

  # 0. convert parameters into useable values (we'll use the names later,
  # for saving files)
  dataset <- get(dataset_name)
  tagset <- get(tagset_name)

  # 1. sum each method type for all schools.
  a1 <- summaryBy(~School, data=dataset, FUN=sum)

  # limit output columns to those in the relevant tagset
sumnames <- paste0(tagset, ".sum")
a1 <- a1[, which(names(a1) %in% c("School", sumnames))]

# save the output
if(remake_figs) {
  filename <- paste0(dataloc, tagset_name, " tagsums by school, ",
                     dataset_name, ", N", nrow(dataset), ".csv")
  write.csv(a1, file=filename)
}

# 2. count total dissertations for each school
a2 <- summaryBy(Year~School, data=dataset, FUN=length)
names(a2) <- c("School", "DissCount")
head(a2)

if(remake_figs) {
  filename <- paste0(dataloc, "disses by school, ", dataset_name,
                     ", N", nrow(dataset), ".csv")
  write.csv(a2, file=filename)
}

# 3. if possible, load file with school names and lat/lng data, created
# by geocode.R, which is much faster than geocoding anew.
if (!exists("all_schools.geo")) {
  action <- readline("Geocoding data is missing. To load a pre-created
                    file, press L; to geocode now, press G.")
  if (tolower(action) == "l") {
    # Load the existing file
    invisible(readline("Select the geocoding csv file from geocode.R.
                       (Filename is like 'geocoding by school,
                        noexcludes, N2711.csv'; press <Enter> when
                        ready."")
    all_schools.geo <<- read.csv(file=file.choose())
    # trim the first column, which is just the row number added on
    # file save
    all_schools.geo <<-
    data.frame(all_schools.geo[2:ncol(all_schools.geo)])
    head(all_schools.geo)
  } else if (tolower(action) == "g") {
    # Do the geocoding now
    if(!exists("geoCodeAll", mode="function")) {
      source(file="geocode.R")
    }
    all_schools.geo <<- geoCodeAll(dataset_name)
  } else {
    warning("Selection for geocoding action not understood;
            trying default for this dataset.")
    filename <- paste0(dataloc, "geocoding by school, ",
                       "school, ", dataset_name, ", N", nrow(dataset),
                       ".csv")
    write.csv(a1, file=filename)
  }
}
```r
dataset_name, "N", diss.count,.csv)
all_schools.geo <- read.csv(filename)

# trim the first column, which is just the row number, added on
# file save
all_schools.geo <-
data.frame(all_schools.geo[2:ncol(all_schools.geo)])
head(all_schools.geo)

# get clean column names
names(all_schools.geo) <- c("School", "Lat", "Lng", "City", "State",
                          "City.State")
head(all_schools.geo)

# 4. stitch together steps 1-3, inner join to eliminate schools left over
# from false positives. this should give us a geocoded index of schools
# with columns for total disscount and for counts of each tag in the
# tagset.
a4 <- merge(all_schools.geo, a1, by="School")
a4 <- merge(a4, a2, by="School")

# 5. return the merged table, since that should be enough to make maps.
# give some sign of success.
print(head(a4))
return(a4)
```

require(mapdata)
require(mapprots)
require(maptools)
require(scales)

# Retrieve variables from other map-related scripts
source(file="carnegie 1 (setup).R")

if(!exists("maptags1", mode="function")) {
  source(file="map by school 1 (setup).R")
}
schools.geo <- maptags1("noexcludes")
consorts.geo <- maptags1("consorts")
head(cdoc2010.geo)
head(consorts.geo)

bins <- cut(schools.geo$DissCount, c(1,2,5,10,50,100, 1000), right=FALSE)

# fix weirdness in schools reported as having doctoral programs but no
# doctorates awarded
  disses.all.fields <- cdoc2010$PROF_D + cdoc2010$SOC_D + cdoc2010$STEM_D +
                      cdoc2010$HUM_D
  cc.doctotal <- cdoc2010$DOCTOT
  argh <- which(disses.all.fields < cc.doctotal)
  data.frame("Sum of disses by field" = disses.all.fields[argh],
             "Reported diss total" = cc.doctotal[argh])
  disses.all.fields[argh] <- cc.doctotal[argh]

  argh <- which(disses.all.fields == 0)
  cdoc2010 <- cdoc2010[-argh,]
  cdoc2010.geo <- cdoc2010.geo[-argh]
  disses.all.fields <- disses.all.fields[-argh]
  rm(argh)

cc.bins <- cut(disses.all.fields, c(1,2,5,10,50,100, 1000), right=FALSE)

# set up color ramp for schools.geo (comp/rhet data)
grays <- gray(length(levels(bins)):0 / length(levels(bins)))
realcolors <- c()
for (i in 1:length(bins)) {
  bin.index <- which(levels(bins)==bins[i])
  realcolors <- c(realcolors, grays[bin.index])
}

# set up color ramp for cdoc2010 (schools in carnegie doctoral classes)
cc.realcolors <- c()
for (i in 1:length(cc.bins)) {
  bin.index <- which(levels(cc.bins) == cc.bins[i])
  cc.realcolors <- c(cc.realcolors, grays[bin.index])
# First graph: superimposing all 10 years of C/R data onto 2010 Carnegie schools
if(remake_figs) {
    filename <- paste0(imageloc, "comp-rhet schools superimposed ",
                       "on carnegie2010 doctoral schools ", Sys.Date(),
                       ".pdf")
    pdf(file = filename)
}

par(mfrow = c(1,1))
# set up background map
map("worldHires", c("usa","Canada"),
xlim = c(-135,-53),
ylim = c(23,58),
col = "gray40",
fill = FALSE)
map("worldHires", c("Mexico"),
xlim = c(-135,-53),
ylim = c(23,58),
col = "gray70",
fill = FALSE,
add = TRUE)
map("state",
    boundary = FALSE,
col = "gray70",
add = TRUE)

# CC doctoral institutions as of 2010, as upward-facing triangles
points(x = cdoc2010.geo$Lng,
y = cdoc2010.geo$Lat,
pch = 24,
col = "gray10",
bg = cc.realcolors)
# Rhet/comp Consortium, as downward-facing triangles
points(x = consorts.geo$Lng,
y = consorts.geo$Lat,
pch = 6,
cex = 1,
col = "black")
# Dissertation dataset as circles
points(x = schools.geo$Lng,
y = schools.geo$Lat,
col = "gray10",
pch = 21,
bg = realcolors)

# Add labels
legend(x = "bottomright",
  title = "Dissertations \n per school, \n2001-2010",
  legend = c("1", "2-4", "5-9", "10-49", "50-100", "100+")
)
fill = grays,
bt = "n"
legend(x = "bottomleft",
  legend = c("Doctoral programs", "R/C dissertations",
  "Consortium of R/C"),
pch = c(24, 21, 6),
bg = alpha("white", 0.3),
box.lty = "blank"
)
title(main = paste("Most doctoral proogram in the US",
  "\n now have some rhet/comp dissertations"),
sub = paste("List of doctoral programs from Carnegie",
  "classification, IPGRAD2010 > 11")
)
if(remake_figs) {
  dev.off()
}
#############################################################
# method tag array.R
#
# GOAL: given method terms in one column, create and append an array of tag
# labels, 0 or 1, and append columns for Method Count and Exclude Level
# (0=keep, 1=maybe throw out, 2=throw out). Note that this used to be done in
# GoogleRefine, but I want it more automate-able.
#
# This file is sourced during `dataprep 2 - load data.R`
####
parse_tags <- function(data) {
  # Check that the columns we're adding don't already exist
  while(any(names(data) %in% tagnames)) {
    c <- readline(paste("Looks like data has already been parsed.",
                       "\nOverwrite (O) or Abort (A)? \nparse_tags > "))
    if(c == "A") {
      warning("Parse_tags not applied; data already parsed.")
      return()
    } else if (c == "O") {
      break
    } else {
      print(noquote("I do not understand your response. Try again?"))
    }
  }
  # Create a data frame to hold the updated info; we'll merge later.
  tags <- data.frame(
    "Pub.number" = data["Pub.number"],
    "Clin" = 0,
"Crit" = 0,
# "Cult" = 0,
"Disc" = 0,
"Ethn" = 0,
"Expt" = 0,
"Hist" = 0,
"Intv" = 0,
"Meta" = 0,
"Modl" = 0,
"Phil" = 0,
"Poet" = 0,
"Prac" = 0,
"Rhet" = 0,
"Surv" = 0,
"Othr" = 0,
"Method.Count" = 0,
"Exclude.Level" = 0
)

# For each method tag, deduce from Method.Terms what tags are present.
mt <- data[, "Method.Terms"]
searchterms <- c("Clinical", "Hermeneutical",
    # "Cultural",
    "Discourse", "Ethnographic", "Experimental",
    "Historical", "Interview", "Meta-Analy", "Model",
    "Philosophical", "Poetic", "Practitioner", "Rhetorical",
    "Survey", "Other")
searchresults <- lapply(searchterms, FUN=function(x) {
    grep(x, mt, ignore.case=F) })

## bug-hunting
# grep("Clinical", a[,"Method.Terms"], ignore.case=F)
for (i in 1:length(searchresults)) {
    tags[searchresults[[i]], i+1] <- 1
}

# Populate Method.Count by summing across each row
for (i in 1:nrow(tags)) {
    tags[i,"Method.Count"] <- sum(tags[i,tagnames])
}

# Account for Method.Count==0, which means that the only tag was excluded
# above
print(noquote(paste("Converting","dissertations with solo tags now excluded from the 
    schema to solo 'Other'")))
zeroindex <- which(tags$Method.Count==0)
tags[zeroindex, "Othr"] <- 1
tags[zeroindex, "Method.Count"] <- 1

# Populate Exclude.Level
el <- grep("xclude", mt, fixed=T)
tags[el, "Exclude.Level"] <- tags[el, "Exclude.Level"] + 2
cbind(bigarray[el,"Method.Terms"], tags[el,"Exclude.Level"])

e12 <- grep("xclude ?", mt, fixed=T)
tags[e12, "Exclude.Level"] <- tags[e12, "Exclude.Level"] - 1

# make sure it worked
head(data.frame(
    "Method.Terms" = data[which(tags$Exclude.Level > 0), "Method.Terms"],
    "Exclude.Level" = tags[which(tags$Exclude.Level > 0), "Exclude.Level"]
), 30)

# # Explore the data
# table(tags["Exclude.Level"])
# table(tags["Method.Count"])
# a <- which(tags$Method.Count == 1)
# lapply(tags[a,tagnames],sum)
# a <- which(tags$Cult == 1 & tags$Method.Count ==1)
# noexcludes[a,c("Title","ABSTRACT")]
# cbind(noexcludes[a,c("Method.Terms","Method.Count")],
# tags[a, "Method.Count"])
# head(tags)
# head(mt)

