A Mathematical Theory of Authorial Intention

Michael Gavin

University of South Carolina

Summer 2019

1 Preface

Computational literary theory serves two interlocking aims: to develop key concepts for quantitative research in literary studies and to debate how these concepts relate to already existing qualitative ideas. In this essay I pursue these goals through focus on the long-standing theoretical problem of authorial intention. What is an intention, and how can intentions be inferred through readings of texts? What is the relation between an intended meaning and some more general significance? How can we identify and measure the effect of human volition on the distribution of meaning? The theory presented below adapts insights from corpus linguistics and information theory to provide a general framework for answering these questions.

I'll introduce two primary metrics. Given a document and a corpus that contains that document, *conceptual work* measures the force with which a text pushes against normal word-use patterns. In randomly generated text, all words are used in ways that regress toward common norms, but intentionally written documents almost always deviate from those norms: intentions use words in weird ways, pushing them out of the flow of statistical probability. *Semantic transparency* describes the entropy of conceptual work in a document. In some texts, semantic deviation is intensely focused among highly significant terms that stray far from general norms; in others, the word-use patterns are more evenly and more normally distributed. Combining these metrics, the mathematical theory of intention presented here involves a dialectic between deviation and conformity, between the individual case and the lexical setting from which texts draw and against which they intervene.

After laying out these basic definitions, I'll show how to produce summarizing readings of individual documents, how to generalize over an author's career, and how to trace intentions expressed through nonhuman aggregating agents like times and places. In the last third of the essay, I'll situate this theory in the narrower frame of twentieth-century literary criticism, and I'll argue that distributional semantics disrupt notions of intentionless meaning by enabling a kind of mindless reading performed by computers. I'll conclude that semantic models contribute to information theory in ways that couldn't have been foreseen in mid-century and that illuminate complementary concerns held by information theorists like Claude Shannon and New Critics like Cleanth Brooks.

As this overview suggests, the argument pursued here will likely seem alien to most literary scholars, even to those sympathetic to the digital humanities. Usually computational methods are offered to readers on more practical grounds through applications that promise access to larger scale analyses. By contrast, this essay will focus on the measurements themselves, exploring their mathematical foundations and inviting readers to join me in experimenting with new ways of thinking about textuality and historicity. My core claims rest in the metrics and in the theoretical perspective those metrics imply.

The form of this essay is key to its argument. I follow a *theory-application-discussion format* that foregrounds the main ideas in general, then shows how those ideas can be applied as methods, and only later situates the theory against prior commentary. The merit of this format is that it highlights the actual argument instead of burying it under needless exposition. The downside is a risk of genre confusion. Although I do offer some examples, this isn't really an exercise in distant reading. Readers accustomed to the empirical rigor of linguistic computing will be bothered by the lack of a proper data and methods section, and they'll perhaps be frustrated that my conclusions don't depend on any particular findings derived from any particular corpus. A larger group of readers will prefer that I begin with a discussion of well-known debates in literary theory, and they'll expect me to identify a problem in the intellectual history of intention that quantitative methods might help resolve. This expectation, too, will be frustrated. My argument does not hinge on any particular reading of, say, the New Critics or Michel Foucault. Instead, it comes from years spent learning quantitative methods and slowly coming to grips with how those methods have transformed my thinking. In the last third of this essay I'll provide a brief overview of those

twentieth-century debates, but my perspective is motivated by the mathematical theory, not the other way around. Many other readers will be hoping for polemical reflections about the role of quantification in the humanities, but they'll find little of immediate interest until, perhaps, the epilogue, which offers a personal reflection about how it feels to do this kind of work under their vigilant gaze.

Although I will perversely insist on referring to my subject as "authorial intention," readers will find that my use of the word "authorial" and my use of the word "intention" will, in the end, correspond only loosely to whatever ideas they associate with those terms. All readers will find themselves, at one point or another, responding with some form of a critique that begins, "Well, that's all fine, but to really get at intention you'd have to ..." Because different scholars bring so many different and often contradictory theoretical commitments to bear on the problem of intention, the precise objection will vary from reader to reader, but nonetheless it must be addressed at the outset. While most scholarship in literary computing presupposes some known object of inquiry (the nineteenth-century novel, the plays of Shakespeare, etc.) this essay is very different. It asks how our conception of a critical topic might change when that topic is considered as a subject of mathematical study. Accordingly, my argument will unfold through successive iterations and will seem to change as it goes. In the beginning, the concept of intention will seem more or less familiar, if somewhat underdetermined, as I describe my efforts to approximate it mathematically. As I proceed through my case studies, however, authorship and intention will be pushed very far, perhaps beyond the threshold of recognizability. In my analysis of John Locke's *Two Treatises of Government*, we'll see that intentions identified in a document will vary drastically based on the corpus chosen for statistical comparison. In my discussion of Aphra Behn's works, the concept of the author will be radically denaturalized and repurposed as an organizing principle for data aggregation. The final case study will explore one consequence of that denaturalization, showing how historical time is similarly deformed. Across these case studies, the calculations, too, will change, and it will become clear that the mathematical theory does not name any single sequence of computational procedures but rather a set of general principles under which such procedures might be devised. When my discussion turns to the intellectual history of intention, I'll introduce a qualitative theoretical framework marked by interpretive flexibility and ontological pluralism. I'll define as *authorial* any metatextual attribute that has been assigned responsibility for a text's

meaning, in whole or in part, and I'll define an *intention* as any meaning so attributed. My hope is that such open-ended terms will prove capacious enough to support not only the quantitative theory but also the competing metaphysical schemes of twentieth-century literary critics.

Throughout, my goal will be to share with readers the exhilarating and uncanny experience of allowing a fundamental term of literary theory to be transformed through computation. What happens to the author when texts are paraphrased as data? What happens to meaning when measured? These questions have nagged me from the beginning of my work as a digital humanist. I offer my answers below.

2 Theory: deviance, work, transparency

2.1 Deviance.

Evaluating documents for intention requires identifying the local meanings of statements and characterizing how those statements either conform to or deviate from common patterns. In the fields of information retrieval and language processing, scholars have worked long and hard on this problem.[1] If you say, "The rain in Spain falls mainly on the plain," speech-recognition software is unlikely to mistake you for saying, "The reign in Spain falls mane Lee on the plane." If you type "nirvana" into a search engine, you'll find web pages about a late-twentieth-century rock band and others about the ultimate state of being. If you type "nirvana nevermind," the search results will be narrowed. These software applications work as well as they do because they're informed by semantic models that automatically disambiguate word senses to separate the intended meaning from a field of possible meanings. Without ever engaging the question of intention as it's debated in literary theory, information scientists have developed a robust set of techniques for distinguishing meanings intended in context from those that pertain generally.

How do these work? Semantic models are built on matrices of word

[1] The intellectual history of information science is not included in literary theory curricula and so is generally unknown to literary scholars. For a review of the field that focuses most directly on issues that pertain to this essay, see Nancy Ide and Jean Véronis, "Word Sense Disambiguation: The State of the Art," *Computational Linguistics* 24, 1 (1998): 1-40.

counts.[2] The simplest format is the term-document matrix, which takes a fixed vocabulary for its rows and, in the columns, records how often each term appears in each document. Slightly more complicated are term-context matrices, which take keywords as their columns and record how often terms appear near those keywords. "Context" refers here to the words that surround any particular instance of a term in a corpus. It can be measured in different ways, but most studies use a "context window" method that counts all words appearing within some narrow range. Term-context models can be highly sensitive to variations in word meaning because, even though they aren't usually organized by punctuation, they often capture collocation at the sentence level: words that tend to appear in similar sentences tend to be semantically related.[3] Term-context matrices represent the historical (and, perhaps, cognitive) process of concept selection and transformation.[4] The term-*document* matrix measures how words are distributed through a cor-

[2] For a humanistic overview of vector-space models, see Michael Gavin, "Vector Semantics, William Empson, and the Study of Ambiguity," Critical Inquiry (2018). A more detailed but still accessible review of the underlying mathematical concepts can be found in Dominic Widdows, *The Geometry of Meaning* (CSLI Publications, 2004). The standard survey of the field is Peter D. Turney and Patrick Pantel, "From Frequency to Meaning: Vector Space Models of Semantics," *Journal of Artificial Intelligence Research* 37 (2010): 141-188. Stephen Clark reviews much of the same material, but from the perspective of linguistics, in "Vector Space Models of Lexical Meaning," in *The Handbook of Contemporary Semantic Theory*, ed. Shalom Lappin and Chris Fox (John Wiley & Sons, 2015), 493-522.

[3] This idea is known as the "distributional hypothesis." Turney and Pantel describe it as claiming that "words that occur in similar contexts tend to have similar meanings" ("From Frequency to Meaning," 143). Turney and Pantel cite many scholars who worked on this idea, but the distributional hypothesis is conventionally attributed to linguist Zellig Harris, who first proposed the idea in "Distributional Structure," *Mind* 10 (1954): 146-62.

[4] The possibility that mathematical models of semantics can be used for the study of cognition goes back to the work of psychologist Charles E. Osgood. See "The Nature and Measurement of Meaning," *Psychological Bulletin* 49, 3 (1952): 197-237. Osgood did not work from corpus data but instead developed a questionnaire that asked respondents to rank word meanings along various binary axes (good/bad, big/small, etc.). Osgood used matrix factor analysis to collapse these metrics to a few key lines of conceptual differentiation. The mathematical procedures Osgood used share much in common with singular-value decomposition, which when performed over word-document matrices is the central procedure of latent semantic analysis. Thomas K. Landauer, also a psychologist, has argued that corpus-based semantics studied in this way provide useful surrogates for or indices to conceptual formations that facilitate cognition. See "LSA as a Theory of Meaning," in *Handbook of Latent Semantic Analysis*, ed. Landauer et al. (Lawrence Erlbaum Associates, 2007), 3-34.

pus; the term-*context* matrix measures how words appear locally in relation to each other.

This paper proposes an alternate data format that represents documents as individual term-context matrices.[5] Given a vocabulary of words, W, and a shorter list of high-frequency features, K, a document can be represented as a matrix, $D(W, K)$, that records how frequently each word appears in the context of each feature. In all examples below, I use a conventional keyword-in-context measurement for the feature space, using a window of five word tokens before and after each instance of a keyword, but scholars could use different data structures. The only requirement is to ensure that K can be measured consistently across all documents. The total corpus, $V(W, K, D,)$, is represented by a three-dimensional array that stretches the traditional term-context matrix over a third dimension of length n (the number of documents), resulting in a very sparse rectangular prism of word counts. (See Figure 1.) Breaking up the term-context matrix in this way supports a variety of computations that evaluate word-use patterns. Every word can be represented by taking a horizontal slice through the array, resulting in the matrix, $W(K, D)$, with keywords for rows and documents for columns. Words and documents can be combined into groups by taking the sum or mean of their elements. Words can be separated into topical clusters and collapsed into an aggregate matrix. Documents can be sorted into groups of any kind. The corpus itself is the largest possible group, and the traditional term-context matrix, $V(W, K)$, can be recovered by simply taking the element-wise sum over the whole, such that

$$V(W, K) = \sum_{D=1}^{n} D(W, K) \tag{1}$$

for all documents in the corpus. Smaller subsets can be taken by coordinating matrix composition with features in the metadata. For example, the analyses below build composite matrices that group documents by theme, author, and year. The immediate goal is to compare patterns of word use in individual documents with patterns that hold over larger composite entities. The most

[5] Vectors of word frequencies denoted with italicized lower-case letters; matrices with italicized capital letters; and 3D tensors with capital italic script. The subscript that follows a matrix notation provides the dimensions, so $D(W, K)$ denotes a document (D) measured over a rows of words (W) each of which is described over columns of features (K).

common calculation compares a vector of collocation values of a given word in a given document, $w(D)$, to the corresponding vector, $w(V)$.

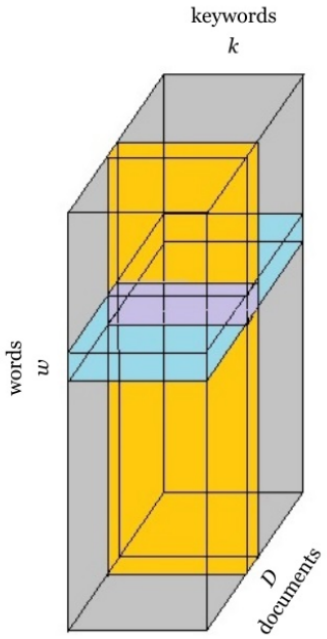

Figure 1: The vector-space semantic model.

Such comparisons conventionally take the form of a dot product or cosine measurement between vectors. For search engines and machine translation, word-sense disambiguation evaluates short statements (often no more than a few words long) by combining their corresponding vectors and measuring the cosine distance that separates the use-in-context from each word taken individually. Imagine typing the word "plant" into translation software. Should it return "la plante" or "l'usine"? The presence of contextual terms like "flowers" or "manufacturing" will in most cases indicate the intended meaning: "the plant manufacturing" will be translated as "l'usine de fabrication" while "the plant flowers" will return "les fleurs des plantes." Thus, in the fields of information retrieval and machine translation, *intention is measured by taking the cosine distance that separates individual uses of words from their uses over a corpus.* Intention is represented as a numerical gap – as a kind of fissure in the space of meaning – that separates the vector implied by a document from the vector recorded in a corpus. Intentions select meaning

from a system of possible meanings. To read for intention involves retracing this process of selection by measuring how much distance separates a word, as used in a document, from the same word, as used generally over the corpus as a whole. Under this framework, an intention is represented by a difference between a document and a corpus.

Formally, the deviation between a word's document-level meaning and its corpus-level meaning is defined as the cosine distance that separates their corresponding vectors, such that

$$\Delta w(D, V) = 1 - cos(w(D), w(V)) \tag{2}$$

which provides a single numerical score for each word in each document, ranging between 0 and 1, where 1 indicates a total disparity (and thus maximum distance) between the document and the corpus, while 0 indicates a total equivalence. Because each document is part of the corpus, its uses are included in the total vector, so an absolute distance of 1 is impossible.

2.2 Work.

Cosine distance is indifferent to scale, but nonetheless words that are used more frequently tend to be used more conventionally. If a word appears only once or twice, it might exhibit a high level of variation. I can use the words "tomato" and "carburetor" in a sentence where they're completely out of place. But words that are used dozens or hundreds of times tend to regress toward their general form. Virtual pseudo-documents, generated randomly using a Markov-chain process, follow this pattern quite smoothly. Notice in Figure 2 the logarithmic shape of the curve, which reaches toward 0 as word-count increases: high-frequency words often become semantically indistinguishable from the normal patterns of the corpus. Yet, in actual documents there are usually words that push against this general dynamic by invoking persistently improbable connections to other words. (See Figure 3.) In real documents written by actual, intending persons, intentions organize language in ways that distort statistical patterns. Significant words hover above the arc of semantic probability.

