The Limits of Distinctive Words: Re-evaluating Literature's Gender Marker Debate

Sean G. Weidman, Pennsylvania State University
James O'Sullivan, University College Cork

Introduction

Armed with some recent VIDA numbers[1] and a well-intentioned skepticism of the current state of gender equality in the literary marketplace, Andrew Piper and Richard Jean So decided to analyze over 10,000 book reviews—15 years' worth—from the *Sunday Book Review* and *The New York Times*. With an investigative goal of determining, if any, the level of gender bias in reviews,[2] they pose a simple question: do book reviewers write about male and female literatures differently? Reporting on the results of their stylistic analysis, the two scholars conclude:

> *New York Times* book reviews overwhelmingly suggest that women tend to write about domestic issues and affairs of the heart, while men thrive in writing about "serious" issues such as politics. It's not that women don't write about politics or men don't write about feelings and families. It's just that there is a very strong likelihood that if you open the pages of the *Sunday Book Review*, you will be jettisoned back into a linguistic world that more nearly resembles our Victorian ancestors. […] Women writers are still being defined by their "sentimental" traits and a love of writing about "maternal" issues, while men are most often being defined by their attention to matters of science and the state.

Though primarily concerned with critical bias in book reviewing, Piper and So also provide an important reminder: that how we read, write, and think about the work of male and female authors—as fundamentally opposed to one another in, for example, "sentimentality"—betrays and then reinforces gender biases we may already hold. The result is discriminatory to be sure, and as Piper and So suggest, likely indicative of various "latent, subtle, and perhaps unconscious attitudes about the idea of women writing books." But a couple of questions—ridiculous as they may at first seem—have been left unasked and unaddressed: Is there any evidence to suggest women *are* writing more about "marriage," "love," and "beauty" than men as indicated in the reviews? If this stylistic bias is consistent in reviews from the last 15 years and are, as Piper and So point out, showing no signs of improving or changing, is it possible that women writers *are* in fact writing with a different literary style than men, one that reviewers happen to notice?

A number of studies in the past 15 years have already engaged with similar questions. Studies that have included multiple genres or mediums in their stylistic analyses and text classifications have concluded that stylistic features often depend as heavily upon gender as genre (Argamon et al. 2003; Janssen and Murachver 2004; Herring and Paolilo 2006, Sarawgi et al. 2011). Other recent studies have highlighted the correlation between distinctive stylistic markers and authorial gender across many genres, including poetry (Hoover 2013), political and legislative speeches (Dahllof 2012; Yu 2014), scientific

---

[1] *VIDA: Women in Literary Arts*, is an organization that compiles data about publication and gender (and race) in an effort to uncover biases, imbalances, or inequalities in top tier publications from around the world. Piper and So provide as their exigence a handful of links to the past few years of VIDA survey numbers, and a concern with "current claims around gender inequality in book publishing; things are bad but *they are getting better*."

[2] Another, even more recent (though ostensibly yet unpublished) study from Julieanne Lamond and Melinda Harvey suggests that, at least as far as Australian book reviews (from 1985-2013) are concerned, "around two third of published authors in Australia are women, but two thirds of the books being reviewed are by men […and] this ratio has remained largely the same for 30 years."

papers (Sarawgi et al. 2011), formal and informal verbal communication (Singh 2001; Yaguchi et al. 2004; Iyeiri et al. 2005 and 2011) and, of course, written word (Argamon et al. 2003; Burrows 2004; Schler et al. 2005; Mikros 2013). In fact, since Lakoff's oft-cited 1973 study brought the question of "language and woman's place" to mass critical attention in linguistics, significant research has shown that linguistic gender markers not only point to differences in communication, but also to how we conceive of gender, from childhood to adulthood (Fine 2010; McHugh and Hambaugh 2010; Baker 2010; Macalister 2011; Pennebaker 2011; Moon 2014). The results of these studies suggest that, for whatever reason, the majority of male and female authors have writing styles and distinct linguistic markers unique to their respective gender.[3] In other words, the question of whether or not men and women tend to speak, write, blog, tweet, and generally communicate through language differently has reached a rather one-sided consensus.

The much more interesting line of questioning, and the one which our research attempts to fundamentally engage in (following Piper and So), is what, if anything, might these differences tell us? Our approach to this question differs slightly from previous research;[4] although many of these studies have consistently shown that stylistic and thematic qualities are associated with male and female language application, noticeably fewer have attempted to address these gender signals as they arise in literature. James Pennebaker's *The Secret Life of Pronouns* (2011) is perhaps the most well-known example, but more recently Paul Baker's *Using Corpora to Analyze Gender* (2014) undertakes a similarly thorough stylometric project.[5] That said, we would be remiss if we did not point out that part of the impetus of this project is one study that does deal with literary texts: the frequently cited gendered wordlist found in David L. Hoover's illustrative article, "Textual Analysis" (2013). In his essay, Hoover presents a list of the one hundred most distinctive words in the works of twenty-six contemporary poets, equally split between male and female authors. His results are provocative for a variety of reasons, not least because, as Hoover remarks, some aspects of his findings are "almost stereotypical," with "[f]emale markers like *children* and *mirrors* and male markers like *beer* and *lust*." While there are some surprises in the list, and Hoover offers some fascinating concordances, he himself has acknowledged that "[a]ny study of the vocabulary of male and female poets would benefit from larger numbers of poets and larger samples, and many other configurations that address different contrasts are possible (e.g., nationality or historical period)." A recent article by Jan Rybicki, "*Vive la difference*: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies" (2015), presents a considerable step in the right direction. Rybicki confronts a number of the stylometric questions Hoover poses, employing multivariate statistical

[3] A note about terminology is warranted here. First, we define the term "markers" as unique words generally employed by one group over another, and in a stylistic vacuum, thus generally characteristic of an author from that respective group. Second, throughout this essay, when we address writing styles and stylistic markers that tend to cluster along traditional gender lines, we employ both traditional terminologies of gender (wo/men) and sex (fe/male). This was a conscious choice, though we realize there are limitations to this type of loose classification; however, as we shift between terms throughout the essay, our intention is to avoid reifying any kind of gender/sex essentialism that could be read into a project of this sort (and we detail as much in our concluding section). In fact, we believe this project attests to the opposite: that the evolving separation between the literary styles of the genders suggests that the effects of gender performativity have uniquely influenced the stylistic markers of each gender over time. More on that in our conclusion.

[4] Though one thing that all of these studies have in common is that they, like us, see the meaning of "style," and what we can learn from it, as having shifted in contemporary contexts (see Herrmann et al., 2015).

[5] With part of his project, Baker (2014) uniquely argues for the usefulness of "corpus approaches [that] enable perspectives other than one based on gender comparison […]. Such an approach frees us from thinking in terms of gendered 'over-use' and 'under-use', does not pit the sexes against one another or implies [*sic*] that certain groups are deviating from the 'norm' in some way" (202). Baker effectively suggests that there is enough "variation *within* the sexes" (original emphasis) to warrant dealing with them individually—or, if nothing else, to warrant special care when dealing with them in concert, a suggestion we hope to have met in this study. For various other methodological approaches to gender studies and computational stylistics, see Baker's 2012 chapter, "Corpora and Gender Studies" and Partington et al.'s 2013 chapter in *Patterns and Meanings in Discourse*, both of which are cited in the bibliography.

analysis for the purposes of gender-based authorial attribution in a corpus of 18th- and 19th-century English fiction. Importantly, Rybicki goes on to analyze a large corpus of literary works from the 20th and 21st centuries, comparing them to the 18th- and 19th-century corpus not only "in terms of their usefulness for gender identification in literary texts" (2016, p. 746), but also with respect to questions of canonicity and the evolution of gender markers over time. But even with some "honest and well-founded cherry-picking" (2016, p. 751), Rybicki admits that despite the results of prior studies that show a distinct, contemporary separation of male and female writers based on gender markers (e.g. Pennebaker 2011), his results suggest distinctive gender markers may begin to fade over time. Rybicki comments that, as the "rigidly demarcated gender roles of the 18th century have been increasingly transgressed," in part because of and in part in spite of literature, "the gender signal is no longer discernible in the later centuries" (2016, p. 755). His observation, in other words, is that authorial gender becomes less differentiable in the noise of linguistic change. Rybicki derives this conclusion at least in part from the difficulties of creating "a stable 'canon' of male and female keywords that would survive a change of corpus or shifts in literary evolution"—and indeed, as he rightly points out, "the notions of 'writing like a man' and 'writing like a woman' are in such constant flux that they risk becoming […] quite problematic" (2016, p. 759) as a question addressed through stylometric analysis at all.

To these studies that have started addressing the intersection of style and gender in literature, we offer an important iteration, one that attempts to address literary style and gender in terms of their evolution in and alongside literature. Bearing in mind Rybicki's results and his concern with the skewing properties of literary change, we make no effort to uncover some kind of gendered, authorial time(less) signature, or suggest that there may be a universally applicable wordlist of literature's distinct gender markers. Any such project may indeed, like Rybicki remarks and as the title of Burrows's related 1996 essay suggests, "tiptoe into the infinite." Instead, we begin with a foundational but partitioned concern for time by setting out with periodization in mind, by analyzing slices of literary history rather than the whole of it at once. While this approach certainly has its limitations—most of which we will attempt to address shortly—if nothing more we hope this may provide stepping-stone insight for future considerations of the relationships between modernity, gender, and the evolution of language and literary style.

Methodology

Accepting that one can never have too large or robust a dataset for this type of macroanalytic case study, we attempt to build on the foundations set down by Hoover and Rybicki, analyzing gender markers across a selection of male and female authors, and doing so crucially with a concern for the evolution of gender markers over specified literary periods. We elected to use fiction (novels and short stories) as opposed to poetry, and we gathered corpora for 54 authors (see Table 1) for a combined 236 novel-length texts. Attempts were made to use equal amounts of text from each author so as to ensure that larger or longer oeuvres did not skew the results toward particular individuals and styles.

Our goal in limiting the mode of our corpus to novel-length fiction is to refine prior studies with more generalized scopes (e.g. Koppel et al. 2002; Argamon et al. 2003) and place our results in the realm of literary application. Thus, while we follow in the methodological footsteps of many prior studies, we have shifted the focus of our investigation away from style, in the macro-analytical sense, to period and its relation to gender-differentiable terminology. Akin to Rybicki and Hoover, Craig's Zeta (Burrows 2006; Hoover 2008) was the primary methodology we deployed to assess our corpus of texts, and the Zeta analyses were conducted using the *Stylo* R package (Eder *et al*. 2013).[6] Initially, we also conducted a series of cluster analyses using Delta (see Argamon 2008), and then for a more robust analysis, employed

[6] For the sake of reproducibility, using *Stylo*, we employed Zeta with a text slice length of 5,000, text slice overlap of 2,500, and an occurrence and filter threshold of 2 and 0.1, respectively. Our original Delta analysis was run at 100 MFW, without culling, and our consensus trees were run at strength of 0.5 and iterations between 100-1000 most frequent words (in increments of 100 words), culled at 0-50% (in increments of 10%).

Stylo's bootstrap consensus tree function. Eventually, by combining the results of our Zeta with a Delta analysis and placing these two statistical methodologies "in conversation," we can learn a bit more about the nature of those results, a matter to which we will shortly return.

It is worth noting briefly that a Zeta analysis introduces its own inherent limitations, forming as it does a list of words preferred and avoided by one dataset (e.g. female authors) insofar as they relate to another dataset (e.g. male authors). It is, in essence, an inherently dichotomous methodology—it will *always* find difference because it is designed to extract the distinct words from comparative sets of texts; the significance of this comparison depends on the legitimacy of that textual opposition, and how a critic might interpret its results. Zeta will always find difference, because it compares one corpus to the other, regardless of what their content might be. Delta, as pointed out once, functions much like a statistical democracy, weighting z-score differences toward a mean that allows the data to "speak for itself" (Eder 2012); but in Zeta the separation is partly a result of dataset preparation, akin to, say, a two-party republic. Rather than the data establishing its own distinctions on who should win—like everyone gets an open, undirected vote in a democracy—in a way we privilege the split in our separation of the datasets, and the model finds differences and sorts the results between the two items with which we present it—i.e. it effectively only has two choices, and it votes for one or the other. The analogy is not perfect, of course. A simple Delta analysis produces a list of the most frequent words of a corpus, which are inevitably common "function" words (articles, conjunctions, pronouns, prepositions, etc.), while Zeta, in its effort to distinguish between two datasets, produces a list of the distinct words, which tends to include both "function" and "content" words.

Other conventional caveats apply: as with any textual analysis assisted by computational methods, scholars are restricted by their access to a reliable corpus of digitized materials.[7] But this, of course, is the constant nature of the computational beast, and we acknowledge the ultimate need for a larger corpus and a more refined dataset. With that in mind, as a kind of literary case-study of the periodization of gender markers, we identified three distinct literary periods of interest—Victorian, modernist, and contemporary—and gathered a selection of fiction texts written in English from across the canonical authors of these eras. Fully aware of the current critical pushback against periodization in general, our period distinctions admittedly follow problematic conventions, and we understand the relative arbitrariness of distinguishing between ad-hoc literary periods, and filling them with a relatively eurocentric corpus. The publication dates of our Victorian era selections range from 1826-1913, our modernist texts range from 1899-1969, and our contemporary works range from 1947-2013—in other words, we included ranges with overlap in order to include particularly well-known authors.

Male	**Female**
Victorian Authors	
Collins, Wilkie (1860-88)	Braddon, Mary (1862-83)
Doyle, Arthur Conan (1892-1913)	Bronte, Anne[8] (1846-48)

[7] Most notably, we found that finding reliable, accurate texts for our female corpora was extraordinarily challenging, and our corpus reflects this difficulty, failing to provide an equal amount of female and male authors from the Victorian and contemporary periods. The difficulty in acquiring sufficient female texts goes to show, we think, that the kind of sexism addressed in the reviewing world may very well exist in the digitization world as well, a result, perhaps, of the (relatively) recent status of many deserving female names in the literary canon, and the latent sexism that persists in many digital literary archival efforts as a result.

[8] As Jockers once points out in *Macroanalysis*, the Brontë sisters, stylistically speaking, write fairly similarly to one another (p. 64). Although in our analyses the sisters did, on the whole, cluster closer together than they did apart, they did so separately, distinct in their individual styles.