## Satisfied that the foregoing worked, let's merge
data[, names(tags)] <- tags
return(data)

### end of wrapper function parse_tags()

#############################################################################
# methodcount barplot.R
#
# Produces a bar plot of method-tag counts per dissertation
#
methods.barplot <- function(dataset_name="noexcludes",
tagset_name="tagnames")
{
    if(tagset_name=="tagnames" || tagset_name=="tagnames.long") {
        data <- get(dataset_name)$Method.Count
    } else if(tagset_name=="tagnames.simple") {
        data <- get(dataset_name)$Counts.simple
else {
  stop("Error: I don't know tagset ", tagset_name, ":")
}

data.t <- table(data)
rows <- length(data)

if (remake_figs) {
  filename <- paste0(imageloc, "method count barplot, ",
                      dataset_name, " (N ", rows, ").pdf")
  pdf(file = filename)
}

# basic barplot
barplot(data.t,
        main = "Most Dissertations use Multiple Methods",
        sub = dataset_name,
        xlab = "Method Tags Assigned",
        ylab = "Dissertations",
        las = 1 # labels always horizontal
      )

# add labels inside tall bars but above short bars
for(i in 1:length(data.t)) {
  if(data.t[[i]] > 25) {
    text(1.2*i-0.5, data.t[[i]] - 20, data.t[[i]])
  } else {
    text(1.2*i-0.5, data.t[[i]] + 20, data.t[[i]])
  }
}

data.mean <- mean(data)
data.sd <- sd(data)

# add some stats
mtext(side = 4,
      las = 1,
      adj = 1,
      text = paste("mean =", round(data.mean,2), "\n",
                   "sd =",round(data.sd,2), "\n",
                   "N =", rows)
    )

if (remake_figs) { dev.off() }

} # end of wrapper function methods.barplot()

# (noautorun) {
  remake_figs
  methods.barplot("noexcludes")
methods.barplot("consorts")
methods.barplot("nonconsorts")
methods.barplot("consorts.plus")
}

# Code to control command-line Mallet from within R.
# original by Ben Marwick (https://gist.github.com/benmarwick/4537873)
# forked for MacOS by Jeremiah Ory
# (https://gist.github.com/drlaboratory/6198388)
# forked again, and currently, by Ben Miller
# (https://github.com/benmiller314)
###
if(autorun) {
  # Load required libraries
  library(foreach)

  ## Step 1. Set up parameters we might want to change often ##
  # 1a. Which dataset to examine
  datasets <- c("top.nonconsorts", "noexcludes", "nonconsorts")

  # 1b. How many topics? Set kseq to a sequence to try several options.
  kseq <- c(100, 150, 200, 500)

  # 1c. optimisation interval for MALLET to use
  # (These choices from Mimno's library(mallet))
  optint <- 20
  optburnin <- 50
  numiterations <- 250

  ## Step 2. Set up stable elements of the working environment
  # 2a. Let's assume we're typically going to need more Java heap space;
  # this sets the maximum allocation
  heap_param <- paste("-Xmx","2g",sep="")
  options(java.parameters=heap_param)

  # 2b. Configure variables and filenames for MALLET
  # where is MALLET, and what is the command that runs it?
  MALLET_HOME <- "/Users/benmiller314/mallet-2.0.7"
  mallet_cmd <- paste0(MALLET_HOME, "/bin/mallet")

  # Loop through each dataset and (3) import instances then
  # (4) build models w/varrying numbers of topics.
foreach(dataset_name = datasets) %do% {
  # 3a. Locate the folder containing txt files for MALT to work on.
  importdir <-
  paste0("/Users/benmiller314/Documents/fulltext_dissertations/clean_",
          dataset_name, "_only")

  # 3b. Import the instance list. This will be stable for a given dataset,
  # regardless of the number of topics.
  output <- paste0(MALLET_HOME, "/", dataset_name, "_instances.mallet")

  # Check to see if the instance list has already been imported. If so,
  # then system(scope) will return 0; otherwise, run the import script now.
  scope <- paste("cd", shQuote(sourceloc),
                  "; cd 'Shell scripts and commands'; ls ", output)

  if (system(scope)) {
    import <- paste(mallet_cmd, "import-dir --input", importdir,
                    "--output", output,
                    "--keep-sequence --remove-stopwords")
    go <- readline(paste("About to import instance list.",
                          "Is that what you meant to do? (Y/N)\n"))
    if(tolower(go) != "y") {
      stop("Never mind, then.")
    }
    print("Beginning import now...")

    # If successful, report back.
    if(! system(import)) {
      print("Done."")
    }
  } else {
    print("Oh, good, the instance file exists. Moving on...")
  }

  # Train the model. Topic-number dependent.
  # 4a. Start looping for each number of topics.
  # kseq is defined at the top of this file.
  foreach(k = kseq) %do% {
    ntopics <- k

    # 4b. File names for output of model (extensions must be as shown)
    outputstate <- paste0(dataset_name, "k", k, "_topic_state.gz")
    outputtopickeys <- paste0(dataset_name, "k", k, "_keys.txt")
    outputdoctopics <- paste0(dataset_name, "k", k, "_composition.txt")
    wordtopics <- paste0(dataset_name, "k", k, "_wordtopics.txt")

    # 4c. String together command to send to MALT via the shell
    train <- paste(mallet_cmd, "train-topics --input", output,
                   "--num-topics", ntopics,}
"--optimize-interval", optint,
"--optimize-burn-in", optburnin,
"--output-state", outputstate,
"--output-topic-keys", outputtopickeys,
"--num-iterations", numiterations,
"--output-doc-topics", outputdoctopics,
"--word-topic-counts-file", wordtopics)

# 4d. Run the command in the shell.
  system(train)
}

# close the loop of datasets
} else {
  message("Autorun is FALSE, so no action was taken.
  message(paste("If you wish to create new topic models,
                check configuration, then set autorun to TRUE."))
}

#########################################################################
# simplifying the schema.R
#
# Goal: Group tags to pool influence of (e.g.) quantitative approaches.
#
# Possible groups: Aggregable (disc, expt, surv, meta), Phenomenological
# (case, ethn), Dialectical (crit, hist, modl, phil, rhet), Craft-Based
# Drop cult and intv, move meta to Agg. As an alternative pool, cf. Michael
# Carter's "Ways of Knowing and Doing in the Disciplines."
####

# define shortcut for new tag names
  tagnames.simple <- c("Aggreg", "Phenom", "Dialec", "Crafty", "Pract")
  tagnames.simple.long <- c(
    "Aggregable",
    "Phenomenological",
    "Dialectical",
    "Craft-based",
    "Practitioner / Teacher Research")

# wrapper function to add these tags to existing tag array
  short_schema <- function(data) {
    # Check that the columns we're adding don't already exist
    while(any(names(data) %in% tagnames.simple)) {
      c <- readline(paste("Looks like data has already been parsed.
                        Overwrite (O) or Abort (A)? \nshort_schema > "))
      if(c == "A") {
        warning("Short_schema not applied; data already parsed.")
      } else {
        # Check that the columns we're adding don't already exist
        while(any(names(data) %in% tagnames.simple)) {
          c <- readline(paste("Looks like data has already been parsed.
                              Overwrite (O) or Abort (A)? \nshort_schema > "))
          if(c == "A") {
            warning("Short_schema not applied; data already parsed.")
          } else {
            # Add the tags to the data
            data <- cbind(data, tagnames.simple)
          }
        }
      }
    }
  }
  short_schema(train)
return()
} else if (c == "O") {
    break
} else {
    print(noquote("I do not understand your response. Try again?"))
}

# Create a data frame to hold the updated info; we'll merge later.
simple <- data.frame(  
    Pub.number = data["Pub.number"],
    Aggreg = -1,  # Aggregable
    Phenom = -1,  # Phenomenological
    Dialec = -1,  # Dialectical
    Crafty = -1,  # Craft-Based
    Pract = -1,  # Practitioner
    # (That last one is a little redundant, but
    # it makes `simple` simpler.)
    Counts.simple = -1
)
head(simple)

# For each data row, record if it's tagged with any member of each group
for (i in 1:nrow(data)) {
    # Aggregable
    a1 <- as.integer(data[i,"Disc"])  
    a2 <- as.integer(data[i,"Expt"])  
    a3 <- as.integer(data[i,"Surv"])  
    a4 <- as.integer(data[i,"Meta"])  
    a <- max(a1, a2, a3, a4)  
    ag <- simple[i,"Aggreg"] <- a

    # Phenomenological
    a1 <- as.integer(data[i,"Clin"])  
    a2 <- as.integer(data[i,"Ethn"])  
    a <- max(a1, a2)  
    ph <- simple[i,"Phenom"] <- a

    # Dialectical
    a1 <- as.integer(data[i,"Crit"])  
    a2 <- as.integer(data[i,"Hist"])  
    a3 <- as.integer(data[i,"Modl"])  
    a4 <- as.integer(data[i,"Phil"])  
    a5 <- as.integer(data[i,"Rhet"])  
    a <- max(a1, a2, a3, a4, a5)  
    di <- simple[i,"Dialec"] <- a

    # Craft-Based
    a1 <- as.integer(data[i,"Poet"])  
    # a2 <- as.integer(data[i,"Prac"])
a3 <- as.integer(grep("tool-building", data[i,"Method.Terms"],
    ignore.case=TRUE))

a <- max(a1,
    # a2,
    a3)

CR <- simple[i,"Crafty"] <- a

# Practitioner
pr <- simple[i, "Pract"] <- as.integer(data[i,"Prac"])

# Now look for multi-modality across these broad categories
simple[i, "Counts.simple"] <- sum(ag, ph, di, cr, pr)

# # Clean up the workspace (not needed after testing)
# rm(a, a1, a2, a3, a4, a5, ag, c, ph, di, cr, pr)

data[, names(simple)] <- simple
return(data)

}  # end of wrapper function short_schema()

## Confirm the function works properly
# data <- head(bigarray)
# data
# data <- short_schema(data)
# data
# data[, tagnames.simple] <- -1
# data
# data <- short_schema(data)
# data
# rm(data, simple)

## Explore the newly configured data
# table(simple$Counts.simple)
# mean(simple$Counts.simple)
# noexcludes[which(simple$Counts == 0),c("Method.Terms",tagnames)]
# names(noexcludes)
# colSums(simple[2:(ncol(simple)-1)])

