Pamphlet 4
→ May 2012

Literary Lab

A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method

Ryan Heuser

Long Le-Khac

Pamphlets of the Stanford Literary Lab

ISSN 2164-1757 (online version)

Ryan Heuser

Long Le-Khac

A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method

Introduction

The nineteenth century in Britain saw tumultuous changes that reshaped the fabric of society and altered the course of modernization. It also saw the rise of the novel to the height of its cultural power as the most important literary form of the period. This paper reports on a long-term experiment in tracing such macroscopic changes in the novel during this crucial period. Specifically, we present findings on two interrelated transformations in novelistic language that reveal a systemic concretization in language and fundamental change in the social spaces of the novel. We show how these shifts have consequences for setting, characterization, and narration as well as implications for the responsiveness of the novel to the dramatic changes in British society.

This paper has a second strand as well. This project was simultaneously an experiment in developing quantitative and computational methods for tracing changes in literary language. We wanted to see how far quantifiable features such as word usage could be pushed toward the investigation of literary history. Could we leverage quantitative methods in ways that respect the nuance and complexity we value in the humanities? To this end, we present a second set of results, the techniques and methodological lessons gained in the course of designing and running this project.

This branch of the digital humanities, the macroscopic study of cultural history, is a field that is still constructing itself. The right methods and tools are not yet certain, which makes for the excitement and difficulty of the research. We found that such decisions about process cannot be made a priori, but emerge in the messy and non-linear process of working through the research, solving problems as they arise. From this comes the odd, narrative form of this paper, which aims to present the twists and turns of this process of literary and methodological insight. We have divided the paper into two major parts, the development of the methodology (Sections 1 through 3) and the story of our results (Sections 4 and 5). In actuality, these two processes occurred simultaneously; pursuing our literary-historical questions necessitated developing new methodologies. But for the sake of clarity, we present them as separate though intimately related strands.

1. How to Build a Semantic Field, or Learning to Define Objects in the Quantitative Study of Literature

The original impetus for this project came from Raymond Williams's classic study, *Culture and Society*, which studies historical semantics in a period of unprecedented change for Britain. We took up that study's premise that changes in discourse reveal broader historical and sociocultural changes. Of course, Williams's ambitious attempt to analyze an entire social discourse, astonishing as it is, lacked the tools and corpora now available to digital humanities scholars. We set out, then, to build on Williams's impulse by applying computational methods across a very large corpus to track deep changes in language and culture. A key promise of such methods is scale. Digital humanities work opens up the study of language, literature, and culture to a scale far larger than is accessible through traditional methods, even those of a scholar as widely read and deeply learned as Williams.

This promise though remains just that until methods in the quantitative study of culture become fleshed out, tested, and refined. In these early stages, consistent reflection and evaluation are imperative. Much rests on exactly how the methods are applied. With the Google Books project, the mass digitization of text from historical and contemporary archives, and the advancement of natural language processing, there has been a surge of interest in the data-driven study of culture (Borgman; P. Cohen; Manovich). About ten months into our project, quantitative historical semantics was given a boost in visibility from the introduction of Google's N-gram viewer.[1] The "buzz" only increased with the publication of Michel and Aiden's "Culturomics" study in Science in December 2010 and Dan Cohen's n-gram based study of the Victorian period.[2] This is exciting work with some tantalizing results thus far.

If Williams were here today, what would he think? Faced with n-grams and the possibility of studying millions of texts at a time... would he be tempted to look up that keyword, "culture"? And what would he find if he did?

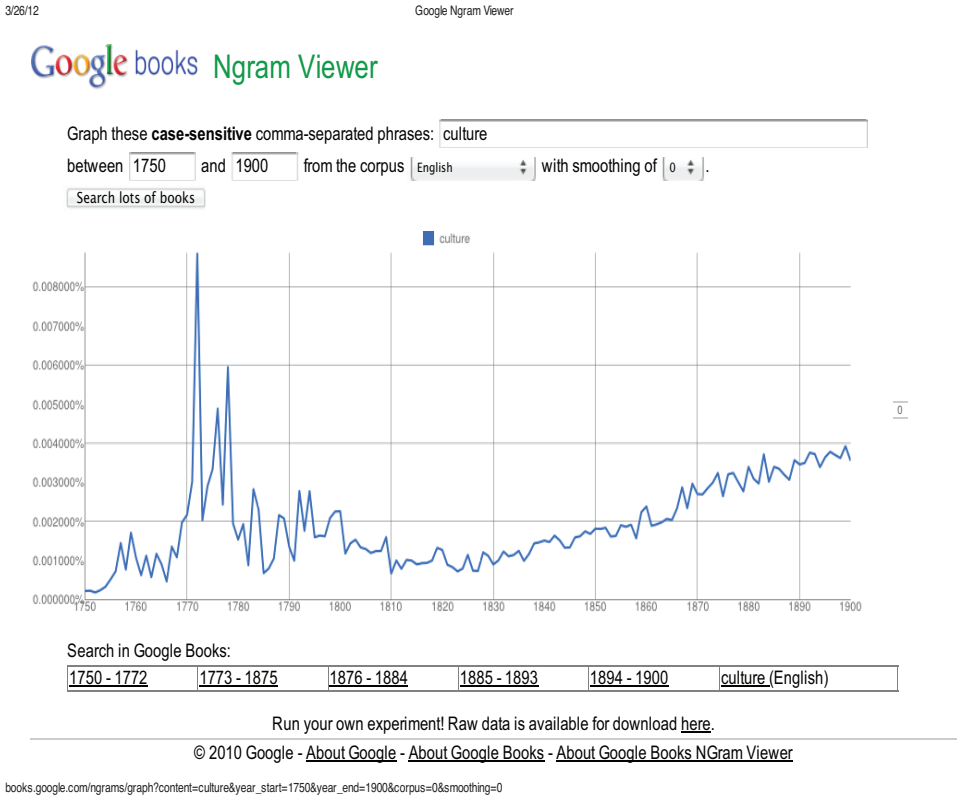

Figure 1: Plot of the term frequency behavior of "culture" in the Google Books corpus, 1750-1900. Source: *Google Books Ngram Viewer*. Google. Web. 1 May 2011.

When we explore word frequency behaviors—something computers readily crunch—as a window into cultural trends—something computers can't understand, the results, like this plot, can be simultaneously intriguing and frustrating (see **Figure 1**). As we look at this plot of the word "culture," there are many questions: What does it mean that the use

1 The N-gram Viewer is an online tool that allows one to trace the historical frequency of any word through the Google Books corpus. It can be found at http://books.google.com/ngrams/.

2 See Michel, et al. and D. Cohen.

of the word "culture" rose dramatically in the 1770s and once again in the 1790s? What can this tell us about changes in the idea of culture? Is this the idiosyncratic behavior of one word or does it reflect a more general trend? More broadly, what is the meaning of changes in word usage frequencies? What do we do with such data? With much current research drawing on word frequencies and other quantifiable aspects of culture, these are big questions. We can see now that the greatest challenge of developing digital humanities methods may not be how to cull data from humanistic objects, but how to analyze that data in meaningfully interpretable ways. To figure this out has been an overarching concern of our research over the past two years, and while we don't claim to have all the answers, we hope to show in this paper that the problem is not intractable.

We chose in our work to focus on the object of the semantic field. A semantic field can be defined as a group of words that share a specific semantic property; while not synonymous, they are used to talk about the same phenomenon (Crystal). If one promise of digital humanities is leveraging scale to move beyond the anecdotal, we wondered, how do we move beyond investigating single words or small groups of words to a more systemic investigation of linguistic changes? Given the semantic richness of language and the diffuseness of cultural trends, it's unlikely that such trends could be isolated by tracking the behavior of a few words. But tracking the frequency behaviors of semantic fields, much wider yet meaningfully related groups of words, had potential. They held out the promise of quantitative results that would more directly reflect changes in big ideas: cultural concepts, values, attitudes. Our gambit was to see what kind of literary history could be done with semantic fields.

1.1 Stage 1: "Seed" Words and Initial Problems

While promising in theory, the practice of building semantic fields soon revealed serious challenges. We based our initial fields on questions raised by prior criticism, but this criticism rarely provided lists of associated keywords. So we immediately ran into a basic problem: how to generate the words to include in a semantic field. For example, we were interested in the literary history of rural and urban spaces. But after quickly exhausting the rural and urban words mentioned in several studies, we turned, awkwardly, to thesauruses and sheer invention to add more.

This was our problem of generation: what practices, principles, and criteria should be used when including words in a semantic field? But after analyzing the frequency trends of some initial fields and their constituent words, we soon realized there was another problem. The frequency behaviors of individual words often diverged wildly. How could we describe the collective behavior of these groups when their behavior was far from collective? For example, we had included the word "country" in our rural field, but, while having the greatest frequency, it trended differently from every other word.

