**WEIDMAN — Gender**

<2 tables; 4 figures>

**Gender Markers: Distinctive Words in Male and Female Authorship**

Weidman, S. G. and O’Sullivan, J.

David L. Hoover’s frequently cited gendered wordlist presents a list of the 100 most distinctive words in the works of 26 poets, equally split between male and female authors. Hoover remarks (2013) that some aspects of his findings are ‘almost stereotypical’, with ‘female 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 ‘any study of the vocabulary of male and female poets would benefit from larger numbers of poets and larger samples, and many other configurations that address different contrasts are possible (e.g., nationality or historical period)’. This paper, while again accepting that one can never have too large or robust a dataset for this type of macro-analysis, attempts to build on the foundation set down by Hoover, analyzing gender markers across a selection of male and female authors.

As with any textual analysis assisted by computational methods, scholars are restricted by their access to a reliable corpus of digitized materials. Hoover drew his dataset from Literature Online, choosing poets who had a substantial sample available (see Table 1).

|  |  |
| --- | --- |
| **Male** | **Female** |
| Ammons, A. R. | Derricotte, Toi |
| Baraka, Imamu Amiri | Gilbert, Sandra M. |
| Brathwaite, Kamau | Glück, Louise |
| Carruth, Hayden | Kumin, Maxine |
| Gunn, Thom | Miles, Josephine |
| Kees, Weldon | Olds, Sharon |
| Kinnell, Galway | Pastan, Linda |
| Reed, Ishmael | Plath, Sylvia |
| Schwartz, Delmore | Rich, Adrienne Cecile |
| Shapiro, Karl Jay | Rukeyser, Muriel |
| Snyder, Gary | Sexton, Anne |
| Stafford, William | Stevenson, Anne |
| Wright, James Arlington | Swenson, May |

Table 1. Authors contributing to David Hoover’s list of gender markers.

In an effort to produce findings that would complement Hoover’s initial study, we took heed of his suggestions to use a larger sample that accounts for historical period. Thus, we identified three distinct literary periods of interest and gathered a selection of typical works from across the canonical authors from each of these eras (see Table 2).

|  |  |
| --- | --- |
| **Male** | **Female** |
| **Victorian authors** | |
| Collins, Wilkie | Braddon, Mary Elizabeth |
| Conan Doyle, Arthur | Bronte, Anne |
| Dickens, Charles | Bronte, Charlotte |
| Disraeli, Benjamin | Bronte, Emily |
| Hardy, Thomas | Eliot, George |
| Kipling, Rudyard | Gaskell, Elizabeth |
| Meredith, George | Kennard, Mary Eliza |
| Thackeray, William | Oliphant, Margaret |
| Trollope, Anthony | Wood, Ellen |
| **Modernist authors** | |
| Beckett, Samuel | Bowen, Elizabeth |
| Conrad, Joseph | Compton-Burnett, Ivy |
| Faulkner, William | Mansfield, Katherine |
| Fitzgerald, Scott F. | Rhys, Jean |
| Hemingway, Ernest | Richardson, Dorothy |
| Joyce, James | Stein, Gertrude |
| Kafka, Franz | West, Rebecca |
| Lawrence, D. H. | Wharton, Edith |
| Nabokov, Vladimir | Woolf, Virginia |
| **Contemporary authors** | |
| Amis, Martin | Atwood, Margaret |
| McCarthy, Cormac | Barker, Pat |
| Naipaul, V. S. | Mantel, Hilary |
| Pynchon, Thomas | Morrison, Toni |
| Roth, Philip | Munro, Alice |
| Rushdie, Salman | Oates, Joyce Carol |
| Salinger, J. D. | Smith, Zadie |
| Updike, John | Walker, Alice |
| Vonnegut, Kurt | Marilynne Robinson |

Table 2. Authors used in our analysis.

We accept that there are issues with choosing canonical authors, but it seems the only feasible way to structure a study of this nature. We also elected to use novels and short stories, as opposed to poetry. We gathered corpora for 54 authors (see Table 2), for a combined total of 212 novels and collections of short stories. Attempts were made to use an equal amount of text from each author to ensure that larger oeuvres did not skew our results.

