# ALGEE-HEWITT ­— Discourse

# <2 figures>

# Discourse, Design, and Disorder: Digital Models for Aesthetic Theory

Algee-Hewitt, M.

As quantitative textual analyses are based primarily on words, which are lexical objects that, by definition, carry meaning and information, these methods should offer a clear path that leads from analysis to interpretation. If two groups of texts are well differentiated by author (Burrows, 2003), genre (Jockers, 2013), or period (Hughes et al., 2012) based on word correlations or frequencies, it is reasonable to assume that those words that differentiated them should offer us a straightforward approach to understanding why *those* words in particular separated *these* texts. Yet this is not the case in most digital humanities findings. In authorship attribution, for example, high-frequency words are valued because they are “content-free” and thus are used ‘unconsciously’ by an author (Burrows, 2003).1 Even when content-laden words are used as the features, it is often the case, as in the Stanford Literary Lab’s first pamphlet, ‘Quantitative Formalism’, that some complex mixture of proportions of many words does the work at separating something like genre (Allison et al., 2011). All too often, reading these kinds of lists of words gets us no closer to understanding *why* two corpora are separated into groups. Even topic modeling, which promises results based on interpretable ‘topics’, only adds to this complex problem because the clusters of words that underlie the model are often only subjectively ‘named’ by the analyst.2

At the root of this problem is the way in which much digital humanities work treats words as feature sets, rather than objects with interpretable weight. This has done much to uncover hidden structural (or stylistic) features of textual corpora, but has required an interpretive leap, from data to metadata, to explain these structures. The actual features that articulate the differences between texts are not easily integrated into the interpretive process. Instead, they establish the *fact* of a differentiation whose *meaning* is often explained by extratextual or a priori knowledge. In this paper, I argue that we can begin to mitigate this interpretative gap by deriving more complex, higher-order, quantitative metrics through which we can compare textual corpora. The two metrics I develop in this paper, heterogeneity and redundancy, are both based in frequency data, but *meaningfully* translate word frequencies into metrics whose relationship to the features is quantitative but whose relationship to the text as a whole is qualitative—that is, they are metrics that can be derived quantitatively, but that offer information that is immediately relatable to the texts and that can be used to interpret the results in a way that is not only quantitatively rigorous but also grounded in a theoretical method for understanding textual or informational difference.