Dickens, Charles (1839-61)	Bronte, Charlotte (1847-57)
Disraeli, Benjamin (1826-80)	Bronte, Emily (1847)
Hardy, Thomas (1880-91)	Eliot, George (1859-63)[9]
Kipling, Rudyard (1888-1901)	Gaskell, Elizabeth (1853-65)
Meredith, George (1856-79)	Oliphant, Margaret (1863-79)
Thackeray, William (1839-55)	Wood, Ellen (1861-66)
Trollope, Anthony (1866-82)	
Modernist Authors	
Beckett, Samuel (1938-53)	Bowen, Elizabeth (1927-31)
Conrad, Joseph (1899-1915)	Compton-Burnett, Ivy (1911-39)
Faulkner, William (1929-48)	Mansfield, Katherine (1911-22)
Fitzgerald, F. Scott (1920-41)	Rhys, Jean (1939-66)
Ford, Ford Madox (1915-35)	Richardson, Dorothy (1915-20)
Hemingway, Ernest (1926-51)	Stein, Gertrude (1909-33)
Joyce, James (1914-22)	West, Rebecca (1918-56)
Lawrence, D. H. (1911-28)	Wharton, Edith (1905-20)
Nabokov, Vladimir (1947-69)	Woolf, Virginia (1925-41)
Contemporary Authors	
Amis, Martin (1991-2010)	Atwood, Margaret (1985-2009)
Garcia Marquez, Gabriel (1967-94) [10]	Barker, Pat (1991-2001)
McCarthy, Cormac (1985-2006)	Gordimer, Nadine (1974-2012)

[9] Due to having caught a number of unfortunate inaccuracies in our copy of Eliot's *Middlemarch* slightly too late in the study, the text was removed from the corpus entirely. Thus, despite it being the pinnacle of "canonical," the novel does not appear in this study.

[10] Given Marquez's canonical status, we included him despite his being one of two authors in our corpus who wrote primarily in another language. For the other author, Nabokov, we included only those texts written originally in English; but this approach was largely impossible for Marquez, who wrote his novels exclusively in Spanish. Fortunately, a number of studies have shown: (1.) that the stylistic signatures of translators generally "remain more or less invisible" (Rybicki 2013) with respect to the original author's text (Burrows 2002; Rybicki 2006, 2010, 2011, 2013; Eder 2013), and that, in other words, the "original author's signal [is] preserved in translation" (Rybicki and Heydel 2013, p. 713); and (2.) that in the case of a single author, "[w]ord-frequency-based" stylometric methods (like our own), "have shown that they are better at attributing the author of the original than the translator" (Rybicki and Heydel 2013, p. 714). That is all to say, for the purposes of this study, the fact that Marquez's works are translated should have no bearing on how his works cluster with one another relative to other authors.

Naipaul, V. S. (1961-2004)	Mantel, Hilary (1992-2012)
Pynchon, Thomas (1973-2013)	Morrison, Toni (1977-2008)
Roth, Philip (1985-2009)	Munro, Alice (1978-2009)
Rushdie, Salman (1981-2001)	Oates, Joyce Carol (1992-2013)
Salinger, J. D. (1947-61)[11]	Smith, Zadie (2002-12)
Updike, John (1990-2008)	Walker, Alice (1976-2004)
Vonnegut, Kurt (1963-90)	

Table 1. Authors used in our study, with date ranges for analyzed works.

Most conspicuously, our selection of authors is based on a notion even more problematic than period, that being, of course, the concept of an author being "canonical."[12] We have settled, for example, with a generally eurocentric canon of modernist authors despite many important trends in modernist studies to expand that view productively and on a global scale. Admittedly, choosing a set of canonical authors creates the potential for this classical conception of literary history to skew stylometric study, especially if groups of authors are formed into particular canons in part due to sharing a similar style and separating themselves via style from their predecessors. This is to say nothing of the correlation between apparatuses of literary and cultural power like the printing press and their fraught historical relationships with marginal groups, especially gender. In literary criticism the issue of gender and the literary marketplace is heavily studied, and various projects have explored and argued sometimes disparate positions.[13] Some have pointed out that historically male-dominated publishers preferred or only allowed the publication of authors with a conventional (read: "male") style (Yarington and De Jong 2007, pp. 1-34), and others have posited that marketplace pressures necessitated women sacrifice "individual styles" to, among other things, sell their books (Paradise 2001, p. 238). Rybicki's recent macroanalysis provides confirmation that, whatever the reason, canonicity seems to be a variable that factors into stylometric results, and it does so not divorced from the issue of gender. Rybicki's Zeta analysis of some 18th-century "canonical" versus "non-canonical" women writers, in his words, "speaks volumes of the possible mechanism of canonization: a female writer enters the generally accepted canon if she writes more like a man" (2016, p. 754).

[11] Upon discovering that our copy of Salinger's *Nine Stories* was actually eight stories—the absent text being "The Laughing Man"—we subsequently added the missing story to the corpus as its own text.

[12] An alternative approach might have been to use a less canonical and more diverse set of authors, but in choosing canonical writers, we are hoping that our dataset will be seen as being particularly aligned with the era in question—if we are to divide this study by literary "era," then we should use the authors who best signify the stylistic principles of such era.

[13] Although somewhat dated, see, for example, Gallagher's *Nobody's story: The Vanishing Acts of Women Writers in the Marketplace, 1670-1820* (Berkeley: U of California P, 1994), particularly her introduction, for a rich bibliography of approaches to this issue on historical, social, and economic grounds, and approaches that derive from media theory, sexuality studies, and feminist theory, all of which take as their basic concern women and their often shifting relationship with literature and the marketplace. For further reading, see also Simons and Fullbrook's collection of essays, *Writing: a woman's business: Women, Writing and the Marketplace* (Manchester: Manchester UP, 1998) for a variety of critical accounts of women writers from many eras negotiating the literary marketplace. More specific to our own corpus, see also Jenny McDonnell's *Katherine Mansfield and the Modernist Marketplace: At the Mercy of the Public* (New York: Palgrave Macmillan, 2010).

The results of any stylometric analysis can only be as unbiased as the data one feeds into it, and the question of how to fundamentally and ideally approach a corpus—and define separate historically and socially unique literary epochs—that attempts to capture periodized styles amid social, historical, and cultural "noise" is, regrettably, beyond the scope of this paper. There are limitations to our approach, but there are also a number of benefits, specifically in terms of the questions it allows us to ask. If, as Rybicki and many others have suggested, women and men do have separate literary styles, what happens when we attempt to isolate the gender groupings in a series of period-based analyses? What role does gender play in the evolution of style over those periods? If one might say that women have entered the canon more so in the contemporary period, is it true that their style inevitably—following Rybicki's suggestion—merges with their male counterparts? If so, how? Does a shift in style from one gender to another become more noticeable in a given period? In the focused context of our sample of traditional, "high canon" literature, do men and women indeed write more similarly over time?

Results

Our initial Delta analyses provided some separation between genders across all periods, with the most distinct separation occurring in the Victorian era. As was expected, the works of respective authors clustered most tightly together, without notable exception; individual authorial style[14] across periods still seems the strongest determiner in our clusters akin to convention. However, group separation generally occurred between genders as well, with male and female authors clustering predominantly with other authors of their gender—though, the algorithm's ability to accurately separate the genders seemed to diminish over time (i.e. inaccurate gender separation became more frequent from the Victorian to the modernist set, and then again from the modernist to the contemporary set), a result that supports Rybicki. These initial analyses looked individually at each period and its set of authors, thus isolating period rather than gender as the primary variable. As the results were somewhat inconclusive, and in the hopes of seeing an evolution of gender and style over time, we decided to do the inverse, and run another set of analyses isolating gender rather than period (Figure 1).

[14] What we define as "style" results from *Stylo*'s approach to preparing texts and measuring stylistic difference. When we run a series of texts through *Stylo*, the settings we employ remove textual features like punctuation, spacing, and other unique formatting until only a "bag of words" remains. The resulting "style" that we analyze thus does not take into account things like form, syntax, grammar, or punctuation. We look, instead, at the unique word counts of each author and their texts, counts which result in unique textual and authorial "signatures" that we can compare and contrast accordingly. For more on this method, see (among many others) Burrows (2004), Eder (2012), and Jockers (2013).

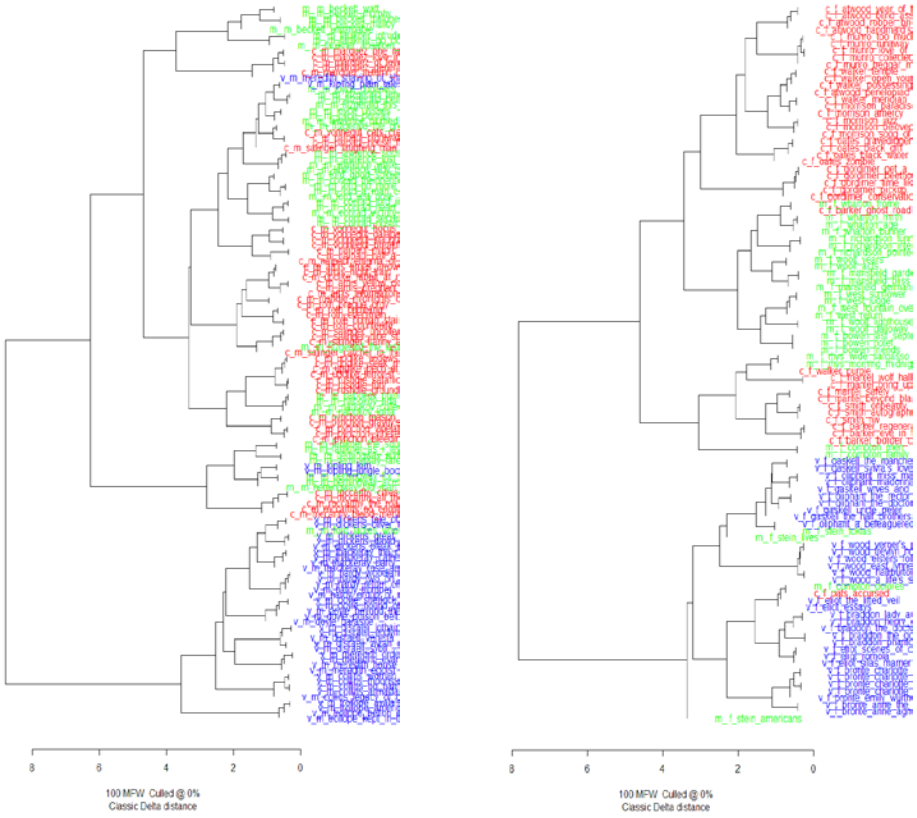


Figure 1. Delta results isolating gender (rather than period). Male authors are on the left, female authors are on the right. Red is contemporary, green is modernist, and blue is Victorian. When tracing a path between two texts along the branches of the dendrogram, the closer those two texts are, the more stylistically similar they are; inversely, the further away those texts are—i.e. the longer the distance one must trace along the branches from one text to another—the less stylistically similar they are.

Having separated the genders, we expected the results to show consistent groupings of the texts based on period, with some minor misattributions for those writers whose styles were particularly unique in comparison to the dominant styles of the period. However, this prediction did not quite hold. The dendrograms of both the male and female sets depict a relatively accurate separation of the Victorian era works from those of the other periods; the Victorian period, unsurprisingly, features the most unique style in the grouping—the one that is less like the others—and in both sets, the modernist and contemporary periods have significantly more overlap in attribution. This is especially true for the male set, the results of which could be said to have as many correct attributions as incorrect. The dendrogram of the female Delta analysis is, of course, not perfect—the function does not really know what to do with Compton, and it is entirely clueless about Stein[15]—but on the whole, their later periods remain more separate and distinguishable than the periods in the dendrogram of the male Delta. The same can only be said for the

[15] We have opted to include Stein as a novelist in this study given her canonical status, but as many critics have noted, the cubist-inspired uniqueness of her style rises well above her literary contemporaries (see, for instance, two canonical studies on this point: Randa Dubnick's *The Structure of Obscurity: Gertrude Stein, Language and Cubism* (U of Illinois P, 1984), and Jayne L. Walker's *The Making of a Modernist: Gertrude Stein from* Three Lives *to* Tender Buttons (U of Massachusetts. P, 1984)). In Stein's case, in other words, one need not a series of computational tools to conclude that her stylistic markers will be outliers when considered among other novelists of the time. Although we tend to agree with that line of thought, her impact on a variety of modernist writers is equally well-documented, and we believe that to omit her would have been to misrepresent her influence on writers in both the male and female sets.

males in the Victorian period—the modernist and contemporary male groups could not be distinguished as neatly or as easily from one another as the respective female groups.

Although these Delta results are interesting in their own right, the Delta MFW lists that were produced only depict the most common function words employed by the authors, and they provide little insight into how (or why) the usages differ—in other words, what we might glean from the different usage rates of certain prepositions between genders and periods is not altogether obvious. However, as a measure designed predominantly to detect distinctive "content" (rather than function) words, Zeta yielded arguably more interpretable data (Figure 2).