I describe this push, this persistence, as the conceptual work documents perform on words. Conceptual work, , is the product of the word's frequency and its deviance, such that

$$C = F\Delta \tag{3}$$

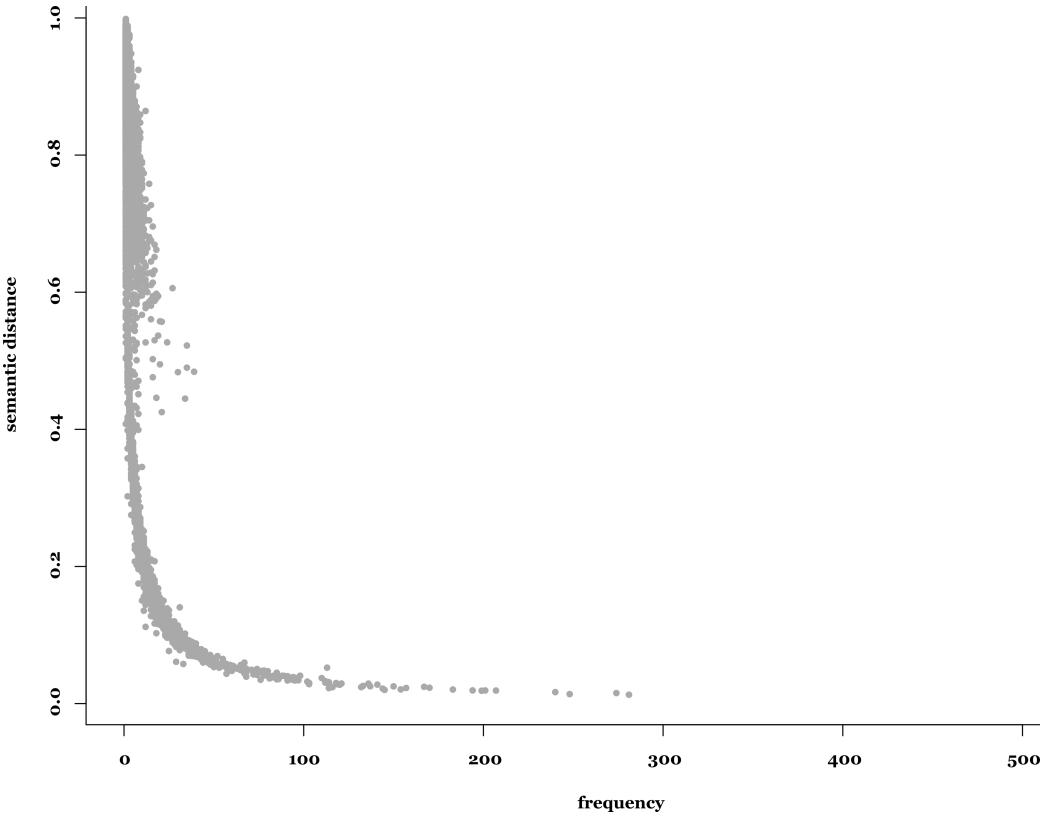

Figure 2: Word deviance and frequency in a simulated document of 50,000 tokens. The small bulge of terms that cluster near the top, as well as between deviance values of 0.4 and 0.6, represent a relatively small amount of "noise" that occurs because the Markov chain is built on a decision tree that looks only at the individual preceding word. To get a sense of this pseudo-document, consider this sample stretch of text: "court opinions humble service next heir king power presence blessed spirit proves lion coward self fed relieved evidence open force power god james crown shillings."

where F is the document-level word frequency and Δ represents, as above, the distance separating the word's vector in D from its vector over V. Figure 3 visualizes a few typical results of this measurement, calculated over a model drawn from the EEBO-TCP corpus, highlighting terms that display the highest conceptual work for each document. In these charts, and in the analyses that follow, $\mathcal{V}$ is a 3D array composed of word-context matrices drawn from roughly 18,000 documents from the EEBO-TCP corpus dated 1640 to 1699;

K is a list of about 2,000 keywords, chosen because they were among the 5,000 most frequent words in every single year during that range; and W is a larger vocabulary of terms that were consistently among the 20,000 most frequent words.

Here, then, is the central claim of this essay: *Words on which a document performs the most conceptual work are most likely to be relevant to a qualitative restatement of that document's intended meaning; that is, they're likely to be relevant to a faithful reading of the text.*

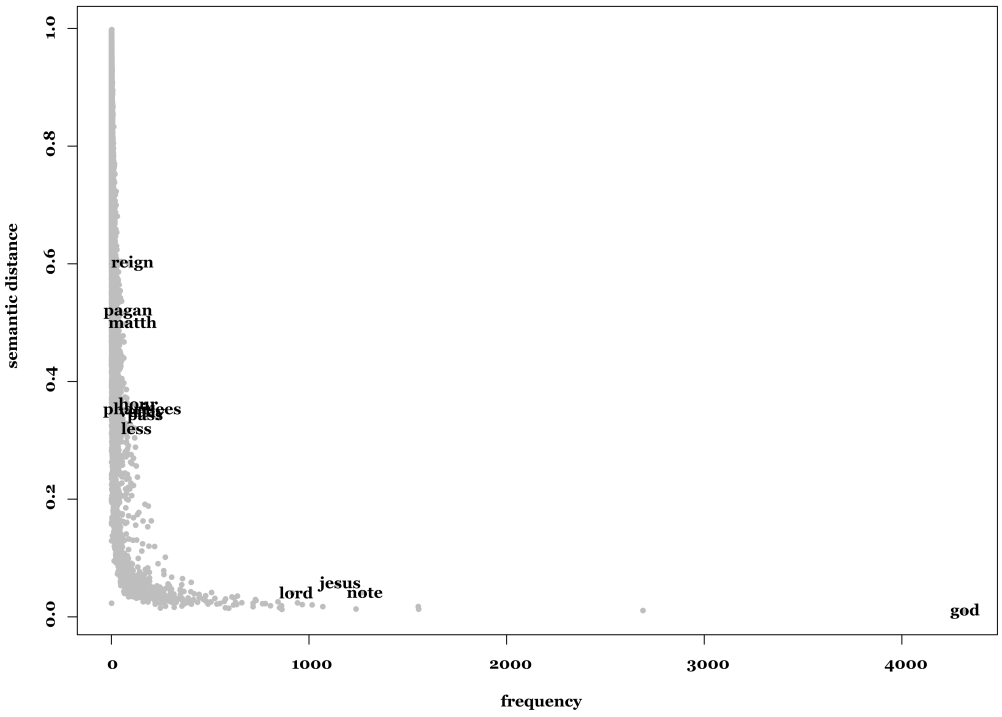

Figure 3: Richard Baxter's Paraphrase on the New Testament (A26981).

2.3 Normalized work.

For interpretive purposes, we often want to distinguish moments when an author invokes commonly shared concepts from moments when the author

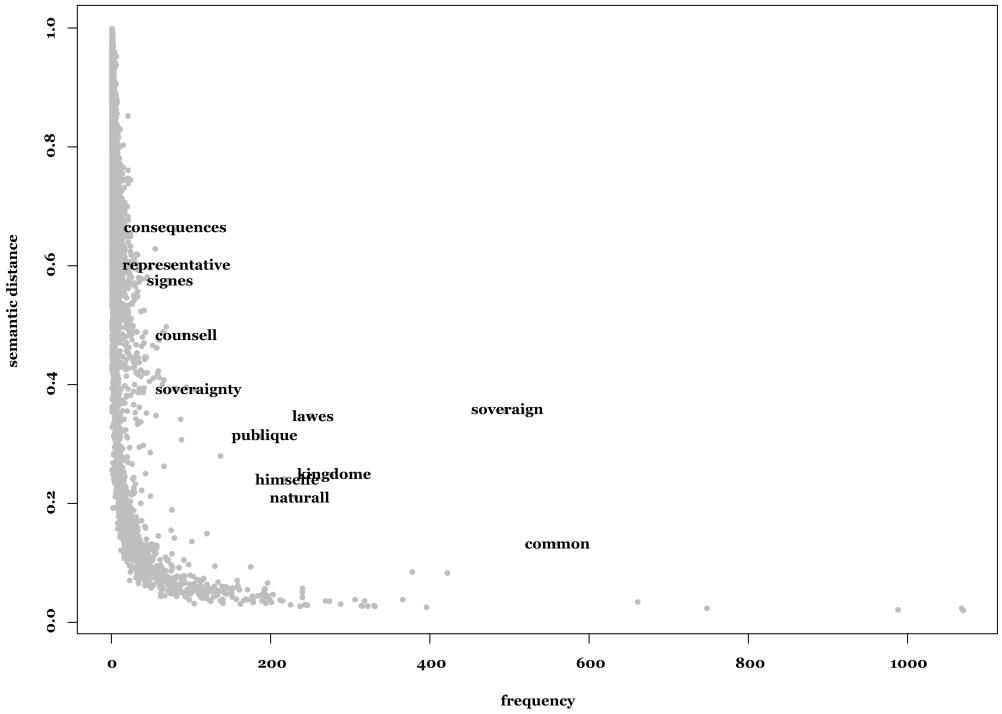

Figure 4: Thomas Hobbes's Leviathan(A43998).

seems to be asking readers to adopt a new way of thinking. Under the theoretical framework of vector-space semantics, this problem can be restated as asking how relevant the corpus-level vector is to the document-level vector. This point requires a bit of elaboration. In information retrieval, words in a search query are represented entirely by vectors trained over a corpus. They aren't assumed to have any special meaning, just a contextual meaning determined by their neighboring words, calculated typically by taking a composite sum, $w(V)+w(V)+...+w(V)$, for every word in the search query, as in the example I mentioned above, which distinguishes *plant + manufacturing* from *plant + flowers* to identify different analogues in French. Notice, though, that every word vector is built over the corpus and is only later used to evaluate searches and documents. Search mechanisms assume that intended meanings in a query select from an already existing space of semantic possibility. In real documents, this assumption often holds. For example,

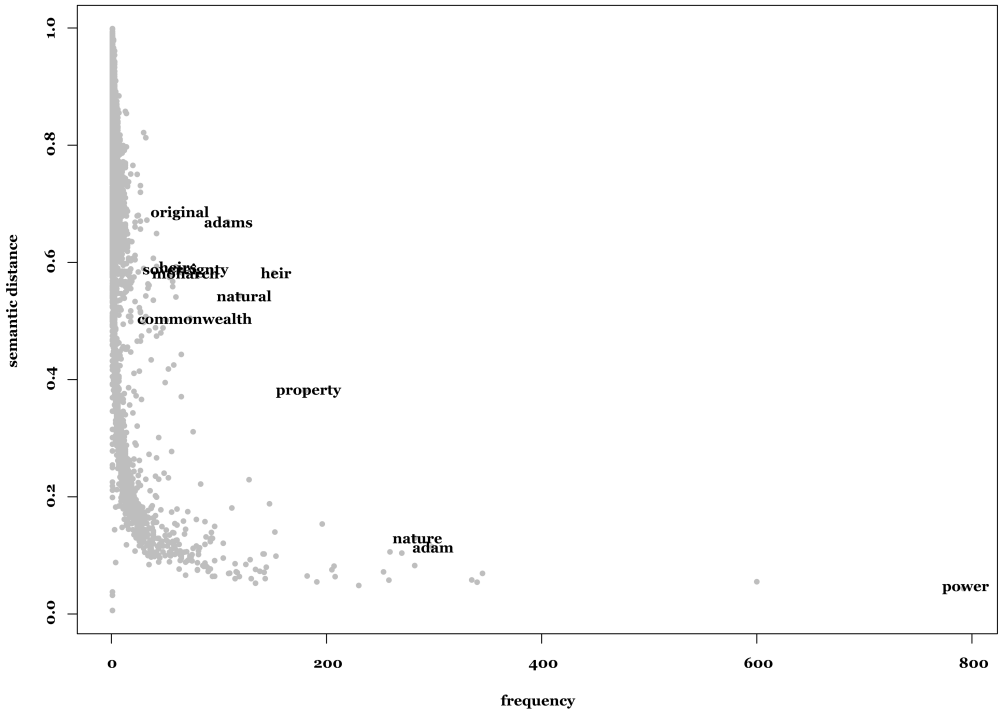

Figure 5: John Locke's Two Treatises of Government (A48901).

the previous sentence did not advance any particularly innovative use of the term "assumption," so in all likelihood a vector trained over a corpus of academic writing will better reflect my meaning than a vector compiled over this essay by itself. However, as we might suspect, and as Figure 3 suggests, the meaning of a word advanced in a text may be (perhaps in some subtle way) unique to that text. In which case, we may prefer to represent such a word using its document-level vector, $w(D)$, rather than $w(V)$. The challenge is to differentiate between these cases and to develop a metric sensitive to this general concern.