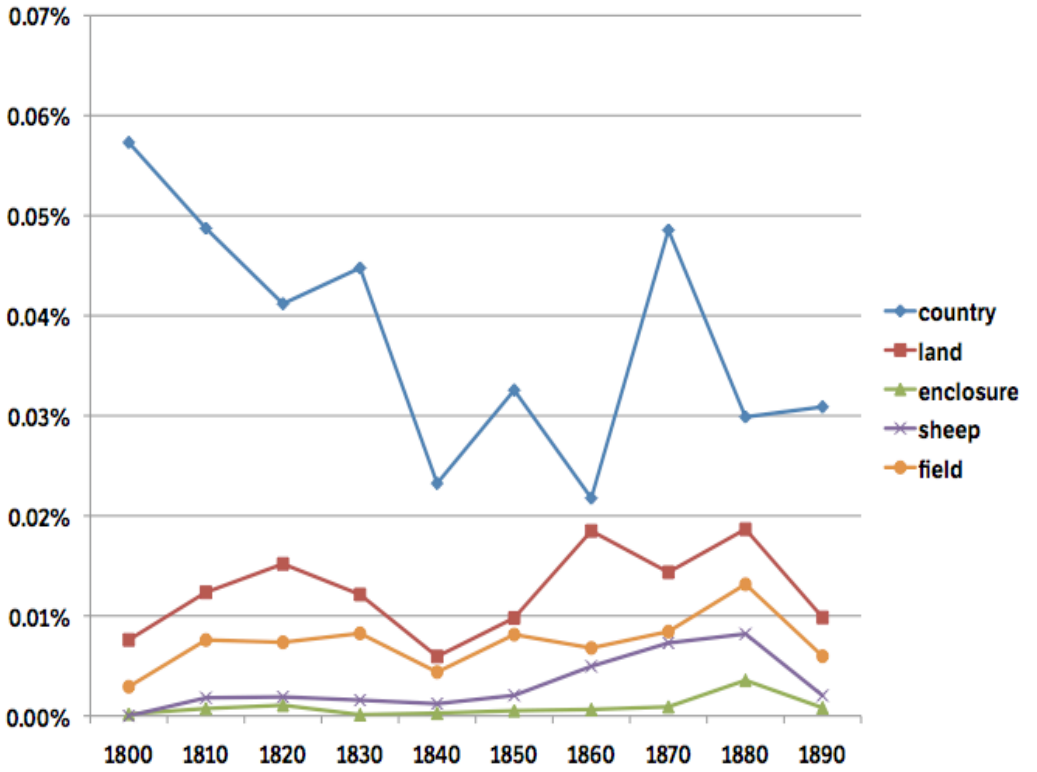

Figure 2: The relative frequency of the word "country" and other rural words across decades of the 19th century. The corpus here is 250 Chadwyck-Healey British novels.

At the same time, the agricultural words in the field ("land," "enclosure," "sheep," "soil," "field") tended to correlate, that is trend in lock-step with each other. Had we only looked at the frequency trend of the field as a whole (its aggregate frequency trend), we would have thought the semantic field of rural spaces behaves not like the correlated agricultural words, but the unrepresentative word "country," whose high frequency dominates the rest of the field.

This second problem could be called one of consistency: given the bluntness of an aggregate frequency trend, which elides the differences in behavior among its constituent word-trends, how could we ensure that our view of the whole was representative of the parts? In response to this problem, we eventually formulated an additional requirement our semantic fields must satisfy. Beyond their semantic coherence, to a certain degree the included words should correlate with each other in their frequency trends. While not conflating semantics and history, this principle required that the semantic link among words reveal itself as a correlation in their historical behaviors. This was a conservative definition of semantic fields (some semantic fields would not meet this criterion), but this conservatism would be useful in the initial stages of our research. Essentially, it would guarantee that our blunt instrument only picked up on highly reliable signals: high precision, if low recall. We would focus on discovering historically consistent semantic fields whose aggregate frequency trends would be representative and meaningful. The question now became: how could we increase our recall, or the number of words in our fields, so that our trends are not only internally consistent, but large enough to describe real, historical trends in novelistic discourse?

1.2 Stage 2: Correlator

Our conservative stipulation that all semantic fields must correlate turned out to be helpful in ways we hadn't even anticipated. It helped propel our project to its next stage. We thought, if ultimately words in a semantic field must correlate with each other, why not simply compute, in advance, the degree of correlation of every word in our corpus with every other word? This way, given certain seed words for a potential field, this computation would reveal correlated words that could be included in that field.

In March 2010, we built such a tool, calling it Correlator. To do so, we made use of a feature of the novelistic database Matthew Jockers had designed: a data-table of the number of occurrences of each word in our corpus. From this, we selected the words that appeared at least once in each decade of the nineteenth century, creating a new data-table of the selected words' frequencies of appearance.[3] We used normalized frequencies—the number of occurrences of a given word in a given decade, divided by the total number of word-occurrences in that decade—to correct for the over-representation of late century texts in our corpus. Then, we built a script to loop through each unique word-to-word comparison, calculate the degree of correlation between the two words' decade-by-decade frequencies, and store this information in a new data-table. As a measure of correlation, we used the Pearson product-moment correlation coefficient, a simple and widely-used statistical measure of the covariance of two numerical series, converted into standard deviations so that differences in scale were ignored[4]. (This scale-invariance was important, as we hoped to find words that behaved similarly despite differences in their overall frequencies.)

Finally, to access this new data, we built a script allowing us to query for the words that most closely correlate with a given "seed" word. For example: of all words in our corpus, which have a historical behavior most like the word "tree"? Correlator answered: "elm," "beech," "shoal," "let's," "shore," "swim," "ground," "spray," "weed," "muzzle," "branch," "bark." And which trend most like "country"? "Irreparable," "form," "inspire," "enemy," "excel," "dupe," "species," "egregious," "visit," "pretend," "countryman," "universal." As a first observation, these results seemed to verify our intuition that "country," given how aberrant its frequency trend was in comparison to those of other rural words, was more often used in its national sense; indeed, Correlator revealed that "country" kept company with words like "enemy" and "countryman."

But beyond this individual verification of the semantic deviance of "country" from the rural field, the very possibility of that verification surprised us. How could Correlator return such semantically meaningful results? Recall that Correlator knew nothing more than the decade-level frequencies of words. Could such coarse historical data really be sensitive to semantics? Querying Correlator for keywords identified through prior criticism, we found a word cohort, as we called the groups of words returned by Correlator, that was mas-

3 This filtering step ensures reliable correlation calculations; null data points can skew correlation coefficients. It also weeds out words with insignificant frequencies. Of course, one casualty of this filter is words invented in the middle of the 19th century, but we felt this drawback was outweighed by the benefits of the filtering step.

4 The Pearson coefficient ranges from +100%, meaning the two numerical series behaved identically, or that the changes in one could predict exactly the changes in the other; to 0%, meaning that no such prediction is possible; to -100%, meaning that changes in one numerical series could predict the changes in the other, by first reversing the direction of those changes. For a sample size of 10 data points (the 10 decades of the nineteenth-century), a correlation of 74% is considered statistically significant with a p-value of 5%. A p-value indicates the probability that the result was reached by chance.

sive and specific in meaning. While "tree" correlated with 333 other words significantly, and "country" 523, the word "integrity" correlated with 1,115, many of which shared a clear semantic relation: "conduct," "envy," "adopt," "virtue," "accomplishment," "acquaint," "inclination," "suspect," "vanity."

Correlator thus proved to be a method of discovering large word cohorts. Already historically consistent, these word cohorts could potentially be refined into semantic fields if we could ensure their semantic coherence. Correlator raised the possibility of generating semantic fields by pruning semantically-deviant words from an empirically-generated word cohort.

1.3 Stage 3: Semantic Taxonomies and Categorization

Having moved through an empirically and historically focused stage of semantic field development, we needed to return to the semantic focus in order to make such purely empirical word cohorts interpretable and meaningful. Our initial approach was to filter through these words for groups that seemed semantically coherent, but this proved too loose and subjective. It had the additional disadvantage of throwing away data in the form of words that correlated historically but seemed not to group semantically with the others. We decided it was irresponsible to decide a priori which words seemed to cohere historically because of a meaningful semantic relation and which words were just statistical noise, coincidences, or accidents. Perhaps these words could share an entirely different, non-semantic kind of relationship.

Abandoning these loose methods of filtering, we sought out semantic taxonomies to help categorize, organize, and make sense of these word cohorts. The database WordNet seemed promising for its clear-cut taxonomy but ultimately was unhelpful because of its idiosyncratic organization and rigid focus on denotation. Finally we turned to the OED. In an amazing stroke of luck, precisely when we needed it, the OED debuted its historical thesaurus, an incredible semantic taxonomy of every word sense in the OED 44 years in the making. It's nearly exhaustive, its categories are nuanced and specific, and it's truly organized around meaning. We used this powerful taxonomy to do two things: first, be more specific in identifying the semantic categories that constituted our word cohorts; second, to expand these word cohorts with many more words.

1.4 Stage 4: Statistical Filtering

With the addition of the historical thesaurus, we arrived at a dialogic method that drew on both quantitative historical data and qualitative semantic rubrics to construct semantic fields with precision and nuance. Correlator pointed us to proto-semantic fields that were then more fully developed using semantic taxonomies. Then, in this final stage, we turned from semantics back to the historical data, filtering these newly-developed semantic fields for two conditions. First, we removed words in the fields that appeared so infrequently that their trends could not be reliably calculated. We set this minimum threshold at 1 occurrence per 1% slice of the corpus, amounting to once every 4 million words, or approximately 11 times per decade. Second, we calculated the aggregate trend for the field, and removed any word that correlated negatively with the trend as a whole. While

turning to semantic taxonomies ensured the semantic coherence of our fields, this final step ensured their historical consistency.

Our ultimate aim in this process was to include as many words in our fields as possible without sacrificing these two requirements. The closer we could get to constructing an exhaustive, semantically tight, and historically consistent field, the closer we would move toward making valid arguments about historical transformations in the broad cultural concepts, attitudes, or values underlying a semantic field. In short, the closer we would get to a method of quantitative literary history.

2. Proof of Concept: The Generated Fields

Following these steps developed our "seed" words into rich, consistent semantic fields that were both semantically and culturally legible. These were the definitive fields that we investigated in the rest of our research. In Sections 4 and 5, we turn to that investigation: examining the fields' historical trends, and interpreting their significance for literary history. Here, we present four examples of the results of our method to demonstrate their legibility, scale, and consistency. These fields developed from a shared, multi-word seed: "integrity," "modesty," "reason," and "sensibility."[5]

Social Restraint Field

Example words: gentle, sensible, vanity, elegant, delicacy, reserve, subdued, mild, restraint

Largest of the fields, "social restraint" includes 136 words relating to social values regarding the moderation of conduct. Words such as "gentle," "reserve," "mild," and "restraint" express the positive valuation of this moderation.