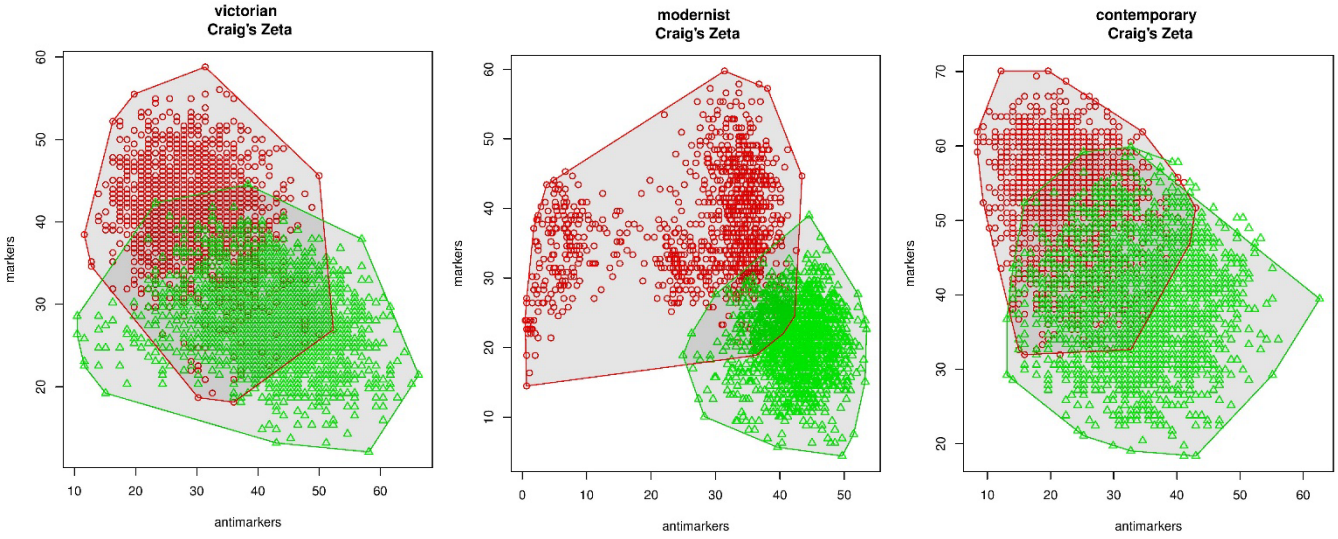


Fig. 2. Zeta results isolating period (rather than gender). Left is the Victorian set, center is the modernist set, and right is the contemporary set. Preferred female "markers" are in red, and avoided female "antimarkers" are in green.

The Zeta graphs depict two groups of markers, one preferred (red, y-axis "markers") and one avoided (green, x-axis "antimarkers") by the female author set when compared to the male author set. The words significantly preferred by female authors (and thus more avoided by male authors) tend toward higher *y* values and lower *x* values, and those words avoided by female authors (and thus more preferred by male authors) inversely tend toward lower *y* values and higher *x* values. The Victorian graph is somewhat unique given its overlap of markers—it is the only graph where $y = x$ lands you in between both sets of markers. The Victorian period, it seems, has caused the algorithm the most trouble in generating a list of words preferred and avoided that does not overlap in a significant way. The male and female authors of the modernist period, conversely, have next to no overlap, and are distinct in their markers. The contemporary period shows some overlap, but overlap that is in part visually deceiving—of the markers shared by the sexes, most tend to be "more" uniquely female (overlapping with higher *y* values than *x* values). The male authors, in other words, have more disparity among the markers they prefer, whereas the female authors have a style that is more focused than ever before, with gender markers that have become significantly more distinct over time.

Compared to their female counterparts, males seem to always write somewhat similarly to other males, especially in the modernist period, where the canonical male authors prefer the same set of markers with few exceptions; but females, in the move from the Victorian period to the modernist period especially, reflect much greater fluctuation in their markers. This trend became slightly more evident when we isolated gender rather than period in a subsequent Zeta analysis (Figure 3). Here, the trends of the genders and their respective markers roughly mirror one another in their greater evolution through the

periods, a sign that perhaps overall, changes in style follow similar patterns, regardless of gender. However, there are also important differences between the period-to-period evolutions of the male and female markers. From the Victorian period to the contemporary period, the female author markers have evolved considerably, spanned by the stylistic explosion of the modernist period—nowhere in the female set do the markers overlap significantly. Each era of women's literature, it seems, has a unique stylistic signature that is clearly distinguishable from the era(s) that came before or after it. The same cannot be said of the males, whose set depicts stylistic markers that evolve much less distinctly; the male contemporary and modernist groups, for example, have a much closer relationship stylistically than do the female groups of the same periods. While there is of course some stylistic evolution in the male sets from the modernist period to the contemporary, current male authors write much more like their modernist counterparts than female authors write like theirs.

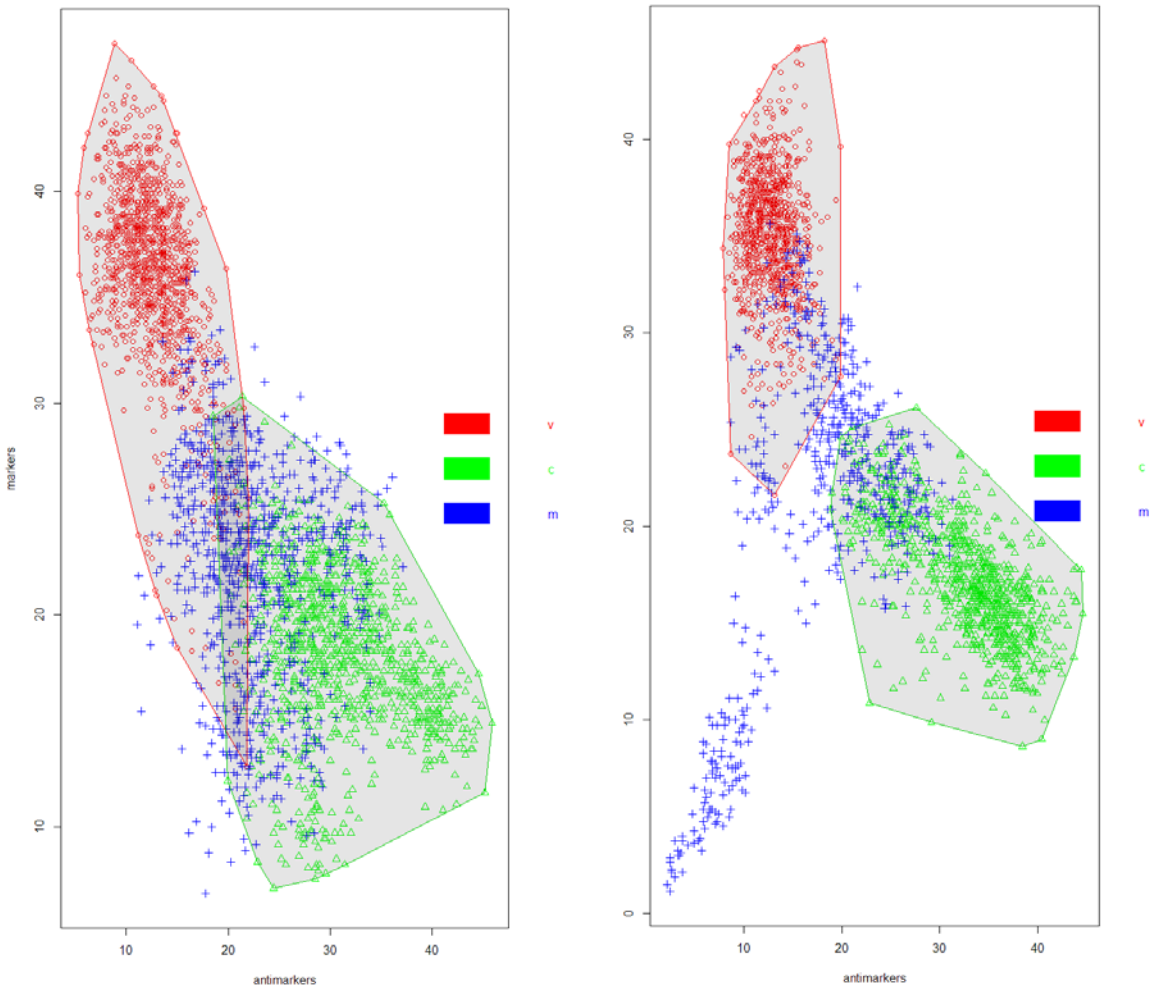


Figure 3. Zeta results isolating gender (rather than period). Male sets on the left, female sets on the right. Red points are Victorian, green are contemporary, and blue are modernist markers.

After seeing this separation, we took the resulting wordlists from each period's Zeta analysis and combined them, using the subsequent MFW lists—composed now of both the words preferred (female) and words avoided (male) for each period—to run another Delta on the three literary epochs with the same settings as our initial analyses. Employing Delta here allowed us to see how accurate our Zeta

analysis wordlists were, and how precisely the previous function created unique wordlists for each period and each gender. Our question became, given the machine-identified list of words preferred and avoided by the females and males of each period, how accurately will a stylistic analysis separate the genders? Any resulting misattributions show us an area where the styles of the genders remain so close as to be functionally unattributable, or identify periods, texts, or authors that deviate from their "suspected" groups. While our resulting dendrograms (Figure 4) show misattributions in all periods, six occur in the Victorian (1 female, 5 male), *many* in the modernist (4 female and 3 male most notably, but more to say on this), but only two in the contemporary (both male). These exceptions hold valuable insight.

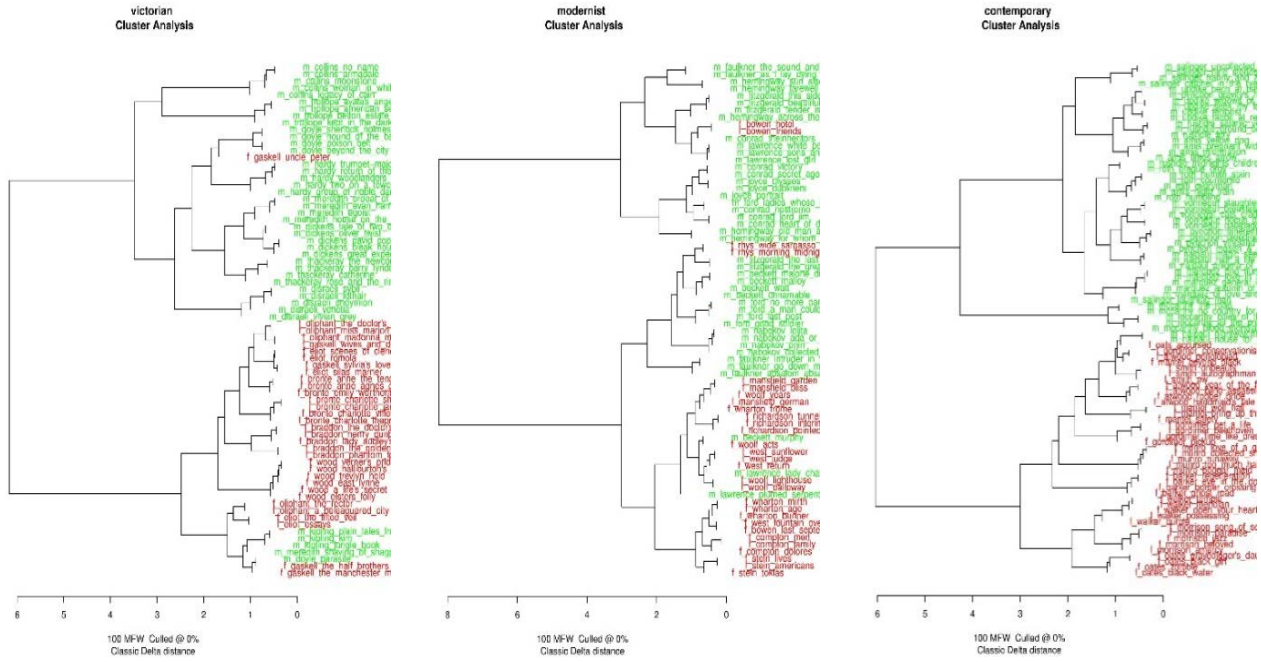


Figure 4. Delta results with the custom wordlists applied. Again, the green are male texts, and the red are female.

The algorithm, for instance, seemed to have trouble clustering the works of Elizabeth Gaskell in the Victorian period. Gaskell's "Uncle Peter" is one of her three included short stories, but the only one written more than five years prior to the rest of her stories (which were otherwise published within a year of one another), and it may indicate a stylistic change within that time. Additionally, Doyle's "Parasite," an 1894 novelette, clusters more closely with Gaskell's remaining two short stories (written 35 years earlier) than his own works (written within 5 years), an association that may have something to do with the non-crime-related theme of Doyle's story of love and the occult. Despite these and the two misattributed exceptions of Kipling and the one of Meredith, the Victorian authors as a whole cluster with themselves; the texts of Dickens and Meredith may group together under a more specific clade, but they do so without sharing or mixing texts among their individual sections. The works of George Eliot—who is, along with the Bronte sisters, writing under a male pseudonym—remain more spread out, though accurately assigned to the female set. Her "Essays" and "Lifted Veil" are both selections of her nonfiction works we slipped in at the last moment as a kind of litmus test of gender and genre; although she does not write as stylistically similar to her own creative work, the essays still cluster with the other works written by women, suggesting that perhaps in the Victorian period, gender is in fact a more important stylistic determiner than genre.

The modernist graph is much messier, and the Delta analysis is much less able to differentiate

clearly between the genders. There are now three distinct clades as opposed to two, with seven total misattributions among them (if we do not count one of these sections as misattributed, as we just as easily might do). Most interesting in this period is the phenomenon of gender clustering but not author clustering: every author except for Stein and Richardson have at least one work that clusters closer to another author than themselves. However, this effect does not hold quite as truly for the males, who have much more consistent clustering throughout their set, with Beckett and D.H. Lawrence the notable exceptions. This is to say nothing of Elizabeth Bowen, who has two of her three novels clustering with the males, or Jean Rhys, whose two novels are also misattributed, and who thus certainly deserves further study in this regard.

However, the macro trend of the dendrograms is even more interesting for our purposes, and as we move from the modernist to the contemporary dendrogram, you may begin to see why. The analysis has gone much smoother here, with only two total misattributions: V.S. Naipaul's novel, *A House for Mr Biswas,* and his short story collection, *The Nightwatchman's Occurrence Book.*[16] And despite a few minor exceptions, the authors' works, more like the Victorian period, now consistently cluster among themselves rather than other authors—though this is, again, exactly what one should expect. With a number of misattributed texts, the Victorian Delta analysis did not perfectly separate the genders, and the modernist Delta failed even more spectacularly. But as the contemporary dendrogram illustrates, we arrive from the stylistic chaos of the modernist period into a remarkably accurate, gender-split "harmony" in the contemporary set, where the two sexes write in distinct styles that are most easily distinguishable from one another. This is significantly more surprising. Contrary to popular belief, it seems males and females write more stylistically distinct in the present day than they did in the Victorian period.

Conclusions and Interpretations

A number of conclusions might be drawn from this analysis, but we want to focus on two broader points. The first details, on a more micro scale, the extent to which gender markers have evolved over these literary periods. The second comments on the macro-evolution of gender markers over the periods and what the stylistic evolution suggests about gender and literary history.