########################################################################

# single topic strength vs rank.R
#
# Goal: Find out the correlation of topic strengths to topic ranks within
# each document, **for a single topic** i.e. how much of the document is the
# top topic? how much is the second? and so on, aggregated over all
# documents, as a scatter plot of contribution (y-axis) vs. topic rank
# (x-axis).
# Rationale: I want to know whether some topics with high overall rank are
# secretly low-but-consistent across lots of docs

####

\[
\text{strength_v_rank <- function(my.topic,} \\
\text{  dataset_name = "consorts",} \\
\text{  ntopics = 55,} \\
\text{  bad.topics = NULL} \\
\text{  )} \\
\]

{ # Exclude non-content-bearing topics
  if(is.null(bad.topics) && dataset_name == "consorts" && ntopics == 55) {
    bad.topics <- c("4", "47", "22", "2", "24", 
                     "13", "50") # bad OCR or ProQuest boilerplate
    # language markers (Italian, Spanish)
  }

  if(my.topic %in% bad.topics) {
    warning(paste("Topic", my.topic, 
                   "has been identified as non-content-bearing"))
  }

  require(data.table)
  dataset <- get(dataset_name)

  # get all topics by document
  if(!exists("get.doctopic.grid", mode="function")) {
    source("get doctopic grid.R")
  }
  grid <- data.table(get.doctopic.grid(dataset_name, ntopics)$outputfile)
  # str(grid)
  # head(grid)

  grid <- grid[, !(names(grid) %in% c(bad.topics, "Pub.number")), with=F] # head(grid)

  # find rank of my.topic within one document
  rankit <- function(row) {
    o <- order(row, decreasing=TRUE)
    ranked.topics <- names(row[o])
    my.rank <- which(ranked.topics == my.topic)
    return(my.rank)
  }

  # apply rankit function across rows of grid (this will be our set of
  # x-values, in order of documents)
  my.ranks <- apply(grid, 1, FUN=rankit)
  # head(my.ranks)
# for y-values, just read down the column of our topic;  
# have to extract values from list with [[subset]]  
my.contribs <- grid[, names(grid) %in% my.topic, with=F][[1]]  
    # str(my.contribs)  
    # head(my.contribs)  

# set up the plot  
maintitle <- paste("Topic Contribution by Topic Rank")  
subtitle <- paste(dataset_name, ntopics, "topics")  
ymax <- paste(dataset_name, ntopics, "topics")  

    # Use get_topic_labels() to retrieve ranks  
if(!exists("get_topic_labels", mode="function")) {  
    source(file="get_topic_labels.R")  
}  
topic.labels.dt <- get_topic_labels()  
    # head(topic.labels.dt)  
overall.rank <- topic.labels.dt[Topic == my.topic, Rank]  
topic.label <- topic.labels.dt[Topic == my.topic, Label]  

legendtext <- paste0("topic: ", my.topic, \\
    "\n", topic.label, \\
    "\noverall rank: ", overall.rank)  

    # plot it  
if(remake_figs) {  
    filename <- paste0(imageloc, maintitle, " ", subtitle, ".pdf")  
    pdf(filename)  
}  

plot(x = my.ranks,  
    y = my.contribs,  
    main = maintitle,  
    xlab = "Topic Rank within Document",  
    ylab = "Topic Contribution within Document",  
    ylim = c(0, ymax),  
    xlim = c(0, ncol(grid)),
    pch=4)  

mtext(subtitle)  
legend("topright", legendtext, bty="n")  
if(remake_figs) { dev.off() }  

invisible(list(ranks = my.ranks,  
    contribs = my.contribs)  
)
} # end of wrapper function strength_v_rank()

# Now run that function on multiple topics
if(autorun) {
  remake_figs

    # Use get_topic_labels() to retrieve ranks; if we have strength_v_rank,
    # we've already sourced `get topic labels.R`.
    topic.labels.dt <- get_topic_labels()
    head(topic.labels.dt)

    # Prepare a list of topics to apply the function to
    bad.topics <- c("4", "47", "22", "2", "24", "13", "50")
    topic.labels.dt <- topic.labels.dt[!(Topic %in% bad.topics)]
    setkey(topic.labels.dt, Rank)
    topics.by.rank <- head(topic.labels.dt[, Topic], 10)

    # Apply the function to our list of topics
    lapply(topics.by.rank, strength_v_rank)

    # Inspect one topic of interest
    strength_v_rank(my.topic=10)

    topic.labels.dt[Topic==10, Label]
}

} # end of autorun section

########################################################################
# subject terms barplot.R
#
# Goal: Extract subject terms, make a frequency table and calculate some
# stats.
# Dependencies: `extract subjects.R` (sourced during `data prep.R`)  
####

subject.barplot <- function(dataset_name = "noexcludes",
    top.many = 30, # Plot this many terms,
    maxsum = 300, # starting from the top
    pct.deep = 0.5 # Set high to avoid "other"
    )
{
  # How far into the list of
  # terms should we go?
  # Default of 0.5 means
  # halfway, i.e. to the
  # median; for all the way,
  # set pct.deep=1.

    # Get the data
dataset <- get(dataset_name)
subj.list <- extract_subjects(dataset$Subject)

# Get the frequency chart.
# maxsum is needed to avoid "othering" half the list.
subj.table <- summary(subj.list, maxsum=300)

# Put the list in descending order by frequency, and chop out the term
# they all share
subj.table <- sort(subj.table, decreasing=TRUE)
subj.table <- subj.table[2:length(subj.table)]

subj.count <- length(subj.table)
subj.mean <- mean(subj.table)
subj.median <- median(subj.table)

if(remake_figs) {
    filename <- paste0(imageloc, dataset_name,
        " subject terms barplot, ",
        "top ", top.many, ", N", nrow(dataset), ".pdf")
    pdf(file=filename)
}

barplot(sort(subj.table[1:top.many], decreasing = FALSE),
    horiz = TRUE,
    main = paste("Top", top.many, "Subject Terms by Frequency"),
    sub = paste0(dataset_name, ",", N", nrow(dataset)),
    axisnames = TRUE,
    width = c(10,10),
    space = 0.4,
    las = 1,
    pty = "m",
    mai = c(5,10,8,5)
)

for(i in 1:top.many) {
    # if(subj.table[[i]] > 20) {
    #     # add a label where x = frequency and y = "device height" or
    #     # something?
    #     text(x = (subj.table[[i]] + 20),
    #         y = (par()$din[2]),
    #         labels = subj.table[[i]],
    #         pos = 4
    #     )
    # } else {
    #     # text(subj.table[[i]] + 20, -i, subj.table[[i]])
    # }
    # }

mtext(paste(nrow(dataset), "theses,", subj.count,
"subjects, median =", subj.median, ", mean =", subj.mean))

if(remake_figs) {
  dev.off()

  ## Now separately plot the remainder of the terms

  filename <- paste0(imageloc, dataset_name,
                     "subject terms barplot, below",
                     top.many, ", above median, N", nrow(dataset), ".pdf")

  pdf(file=filename)
}

# for some unknown reason, top.many+1 still includes the 30th item. \:
barplot(subj.table[top.many+2:length(subj.table)*pct.deep],
        horiz = FALSE,
        main = paste("Frequency of Subject Terms below", top.many,
                      "but above median"),
        axisnames = TRUE,
        width = c(10,10),
        space = 0.4,
        las = 2,
        pty = "m",
        mai = c(5,10,8,5)
)

if(remake_figs) {
  dev.off()
}

} # end of wrapper function subject.barplot()

if(autorun) {
  remake_figs
  subject.barplot()
}

#########################################################################
# tags by school.R#
#
# GOAL: Given a tagged set of dissertation data and a tagging schema,
# aggregate tag frequency and distribution at each school in the dataset.
# After building the function for the analysis, run it on various subsets of
# data and tags.
#
# load required packages
require(doBy)
require(cluster)
require(RColorBrewer)

# make sure we've run dataprep.R
if(!exists("imageloc")) {
  source(file="start here.R")
}

# function for getting data
schoolwise.data <- function(dataset_name="consorts", tagset_name="tagnames") {

  # 0. convert variable names to variables. we'll use the names later in
  # the figure titles.
  dataset <- get(dataset_name, envir=parent.frame())
  tagset <- get(tagset_name, envir=parent.frame())
  tagset.mean <- sapply(tagset, FUN=function(x) paste0(x,".mean"))
  tagset.sum <- sapply(tagset, FUN=function(x) paste0(x,".sum"))

  # 1. remove columns other than method tags and school
  dataset <- dataset[, which(names(dataset) %in% c("School", tagset))]

  # 2. do the summary of each method type for all schools.
  d1 <- summaryBy(~ School, data=dataset, FUN=mean)
  d2 <- summaryBy(~ School, data=dataset, FUN=sum)
  d3 <- summaryBy(~ School, data=dataset, FUN=length)

  return(list("means" = d1, "sums" = d2, "counts" = d3))
}

# function for graphing data
schoolwise <- function(dataset_name="noexcludes", tagset_name="tagnames",
                        agn=TRUE,       # run agglomerative clustering (using agnes)?
                        hcl=TRUE,       # run hierarchical clustering (using hclust)?
                        dia=TRUE,       # run divisive clustering (using diana)?
                        counts=FALSE,   # label each row with the number of
                                          # dissertations per school?
                        agfixedcols=NULL, # optional pre-set order of columns for
                                          # comparison btwn agnes plots
                        difixedcols=NULL, # optional pre-set order of columns for
                                          # comparison btwn diana plots
                        myCol=NULL)      # optional color palette

  {

    # if colors are not provided, default to black and white
    require(RColorBrewer)
    if(is.null(myCol)) { myCol <- brewer.pal(9, "Greys") }

    # 0. convert variable names to variables. we'll use the names later in
```r
# the figure titles.
dataset <- get(dataset_name)
tagset <- get(tagset_name)
tagset <- sapply(tagset, FUN=function(x) paste0(x,".mean"))

# 1-2 call the data-grabbing function
m1 <- schoolwise.data(dataset_name, tagset_name)
m2 <- m1$means

# 3. get more meaningful row names (and a purely numerical matrix, for # heatmapping) Note that the first column will always be the list of # schools because of the query in step 2.
if (counts) {
  row.names(m2) <- paste0(m2$School, " (", m1$counts$School.length, ")")
} else {
  row.names(m2) <- m2$School
}
m2 <- m2[,2:ncol(m2)]
m2 <- data.matrix(m2)
head(m2)

# try this old approach to finding the best sort method
# agn_methods <- c("average","single","complete","ward","weighted");
# agn <- lapply(agn_methods, FUN=function(x) {
#   agnes(m2, diss=F, metric=x)
# })
# agn_best.index <- max(c(agn[[1]]$ac, agn[[2]]$ac, agn[[3]]$ac,
#                        agn[[4]]$ac, agn[[5]]$ac))

# 4. make the heatmap: use pre-determined columns if need be.

# 4b. divisive clustering (diana):
if(dia) {
  filename <- paste0(imageloc,"tags by schools ", dataset_name, ", N", nrow(dataset), ", ", tagset_name, ", diana.pdf")
maintitle <- paste0("Method Tag Averages by school, ", dataset_name, ", ", tagset_name, ", diana")

  if(remake_figs) {
    pdf(file = filename)
  }

  if(!is.null(difixedcols)) {
    di <- heatmap.fixedcols(m2,
      myColInd = difixedcols,
      hclustfun = function(d) { diana(d, metric="ward") },
      scale = "row",
      col = myCol,
      main = maintitle,
```
margins = c(5,10))