**Heterogeneity**

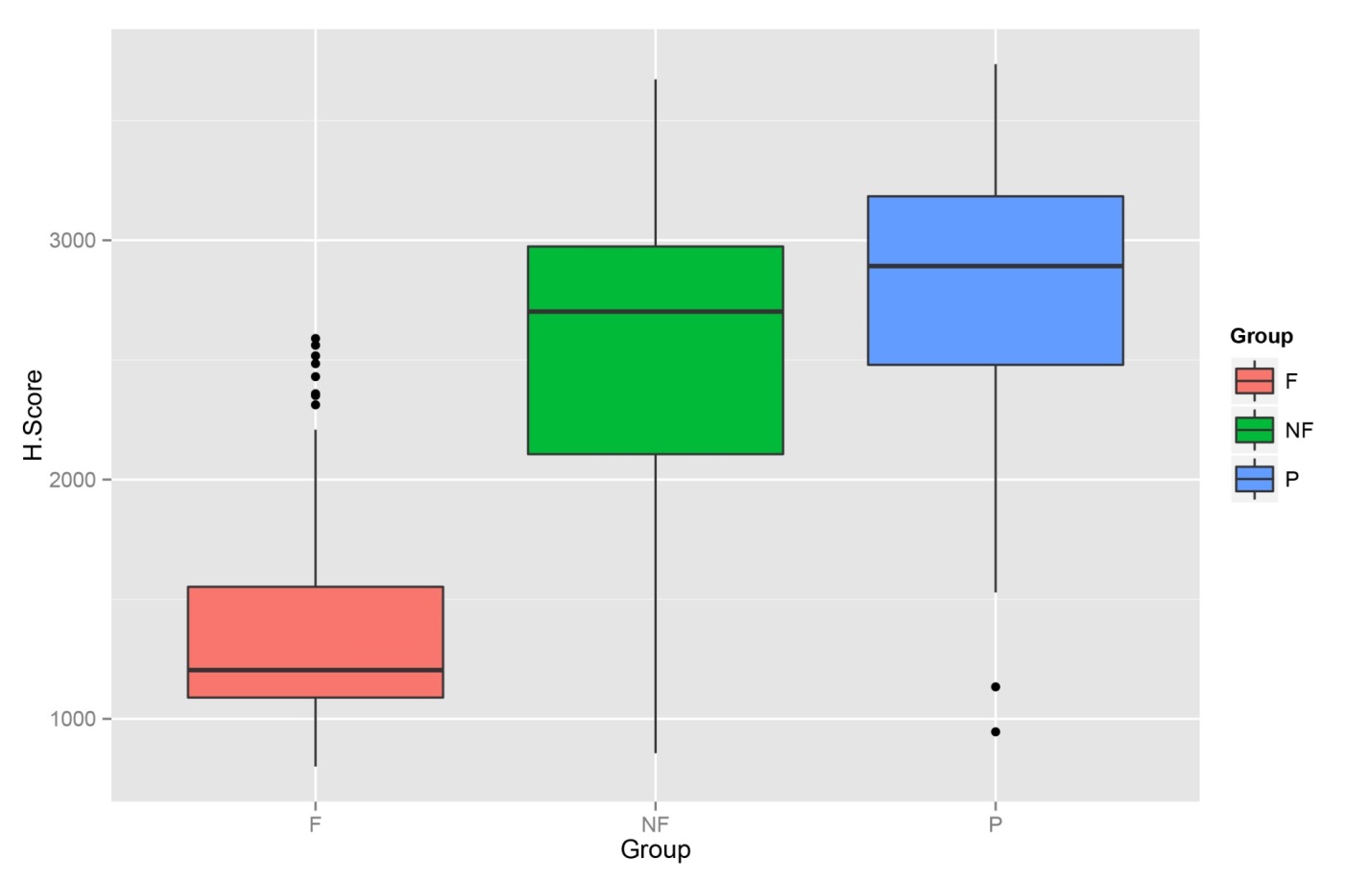
The first metric that I develop in this project draws a connection between the quantitative analysis of texts and literary theory by way of Mikhail Bakhtin’s theory of heteroglossia. In ‘Discourse in the Novel’, Bakhtin proposes that we understand the novel as a genre of accumulation and mixture resulting in a heterogeneous object of multiple and overlapping language (Bakhtin, 1981). While poetry is dominated by a single, unitary voice, prose writing, particularly novelistic prose, is defined by a ‘diversity of voices’, according to Bakhtin: the more ‘novelistic’ the text, therefore, the more competing voices there should be and the less like the rest of the text any one part becomes. While it is useful to parse the distinction between prose and poetry, we can also operationalize the theory of heteroglossia to understand differences between and within different type of writing by measuring the ‘distance’ between the different parts of a text, thereby calculating how internally diverse it is. For this measurement, I divide the text into equally sized chunks and calculate the distance (using both a Euclidean distance metric and a Kullback-Leibler divergence) between every part and every other part. By combining these internal distances into a single score and scaling them logarithmically to account for text length, I am able to derive a single number that represents the amount of linguistic diversity between different parts of the text.3 Texts with a high degree of diversity, or heterogeneity, between their parts would be made up of more internal ‘voices’ following Bahktin. As I show in this paper, this metric of heterogeneity not only separates different kinds of writing, but also writing of different periods, or by different authors or even by different genres. And because the metric is a direct measure of internal diversity, the ways that groups of texts are separated by this metric offers critical information about those groups of texts: texts with a low heterogeneity score are mono-vocal: their patterns of language do not vary across the text. Texts with a high heterogeneity score are highly diverse among their constituent parts and depend upon variety, rather than repetition, to engage their readers.