To begin, we want to address the results of Figure 4, which suggest, contrary to popular critical opinion, that there are not only stereotypical stylistic differences between men and women fiction writers, but that the stylistic separation between genders is greater (or, more easily distinguishable) than ever before. If this is indeed true, what do their respective styles look like? Have these styles progressed or changed over the past couple of centuries?

When close reading the content of the Zeta wordlists from across all periods, a number of stereotypically gender-split patterns emerge in every set.[17] Females generally tend toward the language of place in the private or micro sense—*home, kitchen, church, hallway, school*—whereas males throughout

[16] As Amy Fallon of *The Guardian* tells it, Naipaul was once asked in an interview whether he considered "any woman writer his literary match. He replied: 'I don't think so. […] I read a piece of writing and within a paragraph or two I know whether it is by a woman or not. I think [it is] unequal to me.'" That Naipaul's literary markers actually mirror what he condescendingly deemed the "sentimental" style and "narrow view of the world" of women authors is thus an especially lovely irony.

[17] It's worth noting that we formed these observations and thematic groupings via our own close-reading of the Zeta analysis wordlists, and thus we cannot count out the possibility that a kind of response bias is at work in these "stereotypes" (e.g. literary scholars imbedding previously held conceptions of gender into their classifications). Baker (2014) argues this exact point when he claims "that where gender differences emerge [in digital analysis of corpora], it is likely that they are due to expectations about how males and females should use language, based on the contexts they are in" (203). Although we admit we cannot rule out the possibility—and we encourage critical engagement with our arguments from all readers as a result—as far as anyone is able, we have set out intentionally to avoid such biases and checked personal gender politics at the door.

all periods tend toward greater spaces—*country, earth, city, town, world*. The language of the former is predominantly related to the local and domestic spheres, and the latter follows thematically broader notions of space and location. This distinctive preference between males/females and macro/micro diction even extends to positions, travel, and time. The male set includes directional language—*east, west, south, north, center, upward, beyond, forth*—and terms of travel—*airplane, car, traffic, miles, speed*—that far outnumber comparable occurrences in the female set. Female authors, on the other hand, while generally other-oriented, seem to prefer language of positioning related to the body—*hair, fingers, skin, eyes, heart, face, cheeks, dress, gown*—or immediate, local surroundings—*door, window, chair, stairs, curtains, garden, room.* In this respect the male set of markers are almost always grander in scope than their female equivalents, both in terms of space—*full big, giant, great, grand, deep*—and time—*history, forever, hours, late, century, past, present*.

A not unrelated phenomenon, in our minds, concerns the authors' uses of what we will call "language of certainty." While males throughout all sets have markers distinctly related to language of certainty and confidence—*exactly, absolutely, declared, therefore*—female markers remain throughout all sets trending toward language less confident, of uncertainty (or at least *less* certainty) and doubt—*seemed, perhaps, believed, scarcely, likely*. Perhaps more specifically, along with an implied uncertainty, related female markers—*wished, hoped, wondered, expect, thinking*—intimate a greater focus on interiority than is present in the males' works, the language of which tends to be not only "larger," but also focused on exteriority and the public sphere—*famous, power, everything, film, TV*. Interestingly, compared to their male counterparts, female authors across all periods employ more terminology of relationships—*husband, mother, sister, brother, married, father, God, friends, others, servants, children.* While many of the terms are family oriented, much of the language is more generally relational (i.e. related to "others" and not "self"). To that end, female authors employed more language of interaction than males—*laughed, spoke, saying, talking, exclaimed, paused, thank, welcome, asked, interrupted*—which suggests that when female authors use language that might be considered more "exterior," they do so with a greater focus on community, or a leaning closer to selflessness than selfishness. Here, our results correlate with those presented in Hoover's study.

Of course, to say simply that women still write about "family" and "the home" more than men fails to take into account that much of contemporary women's literature attempts to complicate or subvert many of those stereotypically "female" themes that otherwise seem to be supported by the data. The major limitation of distant reading is that it can remove much of a work's context, and just because we see that men and women tend to use words that indicate a particular topic is no barometer of how they are treating said topic. Indeed, reconciling this thematic data with, on the one hand, a resistance toward essentializing attitudes toward women, and on the other, a sensitivity toward the social hardships women still undergo in male-dominated literary circles, is somewhat of a difficult task, and it is no surprise reviewers might be put at fault for failing to do so adequately. Contrary to the conclusion of Piper and So, our analysis indicates that the gender bias in book reviewing may not be a bias at all. It may be likelier that women authors *are* dealing with a different set of topics in their work than men are in theirs—though, this statement alone does not tell the whole story. Perhaps needless to say, contemporary women writers are no longer composing the stereotyped narratives of the traditional Victorian romance novel; women's literary styles, while certainly engaging with many of those same themes over time, have changed dramatically with regard to how women writers employ what may have once been deemed "classically female" topics. However—and this is vital—we must realize that stylometric methods that rely on word frequencies effectively cannot tell the difference, and report word usage similarities that one might mistake for thematic similarities.

A few illustrative examples may help contextualize this point further. Take, for instance, a couple of well-known Victorian novels by women authors—Elizabeth Gaskell's *Wives and Daughters: An Everyday Story* (1864) and Charlotte Bronte's *Jane Eyre* (1847). The opening description of both novels center around domestic, "micro" spaces. Gaskell's narrator exemplifies this zoomed-in nature of Victorian women's writing in the first sentences of *Wives and Daughters*, where she frames her story's setting: "In a country there was a shire, and in that shire there was a town, and in that town there was a

house, and in that house there was a room, and in that room there was a bed, and in that bed there lay a little girl; wide awake and longing to get up, but not daring to do so." *Jane Eyre* starts in a similarly domestic manner; "out-door exercise," as Jane points out, "now out of the question," we're presented with a description of Jane's cousins "clustered round their mama in the drawing-room: she lay reclined on a sofa by the fireside," while our heroine Jane enters the "breakfast-room […] into the window-seat," where "folds of scarlet drapery shut in my view to the right hand; to the left were the clear panes of glass, protecting, but not separating me from the drear November day." The novels' openings include many of the previously identified "stereotypically female" words and themes. Many of the terms that appear in the opening paragraphs of these two novels also feature prominently in the female Zeta wordlist—*fingers, feet, heart, drawing-room, sofa, mama, crying, children, little, curtain, window, church, housewife.* On a more theoretical scale, in both openings descriptions of interior spaces, relationships, and arguably even senses of doubt or uncertainty are narrowed to the immediate surroundings and bodies of the characters, a style more frequently attributable to women than men.

Now, take instead two well-known novels from the contemporary women's set—Zadie Smith's *NW* (2012), say, and Margaret Atwood's *Year of the Flood* (2009). The initial paragraphs of Smith's first chapter depict a woman describing "Anti-climb paint [that] turns sulphurous on school gates and lampposts" while she lies "in a hammock, in the garden of a basement flat. Fenced in, on all sides." Atwood's *Year of the Flood*, on the other hand, opens with an equally ironic not-so-picturesque rooftop sunrise: "mist rises from among the swath of trees between her and the derelict city." Atwood's novel is the second book of her speculative realist, dystopian trilogy, one both narratively and generically distinct from what one might call Smith's experiment in urban-realism-meets-pseudocomedy. Yet even so, both authors feature a descriptive vocabulary in their opening pages that not only aligns more with one another than someone might expect, but one that also aligns with the preferred language of their Victorian predecessors: *stairs, hallway, window, crying, birds, umbrellas, flower, garden, gate, balcony*.

In all four openings, notably, a female character is described as being trapped in her local circumstances, literally or figuratively, innocuously or harmfully, or some combination of them all. This was, admittedly, pure happenstance, but the similar scope of entry for each of these canonical women authors seems to result in similar language, even despite their period separation. However, the manner in which these settings and narrative contexts crop up have fundamentally changed, and one would imagine that contemporary women authors, no doubt aware that domestically oriented stories and language have become tropes, seem to employ them with the intention of tearing them down, problematizing the assumptions that go along with them, challenging the social and cultural structures that normalized them in the first place, and in the end, effectively rehabilitating their relevance. Smith's "anti-climb paint" and "fenced in" flat, for example, are no doubt symbolically oriented toward gender as much as class. Just as Atwood's horizon, "devoid of life" despite the sunrise, is every bit the dystopian nightmare when placed next to Gaskell's first chapter heading, "The Dawn of a Gala Day," the result of Smith's *NW* is a story—seemingly of startling computational and linguistic similarity to Bronte's—far removed from the drawing-rooms of *Jane Eyre* and the 19th century.

While the stereotypical distinctiveness of the gender markers of men's and women's canonical literatures provides a number of interesting lines of inquiry, it more importantly calls attention to the care necessary when attributing generalized macro-computational findings to individual texts and authors. Gender, we know, does not occur in an essentialist, ahistorical vacuum, and ignoring the social contexts that these works of literature are formed in (and help form) would be a drastic oversight indeed. As Rybicki himself considers, "there are greater powers than [gender] that influence the evolution of literary lexicon […]." Time, as Rybicki postulates, is indeed "much, much more important" (2016, p. 759) to gender and literary style. Taking a brief step back to reexamine what a greater look at the evolution of gender markers in canonical literature might tell us, our previous results (Figures 2-4) again have much to provide.

Although we cannot be sure what exactly has gone on between the modernist and contemporary groups, perhaps the separations of these two periods in the Delta and Zeta analyses indicate that the stylistically radical literary apparatuses of the modernist era provided female authors a much needed

period of transition—or, more precisely, a much needed space for evolution. Our analyses show that female authorship began in the Victorian period as more necessarily imitative of their socially privileged male counterparts, who historically and through various social mechanisms controlled and defined the literary standard. But after the 20th century rolled around and the modernist movement set a precedent for valuing unique, shifting, and marginalized styles, one might note that female authors emerged, really for the first time, as a truly consolidated group with a distinct sense of literary identity.

The levels of authorial pronoun usage in our results are also significant in this regard. Like Pennebaker initially uncovered in *The Secret Life of Pronouns*, in our corpus of canonical literature female authors tended to use more pronouns than males, especially over time—while the pronoun usage of females in the Victorian period is relatively scant, the usage increases in the modernist period, and does so again in the contemporary period. While this growth may simply mirror stylistic trends of the respective times, given the disparate evolution of male pronoun usage, which stays more similar throughout all periods, we believe another association can be made. The origins of Pennebaker's book, as he tells it, started about a quarter century ago when he was sorting through diary entries of individuals who had undergone traumatic experiences. He found that the keeping of a diary, the journaling process, was a process which seemed to help some people overcome their trauma but not others. As he was sorting through and trying to distinguish what was different about those successful cases, Pennebaker discovered that the use of pronouns in particular was the best indicator of improved mental health. Recovery from trauma, he maintains, seems to require some kind of perspective switching, a reflection on the problematic experiences from different points of view—and in language, this capacity derives from the power of pronouns (Pennebaker, p. 13-15).

Extending Pennebaker's findings to the results from our study, one might begin to trace out a similar trauma narrative in literary terms by identifying the correlation between the historical and patriarchal oppressions of women and the moment of evolution in their literary style. In the modernist period especially, one might argue that females most completely stepped away from the otherwise binding, patriarchal conventions of the past—both socially and, as we have seen, in their fiction—and began experimenting with other perspectives. Contrarily, the male authors always remain solidified and distinct in their markers, with, say, a historically consistent social privilege that granted a fixed literary identity. Building on Rybicki's observation "that women become part of the canon if/when they write a little more like men" (2016, p. 759), our results suggest that after the Victorian period, women became part of the canon if/when they wrote entirely *unlike* men, when they developed a style unique from their predecessors, both male and female. When we consider time more heavily, we see that the women authors of the modernist period, who pioneered the means to create various spaces and perspectives for themselves, were the first to create both rooms and literary markers of their own.

It should be noted again, however, that these are claims we can reliably make only about this collection of canonical men and women authors, and it is somewhat debatable given this analysis—limited as it is to this particular dataset—as to whether this trend would continue between various levels of canonicity. Moreover, the fact that women were essential to the project of literary modernism in the early 20th century is by no means a new claim, and a number of projects trace the immense amount of work done by women that allowed the models and movements of modernism to be possible from the first.[18] As an addendum to those rich critical engagements tracking the influence of these women, then,

[18] Outside of their own literary production, many of the included women authors engaged in remarkable systems of literary exchange, running private libraries, bookstores, and independent publishing houses, or working as art collectors, salon hosts, or magazine editors, etc. Numerous critical projects aim, in part if not in full, to resituate these women's roles as essential to the traction of modernism as a literary and cultural movement—to name a few in no particular order: Shari Benstock's *Women of the Left Bank: Paris, 1900-1940* (Austin: U of Texas P, 1986), Dean J. Irvine's *Editing Modernity: Women and Little-Magazine Cultures in Canada, 1916-1956* (Toronto: U of Toronto P, 2008), Angela Kershaw's *Forgotten Engagements: Women, Literature and the Left in 1930s France* (Amsterdam: Rodopi, 2007), Jayne E. Marek's *Women Editing Modernism: "Little" Magazines and Literary History* (Lexington: UP of Kentucky, 1995), Suzanne Clark's *Sentimental Modernism: Women Writers and the Revolution of the Word* (Bloomington: Indiana UP, 1991), and Alice Gambrell's *Women Intellectuals, Modernism, and Difference:*

our project accounts more fully for the stylistic achievement of what we now view as a set of canonical modernist women writers. They were not merely the underlying engine that propelled modernism's social influence and cultural movements with their mediations—they were the most potent stylistic fuel their own modernist machines ran on. Women's collective, macro-contribution to the developing style of modernism has, in this respect, been seriously overlooked.

Acknowledgements

We would like to thank the College of the Liberal Arts and University Libraries, Pennsylvania State University, for supporting our research efforts. Particular thanks to Christopher P. Long, Patricia Hswe, and Barbara Dewey.

An earlier iteration of this research was presented at *Digital Humanities*, Sydney, July 2015.

Biographical Notes

Sean G. Weidman is a PhD student in English at the Pennsylvania State University. His research is predominantly focused in literary modernism—specifically systems of hospitality and sociability in the early 20th century—and he has an ongoing interest in DH methods that offer different means to literary analysis, including computational stylistics, topic modeling, and network analysis.