# Methodology and Results

Craig’s Zeta (Hoover, 2008; Burrows, 2004) was the primary methodology deployed, as was the case in Hoover’s original study. Our Zeta analysis was conducted using Stylo (Eder et al., 2013), with a text slice length of 2,000, text slice overlap of 1,000, and an occurrence and filter threshold of 2 and 0.1, respectively. Initially, we produced standard most frequent wordlists using Delta, with a list of stopwords applied. Comparing these results, we saw little separation between the genders. Zeta, as a measure designed to detect distinctive words, yielded more fruitful data. When comparing authors from across all periods, our results confirm Hoover’s in that there are some stereotypical words that do emerge. On the female side, most notable are ‘household’, ‘smile’, and ‘mother’, while on the male side, ‘country’ and ‘famous’ stand out. ‘America’ is also present in the distinctive male words. The male words are arguably reflective of a dataset that contains slightly more male American authors than it does female American authors—suggesting that Hoover is also correct to point towards nationality as a further potential classifier—but the difference is not so great that it should yield significant skewing. However, while there is a clear presence of stereotypical markers, there is also some significant crossover between authors. This is less pronounced across different periods, but when author sets are considered as a whole (see Figure 1), the macro-separation is not as pronounced as in Hoover’s findings.

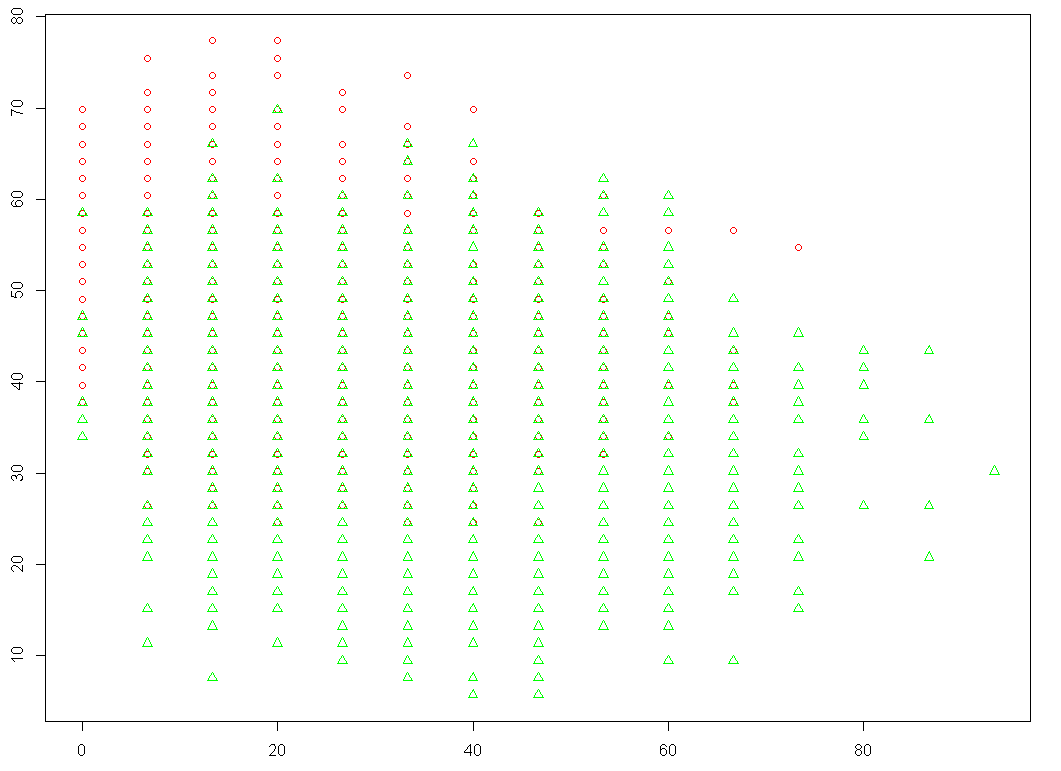


Figure 1. All authors.

As suggested by Hoover, we also analyze authors chronologically, comparing male and female authors across their respective periods. Once again, there are substantial correlations with Hoover’s wordlist, with stereotypes appearing in all sets, across all periods. But, as noted, the crossover between genders fluctuates throughout periods, with contemporary authors sharing the greatest similarities (see Figures 2, 3, and 4). This would suggest that gender distinctions between authors fluctuate across literary epochs.