**Redundancy**

The second metric I describe in this paper is based on Claude Shannon’s work in information theory. While metrics of entropy (or information loss) are common in measuring textual differentiation (the Kullback-Leibler divergence is one such metric), the measure of entropy in these cases is made *between* texts and offers only a way to characterize how different two texts are based on their shared word frequencies. The two interventions that I make in this paper are (a) to measure redundancy, the opposite of the informational entropy described by Shannon, and (b) to make this measurement *within* rather than *between* texts. By slightly altering Shannon’s formula for redundancy (1963) I offer a way to calculate the redundancy of a text that is scaled to the possible combinations of words within the text, rather than the possible combinations of words across the language (as it is traditionally calculated). This again allows me to calculate a single number based on word frequencies (in this case, bigram probabilities) that speaks directly to the information content of the text itself. In other words, the redundancy metric that I describe in this paper not only offers a way to classify different kinds of texts based on their score, but in doing so, it also describes how likely each text is to use repeated word combinations or if a text can be characterized by its tendency to combine words in new ways. This measurement therefore, reveals important information about how the text communicates with the reader that allows me to both describe differences between texts and, more importantly, characterize these differences based on features that are legible at the level of the text.

By combining these two metrics, heterogeneity and redundancy, both of which offer data on how repetition and variety functions within groups of texts, this paper demonstrates how these higher-order features can be used reliably to describe surprising differences between corpora that are uninterpretable through standard digital humanities analyses. In this paper I apply this new combination of metrics to two case studies in long 18th-century literature: genre and canonicity. As I demonstrate how these metrics together can differentiate between different kinds of writing, for example—poetry, prose, and nonfiction (Figure 1) as well as between canonical and noncanonical writing (Figure 2), I simultaneously articulate a theory of how the information content of the texts and their internal repetitions can help explain these groupings. This work also reveals a clear historical pattern among these different groups that is not apparent in any other quantitative analysis of the corpora that I use. These two ways of understanding textual difference therefore not only provide new information on literary history and criticism, but they also offer a way to bridge the gap between the quantitative calculation of clusters and the meaning behind these clusters. Despite their apparent simplicity (both redundancy and heterogeneity use a single number to describe a text), I argue that their actual complexity and their derivation from literary theory and information theory, respectively, allow us to understand the texts that they describe with a much greater nuance and in more critical detail than what we would be able to do from reading the texts, let alone deriving the clusters of the texts directly from word frequencies. What I argue that these two measurements provide is a quantitative aesthetic theory: one that articulates both a quantitative and theoretical way of understanding how texts are grouped and, more importantly, the critical significance of and meaning behind these historical, generic, and textual formations.

**Figures**



p = <2.2e--16

Figure 1. Difference between eighteenth-century genres derived from heterogeneity scores.