Because semantic deviance is in most cases so strongly dependent on term frequency, it is necessary to control for this effect. In the simulated, Markov-chain document described above, more common words tended to gravitate toward corpus-level meanings. Actual documents also reflect this correlation. For example, in the EEBO-TCP model, any time a document uses any word

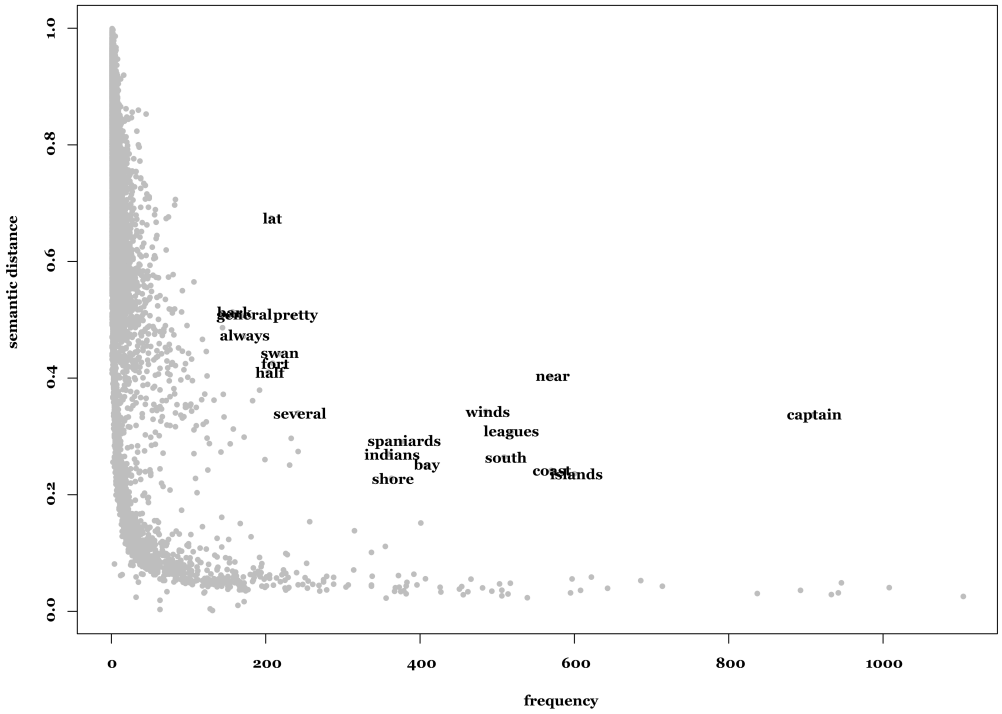

Figure 6: William Dampier's, A New Voyage Round the World (A36106).

just once, the average semantic deviance is 0.81. At 50 times, the average deviance is just 0.19. (See Figure 4.) This dependence can be modeled and specified by taking a logarithmic regression of the average deviance, δ, for words used at a given frequency, and the variation of that deviance, ϕ, such that

$$\delta(f) = \alpha(log\overline{\Delta}) + \beta \tag{4}$$

and

$$\phi(f) = \alpha(log\sigma) + \beta \tag{5}$$

for all unique values, f of F, where σ denotes the standard deviation of Δ for each word of frequency f. These metrics combine into a single threshold, Θ, such that

$$\theta = \delta + \mu\phi \tag{6}$$

where the coefficient μ is set by the researcher. If $\mu = 0$, the threshold simply equals the mean cosine distance, and so any word that is more semantically deviant than average will pass the threshold. Somewhat more stringently, μ can be generated empirically such that

$$\mu = \frac{1 - \delta_1}{\phi_1} \tag{7}$$

which results in $\theta_1 = 1$, placing under the threshold all instances in which a word is used just once.

Using these baseline metrics, conceptual work can be modified to differentiate words that remain within a margin of typical variation from those that more strongly push against semantic norms. With normalization,

$$\Delta' = \Delta - \theta \tag{8}$$

and therefore

$$C' = F\Delta' \tag{9}$$

These adjustments cause negative values in conceptual work for words below the threshold. Words with the largest negative values will usually be the most frequent terms, while the largest positive values will be semantic outliers. At $\mu = 0$, this differentiation characterizes the relationship between a document and a corpus, supplementing the simple representation of conceptual work. When μ is set empirically as above, it aggressively identifies only the most distinctive words. Low values of μ facilitate paraphrase; high values target outliers.

I'll provide more detailed applications of these ideas below, but a preliminary example may be useful here. In John Milton's *Paradise Lost* (A50919), when compared against this corpus, the word "satan" is used in highly unusual ways, while "eve" sits well under the threshold. (See Figure 5.) In John Dryden's *Macflecknoe* (A36643), "dulness" is used pretty distinctively, but "wit" isn't. This suggests that Milton discussed Eve in ways not too far outside the mainstream of seventeenth-century thought but that he was advancing an altogether new conception of Satan. Similarly, Dryden had little original to say about wit but perhaps something very original to say about wit's reciprocal opposite. Of course, the rhetoric of originality I'm casually deploying here can be dropped. The goal of setting these thresholds is not to glorify an author's inventiveness but to provide an index of typicality. How typical or atypical are features of a given document, measured

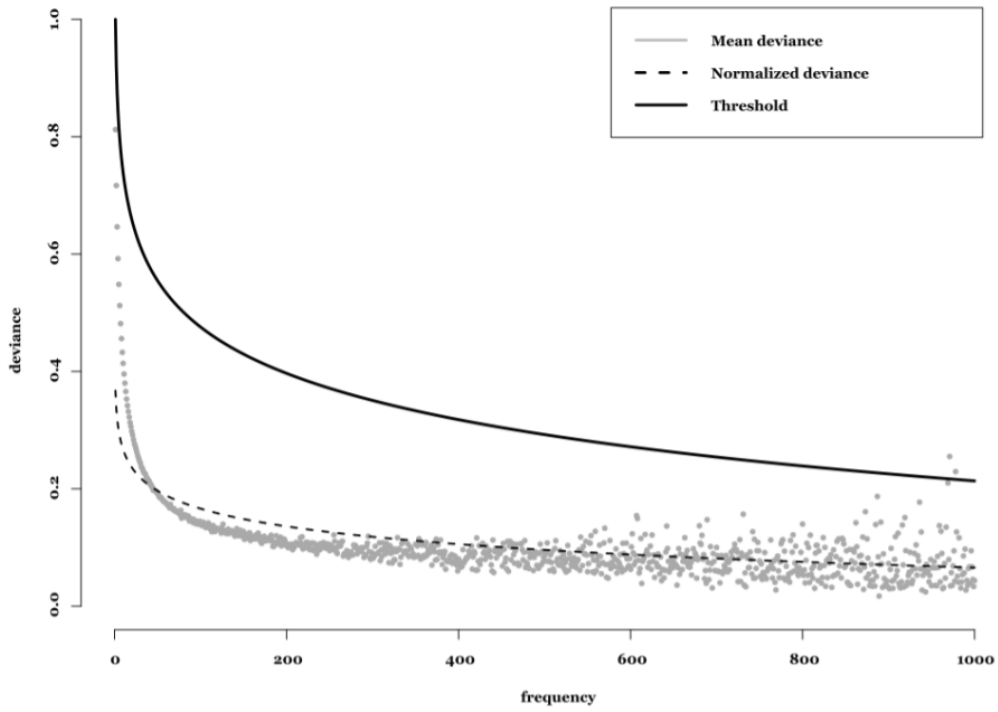

Figure 7: Word deviance and frequency over EEBO-TCP corpus. Notice that as values of F increase above 50, the average semantic deviance in this corpus tends to swing wildly, because the sample size becomes smaller and smaller. There are only so many documents that use any word exactly 637 times, for example, so this model relies on a logarithmic regression to estimate a normalized deviance, δ. The coefficients for the normalized deviance pictured here are $\alpha = -0.043659$ and $\beta = 0.367250$. The coefficients for the normalized "deviation of deviance," ψ, (not pictured) are $\alpha = -0.035277$ and $\beta = 0.317968$. These curves are combined into a threshold value, Θ, that can be used to filter out less distinctive words. As pictured here, $\mu = 1.989$, so words must be two standard deviations above average to pass.

against a given corpus? Restated in terms more directly relevant to literary history: At which points does an author transparently reflect contemporary discourse, and where does the author intervene in that discourse?

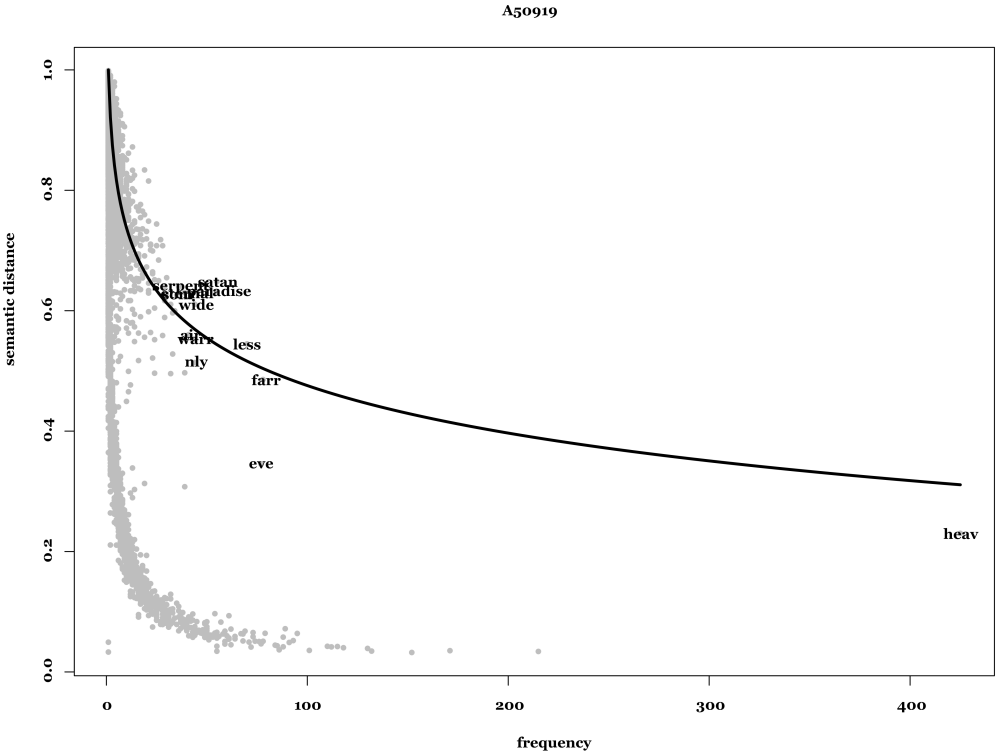

Figure 8: Word deviance and frequency for John Milton's Paradise Lost (A50919).

2.4 Semantic transparency.

In many situations it may be useful to have a single metric that characterizes a text's overall level of semantic deviation. Several possible measures immediately propose themselves. One might take the total semantic deviance, or the average deviance per word type, or the average conceptual work per word type. Unfortunately, these measurements would all be biased, in one direction or another, towards longer or shorter works. (Short documents like *Macflecknoe* almost always have a comparatively high deviance per word.) One might correct for this problem by using the normalized deviance in any of the above calculations, but the underlying tendency would continue to drive results. What's needed is a general way of describing texts' semantic chaos, so to speak. Looking back over the graphs in Figure 3, some seem

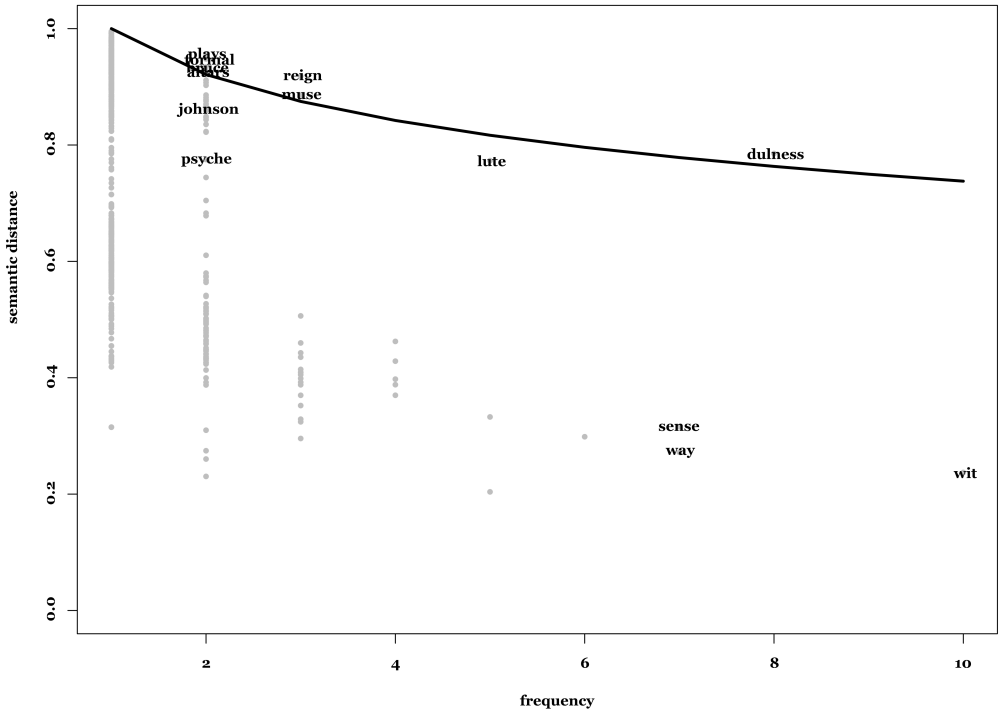

Figure 9: Word deviance and frequency for John Dryden's Macflecknoe (A36643).

to follow a fairly neat curve while others seem to bounce around willy nilly. Some texts use words in highly focused, atypical ways, while others seem content to ride the wave of statistical probability.

The mathematical operation for making this kind of distinction is entropy. Formally, the semantic transparency, H, of a document is the entropy of the work performed on its words, such that

$$H = -\sum p(C) log_n p(C) \tag{10}$$

where n equals the number of unique word in D and p denotes a probability function that treats each value of C relative to the whole, such that $p(C) = C \sum C$ for each word. This definition bears an obvious affinity to information entropy, as first defined in communication theory by Ralph Hartley

and later popularized by Claude Shannon.[6] However, two differences must be mentioned. First, the logarithm is taken to base n, which adjusts the scale to each document based on its number of unique word forms, thus placing all values of H within a common range of 0 to 1, regardless of document length. Second, information entropy as Shannon described it looks exclusively at the relative frequency of characters or words in a document. Because of how conceptual work is defined above, semantic entropy controls frequency with semantic similarity. Thus, the relation between document and corpus is embedded in the calculation. Documents that distribute conceptual work evenly over their vocabulary – those with smoother looking curves – more transparently communicate meanings of the corpus. This greater transparency is captured by a higher entropy score.

At first glance (or even at second and third glances), it may seem strange that smoother looking curves associate with a higher entropy. For example, compare Figures 3 and 6. Richard Baxter's *Paraphrase on the New Testament* (A26981) displays patterns reminiscent of a simulated document, while William Dampier's *New Voyage* uses a number of significant words in highly distinctive ways. Part of this apparent difference can be attributed to distortions in the visualization because the x-axis on Baxter's *Paraphrase* is longer, but nonetheless one senses that Dampier's *Voyage* displays more variation. So, shouldn't its entropy be higher? No. Why? Precisely because conceptual work in Dampier is powerfully focused on a relatively small group of words: "captain," "coast," "islands," "winds," "spaniards," "indians," etc. Those words are responsible for a comparatively high proportion of the total work performed by the document. In Baxter's *Paraphrase*, that semantic responsibility is much more evenly divided among its words. Thus, for Baxter the value of H is 0.96, and for Dampier it is 0.92. Across the EEBO-TCP model, most transparencies range between 0.88 and 0.95, which makes intuitive sense if we analogize this metric as an answer to the question: What proportion of an author's total meaning is inherited from its historical setting, as represented by the corpus? However, if the purpose is to compare documents along a spectrum of transparency, it may be useful to expand the

[6] See Ralph V. L. Hartley, "The Transmission of Information," Bell Labs Technical Journal 7, 3 (1928): 535-563; and Claude Shannon and Warren Weaver, The Mathematical Theory of Communication (University of Illinois Press, 1949), especially 48-53. For a more accessible overview of the mathematical concept of entropy and its application to the study of language, see Thomas M. Cover and Joy A. Thomas, Elements of Information Theory (John Wiley & Sons, 2012), chapter 2.

metric such that its variation is not confined to so narrow a band, in which case semantic entropy might be calculated over normalized work, C' rather than C, disregarding all negative values. As discussed above, normalizing work filters out and emphasizes semantic outliers. Using normalized work as a basis for comparing Baxter and Dampier returns H values of 0.83 and 0.61, respectively.