James O'Sullivan is Lecturer in Digital Humanities at University College Cork. He has previously held faculty positions at the University of Sheffield and Pennsylvania State University. He has been published in a variety of leading international DH journals, including *Digital Humanities Quarterly*, *Digital Scholarship in the Humanities*, and the *International Journal of Humanities and Arts Computing*. He co-edited the co-editor of *Reading Modernism with Machines* (Palgrave Macmillan 2016) alongside Shawna Ross. More on James and his research interested can be found at josullivan.org.

References

Argamon, S. (2008). "Interpreting Burrows' delta: Geometric and probabilistic foundations." *Literary and Linguistic Computing*, 23(2), 131-147.

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). "Gender, genre, and writing style in formal written texts." *Text*, 23(3), 321-46.

Baker, P. (2010a). *Sociolinguistics and corpus linguistics.* Edinburgh: Edinburgh UP.

Baker, P. (2010b). "Will Ms ever be as frequent as Mr? A corpus-based comparison of gendered terms across four diachronic corpora of British English." *Gender and Language*, 4(1), 125-49.

Baker, P. (2012). "Corpora and gender studies." *Corpus Applications in Applied Linguistics*, edited by Hyland, K., Meng Huat, C., and Handford, M., eds. London: Continuum. 100-16.

Baker, P. (2014). *Using corpora to analyze gender.* London: Bloomsbury.

Burrows, J. (2004). "Textual analysis." *A companion to digital humanities.* Ed. Susan Schreibman, Ray

Transatlantic Culture, 1919-1945 (New York: Cambridge UP, 1997).

Siemans, and John Unsworth. Oxford: Blackwell.

Burrows, J. (2006). "All the way through: Testing for authorship in different frequency strata." *Literary and Linguistic Computing*, 22, 27-47.

Eder, M. (2012). "Distance measures and nearest neighbor classifications." *Workshop on Stylometry at the Leipzig European Summer School in July 2012.*

Eder, M. (2013). Computational stylistics and Biblical translation: how reliable can a dendrogram be? *The translator and the computer*, edited by T. Piotrowski and Ł. Grabowski. Wrocław: WSF Press, pp. 155-70.

Eder, M., Kestemont, M. and Rybicki, J. (2013). "Stylometry with R: a suite of tools." D*igital Humanities 2013: Conference Abstracts*. Lincoln (NE): University of Nebraska, Lincoln. 487-89.

Fallon, A. (2011). Interview with V.S. Naipaul. "VS Naipaul fins now woman writer his literary match – not even Jane Austen." *The Guardian*, 1 June.

Fine, C. (2010). *Delusions of gender: How our minds, society, and neurosexism create difference.* New York: W.W. Norton & Company.

Herring, Susan C. and Paolilo, John C. (2006). "Gender and genre variation in weblogs." *Journal of Sociolinguistics*, 10(4), 439-59.

Herrmann, J.B., van Dalen-Oskam, K. & Schöch, C. (2015). "Revisiting Style, a Key Concept in Literary Studies." *Journal of Literary Theory*. 9(1), 25–52.

Hoover, David L. (2008). "Quantitative analysis and literary studies." *A Companion to Digital Literary Studies*, edited by Schreibman, S. and Siemens, R. Oxford: Blackwell Publishing. 517-33.

Hoover, David L. (2013). "Textual analysis." *Literary Studies in the Digital Age*, edited by Kenneth M. Price and Ray Siemens. Modern Language Association of America. http://dlsanthology.commons.mla.org/textual-analysis/.

Iyeiri, Y., Yaguchi, M., and Baba, Y. (2011). Principal component analysis of turn-initial words in spoken interactions. *Literary and Linguistic Computing*, 26(2), 139-52.

Iyeiri, Y., Yaguchi, M., and Okabe, H. (2005). "Gender and style: The discourse particle like in the corpus of spoken professional American English." *English Corpus Studies*, 12, 37-54.

Janssen, A. and Murachver, T. (2004). "The relationship between gender and topic in gender-preferential language use." *Written Communication*, 21, 344-67.

Jockers, M. (2013). *Macroanalysis: Digital methods and literary history*. Urbana: University of Illinois Press.

Koppel, M., Argamon, S., and Shimoni, A. R. (2002). "Automatically categorizing written texts by author gender." *Literary and Linguistic Computing*, 17(4), 401-12.

Macalister, J. (2011). Flower-girl and bugler-boy no more: changing gender representation in writing for children. *Corpora*, 6(1), 25-44.

McHugh, M. and Hambaugh, J. (2010). "She said, he said: Gender, language, and power." *Handbook of Gender Research in Psychology. Volume 1: Gender Research in General and Experimental Psychology*. 379-410.

Mikros, G. K. (2013). "Systematic stylometric difference in men and women authors: a corpus-based study." *Issues in Quantitative Lingustics 3*, edited by Kohler, R. and Altmann, G. Ludenscheid: RAM-Verlag. 206-223.

Moon, R. (2014). "From gorgeous to grumpy: adjective, age, and gender." *Gender and Language*, 8(1), 5-41.

Mukherjee, A. and Liu, B. (2010). "Improving gender classification of blog authors." *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*: 207-17.

Rybicki, J. (2006). "Burrowing into translation: character idiolects in Henryk Sienkiewicz's trilogy and its two English translations." *Literary and Linguistic Computing*, 21(1), 91-103.

Rybicki, J. (2010). "Translation and delta revisited: When we read translations, is it the author or the translator that we really read?" Conference Paper. London: *Digital Humanities 2010*.

Rybicki, J. (2011). "Alma Cardell Curtin and Jeremiah Curtin: the translator's wife's stylistic fingerprint." Stanford: Digital Humanities 2011.

Rybicki, J. (2013). "The translator's other invisibility: Stylometry in translation." *SLE 2013 Annual Meeting*. Croatia, Split University, September 2013.

Rybicki, J. (2016). "*Vive la difference*: Tracing the (authorial) gender signal by multivariate analysis of word frequencies." *Digital Scholarship in the Humanities*, 31(4), 746-61.

Rybicki, J. and Heydel, M. (2013). "The stylistics and stylometry of collaborative translation: Woolf's *Night and Day* in Polish." *Literary and Linguistic Computing*, 28(4), 708-17.

Paradise, N. (2001). "From poet to novelist: Women writers and the literary marketplace." *Eighteenth-Century Women: Studies in Their Lives, Work, and Culture*, 1, 237-62.

Partington, A., Duguid, A., and Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS).* Philadelphia: John Benjamins Publishing Company.

Pennebaker, J. (2011). *The secret life of pronouns: What our words say about us.* London: Bloomsbury Press.

Piper, A. and So, R. J. (2016). "Women write about family, men write about war." *New Republic*, 8 April.

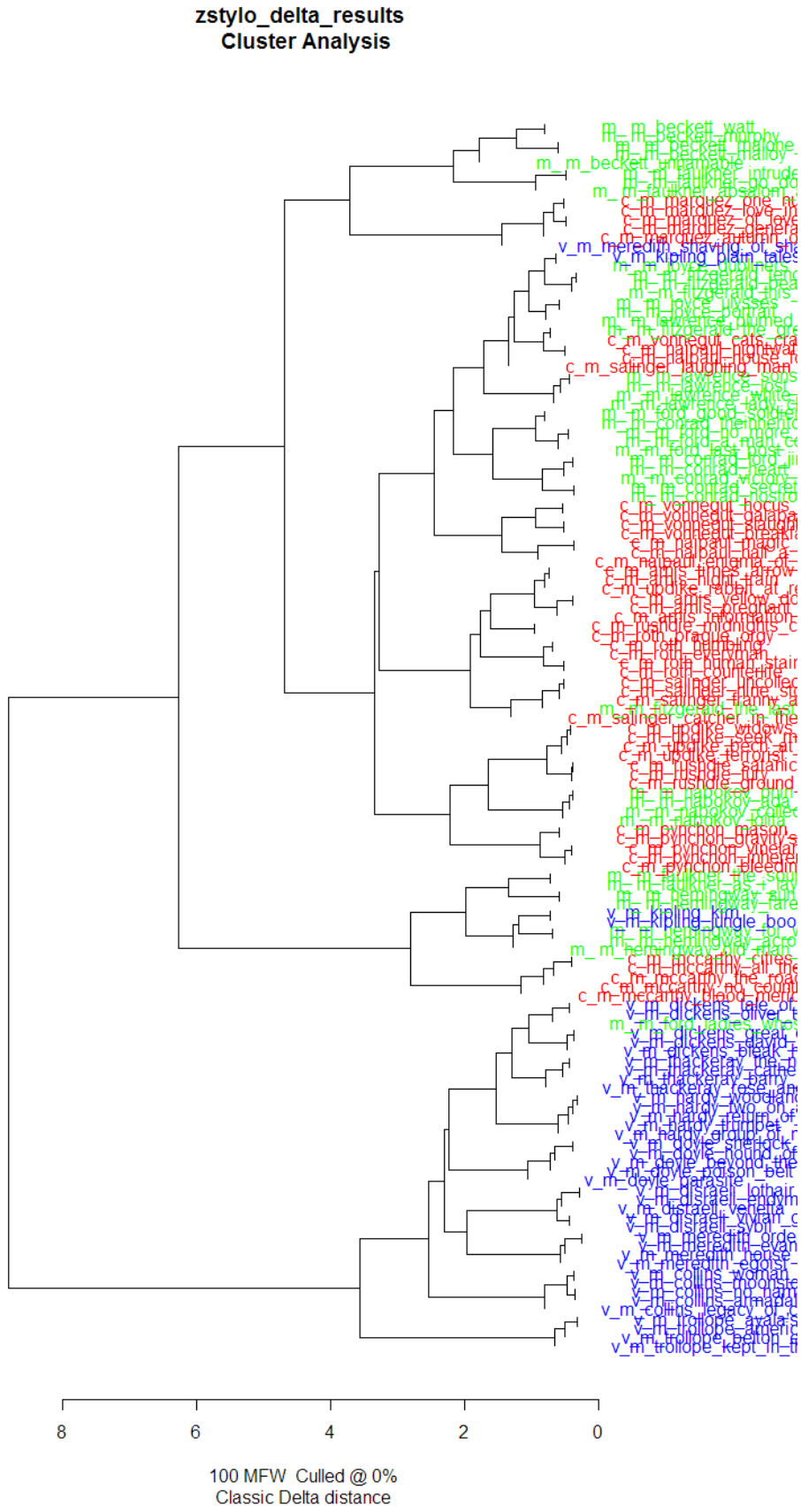
Sarawgi, R., Gajulapalli, K., and Choi, Y. (2011). "Gender attribution: Tracing stylometric evidence beyond topic and genre." *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 78-86.

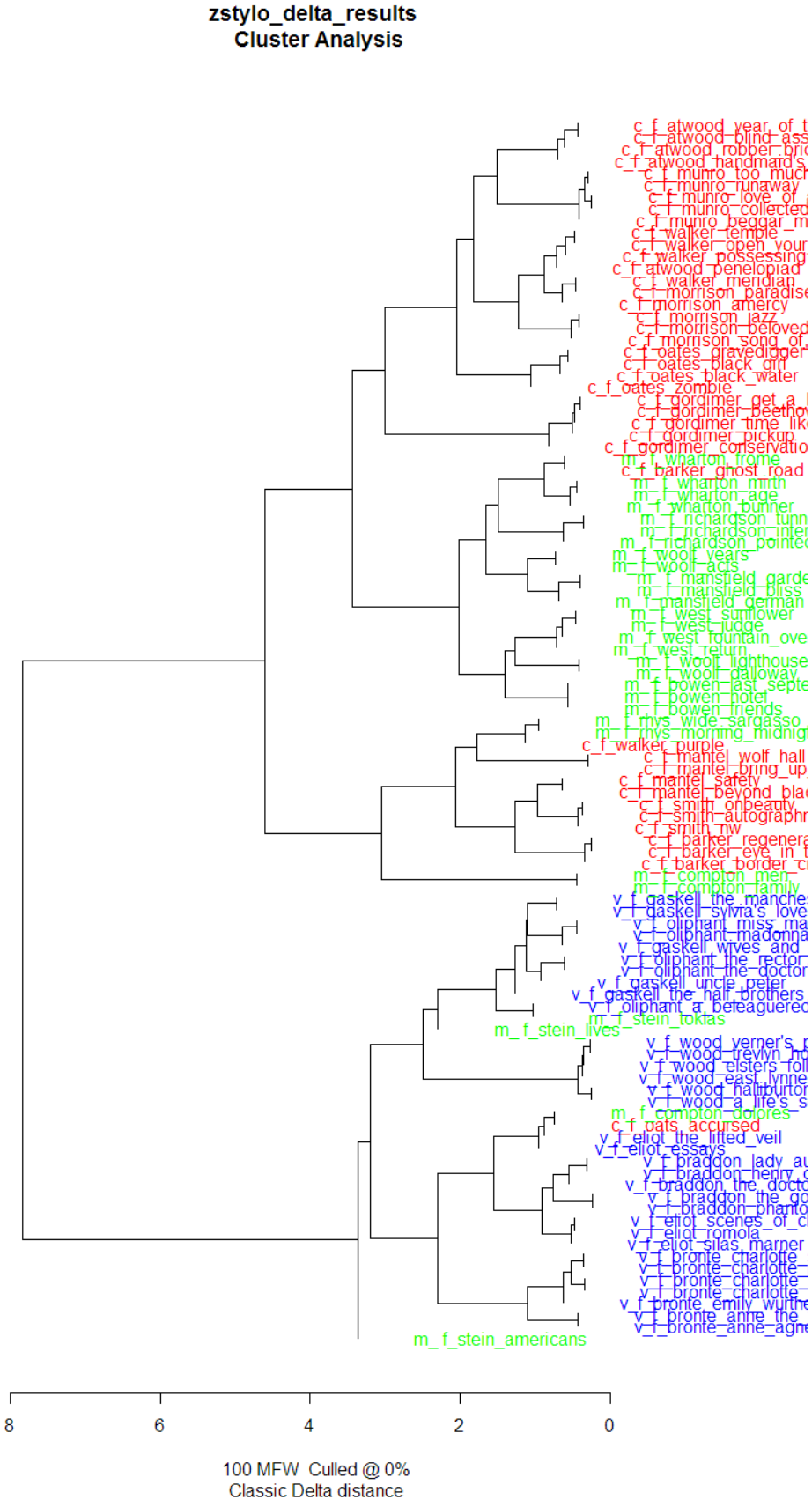
Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2005). "Effects of age and gender in blogging." Symposium Paper. *American Association for Artificial Intelligence*.