Moral Valuation Field

Example words: character, shame, virtue, sin, moral, principle, vice, unworthy

Like the "social restraint" field, the "moral valuation" field relates to values of behavior, but this set of 118 words concerns the ethical evaluation of such conduct.

Partiality Field

Example words: correct, prejudice, partial, disinterested, partiality, prejudiced, detached, bias

With only 20 words, the "partiality" field is a small but semantically distinct group of words relating to values of disinterestedness.

Sentiment Field

Example words: heart, feeling, passion, bosom, emotion, sentiment, ardent, coldly, callous, pang

The "sentiment" field is semantically the most deviant from the other three fields, populated not with values per se but with words relating to emotion and sentiment. The 52 words in this field lay out a wide spectrum of emotional expression and implicitly value a range of healthy or proper emotionality.

5 For a full list of the words included in these and other of our semantic fields, please see Appendix B.

Beyond their semantic tightness and legibility, the fields' scale and historical correlation were considerable, as the data in table 1 shows.

Field	[A] Percent of words in corpus	[B] Number of words after OED (stage 3)	[C] Number of words after filtering (stage 4)	[D] Average correlation coefficient	[E] Median correlation p-value
Social Restraint	0.19%	155	136	91%	.00231%
Moral Valuation	0.24%	124	118	92%	.00229%
Sentiment	0.17%	116	52	77%	.157%
Partiality	0.01%	34	20	92%	.0232%
Collectively	0.61%	429	326	88%	.0411%

Table 1: Magnitude, number of words, and correlation values in four semantic fields. **Column A** indicates the percentage of the words in our corpus belonging to the respective field. **Column B** shows the number of words in the field after the initial word cohort was developed with semantic taxonomies, in other words, after stage 3 of our process. **Column C** shows the number of words remaining in the field after the statistical filtering of stage 4, which represents the final version of the field and is the basis for all further results. **Column D** indicates the average correlation coefficient for these words with the aggregate trend, while **Column E** indicates their median correlation p-value.[6]

3. Methodological Reflection: The Semantic Cohort Method

In this strand of our research, we focused on developing methodologies for computational historical semantics that would allow study on a scale far larger than that accessible through traditional methods of literary and cultural study. In doing so, we built on current n-gram-based research by moving from tracking individual words or hand-selected word groups to tracking macroscopic patterns of linguistic change. We aimed in defining our objects of study not to sacrifice the conceptual richness and cultural specificity that are among the great strengths of traditional methods. Our initial successes in identifying large-scale, culturally interpretable semantic fields suggest that indeed there are ways of scaling up such study.

As we conclude this first part tracing the development of our methodology, it's worth stepping back to collect the lessons we learned in the process. We learned that neither a purely semantic nor a purely quantitative approach is adequate to track historical changes in language. Because no simple relationship between the historical behavior of words and their meaning could be assumed, we adopted a dialogic approach that oscillates between the historical and the semantic, between empirical word frequencies that reveal the historical trends of words and semantic taxonomies that help us identify the meaning and content of those trends. This dialogic method emerged as a pragmatic response to the problems of generation, consistency, and interpretability. It ensures two things: first, that our results are semantically and culturally interpretable; second, that the aggregate data we collect on these large language patterns are reliable measurements of what's actually happening within them. In a way, fulfilling these two goals means limiting our object of study. Strictly speaking, the methods developed here are not looking at word cohorts,

6 See footnote 4 for an account of Pearson correlation coefficients and p-values; a value of above 74% is considered statistically significant, with a p-value of 5%.

which have historical consistency but may lack semantic coherence, or semantic fields, which have semantic coherence but may have an ahistorical relationship. The real object of study is a hybrid one that satisfies both requirements, something that could be called a *semantic cohort*, a group of words that are semantically related but also share a common trajectory through history.[7] This pragmatic limitation of our object of study generates a kind of data that lets us make broad historical arguments of the following type: the large semantic cohort of words sharing semantic property A underwent collective historical trend B in period C. This suggests D...

Given our original goals of finding ways to track historical shifts in semantics, it's fitting that we arrived in the end at a concept like the semantic cohort. The dual character of historical coherence and semantic consistency embedded in this concept succinctly characterizes our methodology: a semantic cohort method of discovering, analyzing, and interpreting large-scale changes in language use.

In learning to define our methodology, a broader lesson emerged that was less about the relation of history and semantics than about the disciplinary models that are complicated when doing this kind of research. Indeed, doing large-scale historical semantics requires a dialogue of the quantitative and the humanistic. The interdisciplinarity of our methods was less an a priori principle that directed our research than a necessity that emerged from the methodological complexities of investigating large-scale cultural and linguistic change. As we move on to present the results of our research, this point will emerge again and again. We hope by the end of this paper to make a case that quantitative methods do not supplant or even simply complement humanistic methods but actually depend on those methods as a partner if they are to take seriously the study of language and culture as their object.

4. Results: Major Shifts in Novelistic Language

Developing methods to generate semantic fields of course was only one part of the overarching project of tracking literary and cultural change at large. Now that we've shown that it's possible to isolate linguistic objects large enough to approach the scale of cultural change, we can move to the payoff: examining those changes, the trends these fields undergo, and what they might mean for literary history. The sequence of results we discovered was indeed striking: quantitative evidence of pervasive and fundamental transformations in the language of the British novel over a crucial period of its development, 1785-1900.[8] This was data from close to 3000 novels, a corpus stretching far beyond the canon and approaching the magnitude of a comprehensive set of British novels in this period. In the rest

7 The term "semantic cohort" is also used in the field of educational psychology when speaking of bilingual language development, but it is a rather loose use of the term to mean essentially a semantic field. Our use of the term is more specific; semantic cohorts are not simply semantic fields, but a subset of semantic fields that share historical trajectories. We include "cohort" in the term to designate the contemporaneous relation of the words in a semantic cohort. Through the rest of this paper we will occasionally use the term semantic field interchangeably with the term semantic cohort, though it should be clear that the semantic fields we constructed have been filtered for historical coherence.

8 We've talked thus far of the nineteenth century, but our project focused on a version of the "long nineteenth century." We started with the bounds of 1800-1900, but seeing that our data showed dynamic changes in the 1780s and 90s, we extended this boundary backward to capture a fuller picture.

of this paper, we will describe these findings and extract their implications for the literary history of the British novel. This will require some retelling of the story of our research, but with the focus on the results rather than the methods. Continuing in a narrative mode seems the most natural way to present our findings given the process of discovery in the digital humanities, which often feels like taking two steps backward for every wandering step forward.

4.1 Discovery, Part 1: Abstract Values Fields

As described in Section 1.1, our investigation of semantic fields was rooted in existing literary scholarship. We turned to these sources for words that might be the "seeds" of historically important semantic fields. One attractive possibility was tracing linguistic changes reflecting the shift from rural to urban life, a defining social transformation of this period. We hadn't yet developed our dialogic methods of historical correlation and semantic taxonomy, so filling out these country and city words into large semantic fields was very difficult. Yet, we did produce some data, still loose and messy, but results nonetheless. It was in trying to make sense of the aggregate frequency trends of these proto-semantic fields that the idea behind Correlator was born. With historical correlation as an added criterion for the semantic fields we would track, we were able to be more rigorous in determining the validity of these early aggregate results. More importantly, Correlator gave us a powerful method of filling out semantic fields by empirical means, a big step beyond the imprecise practice of hand populating these fields.

Working with this tool led us to our first major discovery. Among the groups of potential seed words we had been considering in our early stages of field construction was a group of words related to values and behavior. These seed words—"integrity," "modesty," "sensibility," and "reason"—when inputted into Correlator produced some astonishing results. In our initial trials with Correlator, we had found some relatively large word cohorts, with dozens of words in each, that demonstrated significant, if not strong, correlations in their historical behaviors. The word cohort that emerged from these seed words, however, included almost 900 words with a very high degree of correlation (see **Figure 3**).