3 Application: texts, authors, times

3.1 Overview

Before turning to specific applications, it's worth pausing to review the most relevant aspects of the ideas presented above. First, the theory sets aside biography and focuses exclusively on patterns of word use. Borrowing from the field of information retrieval – where engineers are tasked with answering the practical question, "What do users mean?" – this theory defines reading for intention as finding a fit (or lack of fit) between words as they're used in a document and words as they're used over a corpus. An intended meaning is selected from a field of possible meanings, but intentions also constitute that very field, building it up piece by piece, always deviating (sometimes only slightly, sometimes wildly) from the general structure that slowly accrues. I define this persistent deviation as the conceptual work of a document, and the main purpose of using this metric is to differentiate features of a document that push against common patterns from those that mostly support or follow them. Ignoring typographical errors and other kinds of mistakes, all meanings in a human-authored document are "intended," in the folk sense of that word, but many meanings are best accounted for, not by attributing them to an individual but by tracing their conditions of legibility. Meanings exist outside of and supervene on statements, not because they are psychological entities originating in the speaker's mind, but because they are inherited from a system of other statements. This implies that intentions exist only under a description of a relation between a text and some larger frame within which the text is imagined to act. It also implies (though this may not be obvious to readers and so I can offer it only as a tentative, perhaps useful, definition) that intentions are aggregate entities composed of semantic simples, and so are subject to various kinds of mathematical description and manipulation.

To demonstrate how these principles might be realized in practice, I

present three case studies (briefly, and without reviewing the relevant scholarship in each case). The first paraphrases an individual document. The second characterizes an author's career. The third incorporates temporality, extracting historical events and tracing change over time. I then suggest a few further avenues of methodological inquiry.

3.2 Analyzing a document.

The vector-space semantic model can be used to support readings by identifying keywords (under various metrics, each of which offers a different perspective) and by situating a text against multiple corpora. I'll focus on John Locke's *Two Treatises of Government* (A48901), pictured above in Figure 5.

Such analysis begins by identifying a subcorpus, $\mathcal{L}$, made of the 100 documents from most similar to Locke's *Treatises*. Because Locke's vocabulary is highly distinctive, the documents in $\mathcal{L}$ form a tightly compact group of texts focused on political discussion. The subcorpus includes not only the *Treatises* but also works by politicians, theorists, and polemicists like Robert Filmer, Algernon Sidney, Thomas Hobbes, Matthew Tindal, and others. Among this group, those that are most semantically transparent (that is, the documents that use words in the most EEBO-typical ways) include works by John Milton and Roger L'Estrange. Only four documents in $\mathcal{L}$ have a lower semantic entropy than Locke's *Treatises* and, of those, two were written by Locke's friend and fellow Whig philosopher, James Tyrrell. These initial results suggest in broad outline a picture that should be immediately familiar: when he published the *Treatises* Locke joined a dense body of political debate while intervening in particular around a set of unusually focused key concepts.

To understand the nature of that intervention, measures of conceptual work can be triangulated between Locke's *Treatises*, the subcorpus of contemporary political discussion, and the larger corpus which stands as a proxy for the period as a whole. Let us first examine the minima and maxima of Locke's conceptual work to see which words he uses most conventionally, and which most deviantly. Paraphrasing over Table 2 (which is itself paraphrase), we can say that the *Treatises* discuss how power and authority operate in a government among men, and that Locke addresses this topic by focusing in particular on the natural or original grounds of authority, as inherited from Adam or acquired through property. Locke joins a field of political discussion that, itself, had long debated sovereignty and legislative power, as they are tested in states of nature and states of war. Against this general discourse

Documents most similar to Locke's Two Treatises	**Distance**
A. Sidney, Discourses concerning government (A60214)	0.24
R. Filmer, The free-holders grand inquest (A41303)	0.27
R. Filmer, Forms of government (A41307)	0.27
H. Parker, The right of subiects (A56187)	0.29
T. Goddard, The state-physician unmaskt (A42895)	0.29
Most semantically transparent	**Transparency**
J. Milton, Sovereign right and power (A50940)	0.984
R. L'Estrange, Two cases submitted (A58674)	0.982
A. Littleton, A sermon (A48734)	0.981
H. Vane, A needful corrective (A65173)	0.98
J. Palmer, The present state of New-England (N00400)	0.98
Least transparent	**Transparency**
J. Tyrrell, The ancient constitution (A64083)	0.924
J. Kettlewell, The duty of allegiance (A47295)	0.925
T. Downes, An examination (A36486)	0.932
J. Tyrrell, Disquisition of the law of nature (A64084)	0.936
J. Locke, Two treatises of government (A48901)	0.938

Table 1: Contextualizing Locke's Two Treatises using a vector-space model.

of natural law, Locke's *Treatises* stand out for their sustained attention to the domestic family as a unit of analysis, especially on the relationships among children and fathers. Over EEBO as a whole, the term "children" is closely associated with education, domestic care, concerns over legitimacy, and echoes of biblical myth. (See Figure X.) Not so in Locke's *Treatises*. There, children are further from Moloch and much closer to Grotius. In Locke's world, children are begotten by parents who confront the question of what might reasonably and naturally be limited, delegated, divided, and owed among them. Locke's children are entities in thought experiments about the distribution of power and property in the commonwealth. Indeed, this semantic reconfiguration of "children" is an important hinge on which his argument turns.

The most persistently conventional terms in Locke.				
power	government	right	authority	men
-25.59	-19.89	-19.74	-18.68	-18.54
Locke's most persistently deviant terms.				
heir	adams	natural	property	original
65.56	54.7	46.19	43.18	29.73
Most semantically distinct terms among all Locke's peers.				
natural	council	soveraign	nature	general
795.56	593.47	467.32	464.38	460.4
Terms where Locke is furthest from the full corpus.				
state	children	adam	father	adams
9.55	8.1	7.8	7.45	6.94

Table 2: Conceptual work in Locke's Two Treatises of Government (A48901). The first two rows display the largest negative and positive values for normalized semantic work (CD' for which D is Locke's Treatises and $\mu = 0$). The third row evaluates the conceptual work (not normalized) over a composite matrix, L, representing all documents in $\mathcal{L}$. The fourth row shows terms with the highest triangulated work (where Δ equals the deviance separating L from V minus the sum of the deviances that separate D from L and L from V; that is, words for which Locke deviates from EEBO more than his peers do.

3.3 Generalizing over an author's career.

Further exploration of Locke's *Treatises* would distract from my purpose here, however. A key feature of the 3D textual model is that it preserves keyword-in-context data while keeping documents separate, allowing researchers to reconfigure any corpus around historically significant subsets. When historicizing Locke, above, this task was performed using the diction of the *Treatises*: the 100 most-similar documents were chosen, regardless of any historical facts about the books or their authors, in order to recover something analogous to the seventeenth-century discourse of natural law. In this section and the next I describe how the model can be re-shuffled according to document-level features recorded in the metadata. The phrase "metadata" refers conventionally to information stored in library catalogues, like author, title, publisher, place and date of publication, etc. One might say that computational methods distort literary history, not by changing its scale of anal-

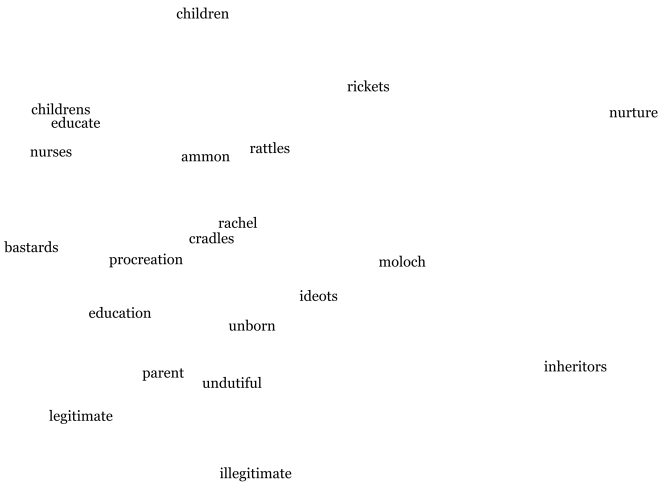

Figure 10: 25 terms most similar to "children," as represented over the EEBO-TCP corpus, V_{wk}. Terms are projected onto a two-dimensional surface and plotted by semantic similarity.

ysis, but placing at its center a new question: How do observable patterns in books relate to observable patterns in bibliographies? To this question I now explicitly turn.

In a corpus, an author is not so much a person as a value of a variable, a datum around which documents can be clustered and ordered. Because for various commonsensical reasons authors are usually presumed to exist prior to and independently of the books they author – that is, they are assumed in literary history to be entities with their own attributes and not to be merely attributes of books, although in bibliographies they are certainly that – authors provide a convenient shorthand for describing the organization of ideas in any collection. To attribute an intention to an author is thus to attribute a paraphrase to the value of an attribute. My goal in this section

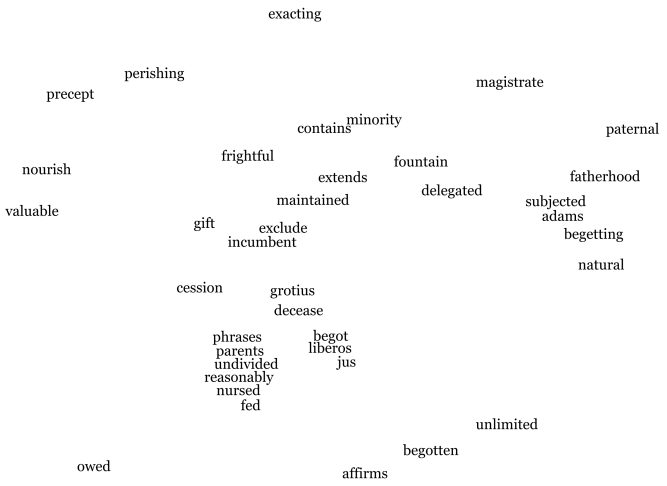

Figure 11: 25 terms most similar to "children" over Locke's Two Treatises D_{wk}.

is to attribute meanings to the value, "Behn, Aphra, 1640-1689."

A subset of $\mathcal{V}$ containing works by Behn is a 3D array, $\mathcal{L}$, of the same structure as $\mathcal{V}$ and $\mathcal{L}$, but composed of 45 matrices drawn from documents where Behn's name appears as first author. From this array is built a composite matrix, B, using element-wise addition, as above. This single matrix represents word-collocation data in all of Behn's texts and can be compared to any of her books individually or to EEBO as a whole. When examining the conceptual work performed by Behn's texts, notice that several of the terms reflect abbreviations of names of characters from her plays, "gall," "tim," and "pat." This is a common artifact of semantic analyses when applied to drama. The other terms, though – "devil," "soft," "vows," and "maid" – point fairly directly to her abiding concern with femininity and sexual ethics. Figure X shows how the terms "vows" and "devil" are dis-

tributed in the semantic space of Behn's works: "vows" sits at the center of amatory entanglements, while "devil" indicates, not a concern with supernatural beings, but a language of masculine banter, common in her plays, that invokes an altogether different perspective on the sexual politics of oath-making and swearing. From the perspective of fiction, "vows" are putatively eternal affective ties too often undone; in her plays, oaths like "devil" are mere ejaculations, a rough and ready language of homosocial exchange that structures male relationships fragmented by sexual rivalry.

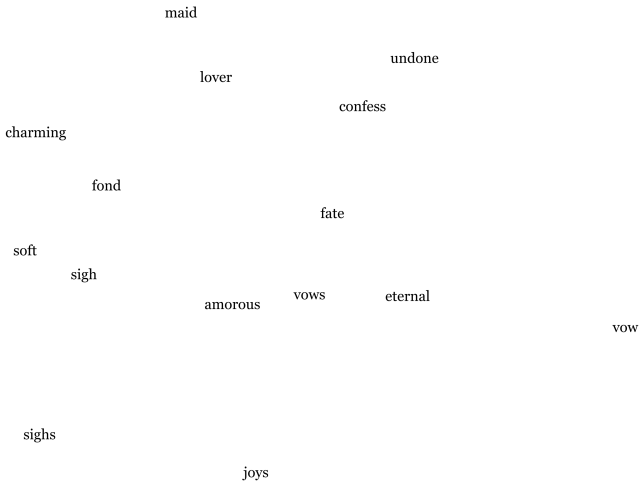

Figure 12: 25 words most similar to "vows" in a corpus of Aphra Behn's works.

The works of an author also provide an alternative backdrop against which individual texts can be compared. Returning to Table 2: In *Love Letters between a Nobleman and his Sister* (A27301), Behn's interests there revolve more centrally on the conditions of romantic address and critical judgment; prominence of the word "madam" suggests that this book, more than and

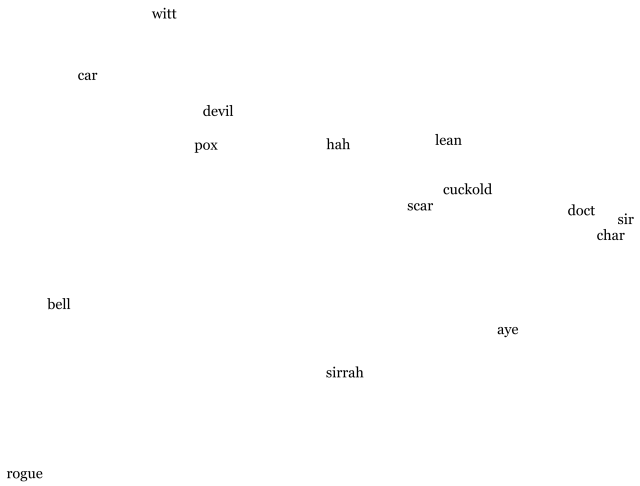

Figure 13: 25 words most similar to "devil" in a corpus of Aphra Behn's works.

differently from her others, ventriloquizes a male address to a female subject, while terms like "lovers" (in the plural), "gay," "charming," and "fancy" show that this address involves persistent evaluation and critique. When triangulated against Behn's works and all of EEBO, terms in *Love Letters* that stand out reflect an overriding concern with time and repetition, as, through the epistolary mode, abstract anchors of the subject ("soul," "sex," "love") are read and pored over thousands of times. Whereas in her other works Behn is concerned primarily with how promises and passions bind lovers into compromised relationships, in *Love Letters* this general concern is complemented with another: we see echoes of how epistolary exchange brings into focus the illusion that sexuality constitutes one's true self, who exists independently of her social setting, who is subject to self-examination, and who transcends mere forms of address.