} else {
  di <- heatmap(m2,
    hclustfun = function(d) { diana(d, metric="ward") },
    scale = "row",
    col = myCol,
    main = maintitle,
    margins = c(5,10)
  )
}
mtext(paste("Each cell gives the likelihood that a given",
  "dissertation from the school in row Y is tagged with the",
  "method in column X.", side = 1))

if(remake_figs) {
  dev.off()
}

} # end of if(dia)

# 4a. agglomerative clustering (agnes):
if(agn) {
  filename <- paste0(imageloc, "tags by schools", ",", dataset_name,
    ", N", nrow(dataset), ",", tagset_name, ",", agnes.pdf)
  maintitle <- paste0("Method Tag Averages by school", ",", dataset_name,
    ",", tagset_name, ",", agnes)

  if(remake_figs) {
    pdf(file = filename)
  }

  if(!is.null(agfixedcols)) {
    ag <- heatmap.fixedcols(m2,
      myColInd = agfixedcols,
      hclustfun = function(d) { agnes(d, method="ward") },
      scale = "row",
      col = myCol,
      main = maintitle,
      margins = c(5,10)
    )
  } else {
    ag <- heatmap(m2,
      hclustfun = function(d) { agnes(d, method="ward") },
      scale = "row",
      col = myCol,
      main = maintitle,
      margins = c(5,10)
    )
  }
mtext(paste("Each cell gives the likelihood that",
  "a given dissertation from the school in row Y",
  "is assigned to the method in column X.", side = 1))

if(remake_figs) {
  dev.off()
}

} # end of if(agn)
if(remake_figs) {
  dev.off()
}
# end of if(agn)

# 4c. agglomerative clustering via hclust:
if(hcl) {
  filename <- paste0(imageloc, "tags by schools, ", dataset_name, 
    ", N", nrow(dataset), ", ", tagset_name, ", hclust.pdf")
  maintitle <- paste0("Method Tag Averages by school, ", dataset_name, 
    ", ", tagset_name, ", hclust")

  if(remake_figs) {
    pdf(file = filename)
  }

  if(!is.null(agfixedcols)) {
    hc <- heatmap.fixedcols(m2,
      myColInd = agfixedcols,
      scale = "row",
      col = myCol,
      main = maintitle,
      margins = c(5,10)
    )
  } else {
    hc <- heatmap(m2,
      scale = "row",
      col = myCol,
      main = maintitle,
      margins = c(5,10)
    )
  }

  mtext(paste("Each cell gives the likelihood",
    "that a given dissertation from the school in row Y",
    "is tagged with the method in column X.", side = 1))

  if(remake_figs) {
    dev.off()
  }
}
# end of if(clust)

if(!exists("di", inherits=F)) di <- noquote("Not run")
if(!exists("ag", inherits=F)) ag <- noquote("Not run")
if(!exists("hc", inherits=F)) hc <- noquote("Not run")

# save the row and column orders to allow for consistent sorting later
return(list("di" = di, "ag" = ag, "hc" = hc))
# close wrapper function schoolwise()

if (autorun) {
    # call the functions for all relevant datasets
    schoolwise("consorts", "tagnames", agn=T, hcl=F, dia=F)
    schoolwise("nonconsorts", "tagnames", agn=T, hcl=F, dia=F)
    schoolwise("top.nonconsorts", "tagnames", agn=T, hcl=F, dia=F)
    schoolwise("noexcludes", "tagnames")

    # schoolwise("nonconsorts", "tagnames", agfixedcols=a$ag$colInd,
    #    difixedcols=a$di$colInd)
    schoolwise("consorts.plus", agn=T, hcl=F, dia=F)

    # next up: re-run with the simplified schema
    schoolwise("consorts", "tagnames.simple")
    schoolwise("nonconsorts", "tagnames.simple")
    schoolwise("noexcludes", "tagnames.simple")

    # schoolwise("consorts", "tagnames.simple", agfixedcols=c$ag$colInd,
    #    difixedcols=c$di$colInd)
    schoolwise("nonconsorts", "tagnames.simple", agfixedcols=c$ag$colInd,
    #    difixedcols=c$di$colInd)

    # explore the data
    # d <- order(c$byschool$Aggreg.mean, decreasing=TRUE)
    # schoolwise("consorts", "tagnames.simple", agfixedcols=d, difixedcols=d)

    # c$byschool[c$di$rowInd, which(names(c$byschool) %in%
    #    sapply(tagnames.simple, FUN=function(x) paste0(x,".mean"))]
    # which(rowsum(c$byschool, row.names(c$byschool)) == 0)
    # ?rowsum
    # remove interim variables
    # rm(m1, m2, m3, noex.by.school.m, nonconsorts.by.school.m,
    #   consorts.by.school.m)

    #------------------------------------------------------------------------
    # thresh.R
    #
    # Provides a function to subset data by threshold number of dissertations
    # in a given timespan
    ######

    thresh <- function(dataset_name = "noexcludes",
                        tagset_name = "tagnames",
                        threshold = 5,
                        since = 2006,
                        until = 2010)
    {

# load required packages
require(doBy)

# 0. convert variable names to variables. we'll use the names later in
# the figure titles.
dataset <- get(dataset_name)
tagset <- get(tagset_name)

# 1. subset the data for the desired years
d1 <- dataset[which((dataset$Year >= since) & (dataset$Year <= until)),]

# 1b. summarize that data by school, counting rows
d2 <- summaryBy(~ School, data=d1, FUN=length)

# 1c. find the schools in that time period with more than threshold
# (default=5)
d3 <- d2[which((d2$Year.length >= threshold), "School"]
thresh.report <- paste0("Found ", length(d3), " schools (out of ",
                      nrow(d2), " schools in ", dataset_name,
                      " from ", since, "-", until, ") with ",
                      threshold, " or more dissertations.")

## 2. get full 10-year tag data for those schools
d4 <- dataset[which(dataset$School %in% d3),]

return(list("thresh.data" = d4,
            "thresh.report" = thresh.report))

#############################################################################
#
# top docs per topic.R
#
# Tools for topic exploration
#
# Provides three functions:
#   * get.doc.composition(dataset, ntopics): retrieves a pre-existing
#     matrix, output by MALLET, with topic proportions for each
#     document in corpus
#   * get.topics4doc(pubnum, dataset_name, ntopics, howmany,
#     showlabels): retrieves top `howmany` topics for a document
#     specified by `pubnum`.
#   * top_topic_browser(...): for a specified topic or range of topics,
#     shows the top `howmany` documents and their method tags, with
#     optional detail view showing top topics for each document at a
#     time. See below for parameters.
#
####

# Step 1. Get the matrix of texts and topics
if(!exists("get.doctopic.grid", mode="function")) {
  source(file="get doctopic grid.R")
}

# 2. Oh, and what were those topics, again?
if(!exists("get.topickeys", mode="function")) {
  source(file="get topickeys.R")
}

# Step 3. Find top 5 docs for each overall top topic
# to get a sense of what's "real" and what's "interesting"

  # Step 4. Find all the top-ranked topics for those docs: maybe that
  # really popular topic isn't actually the main component of the docs that
  # come up.

  # We start with the doc-topic matrix from MALLET:
get.doc.composition <- function(dataset_name="consorts", ntopics=55) {
  # get packages in case we've just restarted R
  require(data.table)

  filename <- paste0(malletloc, "/", dataset_name, "k", ntopics,
                     "_composition.txt")
  doc_topics <- read.delim(filename, header=F, skip=1)
  head(doc_topics)

  # column 1 is an unneeded index; column 2 contains names of identical
  # length, ending with a 7-digit Pub.number followed by ".txt"; final
  # column is empty. Let's simplify.
  doc_topics[, "V1"] <- NULL
  len <- nchar(as.character(doc_topics[1, "V2"]))
  doc_topics[, "V2"] <- substr(as.character(doc_topics[, "V2"]),
                                (len-10), (len-4))
  if (is.na(all(doc_topics[, ncol(doc_topics)]))) {
    doc_topics[, ncol(doc_topics)] <- NULL
  }

  # Get findable column names
  colnames(doc_topics)[1] <- "Pub.number"
  colnames(doc_topics)[seq(2, ncol(doc_topics), 2)] <- paste0("top",
                           seq(1, (ncol(doc_topics)-1)/2, 1))
  colnames(doc_topics)[seq(3, ncol(doc_topics), 2)] <- paste0("wgt",
                           seq(1, (ncol(doc_topics)-1)/2, 1))
  head(colnames(doc_topics))

  # convert to 1-indexed from MALLET's 0-indexed, so everything matches
  doc_topics[, seq(2, ncol(doc_topics), 2)] <-
  (doc_topics[, seq(2, ncol(doc_topics), 2)]+1)
# for some reason, it thinks the topic weights are characters. They're numbers.
doc_topics[, seq(3, ncol(doc_topics), 2)] <-
apply(doc_topics[, seq(3, ncol(doc_topics), 2)], 2,
  FUN=function(x) { x <- as.numeric(x) })

doc_topics.dt <- as.data.table(doc_topics)
setkey(doc_topics.dt, Pub.number)
head(doc_topics.dt)

return(doc_topics.dt)

# Run `get.doc.composition()` when file is sourced, so we don't have to
# recreate this multiple times for the same dataset if we're running
# `top_topic_browser()` using the `for.bind` option.
# TO DO: Make this happen within get.doc.composition() -- i.e. give the
# function the side effect of creating this object -- so it's responsive to
# dataset_name and ntopics.
doc_topics_consorts_55.dt <- get.doc.composition("consorts", 55)

### Helper function: retrieve top five topics for a given Pub.number
get.topics4doc <- function(pubnum, 
  dataset_name = "consorts",
  ntopics = 55,
  howmany = 5,
  showlabels = FALSE)
{
  # get packages in case we've just restarted R
  require(data.table)