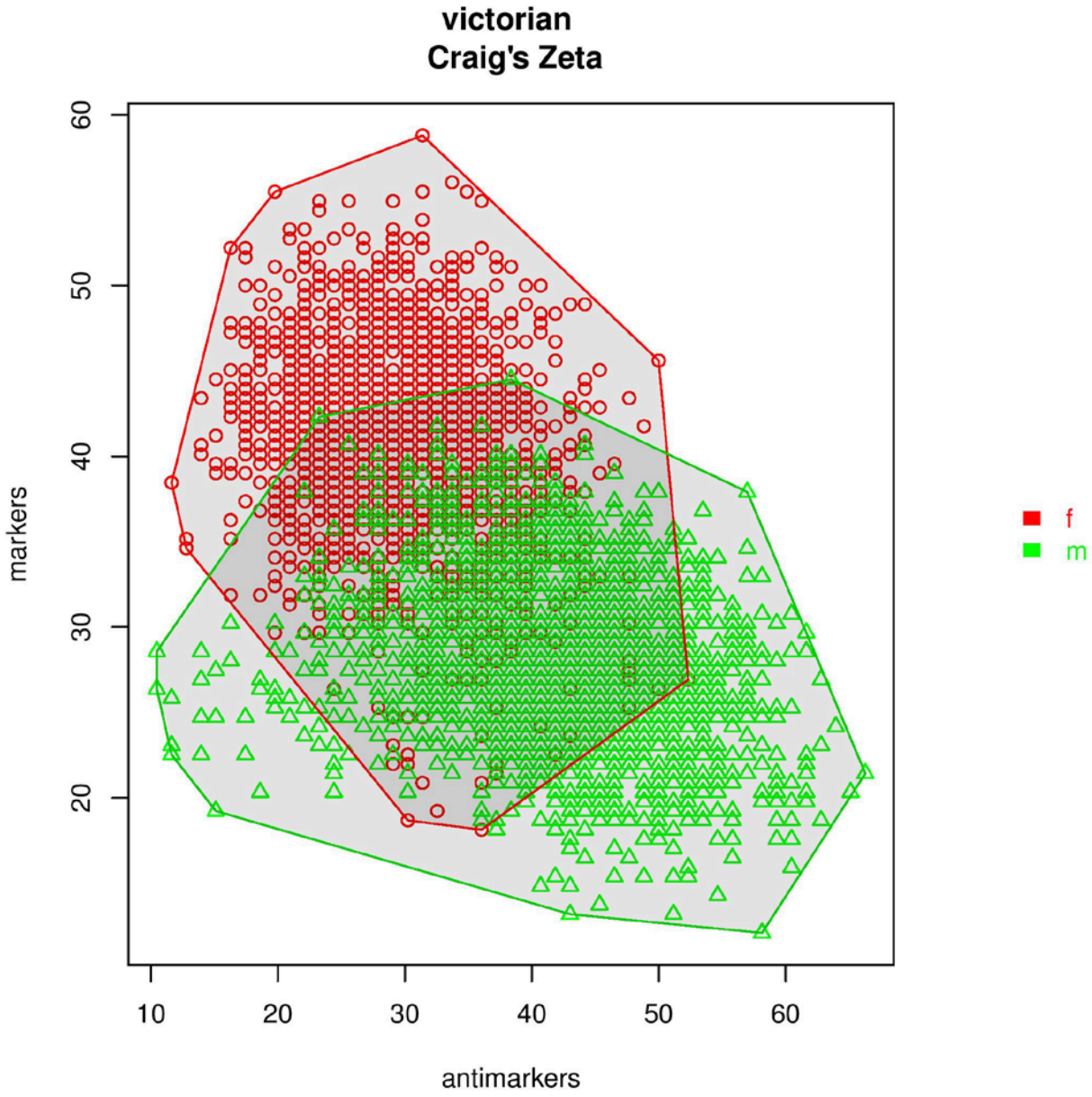
Yarington, E. and De Jong, M. (2007). *Popular nineteenth-century American women writers and the literary marketplace.* Newcastsle: Cambridge Scholars Publishing.

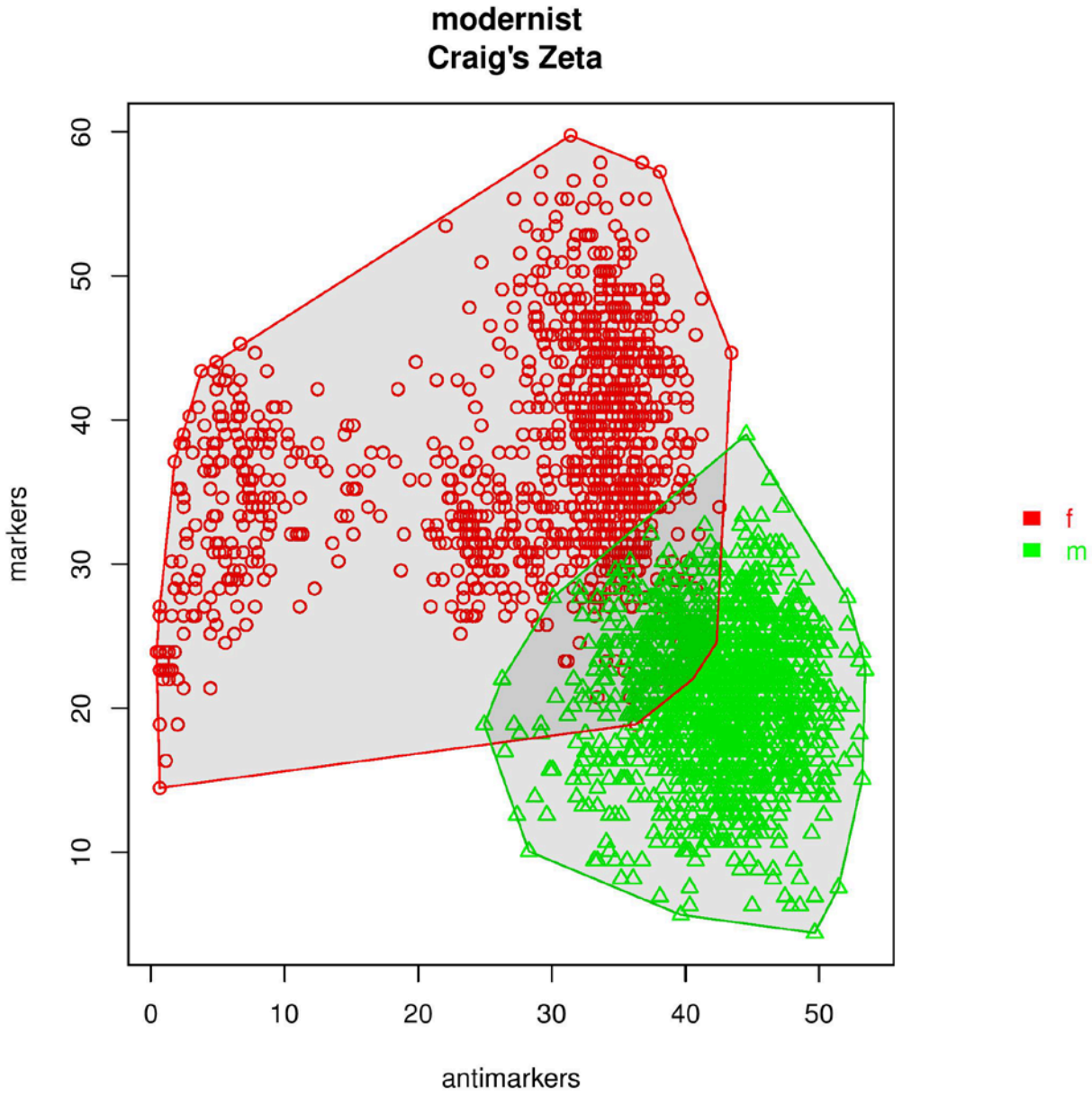
Appendix A: Images

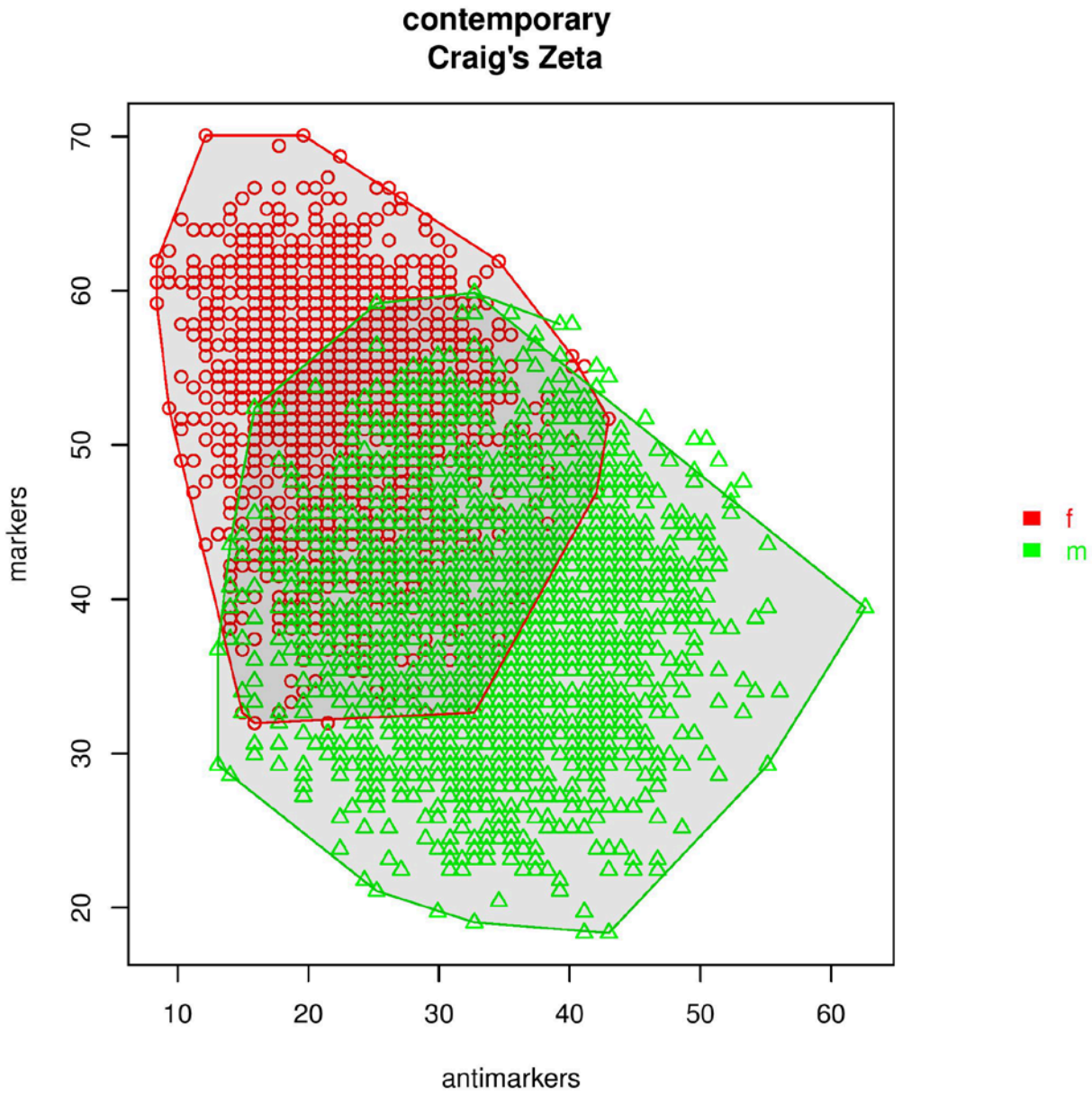
(reproduced at full resolution, in order of appearance in the text)

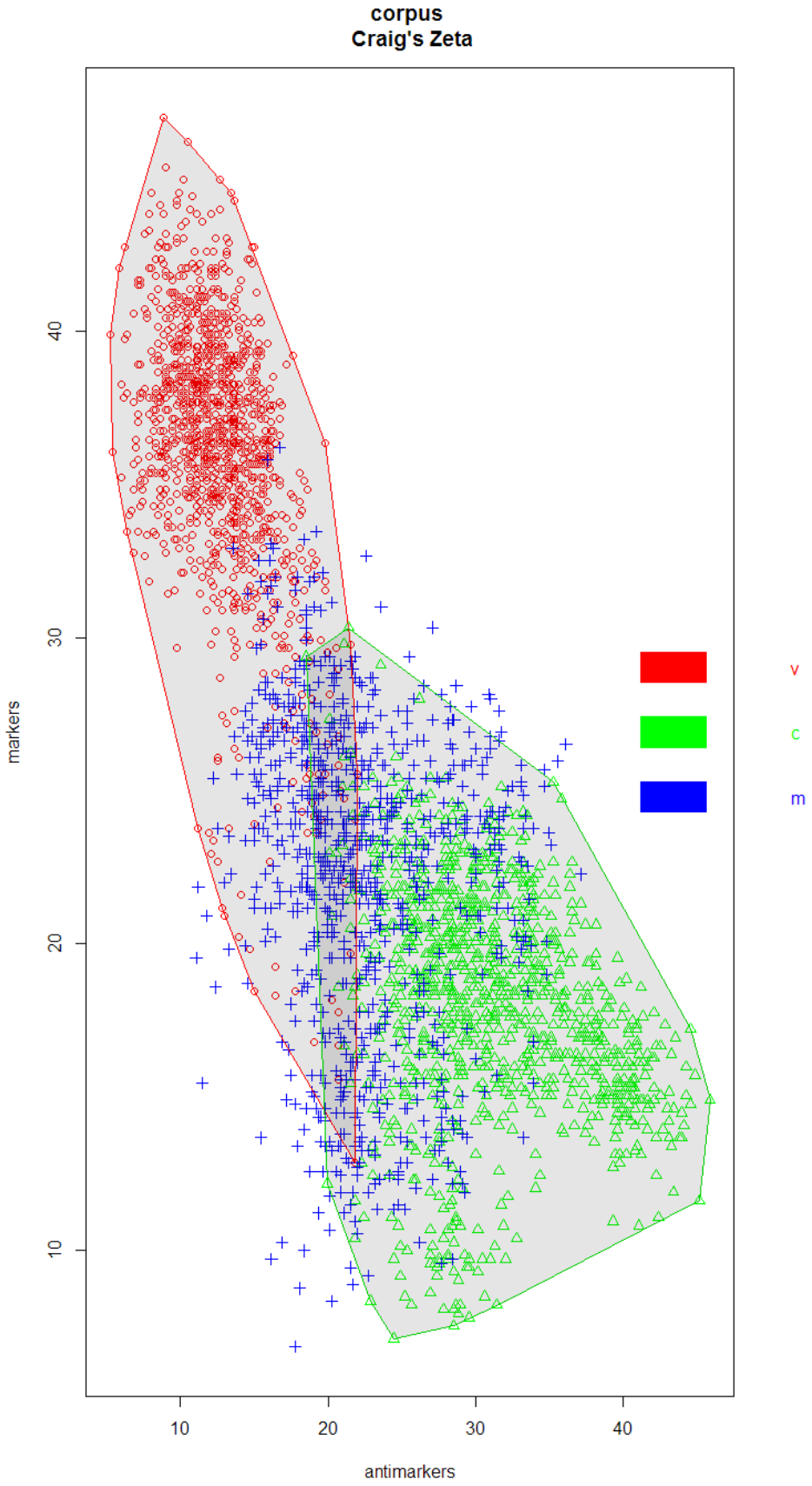

zstylo_delta_results
Cluster Analysis
8
6
4
2
0
100 MFW Culled @ 0%
Classic Delta distance