Persistently deviant terms in Behn's works.						
gall	devil	soft	vows	maid	tim	pat
286.75	248.62	246.11	189.01	184.79	176.63	163.52
Persistently deviant terms in Love Letters.						
madam	business	lovers	gay	maid	charming	fancy
32.59	20.31	18.31	18.13	16.95	16.81	16.79
Love Letters compared to Behn's works and the full corpus.						
times	thousand	soul	sex	read	love	even
4.63	3.62	3.61	2.31	2.29	2.11	2.05

Table 3: Conceptual work in Aphra Behn's works, with emphasis on Love Letters between a Nobleman and his Sister (A27301). The first rows displays the largest values for semantic work over the composite matrix, B, representing all documents in $\mathcal{B}$. The second row evaluates the conceptual work in Love Letters when compared to B. The third row triangulates between Love Letters, Behn's works, and the full corpus; that is, it shows words for which Behn's Love Letters deviates from EEBO more than her other works do.

3.4 Temporality.

Extracting the vocabulary of historical events and tracing change over time requires a more sophisticated manipulation of metadata than does isolating individual authors. However, the two operations share much in common, because a theory of intention that abstracts over many statements by the same author can abstract over collections of any type, including over documents organized temporally. Intuitively, we can think of the word-context matrix representing a document as an answer to the question, "How did the author use words in this book?" In the same way, we can ask, "How did people from a given year use words?" Just like intentions describe the conceptual work performed by authors, so too they can describe the conceptual work performed by chronology. Persons organize meaning in a social field, and dates and timestamps really aren't so different: the totality of statements by Shakespeare might differ from those of Jonson, so too the year 1989 might have something different on its mind than 2016. Such differences are conventionally hypostatized as historical events or conceptual change, and we can think of an event or a change as the *a priori* cause of a concentration in the conceptual work time performs on words.

Unlike the human experience of progressive time, in a corpus all times

exist in simultaneous juxtaposition. Indeed, significant effort is required to mold the shapeless temporality of a corpus into a form that simulates human experience. In practice, this means that the 3D array representing the corpus cannot merely be subsetted, as above. Whereas every author represents a subspace of the corpus – a section that can be taken out while preserving its overall shape – dates organized sequentially constitute a new space with its own internal logic, and so the axis representing documents must be transformed into an axis that represents chronological juxtaposition. (That is, a timeline.) Uses of each word over every year have to be gathered together, and the entire corpus must be re-shuffled accordingly. The array representing temporality, $\mathcal{T}$, has three dimensions: words, keywords, and time. A book-based corpus like EEBO is typically divided by year, so the resulting array has 60 time slots (1640 to 1699), each of which is composed of a matrix which takes the sum of all documents labeled with time t, such that

$$T(W, K) = \sum_{t(D)=t} D(W, K) \tag{11}$$

for all times in the corpus. Against this matrix, semantic deviance can be measured as above. However, because we're looking for words that become themselves in time, rather than documents that veer away from a norm, each word's actual deviance is subtracted from its average (mean) yearly deviance, resulting in positive values when the deviance is unusually low. Frequency also must be represented slightly differently. Because in EEBO most books are short but some are very long, in every year a small handful of long titles tend to distort frequency counts. For this reason, the temporal model disregards raw frequency and calculates conceptual work based on relative document frequency (the proportion of all titles in a given year that include the word). From these two measures of frequency and deviance, conceptual work and semantic transparency are calculated as above. Words with low transparency scores tend to be highly concentrated in small time periods, and so are often associated with historical events, while words with higher transparency rankings tend to be more evenly distributed through time, and so reflect either continuous change or sites of conceptual stability.

Consider the words "oates" and "management," as pictured in Figures X and X. The word "oates" was a fairly common alternate spelling for "oats," without the "e," and so it's used sporadically for the first few decades. In 1679, however, "oates" bursts into the corpus with a new meaning when

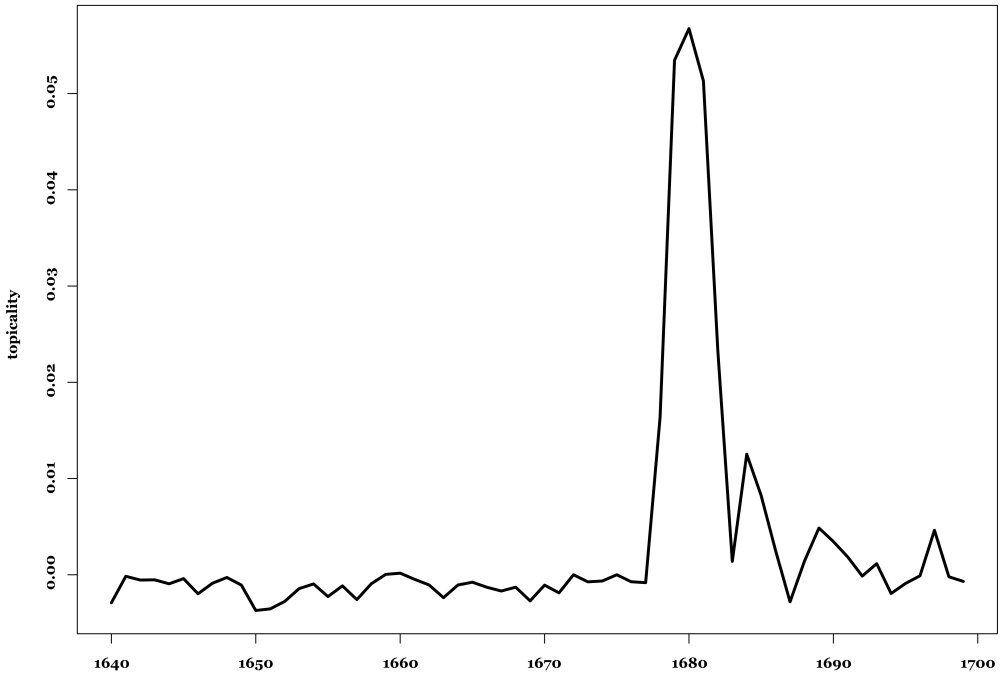

Figure 14: Times series graph showing the topicality of "oates" in the EEBO-TCP corpus, from 1640 to 1699. Topicality is a single metric showing years when a word is used with below-average deviance (years when the word means something closest to its total meaning) and above-average frequency (measured in this case as the proportion of titles published that year that use the word). Titus Oates was a central figure in the Popish Plot, a major political scandal in England from 1679 through 1681. Discussion of Oates's case was tightly concentrated in these years, resulting in a very low entropy of $H = 0.73$.

Titus Oates emerges as the central witness in the Popish Plot. Other terms that follow a similar topical pattern all surround this event; among these include names of key figures, like "godfrey" and "danby," as well as words like "jesuits," "plotters," and "conspirators." Change over time can be found by looking for words with rising or falling conceptual work. For example, the term "management" is among the most modern concepts of the Restoration,

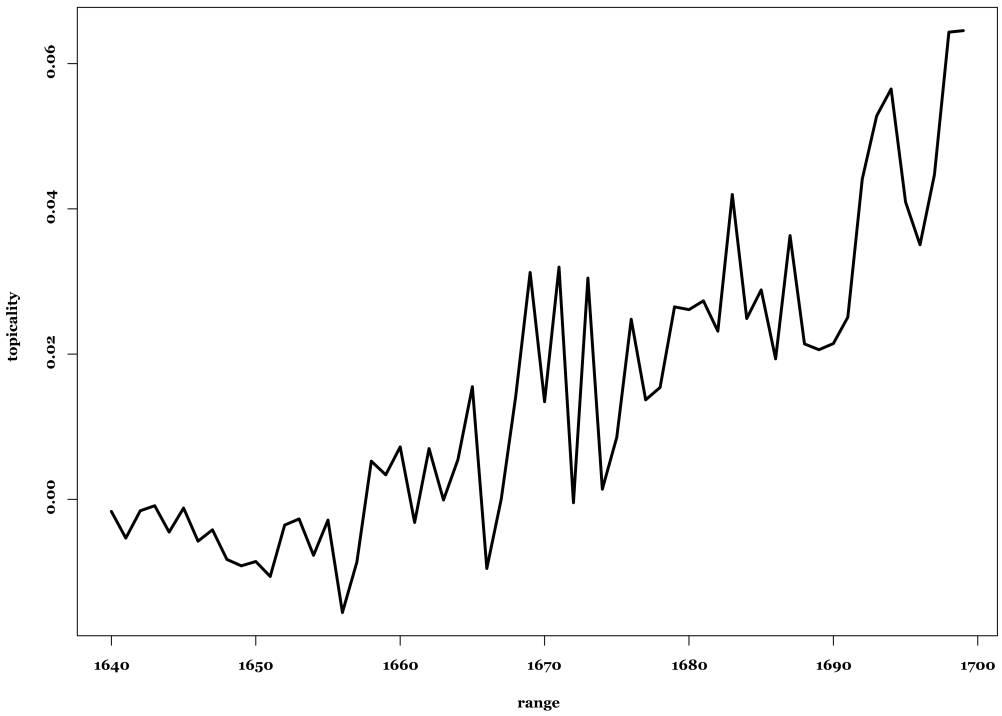

Figure 15: Time series graph showing the topicality of "management," 1640 to 1699. This graph shows an example of a gradual transition that accrues meaning over time. Whereas terms like "oates" are closely connected to specific events, terms like "management" shift through a more diffuse process. Because the change is gradual, conceptual work is fairly evenly distributed over time and the entropy score is comparatively high, at $H = .94$.

gradually increasing in both frequency and coherence over the last three decades. Terms with similar topical arcs include "impartial," "reflections," "methods," "success," and "designs" – all words related to the exercise of critical judgment in an emerging public sphere. Of the concepts inherited by readers and writers in London at the year 1700, these are the new ones.

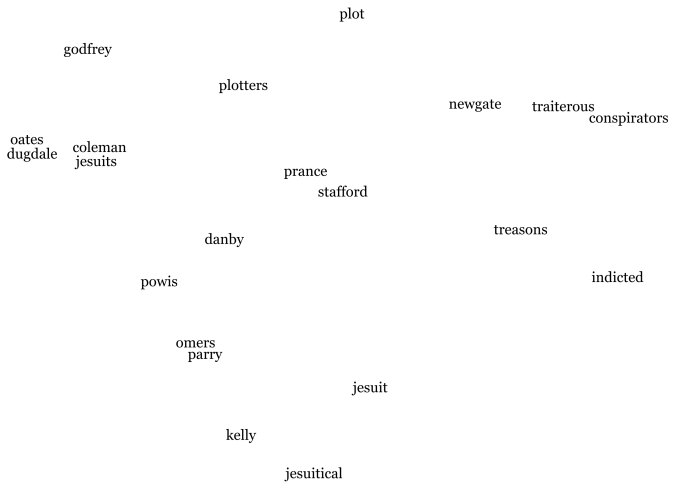

Figure 16: 25 terms most similar to "oates," measured by conceptual work over time. Similarity measurements over this matrix test the distribution of conceptual work over time: these twenty-five words exhibit above-average frequency and below-average semantic deviance during the same years. Notice how strongly these terms cohere around the topic of the Popish Plot.

3.5 Variations and future research.

Finding events and tracing change over time are just two possible extended applications. Once intentions are decomposed into semantic simples, they can be recomposed along any axis of interest. Combined with metadata, vector-space models become powerful explanatory tools that could be applied to almost any question of interest in literary history. For example, social-network models can be used to identify clusters of people and documents, making it possible to trace the distribution of ideas through a complex

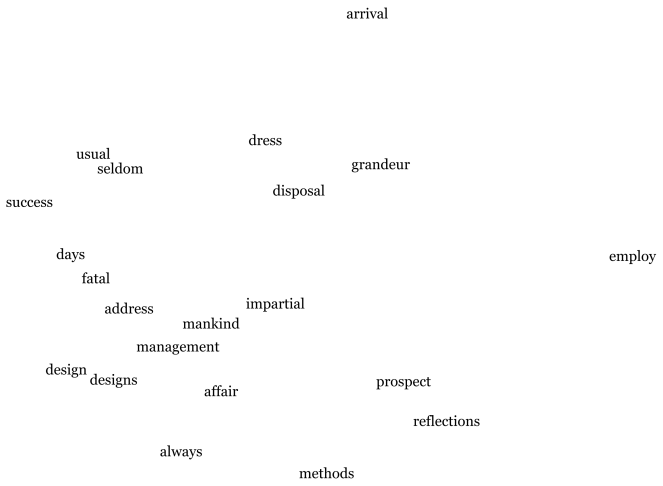

Figure 17: 25 terms most similar to "management" in EEBO, from 1640 to 1699. These words all exhibit a similar pattern of gradually increasing topicality, meaning that they tend to increase in frequency and to consolidate in meaning as time passes. Perhaps surprisingly, such terms also tend to exhibit strong thematic coherence, although not around any specific historical event. Instead, the terms tend to cohere around more general themes, in this case around a secular discourse of critical judgment.

social field over time.[7] Geospatial semantics are another very promising area

[7] In the field of digital humanities, studies that have attempted to combine social-network and textual analysis include David A. Smith, Ryan Cordell, and Abby Mullen, "Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers," *American Literary History* 27, 3 (2015): E1-E15; and Michael Gavin, "Historical Text Networks: The Sociology of Early English Criticism," *Eighteenth-Century Studies* 50, 1 (2016): 53-80.

of research.[8] Just as we measure the distribution of words across documents, authors, or years, we can also measure the distribution of concepts over territory: How does the semantic structure of "congregation" differ in Scotland from England? Historians more interested in specifically literary histories could use this model to discover genres and to characterize the variation of concepts across literary forms.[9] Critics focused on syntax and the history of grammar could begin with a different matrix structure.[10] Rather than organize their collection's vocabulary across columns of keywords, they might choose for values of K a list of canonical grammatical features (prepositional phrases, dependent clauses, etc.) and record the distribution of words across such feature spaces. Further, words themselves could be abandoned as the primary unit of analysis and replaced with other meaning-bearing objects, like phonemes or graphical forms. Nor does the theory depend on any particular notion of historical periodization. Although the examples provided above all treat the later seventeenth century as the primary frame, corpora

[8] See, for example, Ian N. Gregory and Andrew Hardie, "Visual GISting: Bringing together Corpus Linguistics and Geographical Information Systems," *Literary and Linguistic Computing* 26, 3 (2011): 297-314; and Michael Gavin and Eric Gidal, "Scotland's Poetics of Space: An Experiment in Geospatial Semantics," *Cultural Analytics* (November 2017). See also Werner Kuhn, "Geospatial Semantics: Why, of What, and How?" *Journal on Data Semantics III* (2005): 1-24; Angela Schwering, "Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey," *Transactions in GIS* 12, 1 (2008): 5-29; and Krzysztof Janowicz, et al, "Geospatial Semantics and Linked Spatiotemporal Data – Past, Present, and Future," *Semantic Web* (2012): 321-32.