  # pubnum <- "3051708"; doc_tops <- doc_topics.dt    # test values
  if (!is.character(pubnum)) {
    pubnum <- as.character(pubnum)
  }

doc_tops <- get.doc.composition(dataset_name, ntopics)
topic_keys <- data.table(get.topickeys(dataset_name, ntopics))
topic_keys <- topic_keys[as.numeric(doc_tops[pubnum, paste0("top",
    1:howmany), with=F])]
topic_keys[,weight:=as.numeric(doc_tops[pubnum, paste0("wgt", 1:howmany),
    with=F])]
topic_keys <- topic_keys[, list(topic, weight, alpha, top_words)]

  if(showlabels) {
    if (exists("get_topic_labels", mode="function")) {
      source(file="get topic labels.R")
    }
  }
}
topic_labels <- data.table(get_topic_labels(dataset_name, ntopics),
    key="Topic")

topic_keys[, current_label:=topic_labels[topic_keys$topic, Label]]

topic_keys <- topic_keys[, list(topic, weight, alpha, current_label, top_words)]

list("title" = noexcludes.dt[pubnum, c("Title", "Pub.number", tagnames),
    with=F],
    "keys" = topic_keys,
    "abstract" = noexcludes.dt[pubnum, c("KEYWORDS", "ABSTRACT"), with=F] )

# close helper function get.topics4doc
}

### Browse through the top topics and their top-proportioned dissertations
top_topic_browser <- function(
    # assuming we're looping, start where?
    start.rank = 1,

    # alternately, browse one specified topic
    topic = NULL,

    dataset_name = "consorts",
    ntopics = 55,

    # if lots of topics, where to stop?
    cutoff = get("ntopics"),

    # how many docs to show for each topic?
    depth = 5,

    # show current topic labels for indiv. docs?
    showlabels = FALSE,

    # invisibly return results and exit early?
    for.bind = FALSE
)
{
    # get packages in case we've just restarted R
    require(data.table)

    # load the data from the functions defined or imported above
    doc_composition <- paste0("doc_topics_", dataset_name, ":", ntopics, ".dt");

    if(!exists(doc_composition)) {
        doc_topics.dt <- get.doc.composition(dataset_name, ntopics)
    } else {
        doc_topics.dt <- get(doc_composition)
```r

# List the keys for the top N topics, where N = cutoff
len <- min(length(colsums)-1, cutoff)

# list of topics by rank; skip Pub.num
ind <- as.integer(names(colsums.sort)[2:(len+1)])

# If we specified a topic, show just that topic and exit.
if (!is.null(topic)) {
  topic.num <- topic

  # find and display topic rank
  topic.rank <- which(ind %in% topic.num)
  if (remake_figs) {
    print(paste0("Topic of rank ", topic.rank, ":"))
  } else {
    message("\nTopic of rank ", topic.rank, ":\n")
  }
}

# get Pub.numbers for dissertations with the max proportion of that topic
row.ind <- order(outputfile [, which(names(outputfile)==topic.num)], decreasing=TRUE)[1:depth]
diss.ind <- outputfile[row.ind, "Pub.number"]

print(topic_keys.dt[topic.num])

# list of top 1:depth documents for this topic
topdocs <- noexcludes.dt[as.character(diss.ind),
  c("Pub.number", "Title", tagnames), with=F]

  # add a column with the weights this topic has in these docs
doc_tops <- get.doc.composition(dataset_name, ntopics)
weights <- ranks <- c()
for(j in 1:length(diss.ind)) {
  topic.col <- match(topic.num, 
    doc_tops[as.character(diss.ind)][j])
  weights[j] <- doc_tops[as.character(diss.ind)][j, (topic.col+1), with=F]
  ranks[j] <- topic.col/2
}
```
topdocs[, topic_weight:=unlist(weights)]
topdocs[, rank_in_doc:=unlist(ranks)]
topdocs <- topdocs[, c("Pub.number", "Title", "topic_weight", "rank_in_doc", tagnames), with=F]

print(topdocs)

# if we're just looking at one topic, maybe we want to save that list
# of docs and their metadata, and exit.
if(for.bind) {
  return(topdocs)
}

# if we're saving all output, automatically cycle through everything.
# but by default, prompt the user.
if(!remake_figs) {
  a <- readline(paste("Press <enter> for more detail on",
                      "these docs, or S to skip to the next topic\n"))
} else {
  a <- ""
}

while (tolower(a) != "s") {
  for(i in topdocs$Pub.number) {
    print(get.topics4doc(i, dataset_name, ntopics,
                         showlabels = showlabels))
    if (!remake_figs) {
      a <- readline(paste("Press <enter> for next doc,",
                           "D for more details, or",
                           "S to skip to the next topic\n"))
    } else {
      a <- ""
    }
  }
  if (tolower(a) == "s") {
    break
  } else if (tolower(a) == "d") {
    print(noexcludes.dt[i])
    a <- readline(paste("Press <enter> for next doc",
                         "or S to skip to the next topic\n"))
  }
  a <- "s"
}

} else {
  # If we haven't pre-specified a topic, loop through the top topics
  # and their top-proportioned dissertations, optionally showing
  # abstracts and top 5 topics for each of those dissertations
  message("Top ", cutoff, " topics:")
  print(topic_keys.dt[ind])  # top words for each topic
for (i in start.rank:len) {
    # i gives the topic rank
    topic.num <- ind[i]

    # Search outputfile for the dissertations with max proportion of that
    # topic, and get the Pub.numbers
    row.ind <- order(outputfile[, which(names(outputfile)==topic.num)],
                     decreasing=TRUE)[1:depth]
    diss.ind <- outputfile[row.ind, "Pub.number"]

    if (remake_figs) {
        print(paste0("Topic of rank ", i, ":"))
    } else {
        message("\nTopic of rank ", i, ":\n")
    }

    print(topic_keys.dt[topic.num])

    # list of top 1:depth documents for this topic
    topdocs <- noexcludes.dt[as.character(diss.ind),
                              c("Pub.number", "Title", tagnames), with=F]

    # add a column with the weights this topic has in these docs
    doc_tops <- get.doc.composition(dataset_name, ntopics)
    weights <- ranks <- c()
    for(j in 1:length(diss.ind)) {
        topic.col <- match(topic.num,
                           doc_tops[as.character(diss.ind)][j])
        weights[j] <- doc_tops[as.character(diss.ind)][j, (topic.col+1), with=F]
        ranks[j] <- topic.col/2
    }

    topdocs[, topic_weight:=unlist(weights)]
    topdocs[, rank_in_doc:=unlist(ranks)]
    topdocs <- topdocs[, c("Pub.number", "Title", "topic_weight", "rank_in_doc", tagnames), with=F]

    print(topdocs)

    if (!remake_figs) {
        a <- readline(paste("Press <enter> for more detail",
                           "on these docs, or S to skip to the next topic\n"))
    } else {
        a <- ""
    }

    while (tolower(a) !="s") {
        for(i in topdocs$Pub.number) {

print(get.topics4doc(i, showlabels=showlabels))
if (!remake_figs) {
    a <- readline(paste("Press <enter> for next doc, ",
                       "D for more details, or ",
                       "S to skip to the next topic\n"))
} else {
    a <- ""
}
if (tolower(a) == "s") {
    break
} else if (tolower(a) == "u") {
    i <- i-1
} else if (tolower(a) == "d") {
    print(noexcludes.dt[i])
    a <- readline(paste("Press <enter> for next doc or ",
                         "S to skip to the next topic\n"))
}

a <- "s"
}  # end of while loop (of documents)
}  # end of for loop (of topics)
}  # end of if/else for specific topic or all topics

}  # end of wrapper function top_topic_browser()

## Run the big browser function above
if (autorun) {
    dataset_name <- "consorts"
    ntopics <- 55
    if (remake_figs) {
        filename <- paste0(imageloc, "top topics - ", dataset_name, ", K ",
                           ntopics, ".txt")
        readline(paste("About to capture browser output as ", filename,"- 
                      <enter> to continue or <esc> to abort. "))
        capture.output(top_topic_browser(), file=filename)
    } else {
        top_topic_browser()
    }
}

#########################################################################
# top schools by method.R
#
# GOAL: For each method in a given tagset, produce a list of the top X
# schools by either methodological output (number of dissertations using
# that method at that school) or methodological focus (percentage of
# dissertations using that method at that school)
####
# load required packages
require(doBy)

# open wrapper function
toplists <- function(dataset_name = "noexcludes",
tagset_name = "tagnames",

  # How many schools in the list for each method?
  howmany = 5,

  # Set a minimum number of dissertations per school...
  threshold = 5,

  # ... over a specified span of years. (gets passed to
  # thresh() from `thresh.R`
  since = 2006,
  until = 2010,

  # Use methodological focus (T) or raw output (F)?
  rank_by_pcts = TRUE,

  # Display focus and output in one column (T) or two (F)?
  combine = TRUE)
{

  ## 0. convert variable names to variables. we'll use the names later in
  ## the figure titles.
  dataset <- get(dataset_name)
  tagset <- get(tagset_name)

  ## 1. find schools with more than (by default) 5 dissertations in
  ## 2006-2010
  if(!exists("thresh", mode="function")) {
    source(file="thresh.R")
  }

d <- thresh(dataset_name, tagset_name, threshold, since, until)
d1 <- d$thresh.data
subtitle1 <- d$thresh.report

  if (!exists("schoolwise.data")) {
    source(file="tags by school.R")
  }
  a <- schoolwise.data("d1", tagset_name)

  # 2 Star the schools in the consortium
  c <- which(a$counts$School %in% consorts$School)
a$counts$School <- fix_factor(a$counts$School,
  to.add = paste0(a$counts$School[c], "*")
  to.remove = a$counts$School[c])

## 3. for the schools that meet the cutoff, find the "howmany" highest
## real values of each tag

# 3a. Create function to apply to each tag in the tagset
toplist.onetag <- function(a, tag, rank_by) {
  tag.mean <- paste0(tag, ".mean")
  tag.sum <- paste0(tag, ".sum")

  # rank by chosen tag
  if(rank_by) {
    a1 <- order(a$means[, tag.mean], decreasing=TRUE)
  } else {
    a1 <- order(a$sums[, tag.mean], decreasing=TRUE)
  }

  # raw number of disses at top schools
  a2 <- head(a$counts[a1,], howmany)

  # number of disses with chosen tag
  a3 <- head(a$sums[a1,tag.sum], howmany)

  # pct of disses with chosen tag
  a4 <- head(a$means[a1, tag.mean], howmany)

  # cleaner percentage
  a4 <- round(100*a4, 0)

  if(combine) {
    # combine per-tag count and pct
    if(rank_by) {
      a5 <- paste0(a4, "% (" , a3,"))
    } else {
      a5 <- paste0(a3, " (" , a4, ")")
    }