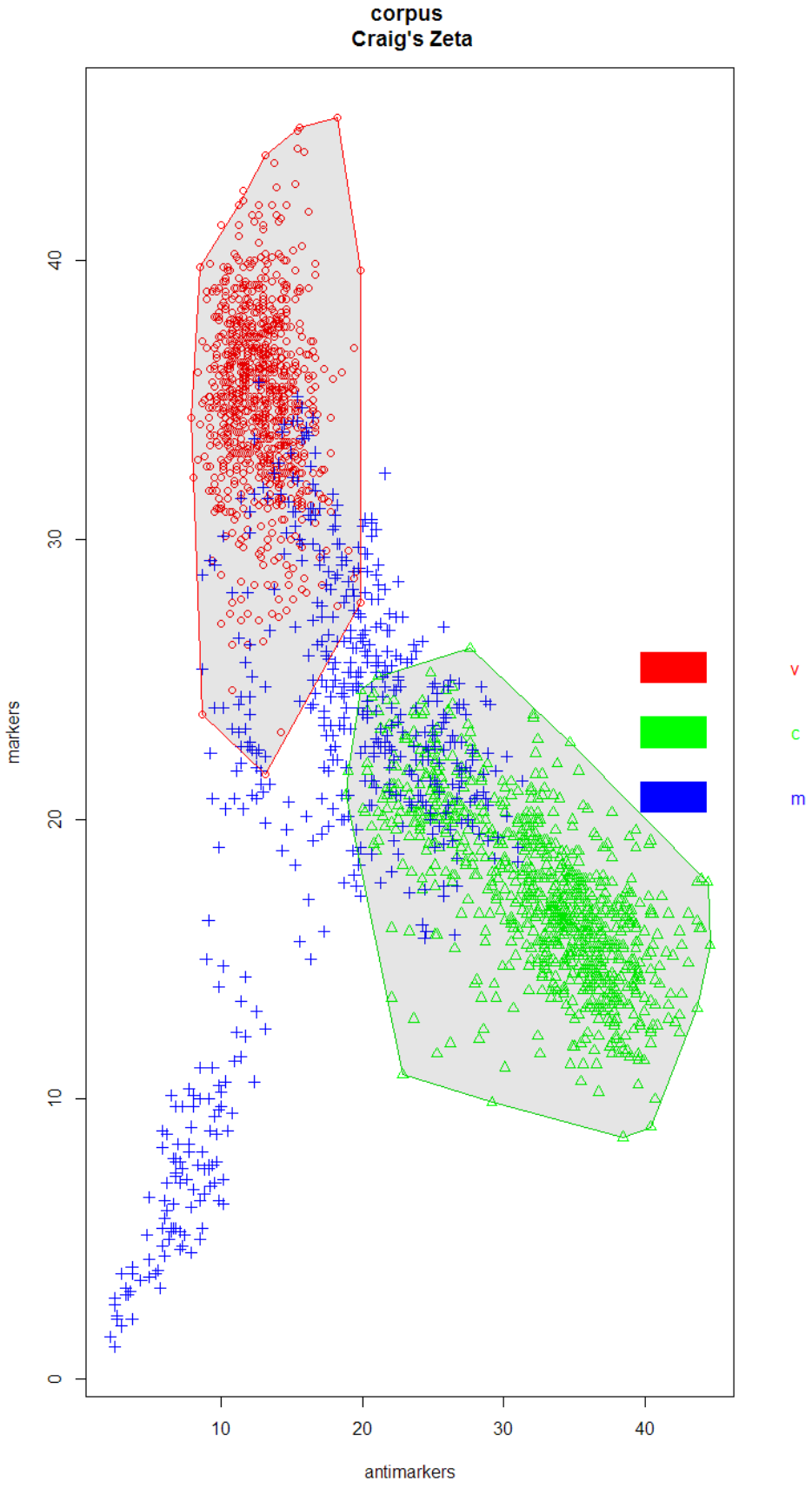

zstylo_delta_results
Cluster Analysis
c_f_atwood_year_of_t
c_f_atwood_blind_ass
c_f_atwood_robber_bric
c_f_atwood_handmaid's
c_f_munro_too_much
c_f_munro_runaway
c_f_munro_love_of_
c_f_munro_collected
c_f_munro_beggar_m
c_f_walker_temple
c_f_walker_open_your
c_f_walker_possessing
c_f_atwood_penelopiad
c_f_walker_meridian
c_f_morrison_paradise
c_f_morrison_amercy
c_f_morrison_jazz
c_f_morrison_beloved
c_f_morrison_song_of
c_f_oates_gravedigger
c_f_oates_black_girl
c_f_oates_black_water
c_f_oates_zombie
c_f_gordimer_get_a_l
c_f_gordimer_beethov
c_f_gordimer_time_lik
c_f_gordimer_pickup
c_f_gordimer_conservatio
m_f_wharton_frome
c_f_barker_ghost_road
m_f_wharton_mirth
m_f_wharton_age
m_f_wharton_burner
m_f_richardson_tunn
m_f_richardson_inter
m_f_richardson_pointed
m_f_woolf_years
m_f_woolf_acts
m_f_mansfield_garde
m_f_mansfield_bliss
m_f_mansfield_german
m_f_west_sunflower
m_f_west_judge
m_f_west_fountain_ove
m_f_west_return
m_f_woolf_lighthouse
m_f_woolf_dalloway
m_f_bowen_last_septe
m_f_bowen_hotel
m_f_bowen_friends
m_f_rhys_wide_sargasso_
m_f_rhys_morning_midnigh
c_f_walker_purple
c_f_mantel_wolf_hall
c_f_mantel_bring_up_
c_f_mantel_safety
c_f_mantel_beyond_blac
c_f_smith_onbeauty
c_f_smith_autographr
c_f_smith_nw
c_f_barker_regenera
c_f_barker_eye_in_t
c_f_barker_border_c
m_f_compton_mern
m_f_compton_family
v_f_gaskell_the_manches
v_f_gaskell_sylvia's_love
v_f_oliphant_miss_ma
v_f_oliphant_madonna
v_f_gaskell_wives_and_
v_f_oliphant_the_rector_
v_f_oliphant_the_doctor
v_f_gaskell_uncle_peter
v_f_gaskell_the_half_brothers
v_f_oliphant_a_beleaguered
m_f_stein_toklas
m_f_stein_lives
v_f_wood_verner's_
v_f_wood_trevlyn_ho
v_f_wood_elsters_foll
v_f_wood_east_lynne
v_f_wood_haliburtor
v_f_wood_a_life's_s
m_f_compton_dolores
c_f_oats_accursed
v_f_eliot_the_lifted_veil
v_f_eliot_essays
v_f_braddon_lady_au
v_f_braddon_henry_d
v_f_braddon_the_docto
v_f_braddon_the_go
v_f_braddon_phantom
v_f_eliot_scenes_of_cl
v_f_eliot_romola
v_f_eliot_silas_marner
v_f_bronte_charlotte_
v_f_bronte_charlotte_
v_f_bronte_charlotte_
v_f_bronte_charlotte_
v_f_bronte_emily_wurthe
v_f_bronte_anne_the_
v_f_bronte_anne_agne
m_f_stein_americans
8
6
4
2
0
100 MFW Culled @ 0%
Classic Delta distance