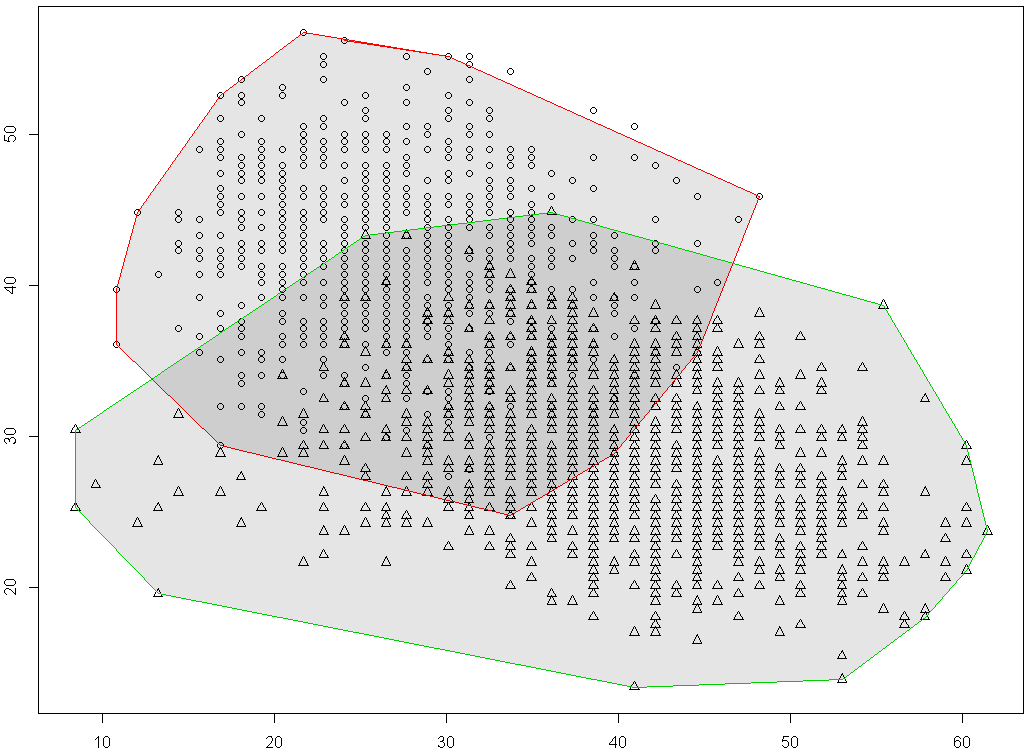


Figure 2. Victorian authors.