[9] See for example essays published in a special issue on genre in *Cultural Analytics* (http://culturalanalytics.org/2017/04/noveltm-special-issue-on-genre/), which includes essays by Ted Underwood, Matthew Wilkens, Matthew Jockers and Gabi Kirilloff, Andrew Piper, and Matthew Erlin. Using an underlying semantic model of each document, as proposed here, would contribute to such analyses by examining the conceptual formations within genres, as well as differentiating among genres.

[10] I know of no study that performs analysis at this level; however, measures of grammatical patterns in documents are common. Just as the 3D tensor structure extends word-context matrices over many documents, a word-feature-document tensor would examine the distributions of words over grammatical patterns over documents, exposing which documents use which words in the most grammatically similar ways. For background on pattern analysis, see Susan Hunston, *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* (John Benjamins Publishing Co, 2000). Dan Shore has become an important advocate in literary studies for maintaining attention to grammatical forms even as the humanities computing is increasingly influenced by the intellectual tradition of information retrieval. See Daniel Shore, *Cyberformalism: Histories of Linguistic Forms in the Digital Archive* (Johns Hopkins University Press, 2018).

could be devised around any category of critical interest.

In any case, the key intellectual challenge is to intuit a trustworthy quantifiable proxy for some qualitative critical concept. Many critics draw hard lines separating literary topics they believe might be studied using data-driven methods from topics that couldn't possibly be quantified. All such boundary-drawing is grossly premature. The expressive capabilities of a mathematical theory of meaning have barely been tested, in this essay or elsewhere.

4 Discussion

Having ended the last section by gesturing toward a handful of ways the procedures of analysis might be extended and transformed, it's possible now to articulate what I take to be the theory's essential idea. Adapting basic techniques of information retrieval, I argue that meanings are built up from semantic simples – that is, any countable instance when one feature of a corpus appears near another. In the examples above, features were keywords and proximity was measured by a context window, but other features could be measured across different metric spaces. Regardless, meanings are approximated by aggregating across such simples. In most cases, the most relevant aggregate is provided by the corpus itself: any individual collocation selects from and combines features, but those features' meanings are usually best approximated by measurements made over the full data. In many cases, however, subsets can be identified where features aggregate in persistently deviant ways, such that their meaning is better approximated, not by the full data, but when restricted to the subset. In these cases, the meaning of the feature is attributed to the subset's characteristic metadatum (typically, a document identifier, a person's name, or, in the case of diachronic studies, a timestamp). Metadata become authorial when they are shown to bear responsibility for such meanings, and I define an intention as any meaning so attributed.

Throughout, I provided examples of each calculation. Readers might have been tempted at some points in the analyses above to express impatience when noticing results they recognize: "Of course Satan is important in *Paradise Lost*! We've always known that 'property' was a keyword in Locke's *Treatise*! There's nothing surprising about that!" However, I trust the sheer variety of topics presented helps to dilute this common knee-jerk reaction and

to direct attention back to the underlying argument. Any technique that provides so many results that feel so intuitive must be right about something. The question I have largely avoided so far is whether that something can reasonably be called "intention."

So, let's talk about intention. At this point, if you have read this far we can speak freely as if between friends. Our conversation might go like this:

You: OK, Mike. I'll admit this sounds cool and it looks like you can do a lot more with computational semantics than I realized was possible. But still, I don't see how you can call it a theory of "intention." It's just a bunch of keywords.

Me: That's right. What a mathematical theory of intention provides are principles for keyword extraction. But if I called the essay "Principles for Keyword Extraction" nobody would read it.

You: So you're just being cynical and misrepresenting your work?

Me: No! Lists of statistically significant terms are important, because they're the primary point of contact between the computer and the scholar, but they're only the tip of the iceberg. When you're identifying the keywords of a document, you're teasing out the words that are most important to its meaning. You're writing a paraphrase, in Cleanth Brooks's sense of that word. But not just any paraphrase. The underlying data is still there and it's all machine readable, which means the computer can scan across thousands of documents to represent individual texts and larger historical trends simultaneously.

You: This sounds more like "distant reading."

Me: Sort of, probably. But that's not a concept that guides my thinking. Forget the word "trend." My point is that ... look ... if you step back to think about how we've talked about intention since Wimsatt and Beardsley, we've been caught in an impasse. Intentions are assumed to be these inaccessible things external to texts, or on the other hand as meanings that just somehow manifest through discourse. The first view captures something right about the basic metaphysical problem, but the New Critics were wrong to blame the author. Once you start describing meanings, you're paraphrasing. The challenge is this: How do you differentiate what's proper to a text from what's proper to a whole textual system? And here's where I think the information theory of Claude Shannon and others might really help. Because they're precisely interested in studying language as a measurable system. Warren Weaver thought Shannon had opened the flood-

gates to a perfect theory that would eventually account for all meanings. At the same time, the late 40's, Cleanth Brooks wanted to talk about how poets push against language to create irreducible meanings in poems. And we finally have the theory and the resources to reconcile these views.

You: And that's why so many of the calculations involve comparing a text to a corpus?

Me: Exactly. It's funny, actually. In the debate over intention, people kept talking about computer poetry. Computer poems were perceived to be intentionless, or at least possibly so. Can you write a poem without intending to write a poem? This was Knapp's and Michaels's question.

You: Words written in sand?

Me: Computer sand, yes. But corpus-based studies present the mirror image of that problem. Instead of automatic poems, they're automatic criticism. Models make possible a new form of intentionless paraphrase. The computer doesn't know it's paraphrasing Locke; it just does it. But can you paraphrase an intention without intending to do so? This is what people get stuck on when they argue about distant reading, because they assume the answer is "no" without even formulating the question to themselves, and certainly without thinking through the underlying mathematical questions, so it all gets tangled up in politics.

You: OK, OK, enough! Walk me through what you're trying to say...

Considered as mental states, the question of how intentions manifest outwardly as action or discourse has been a matter of intense philosophical, critical, and editorial debate.[11] Intentions are not observable in and of themselves; nor are meanings. However, we often want to attribute meanings to intentions. Situations commonly arise when language does not seem to convey itself adequately and so must be supplemented by additional commentary. The author made a mistake, perhaps, or an editor intervened, or the work is so complicated it requires special explication.[12] Further, some meanings might reflect generic conventions or the influence of ideology, rather

[11] For an overview of intention as a problem in the philosophy of human action, see Jonathan Kramnick, *Agents and Objects from Hobbes to Richardson* (Stanford University Press, 2010), who prominently cites G. E. M. Anscombe's *Intention* (Cornell University Press, 1957) and usefully surveys recent debates in analytical philosophy and their prehistory in seventeenth- and eighteenth-century literature.

[12] For intention as a central problem in editorial theory, see G. Thomas Tanselle, *A Rationale of Textual Criticism* (University of Pennsylvania Press, 1989).

than the conscious motives of the author, and these kinds of meaning must be teased out as well. Meaning in such cases might be assigned to other attributes of the text like its genre or historical period, and in these cases meanings are described in the negative, as *unintended* (or at least not necessarily intended). The critical task of reading for intention is thus a three-step process that involves 1) choosing which attributes of a text matter most to one's critical goals, 2) producing a trustworthy paraphrase of that text, then 3) attributing the paraphrase to the specified attributes.

As might or might not be obvious, I'm now speaking at a level of generality that includes and subsumes the mathematical theory while also being roomy enough to accommodate any critical ontology, even those premised on the exclusion or disavowal of intention in the more conventional sense. A key feature of computational literary theory is its ontological pluralism: the decision of what counts as a text and what counts as an attribute is left open.[13] Nor is there any presupposition about which attributes will be considered internal or external to the text. Sometimes you might emphasize the fact that a book was written by a specific person; other times you might want to focus on something else, like ideologies or material objects or formal tropes. Under one ontology, a couplet might manifest the ideas Wordsworth had in mind; under another it might manifest a formal conceit, like paradox. The couplet can be read for intention either way. Although "author" is a conventional bibliographic data field, semantic responsibility can be, and often is, assigned to other attributes, and so we should begin from a more flexible definition of authorship less bound up in personhood and subjectivity and their metaphysical entailments. Authorship can be construed broadly as the condition when any attribute of a text is assigned responsibility for that text's meaning.[14] Thus, intentions are by this definition "authorial"

[13] Sometimes, as here, this openness is deliberate. But like all other fields digital humanities has its share of polemicists and armchair metaphysicians. More often, this openness is an accidental byproduct of computational modeling as a practice, which necessarily involves experimenting with data formats, algorithms, and measurements that, taken all together, could never be reconciled in their various operations to a single, rigorously conceived ontological framework. The line between reality and conceptuality is drawn anew in every for loop!

[14] By describing authorship in terms of responsibility, I mean to invoke Erving Goffman, who defines the "author" as the subject who bears responsibility for a statement, even if he or she may not be the speaker, as when, for example, governments or corporations release statements through spokespeople. However, I also mention Goffman here because I believe that many of the metaphysical debates that concerned critics of the twentieth

even when assigned to nonhuman objects. As a critical trope, to read for intention is not to read for mental states but to declare a fit between a text, a paraphrase, and an attribute.

4.1 Relation to New Criticism.

This account of intention differs widely, in both spirit and execution, from what's offered in previous literary theory. However, it takes on board as guiding principles the metaphysical objections raised by the New Critics and others, so I should say a few words about how my thinking borrows from theirs (without attempting to disentangle their claims from my own admittedly idiosyncratic perspective on their claims).

The New Critics' primary accomplishment was to denaturalize reading by exposing a gap between texts and persons.[15] Traditionally, going back at least to Alexander Pope and David Hume, the golden rule of criticism was to keep in mind the author's intent. Only by setting aside our biases and judging a work by the author's own purpose could we judge fairly.[16] The problem, as Wimsatt and Beardsley pointed out much later, is that minds

century and that surround the practice of digital humanities today can, themselves, be understood in terms of code switching. A change in critical paradigm is like a change in footing: "A change in footing implies a change in the alignment we take up to ourselves and the others present as expressed in the way we manage the production or reception of an utterance." *Forms of Talk* (University of Pennsylvania Press, 1981), 128.

[15] Walter Ong made a similar point, although under his framework this gap is understood in terms of writing versus speech. See his *Orality and Literacy: The Technologizing of the Word* (London: Methuen, 1982). The disconnect between author and text is framed very differently in the field of book history, where the material artifact appears as an intervening explanans, introducing multiple intention-bearing subjects – printers, booksellers, scribal copyists, etc. – who complicate and fragment authorship. I refer here specifically to D. F. McKenzie's essay, "The Book as an Expressive Form," in *Bibliography and the Sociology of Texts* (British Library, 1986), 1-21. For a discussion of the interpretive instabilities that arise, especially in digital settings, see Jerome McGann, *The Textual Condition* (Princeton University Press 1991) and *Radiant Textuality: Literary Studies after the World Wide Web* (Palgrave MacMillan, 2001).

[16] In this sense, criticism invoked the negative ideal of depersonalized reason, characteristic of the public sphere. In "Of the Standard of Taste," Hume explains: "When any work is addressed to the public, though I should have a friendship or enmity with the author, I must depart from this situation; and considering myself as a man in general, forget, if possible, my individual being and my peculiar circumstances." Cited in Michael Gavin, *The Invention of English Criticism, 1650-1760* (Cambridge University Press, 2015), 15.

and poems just seem like fundamentally different things.[17] Authors' intentions don't exist on their own in a way that can be accounted for, really; they have to be recreated in the critical act.[18] Intentions have to be stitched together from the fabric of the very arguments that elevate them as legitimating authorities. Authors aren't really sources of poems; they're conceits of biographically themed essays.[19] Cleanth Brooks took this idea to its next logical step. While historians were building simulations of dead authors' minds, he argued, they were also creating false replicas of the poems themselves. Brooks famously called this the "heresy of paraphrase": by restating what we think an author is trying to say, then seeking a historical explanation for our restatement, we substitute our own presumptively inadequate summary for the poem itself. Meaning becomes invested, not in the reality of the poem, but in our necessarily impoverished model of that reality.

[17] This is not exactly how Wimsatt and Beardsley described the problem, stuck as they were in the shop-talk of criticism. I should mention, too, that they were by no means the first to identify this disconnect. For a better and altogether more accurate account of this issue, see Joshua Gang, "Behaviorism and the Beginnings of Close Reading," *ELH* 78, 1 (2011): 1-25.

[18] That is to say, intentions are built up in critical discourse using what Wimsatt and Beardsley refer to as "external evidence." They explain that biographical "notes tend to seem to justify themselves as external indexes to the author's intention, yet they ought to be judged like any other parts of a composition (verbal arrangement special to a particular context), and when so judged their reality as parts of the poem, or their imaginative integration with the rest of the poem, may come into question" (484).

[19] Nowhere do I find this idea expressed in quite this way, but it's what I understand to be Michel Foucault's contribution to this large and multifaceted field of debate. By identifying the function of authorship as a concept, and the author as a trope, Foucault completes the denaturalization that had begun decades before. See "What is an Author?" in *Aesthetics, Method, and Epistemology*, ed. James D. Faubion (The New Press, 1998), 205-22. Roland Barthes's declaration of the "The Death of the Author" also played an important role, at least in organizing the disposition of literary commentary: the ostensibly cold, scientistic pose of the New Critics could be redescribed as intellectual playfulness, in the broadest and best sense of that phrase. In the author's place, Barthes elevates a different abstraction: the reader. He describes this entity in terms similar to those I'd use to describe a vector-space semantic model: "The reader is the space on which all the quotations that make up a writing are inscribed without any of them being lost; a text's unity lies not in its origin but in its destination. Yet this destination cannot any longer be personal: the reader is without history, biography, psychology; he is simply that someone who holds together in a single field all traces by which the written text is constituted" *Image Music Text*, trans. Stephen Heath (Fontana Press, 1977), 148. In my view, the biggest difference between distributional semantics and poststructuralist theory isn't quantification or method, but the former's willingness to dispense with personifications of this kind.