    # combine raw number with per-tag data
    a6 <- cbind(a2, tag = a5)

    # get cleaner column names
    names(a6) <- c("School", "Total", tag)

    # combine School and Total, then remove Total
    a6$School <- fix_factor(a6$School,
      to.add=paste0(a6$School, " (", a6$Total, ")")
      to.remove=a6$School)

    a6$Total <- NULL
```r
} else {
  # leave per-tag count, per-tag pct, total count, and school as # separate columns
  a5 <- cbind(a2, "P"=a4, "D"=a3, "T"=a2$School.length)
  if(rank_by) {
    a6 <- a5[, c("School", "P", "D", "T")]
    filename <- paste0(imageloc, "Top ", howmany,
        " Schools by Methodological Focus ",
        "(Ranked by Percentage)", ",
        dataset_name, ", ", tagset_name, ".csv")
  } else {
    a6 <- a5[, c("School", "T", "D", "P")]
    filename <- paste0(imageloc, "Top ", howmany,
        " Schools by Methodological Focus ",
        "(Ranked by Number of Dissertations)", ",
        dataset_name, ", ", tagset_name, ".csv")
  }

  # export as tab-delimited
  # TO DO: check if the file exists, prompt to overwrite or abort
  if(remake_figs) {
    # label for file
    names(a6)[1] <- realtags(tag, tagset_name)
    write(t(names(a6)), ncolumns=4, filename, sep=" ",
        append=TRUE)
    write(t(a6), ncolumns=4, filename, sep=" ", append=TRUE)
    write(" ", ncolumns=4, filename, sep=" ", append=TRUE)
    names(a6)[1] <- "School"    # label for screen
  }

  return(a6)
} # end of toplist.onetag()

# 3b. Apply the function to each tag in the tagset
b <- lapply(tagset, FUN=function(x) {
  toplist.onetag(a=a, tag=x, rank_by=rank_by_pcts)
})

if(rank_by_pcts) {
  title <- paste("Top ", howmany,
      " Schools by Methodological Focus (Ranked by Percentage)"")
} else {
  title <- paste("Top ", howmany,
      " Schools by Methodological Output ",
      "(Ranked by Number of Dissertations)"")
}
subtitle2 <- paste("* indicates member of the Consortium of Doctoral",
    "Programs in Rhetoric and Composition")
names(b) <- tagset
```
```r
writeLines(c(title, subtitle1, subtitle2))
print(b)

# close wrapper function toplists()
}

# call function
if(autorun) {
  remake_figs
toplists(rank_by_pcts=T, combine=F)
toplists(rank_by_pcts=F)
}

# TO DO: Write the output to a file for easier porting to Word, Scrivener, # etc.

#########################################################################
# topic cluster reach.R
#
# GOAL: given a cluster of topics identified through frameToD3.R, find out # how many dissertations include at least one topic in that cluster at a # level of over 12% (or whatever).
####

# get all topics by document
cluster.strength <- function (my.topics_name, 
dataset_name = "consorts",
ktopics    = 55,
bad.topics = NULL,
level      = 0.12,
cumulative = TRUE
) {
  # Exclude non-content-bearing topics
  if(is.null(bad.topics) && dataset_name == "consorts" && ktopics == 55) {
    bad.topics <- c("4", "47", "22", "2", "24",
    # bad OCR or ProQuest boilerplate
    "13", "50")
    # language markers (Italian, Spanish)
  }

  my.topics <- get(my.topics_name)

  if(any(my.topics %in% bad.topics)) {
    warning(paste("At least one topic in your list has been",
      "identified as non-content-bearing"))
  }
  if(!exists("get.doctopic.grid", mode="function")) {
```
source("get doctopic grid.R")
}
grid <- data.table(get.doctopic.grid(dataset_name, ntopics)$outputfile)
# str(grid)
# head(grid)

grid <- grid[, !(names(grid) %in% c(bad.topics, "Pub.number")), with=F]
# head(grid)

my.contribs <- grid[, names(grid) %in% my.topics, with=F]

if(!cumulative) {
  individuals <- sapply(1:nrow(my.contribs), FUN = function(x) {
    any(my.contribs[x] >= level) }
  )
  winners <- which(individuals > level)
}
else {
  totals <- sapply(1:nrow(my.contribs), FUN = function(x) {
    sum(my.contribs[x]) }
  )
  winners <- which(totals > level)
}

win.count <- length(winners)
win.pct <- win.count / nrow(my.contribs)

message(paste0("The number of dissertations made up of at least ",
    level*100, " percent of words from this cluster ",
    "of topics (", my.topics_name, ", cumulative=",
    cumulative, ") is ", win.count, " of ",
    nrow(my.contribs), ", or ",
    round(win.pct * 100, 2), ",% of the corpus."))

invisible(list("number" = win.count,
    "percentage" = win.pct))
}

if(autorun) {
  # The Teaching of Writing
  Teaching.of.Writing <- c(1, 32, 30, 3, 9, 39, 41, 40, 45, 6, 25, 8)
  cluster.strength("Teaching.of.Writing")

  # Theories of Meaning-Making
  Theories.of.Meaning.Making <- c(21, 18, 48, 14, 26, 53, 31, 29)
  cluster.strength("Theories.of.Meaning.Making")

  # Audience and Context for Composing
  Audience.and.Context <- c(35, 49, 55, 27, 43, 46, 44)
  cluster.strength("Audience.and.Context")

  # Performative Identities, past and present
  Performative.Identities <- c(23, 10, 16, 33, 15, 11, 7, 37)
cluster.strength("Performative.Identites")

# Politics and Power
Politics.and.Power <- c(36, 20, 28, 54, 17, 52)
cluster.strength("Politics.and.Power")

# other
Other <- c(5, 12, 42, 38, 51, 34, 19)
cluster.strength("Other")

# all together now, more stringent test
sapply(cluster_names, FUN=function(x) cluster.strength(x, level=0.25))

# The Teaching of Writing subcluster that's especially classroom-y
Teaching.of.Writing.1 <- c(1, 32, 30, 3, 9, 39, 41, 40)
cluster.strength("Teaching.of.Writing.1")

# The Teaching of Writing subcluster that's a little more administrative
WPA <- c(45, 6, 25, 8)
cluster.strength("WPA")

# both together now, more stringent test
sapply(c("Teaching.of.Writing.1", "WPA"), FUN=function(x) {
  cluster.strength(x, level=0.25) }
)

# TO DO: make a scatter plot with X-axis = level and Y-axis = cumulative
# cluster strength, and a dataseries for each cluster (all on the same
# graph)

# topics by year.R
#
# GOAL: To graph the rising and falling contributions to the corpus of each
# (or specified) topic over time.
#
# Provides two functions:
#  * topics.by.year(dataset_name, ntopics, ..., per.plot): draws line
#    graphs showing the proportions of the full corpus accounted for by
#    each topic. Allows up to `per.plot` topics to be graphed in the same
#    figure.
#  * topic.variation(dataset_name, ntopics, ...): plots adjacent
#    box-and-whisker plots showing the range of year-to-year proportions
#    of the full corpus for each topic. In a sense, then, compresses all
the graphs produced by topics.by.year into a single graph, allowing for easier comparison of the variability of topic contributions.

```r
topics.by.year <- function(dataset_name = "consorts",
ntopics = 55,
to.plot = NULL,  # any pre-set topics to plot?
do.plot = TRUE,  # should we draw it, or just # return the dataframe?
per.plot = 5     # maximum how many lines per plot?
)
{
  require(data.table)
  require(RColorBrewer)

  # Get topic weights for every document we have
  if(!exists("get.doctopic.grid", mode="function")) {
    source("get doctopic grid.R")
  }
  grid <- data.table(get.doctopic.grid()$outputfile)

  # Get ready to merge
  grid$Pub.number <- as.factor(grid$Pub.number)
  setkey(grid, Pub.number)

  # Merge with noexcludes to add Year data to the topic data
  grid <- merge(grid, noexcludes.dt[, list(Pub.number, Year)], all.x=T)

  # Re-key by year
  setkey(grid, Year)

  # Get some stats for topics within each year
  topic.year.avg <- grid[, lapply(.SD, mean), by=Year]
  topic.year.avg <- topic.year.avg[, !names(topic.year.avg) %in% "Pub.number", with=F]

  # It's a data.table
  df <- as.data.frame(topic.year.avg)

  # Get topic labels, which you've composed elsewhere using `top docs per top.R`, to use in figure legends
  if(!exists("get_topic_labels", mode="function")) {
    source(file="get topic labels.R")
  }
  topic.labels.dt <- get_topic_labels("consorts", 55)
  head(topic.labels.dt)

  # Exclude non-content-bearing topics
bad.topics <- c("4", "47", "22", "2", "24", "13", "50")
topic.labels.dt <- topic.labels.dt[(Topic %in% bad.topics)]
setkey(topic.labels.dt, Rank)
head(topic.labels.dt)

# We'll use plot.me in the other function, so figure it out even if we're
# not actually plotting.
if (!is.null(to.plot)) {
  # any pre-set topics to plot?
  plot.me <- to.plot
} else {
  # plot.me <- c(51, 26, 46, 27, 43) # A range of topics to plot
  # plot.me <- 2:ncol(df)-1 # gives 1:ntopics by topic number

  # gives all topics in order of rank
  plot.me <- topic.labels.dt[order(Rank), Topic]
}

# Set graphing parameters; see
# http://www.statmethods.net/graphs/line.html.
if (do.plot) {
  # X axis will be years
  xrange <- range(df$Year)

  # Y axis will be % of the corpus contributed by topic
  yrange <- c(0, max(df[, !names(df) %in% "Year"]))

  # Use different colors for each plot
  mycol <- brewer.pal(n=per.plot, name="Dark2")

  # Use different symbols for each plot?
  # plotchar <- seq(18, 18+length(plot.me), 1)

  # Nah, use same symbols for each plot
  plotchar <- rep(20, length(plot.me))
  maintitle <- "Average Topic Proportions over Time"

  ## Draw `per.plot` (by default, 5) lines on the same plot, then start
  ## a new plot and repeat.

  # start recording to file if desired
  if (remake_figs) {
    if (is.null(to.plot)) {
      filename <- paste0(imageloc, maintitle, ",", Topics ranked ",
        i", ",", (i+per.plot-1), ",.pdf")
    } else {
      filename <- paste0(imageloc, maintitle, ",", Topics ",
        to.plot[i", ",", to.plot[i+per.plot-1], ",.pdf")
    }
  }
pdf(filename)

# set a target for when to repeat
enough <- seq(1, length(plot.me), by=per.plot)

for (i in 1:length(plot.me)) {
  # Get a per-plot index to rotate through colors
  j <- ((i-1) %% per.plot) + 1

  # After each set of `per.plot` topics, start a new plot
  if (i %in% enough) {
    # make sure we don't print extra nulls at the end
    if((length(plot.me) - i) > per.plot) {
      legend.offset <- (per.plot-i) %% per.plot
    } else {
      legend.offset <- length(plot.me) - i
    }

    if(remake_figs) {
      # close an open file connection from the previous loop,
      # if there is one
      dev.off()

      # update filename
      if(is.null(to.plot)) {
        filename <- paste0(imageloc, maintitle,
                           ", Topics ranked ", i, ", ",
                           (i+legend.offset), ".pdf")
      } else {
        filename <- paste0(imageloc, maintitle, ", Topics ",
                           to.plot[i], ", ",
                           to.plot[(i+legend.offset)], ".pdf")
      }

      # start writing a new file
      pdf(filename)
    }
  }

  # set up a blank plot in a standard size
  plot(x = df$Year,
       y = rep(yrange, length((df$Year))/2),
       type = "n",
       xaxs = "r",
       ylab = "Portion of corpus (scaled to 1)",
       xlab = "Year",
       bty = "n",
       main = maintitle)
# add a legend for up to five values
if(i <= 10) {
    legendloc <- "bottomright"
} else {
    legendloc <- "topright"
}

if(is.null(to.plot)) {
    legend(legendloc, title = paste0("Topics, ranked ", i, "-",(i+legend.offset), " of ", nrow(topic.labels.dt)),
          legend = paste0(plot.me[seq(i, (i+legend.offset), 1)], ": ", topic.labels.dt[Topic %in% plot.me[i:(i+legend.offset)], Label]),
          fill=mycol[j:(j+legend.offset)],
          border=mycol[j:(j+legend.offset)],
          bty="n",
          cex=0.8)
} else {
    legend(legendloc, title="Topics",
          legend = paste0(plot.me[seq(i, (i+legend.offset), 1)], ": ", topic.labels.dt[Topic %in% plot.me[i:(i+legend.offset)], Label]),
          fill = mycol[j:(j+legend.offset)],
          border = mycol[j:(j+legend.offset)],
          bty = "n",
          cex = 0.8)
}

# end new plot + legend

# draw the line and loop back
lines(x = df$Year,
      y = df[,as.character(plot.me[i])],
      type = "l",
      pch = plotchar[j],
      col = mycol[j])

# end of for loop

# now that we're done looping, close the final file connection
if(remake_figs) {dev.off()}

} # end if(do.plot)

invisible(list("df"=df, "rank.order"=plot.me))

} # end of wrapper function topics.by.year()
## Second function: find year-to-year peak variation for each topic