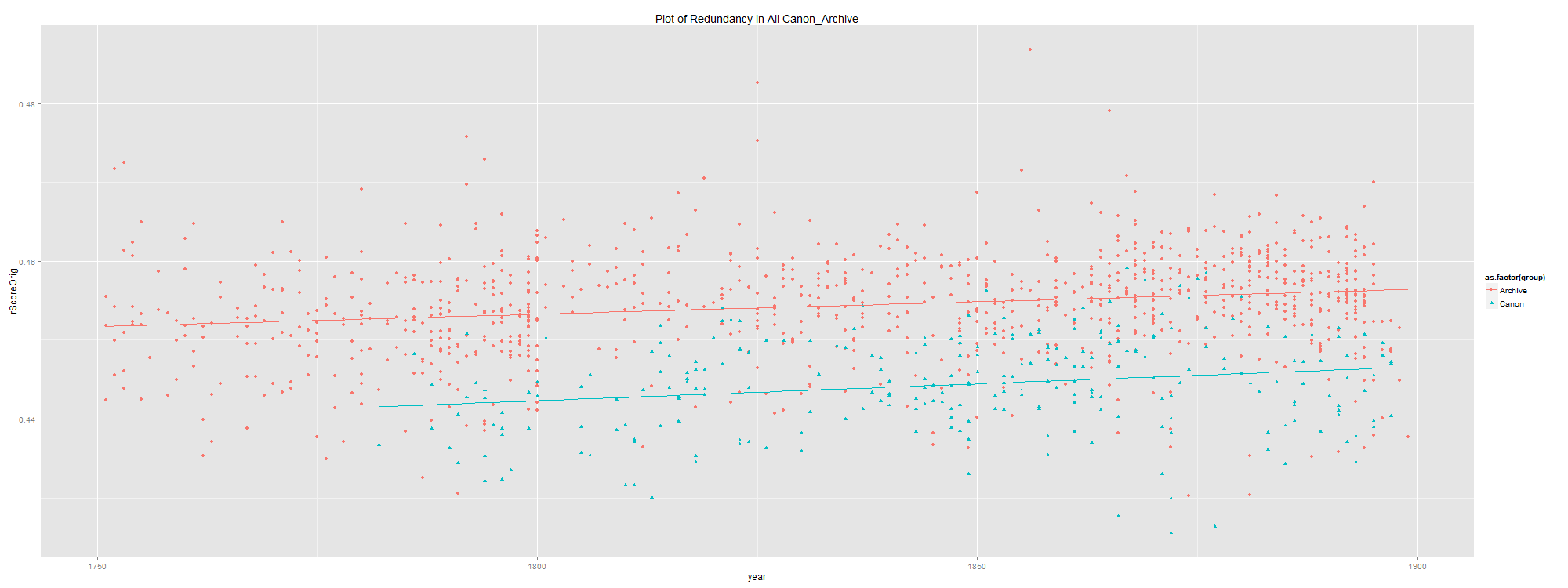


Figure 2. Change in redundancy over time in canonical (blue) and noncanonical (red) texts.

**Notes**

1. This method, particularly the dependence upon ‘content-free’ words, was critiqued by Hoover (2009) in ways that echo the concerns that I raise here.

2. There are notable exceptions to this, particularly studies that begin with a predefined list of meaningful terms and use their frequencies only to differentiate between texts. This is most evident in spatial humanities work that uses predefined geographic names to differentiate between texts (Wilkens, 2011).

3. This measure is related to the Type-Token ratio, a standard metric of lexical diversity within a document, but where the Type-Token ratio measures lexical diversity across the entire document, the metric of heterogeneity that I propose here compares diversity *between* rather than *within* segments of the text, measuring, for example, the degree to which the ending of a text is like the beginning, or the middle.

**References**

**Allison, S., Heuser R., Jockers M., Moretti F. and Witmore W.** (2011). Quantitative Formalism: An Experiment. *Literary Lab Pamphlet*,**1**, 15 January.

**Bakhtin, M.** (1981). Discourse in the Novel. In Holquist, M. (ed.), Emerson, C. and Holquist, M. (trans.), *The Dialogic Imagination*. Austin: University of Texas Press.

**Burrows, J.** (2003). Questions of Authorship: Attribution and Beyond. *Computers and the Humanities*, **37**(1): 5–32.

**Hoover, D.** 2009. “Word Frequency, Statistical Stylistics and Authorship Attribution.” In **Archer, D.** (ed.), *What’s in a Word-list? Investigating Word Frequency and Keyword Extraction.* London: Ashgate, pp. 35–52.

**Hughes, J., Foti, N., Krakauer D. and Rockmore D.** (2012). Quantitative Patterns of Stylistic Influence in the Evolution of Literature. *PNAS*, **109**: 7682–86.

**Jockers, M.** (2013). *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana.

**Shannon, C. and Weaver, W.** (1963). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

**Wilkens, M.** (2011). Canons, Close Reading, and the Evolution of Method. In Gold, M. K. (ed.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.