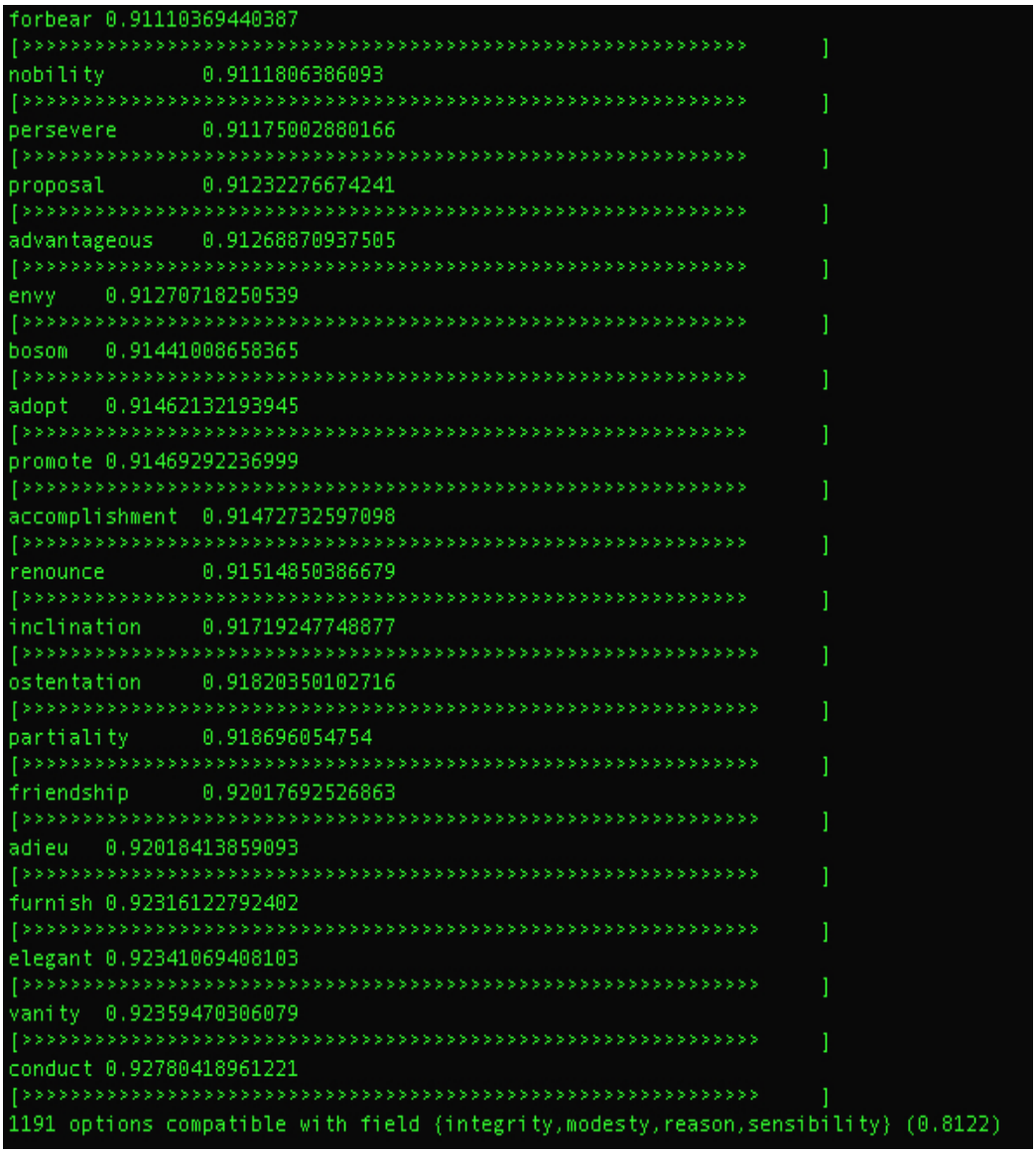

Figure 3: Correlator output from seed words "integrity," "modesty," "reason," and "sensibility." Words with the highest correlation with the seed words are listed at the bottom. The number to the right of each word is that word's correlation coefficient with respect to the seed words.

More importantly, the cohort seemed to have a remarkable degree of semantic coherence. It centered on a semantic field of abstract words used to talk about values and social behavior. For example, among the top 20 most correlated results were the words "conduct," "vanity," "friendship," "partiality," "ostentation," "accomplishment," "envy," and "forbear." These are just a few examples; the cohort contained hundreds of words like these. What we seemed to have discovered was a massive group of words relating to values and social behavior that for some reason followed the same historical trajectory.

The cohort's trend was as striking as its scale and coherence. Initial tests showed that this cohort of hundreds of words declined precipitously in usage across the nineteenth

century. This finding demonstrates some of the key strengths of Correlator. The tool helped us identify a large-scale word cohort with high consistency of historical behavior. And because it compares the usage of words on a decade-by-decade basis, it pointed us to particularly dynamic, century-wide trends. To follow up on these initial findings, we focused on filling out this cohort of social value words to make sure we were catching the full scope of the trend. As mentioned in Section 1.3, our first thought was to comb through the cohort, picking out the words that fit within this emerging social values semantic field. That filtering process gave us a rough field of over 75 word lemmas:

> integrity, modesty, sensibility, reason, talent, conduct, elegant, ostentation, partiality, friendship, accomplishment, character, persevere, vanity, forbear, benevolence, assiduity, understanding, extravagance, zeal, delicacy, firmness, envy, reluctance, excellence, vexation, esteem, virtue, prejudice, unrelenting, accomplish, sincere, nobility, taste, sedulous, admiration, sentiment, rational, brilliancy, falsehood, prudent, excess, superiority, unworthy, malignant, sensible, genius, reflection, pleasure, dignify, artifice, happiness, indolence, principle, discernment, coldness, self-denial, depravity, indulge, infamy, malice, faultless, adherence, perseverance, profligate, aversion, penetration, solicitous, despise, indulgence, ardent, candour, softness, restraint, impatience, insensibility

As interesting as this group of words was, we had two major reservations with this method. First, the semantic coherence of the group was still loose. It was clear that these words were predominantly abstractions and were related to values of social behavior. To make our conclusions specific and relevant to literary and cultural study, though, we needed to identify and categorize the semantic content of this field more precisely. The second major problem was we didn't know how comprehensive these results were in delineating the entirety of a semantic field; were there many other social values and abstractions we weren't seeing? Thus, while developing ways to categorize and specify the semantic content of these results, we also sought to continue expanding this semantic field with more words.

This led us to semantic taxonomies. We used the OED's historical thesaurus to identify the semantic content of the field, break it down into more specific sub-fields, and fill those out with further words from the OED's semantic categories.[9] As mentioned in Section 2, we identified and filled out four sub-fields of abstract values words: words relating to values of social restraint and moderation; words of moral valuation; words relating to sentiment; and words relating to values of objectivity. After using the empirical data to filter out added words that didn't correlate historically, we had four developed semantic cohorts, tighter in their semantic relation and closer to exhaustiveness than before. We finally felt ready to look closely at the aggregate trends of these fields.

With these more focused fields, the trends we found were dramatic. Tracing their behavior over the nineteenth century, we found they exhibit parallel downward trends. For instance, the field of social restraint and moderation words exhibits a steady downward decline

9 See appendix C for more detail on this process: how we used the historical thesaurus and what specific semantic categories we drew on.

(Figure 4) from ~0.30% of all word tokens[10] (about 1 in every 325 words) at the beginning of the century to ~0.15% of all word tokens (about 1 in every 700 words) by the end of the century, a decrease of about 55%.[11]

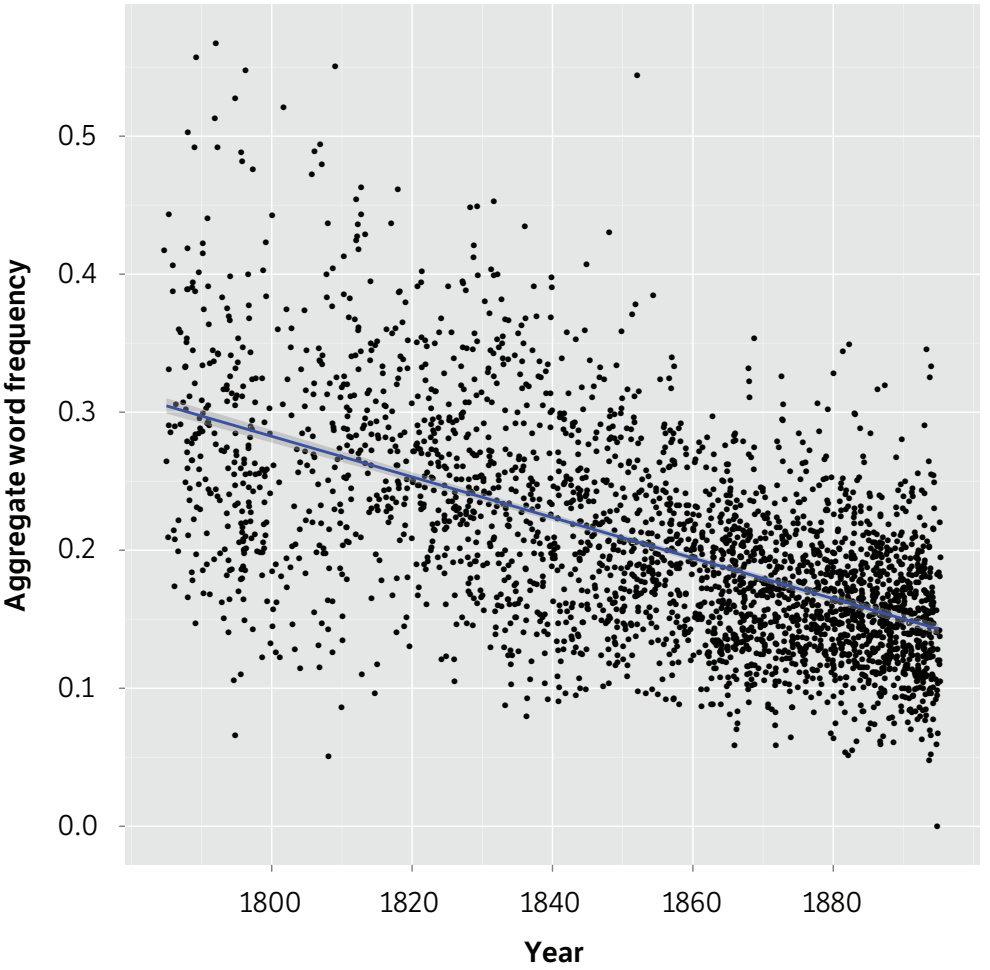

Figure 4: Aggregate term frequencies of the social restraint field in novels, 1785-1900. For all the plots in this section, each point represents the frequency of these words in a particular novel. The X-axis represents the novel's date of publication. The Y-axis represents the percentage of the novel's words that are of the field.

The field of moral valuation words shows a similar trend (Figure 5), declining from ~0.43% of all word tokens (about 1 in every 235 words) at the beginning of the century to ~0.15% of all word tokens (about 1 in every 670 words) by the end of the century, a decrease of about 65%.