```
topic.variation <- function(dataset_name = "consorts", 
                           ntopics = 55, 
                           to.plot = NULL,  # any pre-set topics to plot? 
                           notch = FALSE)
{
  # okay, this is interesting
  df <- topics.by.year(dataset_name, ntopics, to.plot, do.plot=FALSE)
  rank.order <- df$rank.order
  df <- df$df

  maintitle <- paste("Yearly Variation of Topic Proportions", 
                     "Generally Preserves Topic Rank")
  if(dataset_name == "consorts") {
    subtitle <- paste0("Consortium dissertations, N", nrow(grid), 
                        ", years 2001-2010")
  } else {
    subtitle <- paste0(dataset_name, " dissertations, N", nrow(grid), 
                        ", years 2001-2010")
  }

  # Get topic labels, which you've composed elsewhere using 'top docs per 
  # topic.R', to use in figure legends
  if(!exists("get_topic_labels", mode="function")) {
    source(file="get_topic_labels.R")
  }
  topic.labels.dt <- get_topic_labels("consorts", 55)
  head(topic.labels.dt)

  # Exclude non-content-bearing topics
  bad.topics <- c("4", "47", "22", "2", "24", "13", "50")
  topic.labels.dt <- topic.labels.dt[!(Topic %in% bad.topics)]
  setkey(topic.labels.dt, Rank)
  head(topic.labels.dt)

  # draw the plot
  if(remake_figs) {
    filename <- paste0(imageloc, maintitle, ".pdf")
    pdf(filename)
  }

  boxplot(df[!names(df) %in% "Year"], rank.order, 
          main = maintitle, 
          # xlab="Topic Number, Arranged by Overall Rank within Corpus", 
          # cex.axis=0.6, las=2, 
          ylab = "Portion of Corpus (scaled to 1)", 
          xaxt = "n", 
```
notch = notch
)

axis(1,
    at = seq_along(df[!names(df) %in% "Year"], rank.order]),
    labels = topic.labels.dt[Topic %in% rank.order, Label],
    las = 2,
    lheight=0.5
)

mtext(subtitle, side=3)
# abline(v=(0.5+seq(from=5,to=length(plot.me), by=5)), lty="dotdash")

if(remake_figs) {
    dev.off()
}

if(autorun) {
    topics.by.year()
    topic.variation()
}

################################################################################
# variation of topic proportions.R
#
# Goal: Find out the curve of topic strengths within each document, i.e. how
# much of the document is the top topic? how much is the second? and so on,
# aggregated over all documents, as a boxplot of contribution (y-axis) sorted
# by topic rank (x-axis).
#
# Rationale: I want to know at what level to cut off "cotopics": what's a
# realistic scenario?
###

topic.proportions <- function(dataset_name = "consorts",
    ntopics = 55,
    # if default dataset and ntopics are used,
    # use default bad.topics
    bad.topics = NULL,
    # Draw notch in barplot to check for overlap?
    use.notch = FALSE,
    # Use topic browser for outlier dissertations?
    explore.outliers = FALSE)
{
    require(data.table)
if(!exists("get.doctopic.grid", mode="function")) {
  source("get doctopic grid.R")
}

grid <- data.table(get.doctopic.grid()$outputfile, key="Pub.number")
# str(grid)
head(grid)

# Exclude non-content-bearing topics
# If none are set in parameters, use defaults:
if(is.null(bad.topics) && dataset_name=="consorts" && ntopics==55) {
  bad.topics <- c("4", "47", "22", "2", "24", "50", "13")
}

grid.clean <- grid[, !(names(grid) %in% c(bad.topics, "Pub.number")),
  with=F]

print(head(grid.clean))

# decreasing sort across each row -- ignore column (i.e. topic) names
grid.sorted <- t(apply(grid.clean, 1, FUN=function(x) {
  sort(x, decreasing=T)
}))

# each row is a dissertation; we lose topic numbers, but now column 1
# is the weight of the top-ranked topic for that row, column 2 the
# weight of the 2nd-ranked topic, and so on. Let's look at the 10
# top-ranked topics for every dissertation.
print(head(grid.sorted[, 1:10]))

# start empty, build up
stats <- data.frame()
for (i in 1:3) {
  # message(paste0("Stats for ", i,
  # "-ranked topic within dissertations:"))
  stats <- rbind(stats, boxplot.stats(grid.sorted[, i])$stats)
}

# lower whisker, lower 'hinge', median, upper 'hinge', upper whisker
names(stats) <- c("lower", "Lhinge", "median", "Uhinge", "upper")
stats <- cbind("rank of topic within diss"=c(1, 2, 3), stats)

# we'll return the stats data.frame later.

## Time to make the plot
maintitle <- "Variation of Topic Proportions, Top 10 Topics per Document"
subtitle <- paste0(dataset_name, ", N=", nrow(grid))

if(remake_figs) {
  pdf(file=paste0(imageloc, maintitle, ".pdf"))
```r
boxplot(grid.sorted[, 1:10],
  cex.axis = 1,
  las = 1,
  main = maintitle,
  sub = subtitle,
  xlab = "Topic Rank",
  ylab = "Portion of Document (scaled to 1)",
  yaxp = c(0, 1, 10),
  notch = use.notch
)

## mark line covering top three quartiles for the 2nd-ranked topic,
## but only the top quartile for 3rd
# abline(h=0.12)

if(remake_figs) {
  dev.off()
}

## Optionally extract top-topic outliers for further examination
if=explode.outliers) {
  upper.whisker <- boxplot.stats(grid.sorted[, 1])$stats[5]

  # just look at #1 topic
  outliers.index <- which(grid.sorted[, 1] > upper.whisker)
  outliers <- cbind(grid[outliers.index, "Pub.number", with=F],
                    grid.sorted[outliers.index, 1:10])
  outliers <- outliers[order(outliers$V1, decreasing=T), ]

  # boxplot(outliers[, 2:ncol(outliers)])

####
# I have a hypothesis that these are mostly language-based topics.
# Let's look at the top topics represented here. STRATEGY:
# 1. For each Pub.number, get top-ranked topic number by finding the
#     max within that row of `grid`.
# 2. Make a table of these topic numbers.
# 3. Retrieve the labels for each topic in the table.

mytopics <- c()  # start empty and build up
myvalues <- c()  # what are those high percent-of-text values?

for (i in outliers$Pub.number) {
  row <- grid[which(grid$Pub.number==i), 2:ncol(grid), with=F]
  mytopic <- which(row == max(row))
  mytopics <- c(mytopics, mytopic)
  myvalues <- c(myvalues, max(row))
}