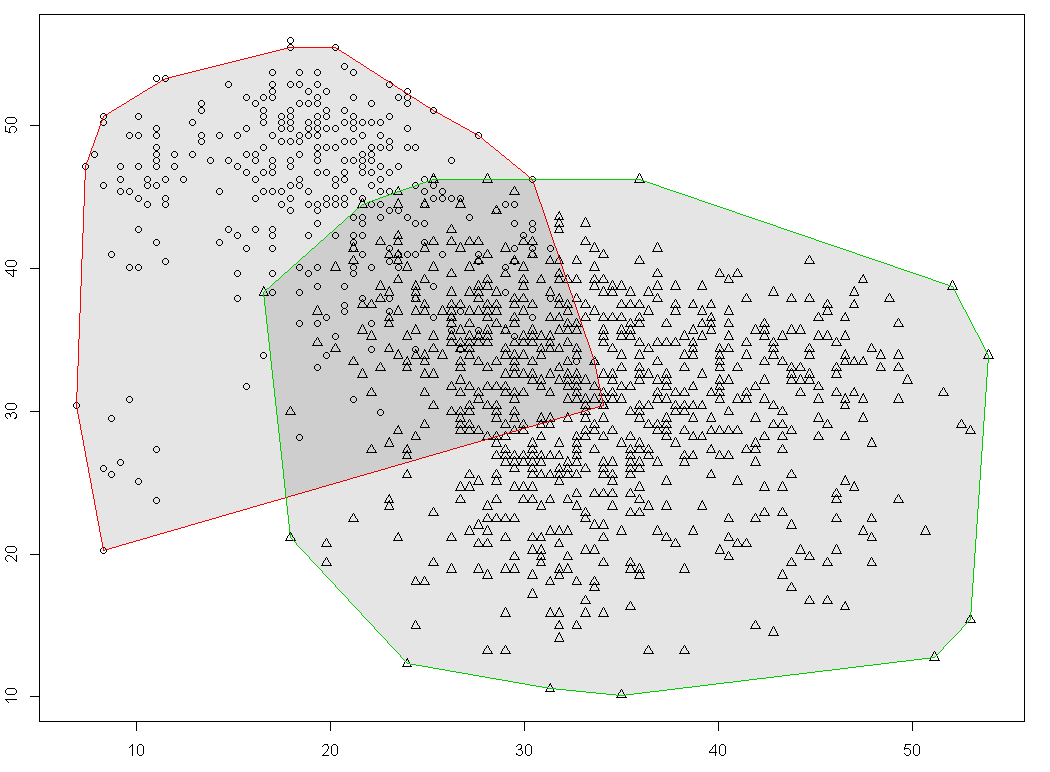


Figure 3. Modernist authors.

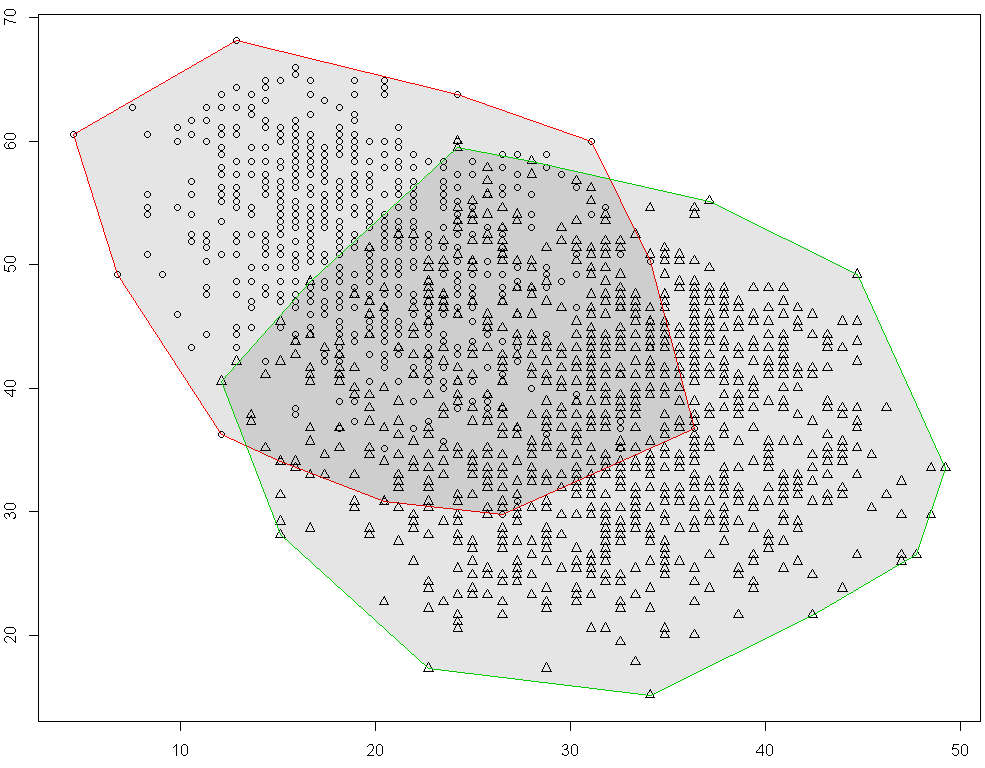


Figure 4. Contemporary authors.

Our methodological implications do require brief commentary. 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 literature) insofar as they relate to another dataset (e.g., male literature). In Delta, the data speaks for itself, whereas in Zeta we use a mode of classification, so the separation we receive is partly a result of our dataset preparation. Rather than the data establishing its own distinctions, we (in a way) privilege and impose the gender split in our separation of the male and female datasets. There are literary and mathematical justifications for this, but it is worth acknowledging, and certainly worth discussing in the context of a gender-specific study. However, taking into account the data from our Delta analysis, the results of which turned up inconclusive, the subsequent Zeta can be seen to merely refine an already existent pattern. That being said, this particularity is something we will focus on further as we fine-tune our analysis for presentation. In doing so, we will computationally compare results across all of the lists, including Hoover’s, to see if any correlations may have been missed in this initial stage, which will ultimately allow us to conduct a closer analysis of the distinctive gender markers.