10 A word token is an occurrence of a word. Token is distinguished from type, which stands for the word itself. For example, if a text has a lexicon of 100 unique words but is 800 words long, we say that that text is composed of 100 types but 800 tokens.

11 These percentages are drawn from the linear regression fit of the data. The graph shows the range of frequency values among the novels in any given period while the regression is a useful articulation of the overall trend. We use "century" here as shorthand for the full historical range of our data, extending from 1785 to 1900.

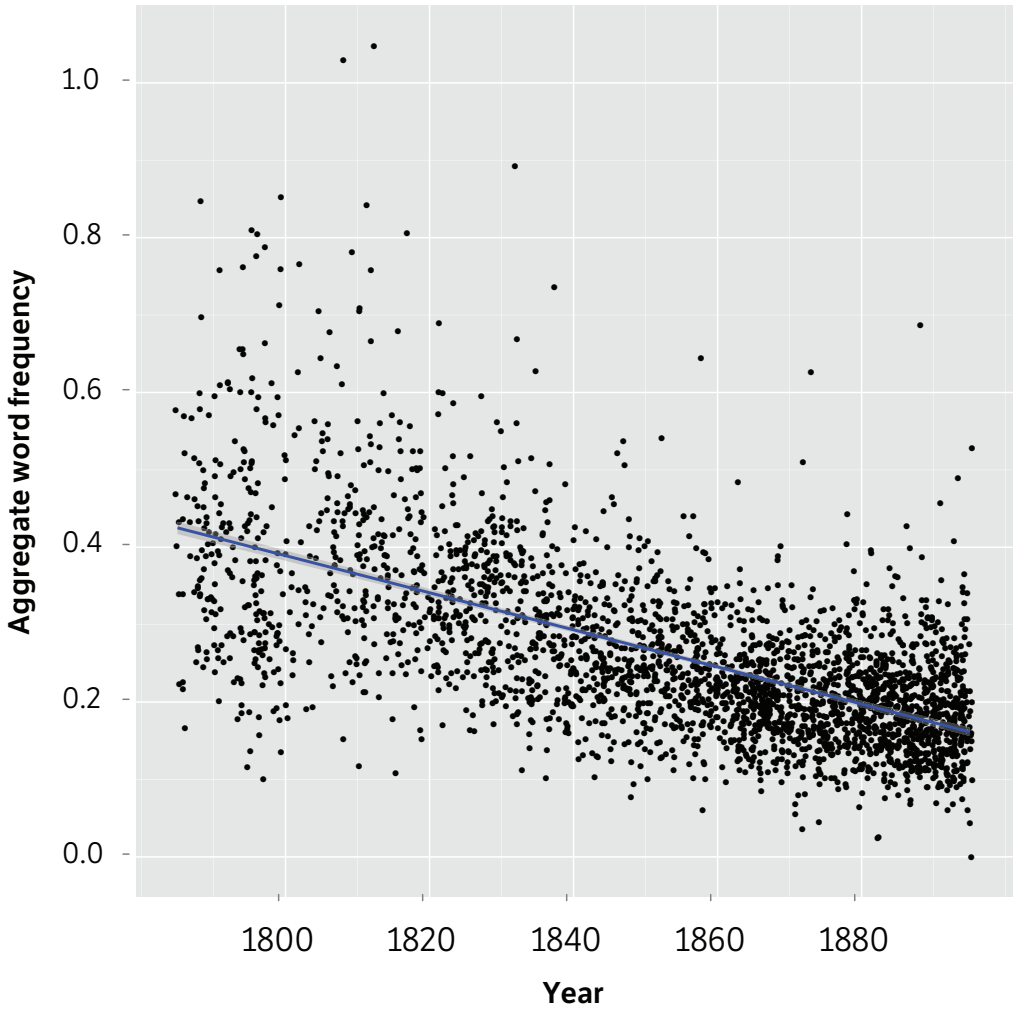

Figure 5: Aggregate term frequencies of the moral valuation field in novels, 1785-1900.

The field of sentiment words shows a steady decline (**Figure 6**) from ~0.25% of all word tokens (about 1 in every 380 words) at the beginning of the century to ~0.14% of all word tokens (about 1 in every 700 words) by the end of the century, a decrease of almost 45%.

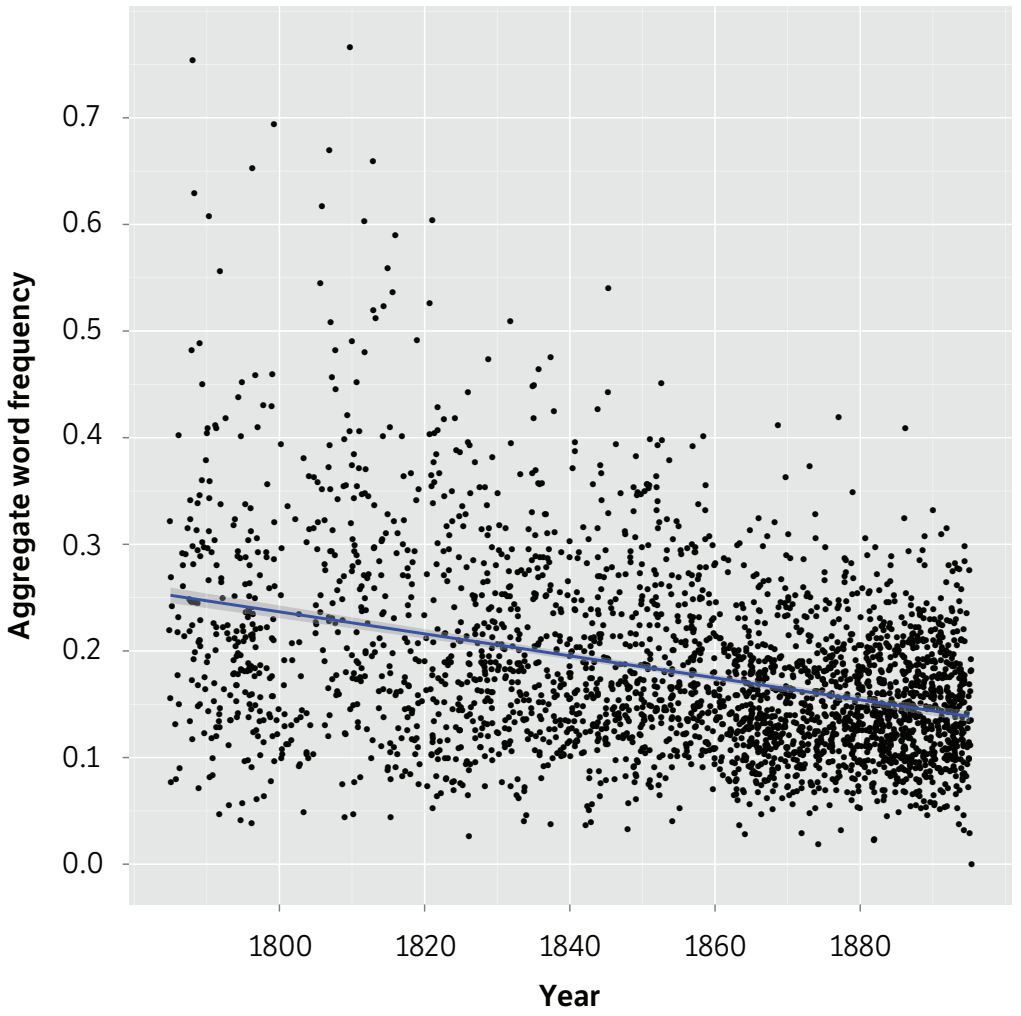

Figure 6: Aggregate term frequencies of the sentiment field in novels, 1785-1900.

The field of partiality words exhibited aggregate term frequencies an order of magnitude lower than the other fields but nevertheless exhibited a parallel trend (**Figure 7**). It declines steadily from ~0.02% of all word tokens (about 1 in every 4500 words) at the beginning of the century to ~0.006% of all word tokens (about 1 in every 17500 words) by the end of the century, a decrease of almost 75%.