Here, then, was the impasse: intentions that rested in the author's mind were alienated from the text and could be represented only through reductive paraphrase, but it was difficult to imagine any alternative. How could anything like meaning emerge from anything like a poem, all by its lonesome?

4.2 Twentieth-century computer poetry.

Coincidentally, the decades during which critics and philosophers of language grappled with this impasse were also the decades when computers entered the lexicon. Computerized, mechanical brains captured many imaginations but had not yet intruded into many daily lives. Like poststructuralist theory, computers were inhuman, cold, unsympathetic, and animated by a vaguely sinister, unknowable intelligence. And, it so happened, one small but important strand of argument over the metaphysics of meaning hinged on, of all things, computers, and on the fantastic possibility of artificial genius.

Monroe Beardsley was the first to introduce computer poems into the debate over intention. Writing in 1970, he cites a snippet of computer-generated verse that had been published a few years before in the *New York Times Magazine*:

> While life reached evilly through empty faces
> While space flowed slowly o'er idle bodies
> And stars flowed evilly on vast men
> No passion smiled.[20]

Beardsley uses this poem to demonstrate his premise that "textual meaning is not identical to authorial meaning."[21] A decade later, P. D. Juhl made

[20] Monroe Beardsley, "The Authority of the Text," in *Intention and Interpretation*, ed. Gary Iseminger (Philadelphia: Temple University Press, 1992), 26.

[21] Beardsley was responding specifically to E. D. Hirsch, Jr. who defines meaning as authorially intended meaning. Hirsch explains, "Almost any word sequence can, under the conventions of language, legitimately represent more than one complex of meaning. A word sequence means nothing until somebody either means something by it or understands something from it. There is no magic land of meaning outside human consciousness." *Validity in Interpretation* (Yale University Press, 1970), 4. "Meaning," Hirsch insists, "is an affair of consciousness" (37). He refines this definition later during a discussion of Edmund Husserl's Logische Untersuchungen: "Verbal meaning is, by definition, that aspect of a speaker's 'intention which, under linguistic conventions, may be shared by others" (218, italics original).

the same point.[22] Because there exists no author of an automated poem, but because its words are nonetheless interpretable, they argue, meaning at some level must exist independently of any author. As Beardsley put it, such a poem "has meaning, but nothing was meant by anyone."[23] Others refused to concede that such metaphysical niceties should guide interpretation. E. D. Hirsch, John Searle, Jerrold Levinson, Steven Knapp, and Walter Benn Michaels all defended an intention-centric view of meaning, although they disagreed among each other about the precise details.[24] Like Juhl, Searle and Hirsch were willing to stipulate that words could have unintended significance, but for them such meanings were a limit case outside their main concerns (criticism, in Juhl's and Hirsch's case, and the analysis of speech acts, in Searle's). Knapp and Michaels were famously dogmatic on this point.[25] For them, the notion of meaning-not-meant was as obviously self-contradictory as a four-sided triangle. The whole idea of unintended meaning, they insisted, rested on a false distinction: meanings and intentions are actually one and the same thing. Words written in sand or lines produced by a computer aren't examples of intentionless discourse; they're just systems manipulating objects that happen to look like words. To be real words, they'd have to be used by a speaker to convey thoughts. Randomly generated sentences, by contrast, are merely "spewed out by computer programs."[26] At best, we might say that the programmer has intentions, which the computer successfully or unsuccessfully realizes, but the poem itself merely reflects possibilities encoded in the system.

Few critics at the time exhibited much curiosity about the actual process of randomly generating a poem.[27] This process can be inferred from

[22] P. D. Juhl, *Interpretation: An Essay in the Philosophy of Literary Criticism*, (Princeton University Press, 1980), 84-85.

[23] Beardsley, "Authority of the Text," 26.

[24] For John Searle's contribution to this debate, see "Literary Theory and Its Discontents," *New Literary History* 25, 3 (Summer, 1994): 637-67; and "Reiterating the Differences: A Reply to Jacques Derrida." *Glyph* 2 (1977): 199–208.

[25] Steven Knapp and Walter Benn Michaels, "The Impossibility of Intentionless Meaning," in *Intention and Interpretation*, ed. Gary Iseminger (Philadelphia: Temple University Press, 1992), 51-64.

[26] This particular phrase comes from Jerrold Levinson's summary of the debate in "Intention and Interpretation: A Last Look," in *Intention and Interpretation*, ed. Gary Iseminger (Philadelphia: Temple University Press, 1992), 223.

[27] For discussions of early computer poems, see C. T. Funkhouser, *Prehistoric Digital Poetry: An Archeology of Forms, 1959-1995* (The University of Alabama Press, 2007); C.

the example cited above. Each line contains seven words: conjunction, noun (singular or plural), past-tense verb, adverb, preposition, adjective, and plural noun. The last line of each stanza begins with an article followed by a noun, a past tense verb, and a period. Given this syntagmatic structure, a loop is executed that runs the first procedure three times and the last once, filling in the slots with words chosen randomly from previously compiled lists for each category.[28] Commenters at the time were preoccupied with characterizing the role of human intention in such processes.[29] Indeed, this concern still surfaces, though nowadays critics like N. Katherine Hayles tend to find a cognitive partnership between human and machine, rather than a conflict between them.[30] To at least some computer poets, such automatism

T. Funkhouser, "First Generation Poetry Generators," in *Mainframe Experimentalism: Early Computing and the Foundations of the Digital Arts*, ed. Hannah B. Higgins and Douglas Kahn (University of California Press, 2012), 243-65; and John Cayley, "Time Code Language: New Media Poetics and Programmed Signification," in *New Media Poetics: Contexts, Technotexts, and Theories*, ed. Adalaide Morris and Thomas Swiss (MIT Press, 2006), 307-33.

[28] Not all computer poems are quite this simplistic. The famous Cybernetic Serendipity art exhibit included computer poems, and presenters were very explicit about the underlying procedures. For explanations of several works and a good sense of how the form was imagined in these early years, see Jasia Reichardt, ed. *Cybernetic Serendipity: the Computer and the Arts* (Frederick A. Praeger, 1969). This collection included works by Marc Adrian, Robin McKinnon Wood and Margaret Masterman, Nanni Balestrini, Alison Knowles and James Tenney, Edwin Morgan, Jean Baudot, and E. Mendoza.

[29] For example, Margaret Masterman writes, "In short, the ultimate creative act, for the computer poet, lies in writing the thesaurus and in filling in the semantic derivatives. Thus the human creative process is pushed on stage further back; and the poet composes a poetic system, which can produce for him any number of poems formed from a given frame, between which he then chooses." "Computerized Haiku," in *Cybernetics, Art, and Ideas*, ed. Jasia Reichardt (New York Graphic Society, 1971), 183. Similarly, Dick Higgins remarks, "We do not communicate with a computer by telling it verbally what to do. Instead we provide a structure on the basis of which it processes whatever data we provide." "Computers for the Arts (May 1968)," in *Mainframe Experimentalism: Early Computing and the Foundations of the Digital Arts*, ed. Hannah B. Higgins and Douglas Kahn (University of California Press, 2012), 292.

[30] See N. Katherine Hayles, *Electronic Literature: New Horizons for the Literary* (University of Notre Dame Press, 2008); Lori Emerson, *Reading Writing Interfaces: From the Digital to the Bookbound* (University of Minnesota Press, 2014); Jessica Pressman, *Digital Modernism: Making it New in New Media* (Oxford University Press, 2014); and Nick Montfort, "Conceptual Computing and Digital Writing," forthcoming in *Postscript: Writing After Conceptual Art*, edited by Andrea Andersson http://hdl.handle.net/1721.1/92876. Mordecai-Mark Mac Low argues, "The removal of the ego from the writing of the work

promised to upend modernist aesthetics, exposing as mere formulae patterns of language believed to be the essence of poetic genius. In his preface to the collection, *Computer Poetry* (1973), R. W. Bailey describes the form in grandiose terms:

> Computer poetry is warfare carried out by other means, a warfare against conventionality and language that has become automatized. Strange as it seems, our finite state automata have become the poet's allies in this struggle, the long historical battle by which mankind pries into the surface of language to reveal its latent mysteries.[31]

I do not mean to suggest that Bailey was correct to assess computer poems as aesthetically revolutionary.[32] The Dadaists had experimented with automatic writing decades before, and their parlor game, called Exquisite Corpse, was based on similar principles: participants would compose short stories, taking turns adding words step-by-step, producing stilted narratives with sharp twists and comical, uncanny juxtapositions.[33] Their work was continued in various directions over the next several decades by the OuLiPo group, which was devoted to experimenting with "potential literature."[34] The most important part of the basic process wasn't randomization but the

---

ultimately was replaced by the presence of the ego in the choice of source and method. However, as with all good experiments, an unexpected result was also obtained: recognizable poetry without a poet writing it, whatever influences may have been exercised behind the curtain." "The Role of the Machine in the Experiment in Egoless Poetry," in *Mainframe Experimentalism: Early Computing and the Foundations of the Digital Arts*, ed. Hannah B. Higgins and Douglas Kahn (University of California Press, 2012), 303.

[31] Robert W. Bailey, ed., *Computer Poems* (Potagannissing Press, 1973).

[32] Bailey himself seems to have been ambivalent on this point, writing in a contemporaneous essay that "Despite such temptations as these to predict a glorious future for computer-assisted poetry, I believe that the technique has no interesting potential." "Computer-Assisted Poetry," in *Computers in the Humanities*, ed. J. L. Mitchell (Edinburgh University Press, 1974), 293.

[33] See André Breton, Paul Eluard, Philippe Soupault, *The Automatic Message, The Magnetic Fields, The Immaculate Conception*, trans. David Gascoyne, Antony Melville & Jon Graham (Atlas Press, 1990). For more on the game and its relation to Surrealist art, see Kanta Kochhar-Lindgren, Davis Schneiderman, and Tom Denlinger, eds. *The Exquisite Corpse: Chance and Collaboration in Surrealism's Parlor Game* (University of Nebraska Press, 2009). Although he doesn't mention its Dadaist origins, John R. Pierce compares narratives generated this way to the Markov chain process in *An Introduction to Information Theory: Symbols, Signals and Noise* (Dover, 1980 [1961]), 110-11.

[34] A detailed and sympathetic history of the *Ouvroir de littérature potentielle* (OuLiPo)

formulaic recombination of paradigmatically similar words over some syntagmatic structure, whether programmed in a computer or intuited by human participants.

Because most commentary was stuck on an anthropomorphic concept of intention, most critics failed to appreciate how computer-generated verse implied a new (or, at least, different) concept of meaning. This concept emphasized, not communication between persons whose intentions might or might not guide interpretation, but a structured process of selection that creates meaning through the dynamic interplay between systematic possibilities and individual instances. The automatic generation of *parole* placed its relation to *langue* under new light.

4.3 Claude Shannon's theory of communication.

Bailey's use of the phrase "finite state automata" harks back to the mathematical theory of Claude Shannon, and, in fact, the procedures developed for writing computer poems were informed by Shannon's work. Based on his wartime experience in cryptography and, after the war, among the engineers at Bell Labs, Shannon understood language as a finite-state machine that produced sequences of symbols.[35] The telegraph was his prime example: the machine produces dots, dashes, and spaces, each of which represents a discrete state, and its total operation can be described as a series of shifts from one state to the next. The task of information theory is to identify statistical properties of language that allow these shifts to be encoded as efficiently as possible. In telegraphy, such optimization is built into Morse code, which uses short sequences for common letters like "a" or "e" but longer sequences for less frequent letters like "z." Considered at a higher order, this approach finds words and phrases that occur redundantly and so can be stored or transmitted more easily. Whereas in computer poetry the transition from one state to the next was programmed through a given set of limited parameters, Shannon showed how statistical knowledge of language could be used to produce simulated discourse through a Markov chain that mimics human discourse step-by-step. The result sounds much like a computer poem or a

is provided in Daniel Levin Becker, *Many Subtle Channels: In Praise of Potential Literature* (Harvard University Press, 2012). Stephen Ramsay directly connects their work to "algorithmic criticism" in *Reading Machines: Toward an Algorithmic Criticism* (University of Illinois Press, 2011), chapter 2.

[35] See Shannon, *The Mathematical Theory of Communication*, 36-64.

game of Exquisite Corpse:

> THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.[36]

Something like meaningful discourse can be achieved through a stochastic process constrained by rules that limit what choices are available. These rules – which are not so much rules as measurable redundancies – enable this approximation by reducing the field of possible word combinations. The total amount of information in a given message is a function of the uncertainty it resolves; that is, of the total system of possibilities from which it draws. A communication device like a light switch conveys very little information, because it selects from only two possibilities, "on" or "off." A television signal, by contrast, selects from many thousands of possible states at every moment. Written language sits somewhere in between. To understand how best to encode such messages does not require interpreting the meaning of any individual transmission. Instead, efficiency requires a detailed representation of the information source that identifies and differentiates among likeliest scenarios.[37]

Shannon was neither linguist nor philosopher nor critic, so he was free simply to set the question of meaning aside. Others, though, could not resist the temptation to find in his theory of information a theory of meaning, or of "semantic information," as some would call it. If information can be measured by the uncertainty it resolves, then we should be able to measure how informative, how meaningful, a statement is. Philosophers Yehoshua Bar-Hillel and Rudolph Carnap explored this idea at length, arguing that the meaningfulness of a statement was the complement of its likelihood: the semantic content of a statement equals one minus the general probability of it being true.[38] Thus, a statement like "the apple is red" has very little semantic value, because apples are very likely to be red, but a statement

[36] Ibid., 44.

[37] Shannon explains, "The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel, by the use of proper encoding of the information" (39).

[38] Yehoshua Bar-Hillel and Rudolf Carnap, "Semantic Information," *The British Journal for the Philosophy of Science* 4, 14 (1953): 147-57.

like "this ticket won the lottery" is very informative indeed. One curious consequence of this theory is that paradoxes become the most informative kinds of statements: declarations of the impossible like "Alice is not Alice" entail the maximum semantic content. In this way, philosophers analogized Shannon's technique for measuring the relative frequencies of symbols to a different concept that gauged the relative likelihood of real-life situations and the statements that described them.