# Conclusions

Worth noting is that the ongoing dispute in literary studies concerned with gender and writing style is wide and varied, and the scope of our research does not immediately affect the debates of gender theory or the potentiality of a distinct form of *écriture feminine*. In the parameters of such literary discourses, however, our project does provide a quantitative means to approach an overwhelmingly qualitative discussion. Specifically, our preliminary analyses lend evidence to the claims that such gender differences are evident in writing across periods, but that these differences are less distinguishable depending on the period in question.

Studies that have included multiple genres or mediums in their stylistic analyses and text classifications have concluded that stylistic features often depend more heavily upon genre than gender (Herring and Paolilo, 2006; Janssen and Murachver, 2004; Argamon et al., 2003). Yet what these studies have consistently shown is that gender and genre both seem to reveal stylistic and thematic qualities historically associated with male and female language application. Our goal in limiting the mode of our corpus to fiction novels (and novel-length collections of short stories) is to refine the results of prior studies with more generalized scopes (e.g., Koppel et al., 2002; Argamon et al., 2003) and place similar conclusions within a periodized context, ideally providing ample opportunity for literary application.

While we follow in the methodological footsteps of such 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. A number of projects (Argamon et al., 2003; Burrows, 2004; Schler et al., 2005; Pennebaker, 2011) have concluded that the use of function words—pronouns (taken customarily as female markers) and determiners and prepositions (taken customarily as male markers)—provide a reliable basis for gender identification in writing. None of them, markedly, has sufficiently addressed the issue of modernity and the evolution of language and its changes over time, though Hoover has suggested that this is the way forward. Our research separates from such work in that we aim to do just that: distinguish how gender differences have evolved over and between selections of specified, canonical literary periods, focusing on distinct, rather than functional, word choices. Admittedly, this is a first step, and akin to Hoover in his initial study, we acknowledge the need for a larger corpus and more refined dataset—this, of course, is the constant nature of the computational beast. But we can at least view this is a next step.

# References

**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.

**Burrows, J.** (2004). Textual Analysis. In Schreibman, S., Siemens, R. and Unsworth, J. (eds), *A Companion to Digital Humanities*. Oxford: Blackwell.

**Burrows, J.** (2007). All the Way Through: Testing for Authorship in Different Frequency Strata. *Literary and Linguistic Computing,* **22**(1): 27–47.

**Eder, M., Kestemont, M. and Rybicki, J.** (2013). Sylometry with R: A Suite of Tools. In *Digital Humanities 2013: Conference Abstracts*, University of Nebraska–Lincoln, pp. 487–89.

**Herring, S. C. and Paolilo, J. C.** (2006). Gender and Genre Variation in Weblogs. *Journal of Sociolinguistics,* **10**(4): 439–59.

**Hoover, D. L.** (2008). Quantitative Analysis and Literary Studies.’ In Schreibman, S. and Siemens, R. (eds), *A Companion to Digital Literary Studies*. Oxford: Blackwell Publishing, pp. 517–33.

**Hoover, D. L.** (2013). Textual Analysis. In Price, K. M. and Siemens, R. (eds), *Literary Studies in the Digital Age*. Modern Language Association of America, http://dlsanthology.commons.mla.org/textual-analysis/.

**Janssen, A. and Murachver, T.** (2004). The Relationship between Gender and Topic in Gender-Preferential Language Use. *Written Communication,* **21**: 344–67.

**Koppel, M., Argamon, S. and Shimoni, A. R.** (2002). Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing,* **17**(4): 401–12.

**Pennebaker, J. W.** (2011). *The Secret Life of Pronouns: What Our Words Say about Us.* Bloomsbury, London.

**Schler, J., Koppel, M., Argamon, S. and Pennebaker, J.** (2005). Effects of Age and Gender in Blogging. Symposium paper for *American Association for Artificial Intelligence*.