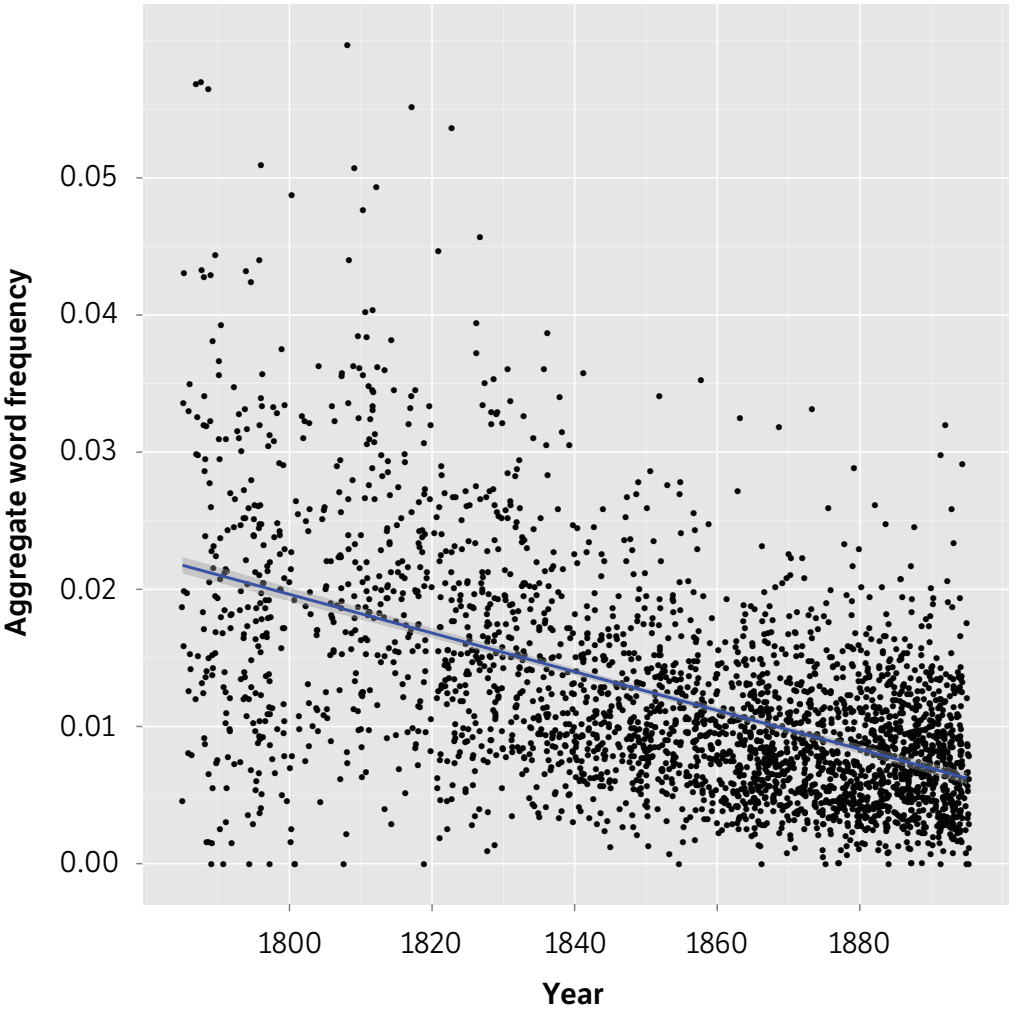

Figure 7: Aggregate term frequencies of the partiality field in novels, 1785-1900.

Collectively, the aggregate term frequency for the fields of abstract values decreases through the nineteenth century (**Figure 8**), from ~1.0% of all word tokens (about 1 in every 100 words) in the period of 1800-1810, to ~0.44% of all word tokens (about 1 in every 225 words) by the century's end, a decrease of about 55%.

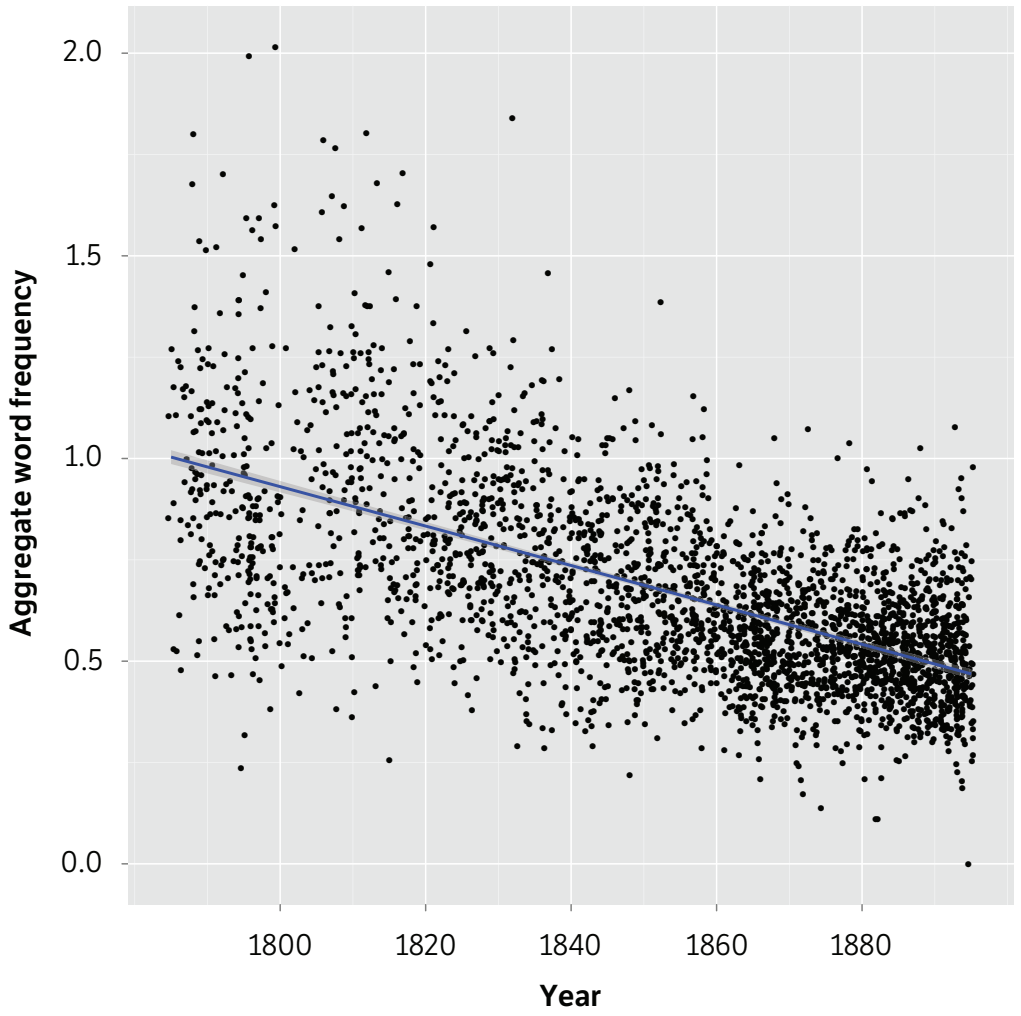

Figure 8: Aggregate term frequencies of the abstract values fields combined in novels, 1785-1900.

Fully interpreting this dramatic declining trend requires looking closely at the shared characteristics of the words in these fields and contextualizing them within the literary and cultural history of the period. This gives us a better sense of what sort of linguistic shift is occurring here and opens the investigation into the reasons behind it. Examining the abstract values words, we can isolate several key characteristics.[12] First, the words are largely abstractions: abstract nouns such as "modesty," "extravagance," or "propriety"; and abstract adjectives such as "elegant," "indecent," or "restrained." More specifically, they form a cluster of abstractions centered on ideas of social normativity and the regulation of behavior. For instance, the "social restraint" field, which includes words like "restraint," "moderation," "self-control," "excess," "indulgence," and "ostentation," clearly delineates a set of social values prescribing the proper limits of personal behavior, the moderate range of conduct considered socially acceptable. Given this emphasis on social norms, it's no surprise to find that these fields are also rich in highly evaluative, highly polarized language. These are words used to articulate specific social values, judge be-

12 See appendix B for the full listing of words in these fields.

havior, and point out lapses and violations. Thus, the fields include many words such as "moral," "virtue," and "decent," but also their opposites, "immoral," "sin," and "indecent." It's worth noting as well that the abstract values words are predominantly Latinate, which makes sense given the dominance of abstractions. What we found then was a massive semantic cohort of abstract, socially normative, evaluative, and highly polarized words that underwent a systemic and significant decline in usage over the century.

What could account for such a shift? What was it about these words that made them subject to such a dramatic change in literary language? We wondered, if these are the types of words that were declining, are there other groups of words, other trends, that might help us contextualize and make sense of this decline? Our first hunch was to try to find a shift in values over the century; if what we had isolated in the abstract values fields were late eighteenth- and early nineteenth-century British values, then perhaps we could find fields of Victorian values that supplanted them. Through many potential seed words though this search proved unfruitful. No major trends in other kinds of values words emerged from the data. We were stuck.

4.2 Discovery, Part 2: "Hard Seed" Fields

It was at this roadblock that having a purely quantitative method, a way to identify proto-semantic fields through historical frequency data alone, opened the road to our next major discovery. After a long series of fruitless seed words, perhaps on a whim, perhaps on a wild hunch about words completely different from the abstract values, we inputted into Correlator the innocuous little word "hard." True serendipity. What emerged in the output of Correlator was a massive cohort of over 400 word lemmas that shared an even tighter historical correlation than the abstract values cohort. We named this cohort "hard seed." The first thing that struck us about "hard seed" was just how different these words were from the abstract values words. Among the top 20 words most correlated with "hard": "smoke," "go," "brush," "look," "rough," "liquid," "back," "come," "face," "ache," "finger." Even more fascinating though was the aggregate trend of this word cohort. In strict contrast to the behavior of the abstract values fields, this cohort showed a dramatic rise over the nineteenth century. We may not have found the expected shift toward Victorian values, but we found something even more interesting, a massive group of words categorically different from the abstract values fields that contextualizes and frames their decline within an even broader movement. It's important to note that finding this other major trend would have been nearly impossible without the quantitative methods at our disposal. When we were searching for semantic fields related to the decline in the abstract values fields, it did not and would not have occurred to us to look toward a group of "hard seed" type words. They are not semantically or culturally related to the abstract values words in any immediately clear way. We might still have discovered the trend for the word "hard," but without Correlator's ability to aggregate word cohorts around trends, we would have had no sense of its significance. It took a computational method of finding language trends to discover this other group of words that, while not semantically related to the abstract values words, are historically related.

After applying to "hard seed" the semantic cohort method of identification, correlation, categorization, expansion, and refinement, we found we had isolated quite an interesting

creature. Instead of a single semantic field tightly organized around a specific semantic property, this highly correlated word cohort comprised a variety of semantic fields and types of words including:

Action verbs: "come," "go," "drop," "stand," "touch," "see"...

Body parts: "finger," "face," "hair," "chin," "hand," "fist" ...

Colors: "red, "white," "blue," "green," "brown," "scarlet" ...

Numbers: "three," "five," "two," "seven," "eight," "four" ...

Locative and directional adjectives and prepositions: "down," "out," "back," "up," "over," "above" ...

Physical adjectives: "hard," "rough," "flat," "round," "clear," "sharp" ...