This scheme survived for decades and was later expanded by Fred Dretske, but it suffered from two major flaws.[39] The first and probably most obvious is that, although the theory is mathematical, its objects are not quantifiable. There's no way anybody has been able to think of to specify how surprising one statement is in comparison to another. Too much depends on the situation of utterance and too many factors are at play, so the elaborate mathematical apparatus Bar-Hillel and Carnap worked out could never be realized as an empirical method. Second and more important is that "informativeness" has little relation to "meaning" in any sense that matters to most people. Rarely do they ask *how much* something means; they ask *what* something means. For this reason, whereas Shannon's move to quantify the amount of information made sense from an engineering perspective, there was no perspective outside philosophy from which the amount of meaning seemed meaningful.

The problem faced by all these writers (linguists, literary theorists, automatic poets, and even Shannon, although his answer was just to acknowledge and disregard it) was how to talk sensibly about the operation of language as a system while accounting for particular cases without stepping outside that system, either by invoking phantom subjects – authors, readers, selves and somebodies all – to whom meaning can safely be attributed, or by surrounding statements with phantom objects – apples, lottery tickets, and Alices – to which meaning can safely be said to refer. Semantics were presumed

[39] Fred I. Dretske, *Knowledge and the Flow of Information* (MIT Press, 1981). For a more recent extension of this line of inquiry, see Luciano Floridi, *The Philosophy of Information* (Oxford University Press, 2011); Pieter Adriaans, "A Critical Analysis of Floridi's Theory of Semantic Information," *Knowledge, Technology & Policy* 23, 10 (June 2010): 41–56; and Pieter Adriaans, "Between Order and Chaos: The Quest for Meaningful Information," *Theory of Computing Systems* 45 (2009): 650–74. A closely related strand of argument advanced by Dan Sperber and Deirdre Wilson combines Paul Grice's theory of "implicature" with Shannon's communication model. See *Grice, Studies in the Way of Words* (Harvard, 1989) and Sperber and Wilson, *Relevance: Communication and Cognition* (Harvard, 1986).

to involve a relation between language and that phantom world. This assumption was so powerful that it guided virtually all discussion, even by New Critics, computer poets, linguists, deconstructionists, philosophers of history, and others who hoped to upend it. But because the questions this assumption raised were fundamentally metaphysical, they were difficult to resolve. Impossible, probably.

In the end, or in what my very partial history takes as its end, the solution came not by thinking the problem through, but by approaching it as if by happy accident from an altogether different angle, by treating the problem of meaning as a problem of search. You can't appeal to apples or Alices when designing a search engine. It doesn't matter what ontological commitments you bring to bear, because the entire imaginary apparatus of reality is irrelevant to the task at hand. All you have are the documents and the metadata. You start with huge, unmanageable collections – libraries full of books, crates full of papers and microfilm, computers and mainframes full of files – and you have to figure out how to find just the right ones. While computer poets were amazed and amused that something like meaning could be simulated by self-contained finite-state machines, librarians faced the mirror-opposite challenge: using machines to organize and navigate a field of already existing meaning. Rather than spew poems, they needed computers to spew bibliographies. And so my story here will end where the theory began, with information retrieval.

4.4 Information retrieval.

"Information retrieval" names a field of scholarship devoted to solving a disarmingly practical challenge: how to sort and search for documents. The advent of mechanical text processing in the 1950s sparked a huge monetary and intellectual investment in developing automated techniques for organizing library holdings. In journals like *American Documentation* (1950-1968) and in textbooks like Joseph Becker's *Information Storage and Retrieval* (1963), scholars, engineers, and technology enthusiasts brainstormed techniques for storing and cataloguing information.[40] I say "brainstormed" because many of their ideas were false starts, and the best ideas would not be realized until much later. Computers at this time were still large and expensive ma-

[40] Joseph Becker and Robert M. Hayes, *Information Storage and Retrieval: Tools, Elements, Theories* (John Wiley & Sons, 1963).

chines that manipulated data on punch cards or film. Memory was costly and inadequate.[41] Nonetheless, engineers like Hans Peter Luhn at IBM and computer scientists like Karen Spärck Jones and Gerard Salton intuited a family of mathematical concepts that would prove enormously influential in the decades that followed.[42]

Most important among these was the theory of semantic space, conventionally attributed to Luhn and Salton, which adapted principles of linear algebra to the study of language.[43] Here, the difference from Shannon is instructive. In Shannon's framework, the precise content of any individual message was not pertinent; instead, he needed to know how language worked in general. The most important quantitative element was a sequence of average word frequencies. Which words are used most frequently overall? To know how much information is in a given document, you compare the distribution of words in that document to the average distribution over the whole. It's just a single line of numbers (the message) compared to another single line (the information source). Even when the vector of probabilities was extended to second- or third-order conditional probabilities, there was nothing in the theory designed to compare documents against each other; there was no conceptual mechanism to transform the space of words into the space of anything else. By contrast, Luhn and others experimented with a data format, the term-document matrix, designed for this express purpose. Under

[41] For this reason, discussion in the 1950s and 1960s centered around mechanical considerations, as librarians debated whether information retrieval was best performed by expensive general computers or by potentially cheaper special-purpose machines built for handling card catalogues. See Becker and Hayes, *Information Storage and Retrieval*, chapter 7. By the 1980s, general purpose computers were less expensive, so Salton's work could focus primarily on the underlying mathematical structures.

[42] The development of information retrieval as a field can be traced through three textbooks: Becker's and Hayes's *Information Storage and Retrieval*; Lauren B. Doyle, *Information Retrieval and Processing* (John Wiley & Sons, 1975); and Gerard Salton, *Automatic Information Organization and Retrieval* (McGraw-Hill, 1983).

[43] The earliest representation of semantic space can be found in H. P. Luhn, "A New Method of Recording and Searching Information," *American Documentation* 4, 1 (1953): 14-16. This theory was further elaborated in Gerard Salton, et al., "A Vector Space Model for Automatic Indexing," *Communications of the ACM* 18, 11 (1975): 613-20. In addition to advances in computer processing, the development of the field depended on changes in the typical college curriculum. In 1963, when describing the work of Luhn and others, Becker and Hayes noted their use of "linear algebra," a phrase they placed in scare quotes (343). By the 1980s, linear algebra was an established theory for describing complex dynamic structures.

their theory, words are not represented simply as a scalar frequency but as a vector of frequencies over all documents in the collection. In the jargon of linear algebra, each matrix represents the transformation from one vector space (of words) into another (of documents).[44] Rather than ask how often words are used, this theory asks where they're used and in what relation. Thus, it becomes possible to distinguish the lexical content of documents and to identify which terms are used most distinctively where.

Under this theory, conventional notions of word meaning are displaced by a simpler notion of semantic similarity, conceived as proximity in vector space. Matrices can be thought of geometrically as multidimensional coordinate systems where words and documents are suspended in mutual relation. Similar words and similar documents appear near each other in this abstract space. Meaning as such resides not in the words themselves, nor in the documents, but in the emergent structure of their collocation. This idea, which Magnus Sahlgren describes as the "geometric metaphor of meaning," has become one of information science's organizing conceits.[45] In principle, it suggests that meaning does not reside in any particular use of language at any time but in the brute, mindless facts of lexical relation.

However, it's worth returning here to computer poetry which, I have argued, presented a similar limit case for literary critics' theories of meaning. Monroe Beardsley, recall, referred to computer poems as having a kind of depersonalized meaning not meant by anyone. Critics debated whether meaning could exist in this way, or whether intentions should be attributed to the computer programmer. We might ask, analogously, where intention resides in the act of search. Is there intention in the results of a library catalogue keyword search? When the items are listed by "relevance," to what are they relevant? Like a computer poem, a search engine returns automatically gen-

[44] In the mathematical theory of intention presented above, this situation is slightly more complex in that it operates over a third-order tensor. (See Figure 1.) In which case, the tensor represents a transformation of a transformation: first wordspace is transformed in the space of keywords, then that matrix is distributed again over the space of documents. Though convenient for descriptive analysis, this format is computationally intensive and not very practical. Information retrieval systems usually separate these processes out, using a word-keyword matrix to represent word-level semantics (for identifying synonyms or disambiguating search words) and another term-document matrix to represent library holdings. See Salton, *Automatic Information Organization and Retrieval*, 122-28.

[45] Magnus Sahlgren, *The Word-Space Model: Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations Between Words in High-Dimensional Vector Spaces*. PhD Thesis. Stockholm University, 2006.

erated language, selected from a predetermined set of possible words. Rather than mimic syntax, search results instead list document titles, selected because of their proximity in vector space. This proximity is measured against a search query entered by the user. The semantic model mediates between the input of the query and the output of the results. If there is meaning or intention in the system, it is revealed through this process of selection whereby documents are matched to corresponding keywords. Under this theory, that match is what meaning is. If you're searching for documents other people wrote, that's what intention is.

Information retrieval begins with a list of words and returns a list of documents that best reflect what the searcher meant to find. *The mathematical theory of intention reverses this process, working from a document and returning a list of words that best reflect what the author meant to say.* This reversal characterizes much work in humanities computing and most of what is called "distant reading."

When humanists first encounter the ideas of distributional semantics they often recoil: it can't be right, they advise, to represent a document as a mere bag of words or to confuse a word with a mere list of its collocates. Haven't you read your Chomsky? Such complaints reflect an impoverished and inadequate understanding of linear algebra as a mode of thought. The key affordance of the theory of finite-dimensional vector space is its ability to toggle dialectically between the "row perspective" and the "column perspective," as mathematician Gilbert Strang puts it, and thereby to show how the observations we make and the variables through which we define those observations are mutually constitutive. Linear algebra is a compact and well-conceived theory for describing the distribution of difference in many systems. Applied to written language, vector spaces represent a field of words and documents that build themselves into reciprocal being.

Further, because documents are attached to metadata – not just to authors but also to times and places and so on – the theory of semantic space becomes simultaneously a theory of historical social space. This redescription of information retrieval's purpose folds questions of subjectivity and historicity into its complex analytic structures. But what happens to the theory when it becomes filled with people? And what happens to people so constituted? The task that now confronts computational literary theory, and which this essay has addressed in a very narrow way, is to reconcile the theoretically rigorous description of textual objects, as articulated in information science, with the theoretically rigorous description of historical significance,

as articulated in the humanities. If you notice that I have failed to accomplish this task, you'll have begun to appreciate the sheer magnitude of the intellectual challenge it presents.

5 Epilogue

When I speak or write about quantitative methods, I often feel like Severus Snape delivering the opening lecture of Potion Making. Hoping to spark the imaginations of a roomful of good Gryffindors as they struggle against a mix of trepidation and boredom, I hear myself promising them power. I admit, computer-based techniques can neither bottle fame nor brew glory – and they certainly put no stopper in death – but nonetheless there are lots of cool things you can do with them. Empowered cool being, as Alan Liu might argue and so unlike Snape himself, the peculiar aesthetic of our neoliberal Slytherin.

The desire for interpretive power is often mythologized in masculine fantasy. Cleanth Brooks's final paragraph in *The Well Wrought Urn* (1947) returns to Pope's *Rape of the Lock*, focusing on the ability of that poem to capture and hypostatize femininity:

> In one sense, Pope's treatment of Belinda raises all the characteristic problems of poetry. For Pope, in dealing with his "goddess," must face the claims of naturalism and of common sense which would deny divinity to her [and] transcend the conventional and polite attributions of divinity which would be made to her as an acknowledged belle. ... The poetry must be wrested from the context: Belinda's lock, which is what the rude young man wants and which Belinda rather prudishly defends and which the naturalist asserts is only animal and which displays in its curled care the style of a particular era of history, must be given a place of permanence among the stars.[46]

In a reductive understanding of Pope's burlesque, Belinda's divinity is either silly (when measured against common sense) or utterly conventional. To read Pope's poem as the transparent expression of a paraphrase is to fit it to a bad model. There's a flow of cultural expectation about what words

[46] Cleanth Brooks, *The Well Wrought Urn: Studies in the Structure of Poetry* (Harcourt Brace, 1947), 214.

mean, and it'd be all too easy to let Pope's words be caught in that flow. To understand the conceptual work his poem performs demands recognizing how he transforms Belinda's lock into a semantically independent object. This perspective of critical authority is what New Criticism offers. Whatever the theoretical or metaphysical arguments on its behalf, what close reading promises in the end is the power to grasp how poems exert power on language and on reality.

Writing his introductory survey to Shannon's *Mathematical Theory of Communication* (1949), Warren Weaver ends with a very different parable:

> An engineering communication theory is just like a very proper and discreet girl accepting your telegram. She pays no attention to the meaning, whether it be sad, or joyous, or embarrassing. But she must be prepared to deal with all that come to her desk. ... Language must be designed (or developed) with a view to the totality of things that man may wish to say; but not being able to accomplish everything, it too should do as well as possible as often as possible.[47]

Warren offers a fantasy of pure transparency. The girl accepting your telegram attaches herself to no meaning in particular because in her role she encompasses all meanings. As a communication channel, she contains and transmits the intentions of all men, "the totality of things that man may wish to say." She's a system of semantic possibility. Not a flawless system, surely, but doing the best she can.

How to reconcile these views without adopting their politics? When writing about computational criticism, it's tempting at times to slip into similar fantasies. In part this is simply emotional. The joys of learning a new skill or mastering a new field of theory bring with them all the joys of discovery and all the ideologies of self that discovery entails. One feels less like Snape and more like Jack the Pumpkin King having wandered into the land of Christmas: "What's this?! What's this?!" But Jack doesn't know to accept the authority of Santa Claus. As an allegory of interdisciplinary inquiry gone awry, *The Nightmare Before Christmas* stages a collision between knowledge systems that turns catastrophic because the masters of one field are unwilling to submit to the masters of another. And here's where the fantasy of empowerment fails. What's disconcerting about computational theory is that it

[47] Shannon and Weaver, *Mathematical Theory*, 27.

gives you power over your ideas only if you submit to others' authority. And why should Cleanth Brooks surrender his power? Why Shannon? Why us?

The reasons are many and, to me, too obvious to state clearly. I would rather be disciplined than discipline, and so I focus on the question of intentionality because it's precisely the kind of interdisciplinary topic over which scholars assert dogmatic control in their various domains. "I hate intentionality," a colleague of mine said just yesterday in our English Department lounge during a friendly debate over gender politics in Octavia Butler's fiction. It's a concept around which our commitment to humanistic explanation is tested and our ambivalences are exposed. Intention remains a soft spot in criticism. Readers will find different sections of this essay opaque, frivolous, wrongheaded, or all three.

Reader,
I feel
your pain.
I want you to feel
me feeling
your pain.
This is intentional.