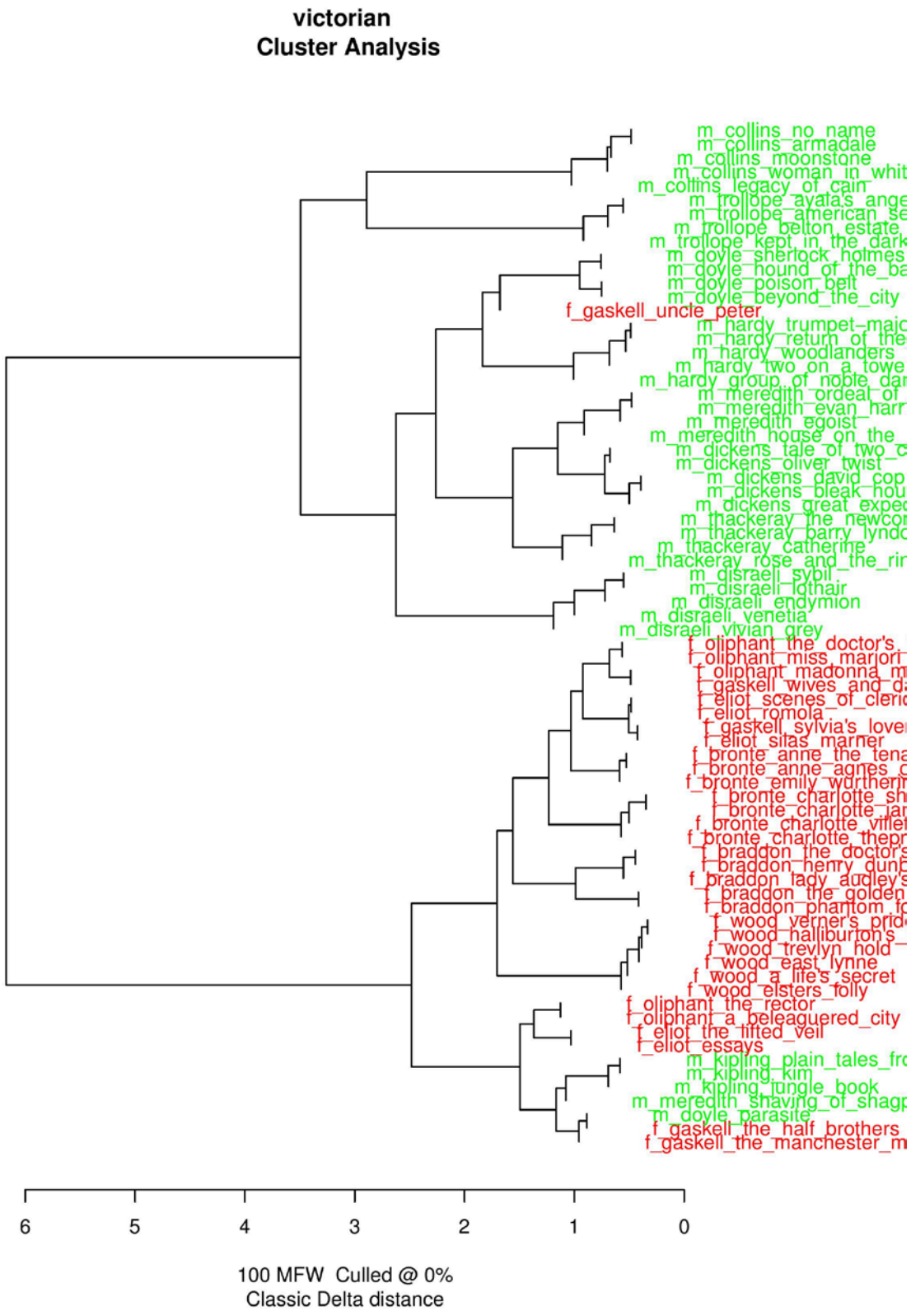

victorian
Craig's Zeta
markers
antimarkers
10
20
30
40
50
60
f
m

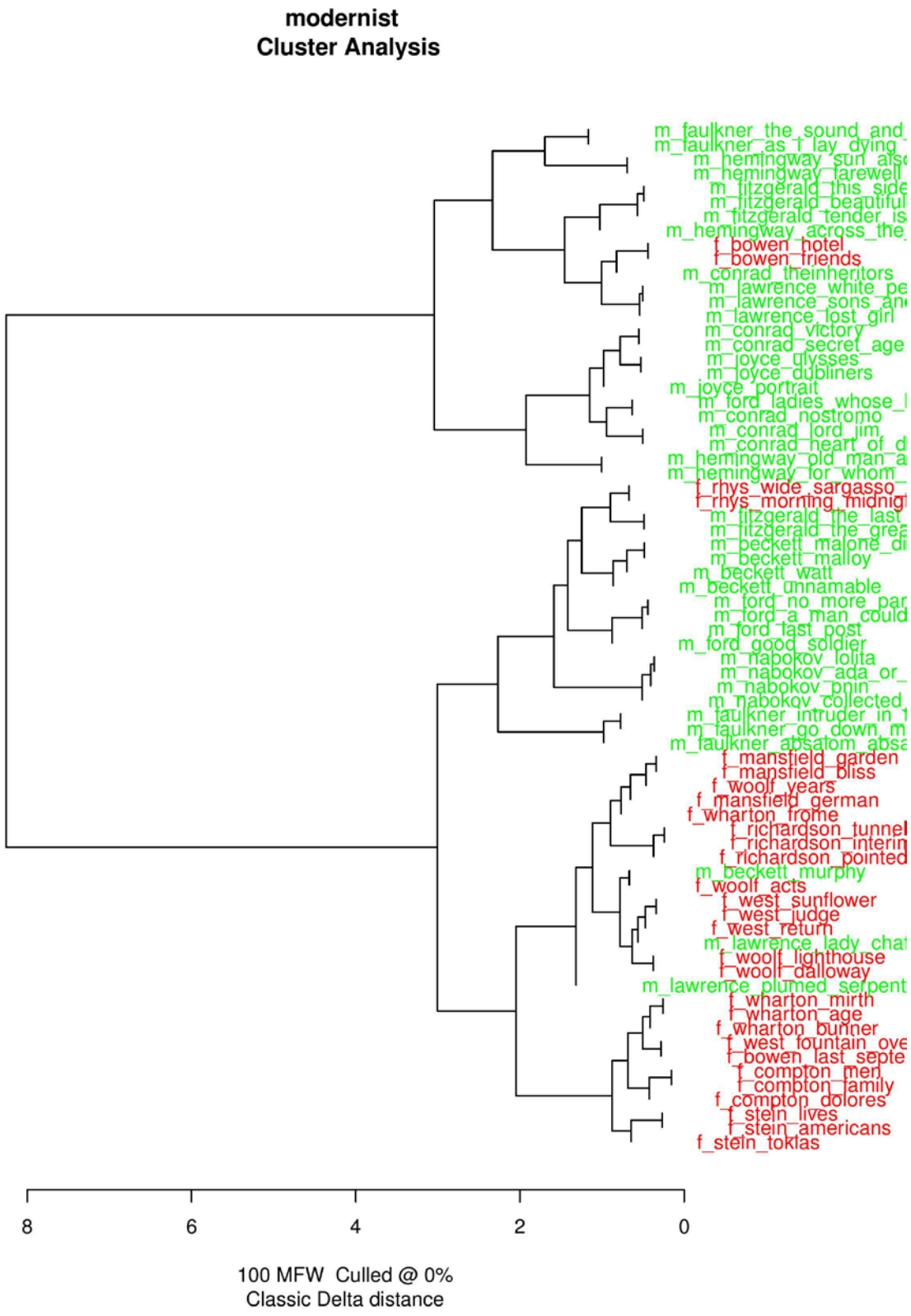

modernist
Craig's Zeta
markers
antimarkers
0
10
20
30
40
50
10
20
30
40
50
60
f
m

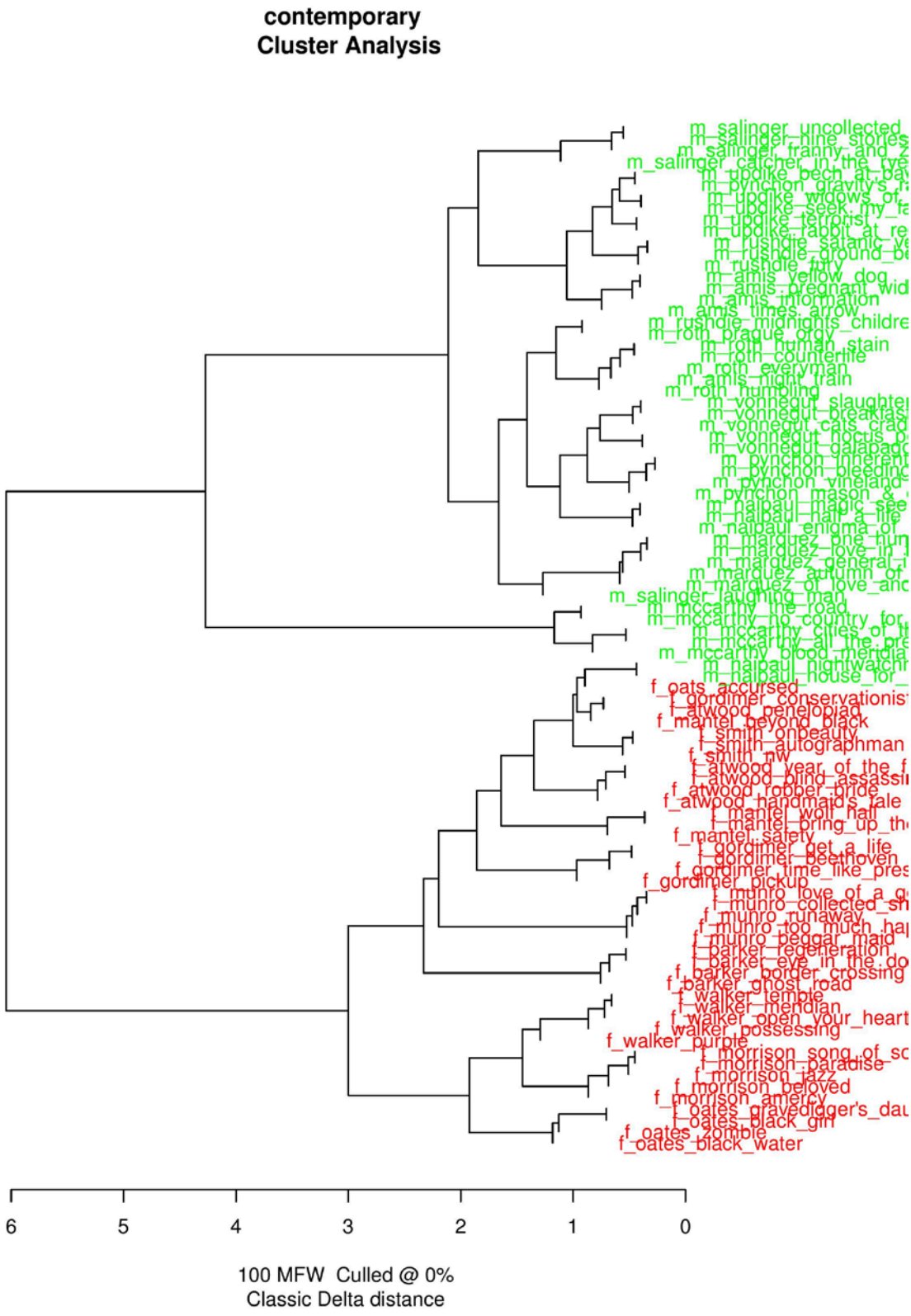

contemporary
Craig's Zeta
markers
antimarkers
10
20
30
40
50
60
20
30
40
50
60
70
f
m


corpus
Craig's Zeta
markers
antimarkers
10
20
30
40
v
c
m


corpus
Craig's Zeta
markers
antimarkers
0
10
20
30
40
10
20
30
40
v
c
m


victorian
Cluster Analysis
m_collins_no_name
m_collins_armadale
m_collins_moonstone
m_collins_woman_in_whit
m_collins_legacy_of_cain
m_trollope_ayala's_ange
m_trollope_american_se
m_trollope_belton_estate
m_trollope_kept_in_the_dark
m_doyle_sherlock_holmes
m_doyle_hound_of_the_ba
m_doyle_poison_belt
m_doyle_beyond_the_city
f_gaskell_uncle_peter
m_hardy_trumpet-majo
m_hardy_return_of_the
m_hardy_woodlanders
m_hardy_two_on_a_towe
m_hardy_group_of_noble_da
m_meredith_ordeal_of_
m_meredith_evan_harr
m_meredith_egoist
m_meredith_house_on_the_
m_dickens_tale_of_two_c
m_dickens_oliver_twist
m_dickens_david_cop
m_dickens_bleak_hou
m_dickens_great_exped
m_thackeray_the_newcom
m_thackeray_barry_lyndo
m_thackeray_catherine
m_thackeray_rose_and_the_rin
m_disraeli_sybil
m_disraeli_lothair
m_disraeli_endymion
m_disraeli_venetia
m_disraeli_vivian_grey
f_oliphant_the_doctor's
f_oliphant_miss_marjori
f_oliphant_madonna_m
f_gaskell_wives_and_da
f_eliot_scenes_of_cleric
f_eliot_romola
f_gaskell_sylvia's_lover
f_eliot_silas_marner
f_bronte_anne_the_tena
f_bronte_anne_agnes_g
f_bronte_emily_wurtherin
f_bronte_charlotte_sh
f_bronte_charlotte_jan
f_bronte_charlotte_villet
f_bronte_charlotte_thepr
f_braddon_the_doctor's
f_braddon_henry_dunb
f_braddon_lady_audley's
f_braddon_the_golden
f_braddon_phantom_fo
f_wood_verner's_prid
f_wood_halliburton's_
f_wood_trevlyn_hold
f_wood_east_lynne
f_wood_a_life's_secret
f_wood_elsters_folly
f_oliphant_the_rector
f_oliphant_a_beleaguered_city
f_eliot_the_lifted_veil
f_eliot_essays
m_kipling_plain_tales_fro
m_kipling_kim
m_kipling_jungle_book
m_meredith_shaving_of_shagp
m_doyle_parasite
f_gaskell_the_half_brothers
f_gaskell_the_manchester_m
6
5
4
3
2
1
0
100 MFW Culled @ 0%
Classic Delta distance


modernist
Cluster Analysis
m_faulkner_the_sound_and_
m_faulkner_as_i_lay_dying_
m_hemingway_sun_also
m_hemingway_farewell_
m_fitzgerald_this_side
m_fitzgerald_beautiful
m_fitzgerald_tender_is
m_hemingway_across_the_
f_bowen_hotel
f_bowen_friends
m_conrad_theinheritors
m_lawrence_white_pe
m_lawrence_sons_an
m_lawrence_lost_girl
m_conrad_victory
m_conrad_secret_age
m_joyce_ulysses
m_joyce_dubliners
m_joyce_portrait
m_ford_ladies_whose_l
m_conrad_nostromo
m_conrad_lord_jim
m_conrad_heart_of_d
m_hemingway_old_man_a
m_hemingway_for_whom_
f_rhys_wide_sargasso_
f_rhys_morning_midnig
m_fitzgerald_the_last_
m_fitzgerald_the_grea
m_beckett_malone_di
m_beckett_malloy
m_beckett_watt
m_beckett_unnamable
m_ford_no_more_par
m_ford_a_man_could
m_ford_last_post
m_ford_good_soldier
m_nabokov_lolita
m_nabokov_ada_or_
m_nabokov_pnin
m_nabokov_collected_
m_faulkner_intruder_in_t
m_faulkner_go_down_m
m_faulkner_absalom_absa
f_mansfield_garden
f_mansfield_bliss
f_woolf_years
f_mansfield_german
f_wharton_frome
f_richardson_tunnel
f_richardson_interim
f_richardson_pointed
m_beckett_murphy
f_woolf_acts
f_west_sunflower
f_west_judge
f_west_return
m_lawrence_lady_chat
f_woolf_lighthouse
f_woolf_dalloway
m_lawrence_plumed_serpent
f_wharton_mirth
f_wharton_age
f_wharton_bunner
f_west_fountain_ove
f_bowen_last_septe
f_compton_men
f_compton_family
f_compton_dolores
f_stein_lives
f_stein_americans
f_stein_toklas
8
6
4
2
0
100 MFW Culled @ 0%
Classic Delta distance


contemporary
Cluster Analysis
m_salinger_uncollected_
m_salinger_nine_stories_
m_salinger_franny_and_z
m_salinger_catcher_in_the_rye
m_updike_bech_at_bay
m_pynchon_gravity's_ra
m_updike_widows_of_
m_updike_seek_my_fa
m_updike_terrorist_
m_updike_rabbit_at_re
m_rushdie_satanic_ve
m_rushdie_ground_be
m_rushdie_fury
m_amis_yellow_dog
m_amis_pregnant_wid
m_amis_information
m_amis_times_arrow
m_rushdie_midnights_childre
m_roth_prague_orgy
m_roth_human_stain
m_roth_counterlife
m_roth_everyman
m_amis_night_train
m_roth_humbling
m_vonnegut_slaughter
m_vonnegut_breakfas
m_vonnegut_cats_crad
m_vonnegut_hocus_po
m_vonnegut_galapago
m_pynchon_inherent
m_pynchon_bleeding
m_pynchon_vineland_
m_pynchon_mason_&_
m_halpaul_magic_see
m_halpaul_half_a_life_
m_halpaul_enigma_of_
m_marquez_one_hun
m_marquez_love_in_
m_marquez_general_i
m_marquez_autumn_of_
m_marquez_of_love_and
m_salinger_laughing_man
m_mccarthy_the_road
m_mccarthy_no_country_for_
m_mccarthy_cities_of_t
m_mccarthy_all_the_pre
m_mccarthy_blood_meridia
m_halpaul_homewrecker
m_halpaul_house_for_
f_oats_accursed_
f_gordimer_conservationist
f_atwood_penelopiad_
f_mantel_beyond_black
f_smith_onbeauty_
f_smith_autographman_
f_smith_nw_
f_atwood_year_of_the_f
f_atwood_blind_assassi
f_atwood_robber_bride_
f_atwood_handmaid's_tale_
f_mantel_wolf_hall_
f_mantel_bring_up_th
f_mantel_safety_
f_gordimer_get_a_life_
f_gordimer_beethoven_
f_gordimer_time_like_pres
f_gordimer_pickup_
f_munro_love_of_a_g
f_munro_collected_sh
f_munro_runaway_
f_munro_too_much_ha
f_munro_beggar_maid_
f_barker_regeneration_
f_barker_eye_in_the_do
f_barker_border_crossing_
f_barker_ghost_road_
f_walker_temple_
f_walker_meridian_
f_walker_open_your_heart_
f_walker_possessing_
f_walker_purple_
f_morrison_song_of_so
f_morrison_paradise_
f_morrison_jazz_
f_morrison_beloved_
f_morrison_amercy_
f_oates_gravedigger's_dau
f_oates_black_girl_
f_oates_zombie_
f_oates_black_water
6
5
4
3
2
1
0
100 MFW Culled @ 0%
Classic Delta distance