As seen in table 2, these fields were even more massive than the abstract values fields, accounting for almost 4.5% of all word occurrences in our corpus.[13]

Field	[A] Percent of words in corpus	[B] Number of words after OED (stage 1.3)	[C] Number of words after filtering (stage 1.4)	[D] Average correlation coefficient	[E] Median correlation p-value
Action Verbs	1.99%	257	248	73%	.742%
Body Parts	0.65%	147	111	71%	.773%
Colors	0.13%	96	46	57%	6.16%
Locative Prepositions	1.09%	28	27	74%	.499%
Numbers	0.37%	46	44	73%	.679%
Physical Adjectives	0.20%	32	32	79%	.227%
Collectively	4.43%	606	508	71%	1.51%

Table 2: Magnitude, number of words, and correlation values for the hard seed fields.

Tracing their behavior over the nineteenth century, we found the "hard seed" fields exhibited parallel upward trends. The field of action verbs, a field of substantial magnitude, exhibits a steady rise (**Figure 9**) from ~0.96% of all word tokens (about 1 in every 100 words) at the beginning of the century to ~2.7% of all word tokens (about 1 in every 40 words) by the end of the century, an increase of over 180%.

13 Please see Table 1 for an explanation of these columns. See appendix B for the full listing of words in these fields.

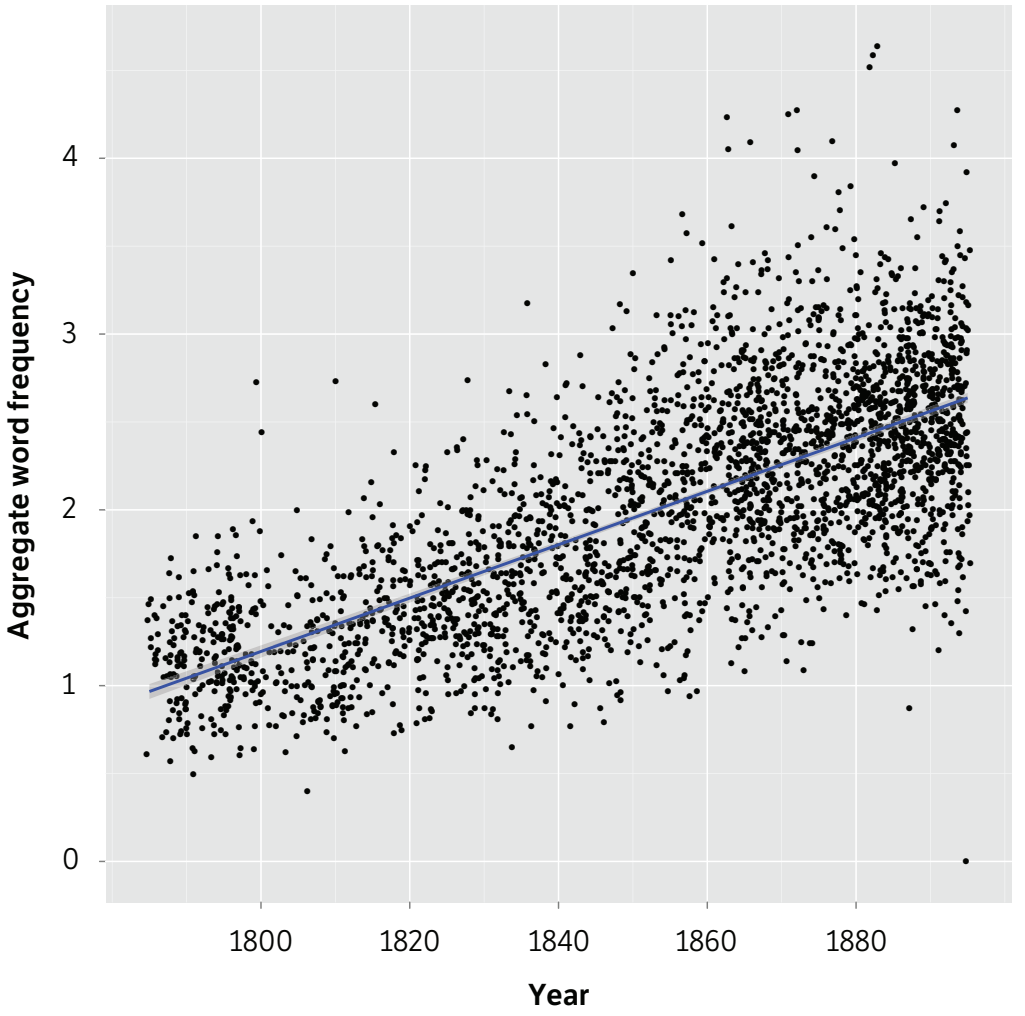

Figure 9: Aggregate term frequencies of the action verbs field in novels, 1785-1900.

The body parts field also shows a rise (**Figure 10**), though of a gentler slope, increasing from ~0.45% of all word tokens (about 1 in every 220 words) at the beginning of the century to ~0.80% of all word tokens (about 1 in every 120 words) by the end of the century, an increase of about 80%.

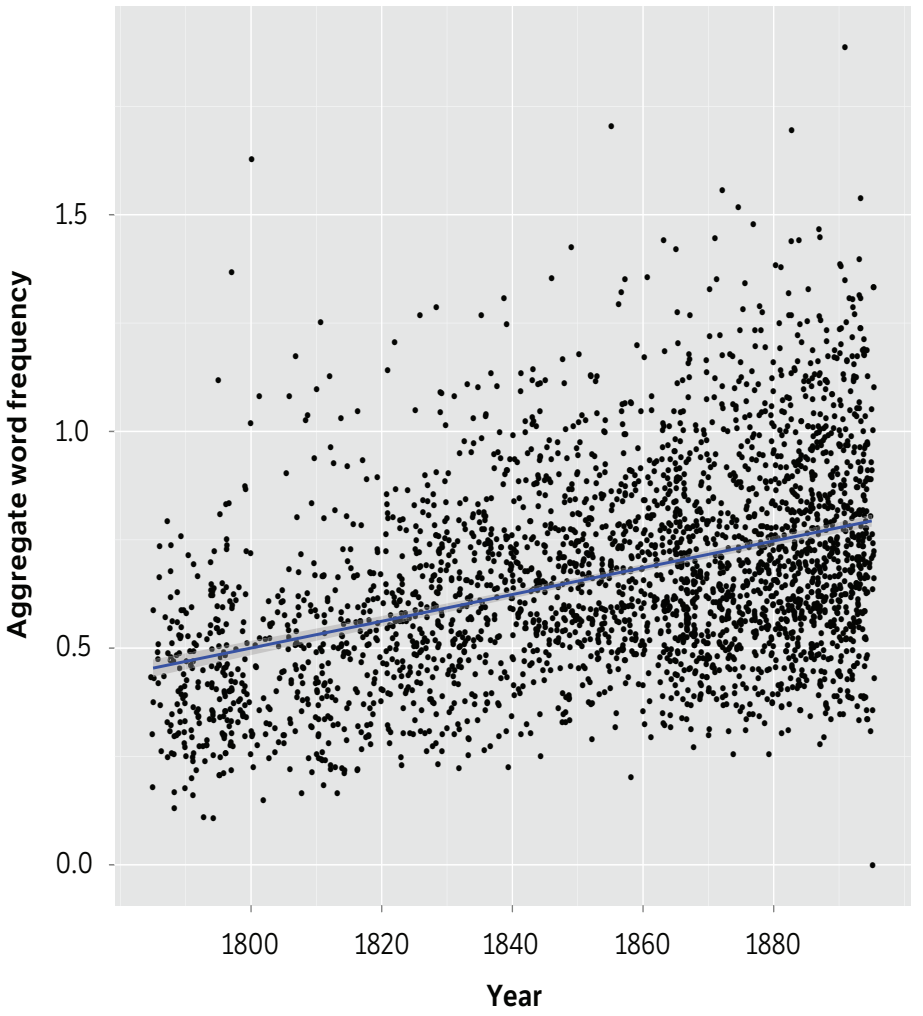

Figure 10: Aggregate term frequencies of the body parts field in novels, 1785-1900.

The colors field shows an even sharper rise (**Figure 11**) from ~0.04% of all word tokens (about 1 in every 2000 words) at the beginning of the century to ~0.19% of all word tokens (about 1 in every 530 words) by the end of the century, an increase of over 290%.