# count 'em up
```
mytopics.t <- table(mytopics)

# get labels
if(!exists("get_topic_labels", mode="function")) {
    source(file="get_topic_labels.R")
}
labels <- get_topic_labels(dataset_name, ntopics)
labels.t <- labels[unique(mytopics), Label, key=Topic]

# merge in the counts
labels.t[, "Outlier Count"] <- mytopics.t

# merge in the values
b <- aggregate(data.frame(mytopics, myvalues), by=list(mytopics),
    FUN=c)
labels.t <- labels.t[b, ][,mytopics:=NULL]

# sort by descending outlier frequency
labels.t <- labels.t[order(mytopics.t[, "Outlier Count"], decreasing=T), ]

# report back
message("Upper outliers for top-ranked topics:")
print(labels.t)
message(paste("Total outliers for top-ranked topic:",
    sum(labels.t[, "Outlier Count", with=F])))

# Okay, my hypothesis is false! All sorts of topics here.
# Interesting. Still, I may want to remove the dissertations with
# top-ranked language topics beforehand, since they do tend to
# dominate their dissertations.

## Browse more details of these outlier dissertations
if(!exists("get.topics4doc", mode="function")) {
    source(file="top_docs_per_topic.R")
}
if (!remake_figs) {
    a <- readline(paste("Press <enter> for more detail on these ",
        "docs, or S to skip to the end\n"))
} else {
    a <- ""
}

while (tolower(a) != "s") {
    for(i in outliers$Pub.number) {
        print(get.topics4doc(i, dataset_name, ntopics,
            showlabels=TRUE))
        if (!remake_figs) {
            a <- readline(paste("Press <enter> for next doc, ",
                "D for more details, or ",
                "S to skip to the end\n"))
        } else {
            a <- ""
a <- ""
}

if (tolower(a) == "s") {
  break
} else if (tolower(a) == "d") {
  print(noexcludes.dt[i])
  a <- readline(paste("Press <enter> for next doc",
    "or S to skip to the next topic\n")
  )
}

a <- "s"

# TO DO: browse low-liers (though my hypothesis there is that they're
# mostly dissertations dominated by bad.topics; in which case the thing
# TO DO is to eliminate those dissertations entirely from the topic
# modeling dataset beforehand, and run the model again.)

} # end if(explor.outliers)

message(paste("Stats for contributions of topics at various ranks",
  "within dissertations:"))
return(stats)
}

if(autorun) {
  remake_figs
  topic.proportions()
  topic.proportions(explor.outliers=T)
}

########################################################################
Appendix G:
Unix Shell Scripts Used to Prepare Text for Topic Modeling

#!/bin/bash  # declare our shell environment

####
# ben_clean_and_consolidate.sh
#
# GOAL: given pdf files from which you want to build a topic model, extract
# the full text and remove boilerplate that could skew your results.
#
# STRATEGY: Read in a list of files in a directory.
# While items remain in the list, run a series of commands on them,
# saving the results into new files to avoid accidental overwriting.
# Based on scripts by Micki Kaufman (https://twitter.com/MickiKaufman).
#
# Four functions are defined and then executed at the bottom: extract, clean,
# combine, spellcount. Each function runs a loop, containing commands to
# apply; the variable `line1` will be read in from ls (the directory
# listing), which causes the loop to execute on each file in the directory.
# The functions are called at the bottom; comment out the ones you don't want
# to run.
####

# Declare some basics: source and destination.
# NB: these will likely change often!
DATASET=$1
PDF="/Users/benmiller314/Documents/fulltext_dissertations/morepdfs"
SRC="/Users/benmiller314/Documents/fulltext_dissertations/morepdfs/as text files"
DST="/Users/benmiller314/Documents/fulltext_dissertations/clean_""$DATASET""_only"

# Store cumulative data in its own directory
CUMUL="/Users/benmiller314/Documents/fulltext_dissertations/cumulative"

## for testing purposes
# line1="3298352.PDF"

## Zeroth function: extract text from pdf. Run in the $PDF folder.
# NB: pdftotext is available for free from http://www.bluem.net/en/mac/packages/
function extract()
{ # Make sure we have a place to output to.
  if ! [ -d "$SRC" ] ; then
    mkdir "$SRC"
  fi

  # Start the loop.
  while read line1; do
    PUB=`printf "$line1" | awk 'BEGIN { FS="." } { print $1; }'`

    # progress report
    printf "Converting $line1 to $PUB.txt ... "

    # convert the file.
    pdftotext "$PDF/$line1"

    # progress report
    if [ $? = 0 ] ; then printf "File made " ; fi

    # move to txt folder.
    mv "$PDF/$PUB.txt" "$SRC/$PUB.txt"

    # progress report
    if [ $? = 0 ] ; then echo "and moved." ; fi

  # Close the loop.
  done

  # Close the function.
}

## First function: Get text that R can read. Run in the $SRC folder.
function clean ()
{
  # Make sure we have a place to output to.
  if ! [ -d "$DST" ] ; then
    mkdir "$DST"
  fi

  # Start the loop.
  while read line1; do

    ## Step 1. Copy the file to a new directory, making changes as it goes

    # progress report
    echo "Cleaning from SRC $line1 to DST $DST/$line1"

    # 1a. Convert text encoding from ISO 8859-1 (Latin-1) to UTF-8 (unicode standard)
    # 1b. Using tr, delete all characters except for line breaks and Western characters

}
# 1c. Using sed, delete the first page added by UMI (which starts in line 1, and usually ends with the zip code)
# 1d. Save to a file in the destination directory.

iconv -f ISO_8859-1 -t UTF-8 "$SRC/$line1" | \
tr -cd '11\12\40-\176' | \
sed "1,-1346/d" > "$DST/cleaned_$line1"

# catch the case where we've stripped too much (i.e. the file has 0 bytes) and do it again without sed
if ! [ -s "$DST/cleaned_$line1" ]; then
  iconv -f ISO_8859-1 -t UTF-8 "$SRC/$line1" | \
  tr -cd '11\12\40-\176' > "$DST/cleaned_$line1"
fi

# Close the loop.
done

# Close the function.
}

## Second function: Combine files into a big cumulative one. Here's how:

function combine () {
  # Step 1. Outside the loop, create an empty file to hold the cumulative output.
  if ! [ -e "$CUMUL/$DATASET_cumulative.txt" ]; then
    printf '' > "$CUMUL/$DATASET_cumulative.txt"
  else
    echo "ERROR: $CUMUL/$DATASET_cumulative.txt already exists; aborting combine step."
    exit 1
  fi

  echo "Making cumulative file:"   # progress report

  # Step 2. Concatenate the cleaned file (after removing line breaks) and append it to the cumulative output file in MALLET-ready format, as per http://mallet.cs.umass.edu/import.php: "The first token of each line (whitespace delimited, with optional comma) becomes the instance name, the second token becomes the label, and all additional text on the line is interpreted as a sequence of word tokens."

  while read line1; do
    printf "adding $line1..."   # progress report
    # 2a. Using awk, strip '.txt' and 'cleaned_' off the filename to get the Pub.number of the diss. We'll use these as instance names in MALLET.

PUB=`printf "$line1" | awk 'BEGIN { FS="." } { print $1; }' | awk 'BEGIN { FS="_" } { print $2; })'`
# 2b. We don't have labels right now, but eventually we could use # methods exported from R.
LAB="placeholder"

# 2c. Using tr, remove all commas so we can get a clean csv, # then replace newlines with spaces (get all text on one line).
CONTENTS=`cat "$DST/$line1" | tr -d ',' | tr -s '\n' " "`

# 2d. String together the instance names, labels, and the file # contents;
echo "$PUB $LAB $CONTENTS" >> "$CUMUL/$DATASET_cumulative.txt"

if [ $? = 0 ]; then
  # progress report
  echo "done."
fi
done

if [ $? = 0 ]; then
  # progress report
  echo "All text saved to $CUMUL/$DATASET_cumulative.txt"
  echo ''
fi

## Third function: Get data toward a conservative estimate of OCR accuracy.
## Strategy: for each file "$line1" in a directory index (produced by ls), # (1) find the wordcount # (2) run a spellcheck, and save errors to a file # (3) count the number of errors in that file # (4) compile into a single file for further processing in R

function spellcount ()
{
  # (step 0 or 4a) Outside the loop, create a placeholder output file
  if ! [ -d "$DST/spellstats" ]; then
    mkdir "$DST/spellstats"
  fi

  if ! [ -e "$DST/spellstats/spellstats.csv" ]; then
    echo 'Pub.Number,WordCount,ErrorCount' > "$DST/spellstats/spellstats.csv"
  else
    echo "spellstats.csv already exists; aborting script to avoid duplication."
  fi
  echo "To append, use new DST folder and concatenate later."
  exit 1
fi

    echo "Counting spelling errors..."

while read line1; do

    ## (step 1) Get wordcount, save to a variable.
    WC=`wc -w "$DST/$line1" | awk '{ print $1; }' -`

    ## (step 2) Find misspelled words, save to file in case we want to analyze
    ## later.
    # NB: apparently this isn't included in OS X 10.7 (Lion). Boo. To download
    # the aspell command, you'll need something like Fink
    # http://www.finkproject.org/download/srcdist.php and Apple Developer Command
    # dictionaries. Once those are installed (no small feat), uncomment and run
    # the commands in file "install aspell dictionary.txt"

    aspell list < "$DST/$line1" > "$DST/spellstats/wordswrong_$line1"

    ## (step 3) count the lines in the wordswrong file; save the numbers in a
    ## variable.
    ERRS=`wc -l "$DST/spellstats/wordswrong_$line1" | awk '{ print $1; }' -`

    ## (step 4) combine files into a cumulative table. Here's how:
    # Step 4a. Outside the loop, create a placeholder output file. (See
    # above.)

    # Step 4b. Strip '.txt' and 'cleaned_' off the filename; this will help
    # us join tables later.
    PUB=`printf "$line1" | awk 'BEGIN { FS= "." } { print $1; }' | awk 'BEGIN {
    FS= "\n" } { print $2; }'`

    # Step 4c. String together the Pub.number, the wordcount, and the
    # errorcount; append to the output file.
    echo "-- checking $line1"
    echo "$PUB, $WC, $ERRS" >> "$DST/spellstats/spellstats.csv"

    ## Close the loop

## Final report: Tell us what we've got!
    echo "Spelling counts saved to $DST/spellstats/spellstats.csv."

    FILECOUNT=`wc -l "$DST/spellstats/spellstats.csv" | awk '{ print $1; }' -`

    let FILECOUNT=$FILECOUNT-1    # account for headers in 1st line
    echo "$FILECOUNT files processed."
    echo ''

    ## Close the function


```bash

# echo "Currently DST folder is $DST"
# echo "and the SRC folder is $SRC"

## Go to the files, and run all the functions.
## IMPORTANT: Comment out those you don't need right now.
CURRENT_DIR=$PWD
# cd "$PDF"           # Go to pdf directory
# ls *.*PDF | extract  # Call 0th function
# cd "$SRC"           # Go to source directory
# ls *.*txt | clean    # Call 1st function
cd "$DST"            # Go to output directory
ls cleaned* | combine  # Call 2nd function
# ls cleaned* | spellcount # Call 3rd function
cd "$CURRENT_DIR"    # Go back where we were
```
Works Cited


DeBoer, Fredrik. Re: Video of Banks’ Talk? 3 Apr. 2015. E-mail.
---. Re: Writing Pedagogy/“the Essay” (was Banks’ Talk and Essay Writing). 28 Mar. 2015. E-mail.


Smit, David. “Stephen North’s The Making of Knowledge in Composition and the Future of Composition Studies ‘Without Paradigm Hope.’” The Changing of Knowledge in


Wright, James. Publishing Dissertations Online. 29 July 2013. E-mail.