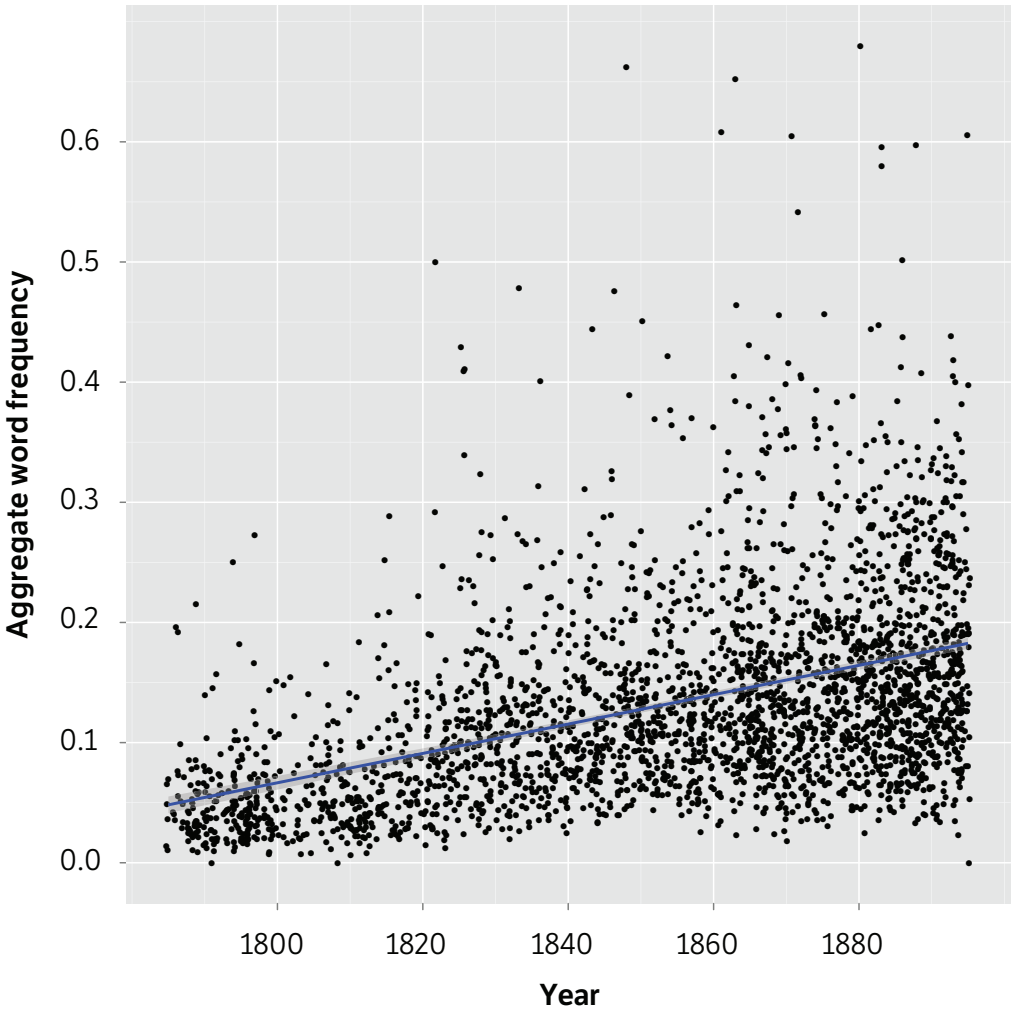

Figure 11: Aggregate term frequencies of the colors field in novels, 1785-1900.

The field of numbers increases (**Figure 12**) from ~0.2% of all word tokens (about 1 in every 470 words) at the beginning of the century to ~0.3% of all word tokens (about 1 in every 300 words) by the end of the century, an increase of about 54%.

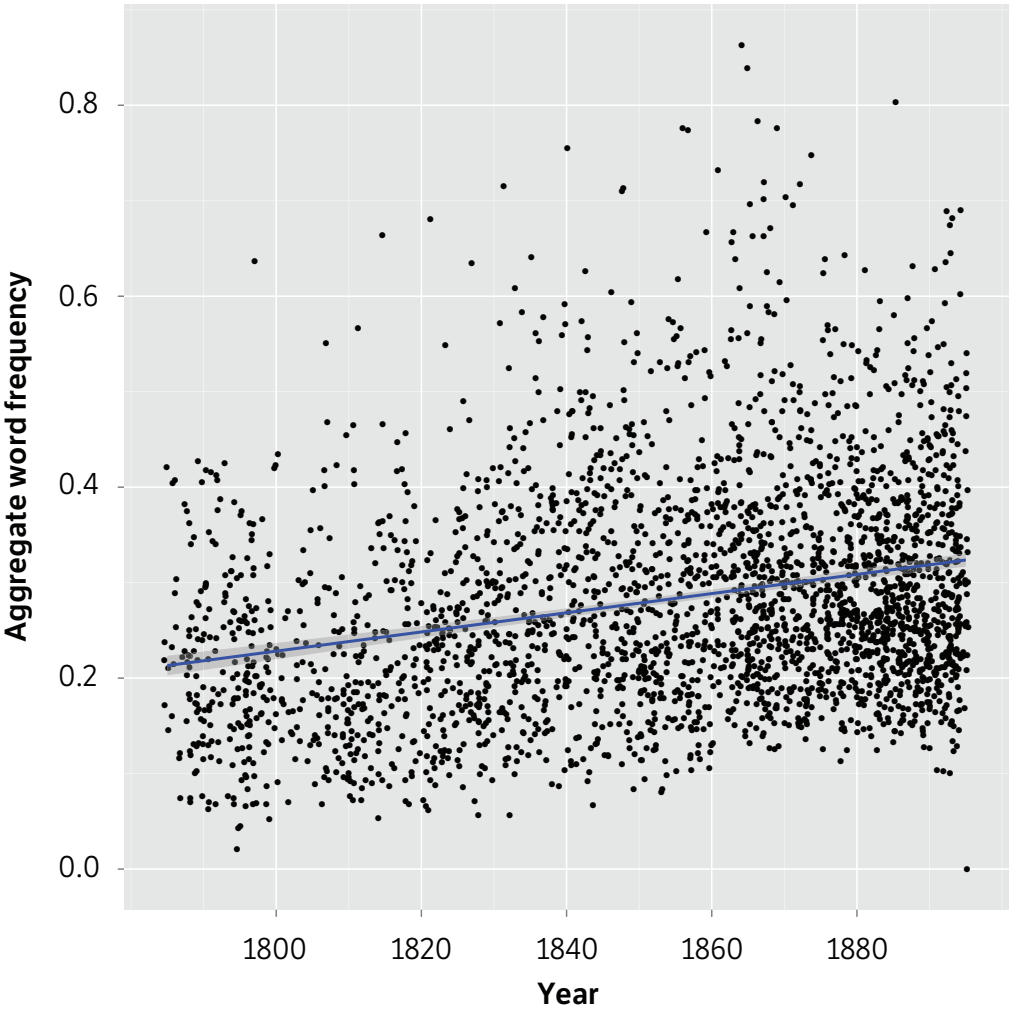

Figure 12: Aggregate term frequencies of the numbers field in novels, 1785-1900.

The field of locative and directional adjectives and prepositions shows a rise (**Figure 13**) from ~0.59% of all word tokens (about 1 in every 170 words) at the beginning of the century to ~1.43% of all word tokens (about 1 in every 70 words) by the end of the century, an increase of over 140%.

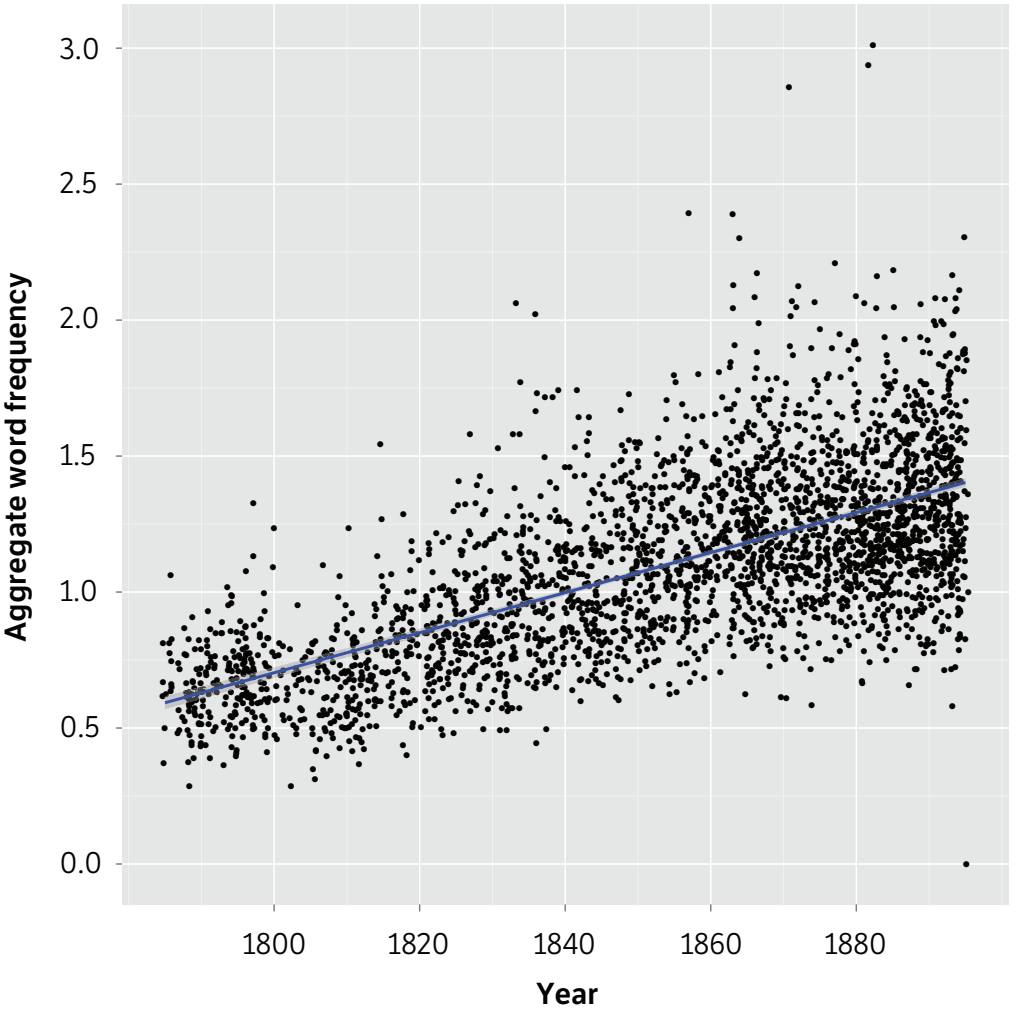

Figure 13: Aggregate term frequencies of the locative prepositions field in novels, 1785-1900.

The physical adjectives field rises (**Figure 14**) from ~0.07% of all word tokens (about 1 in every 1300 words) at the beginning of the century to ~0.28% of all word tokens (about 1 in every 350 words) by the end of the century, an increase of over 280%.

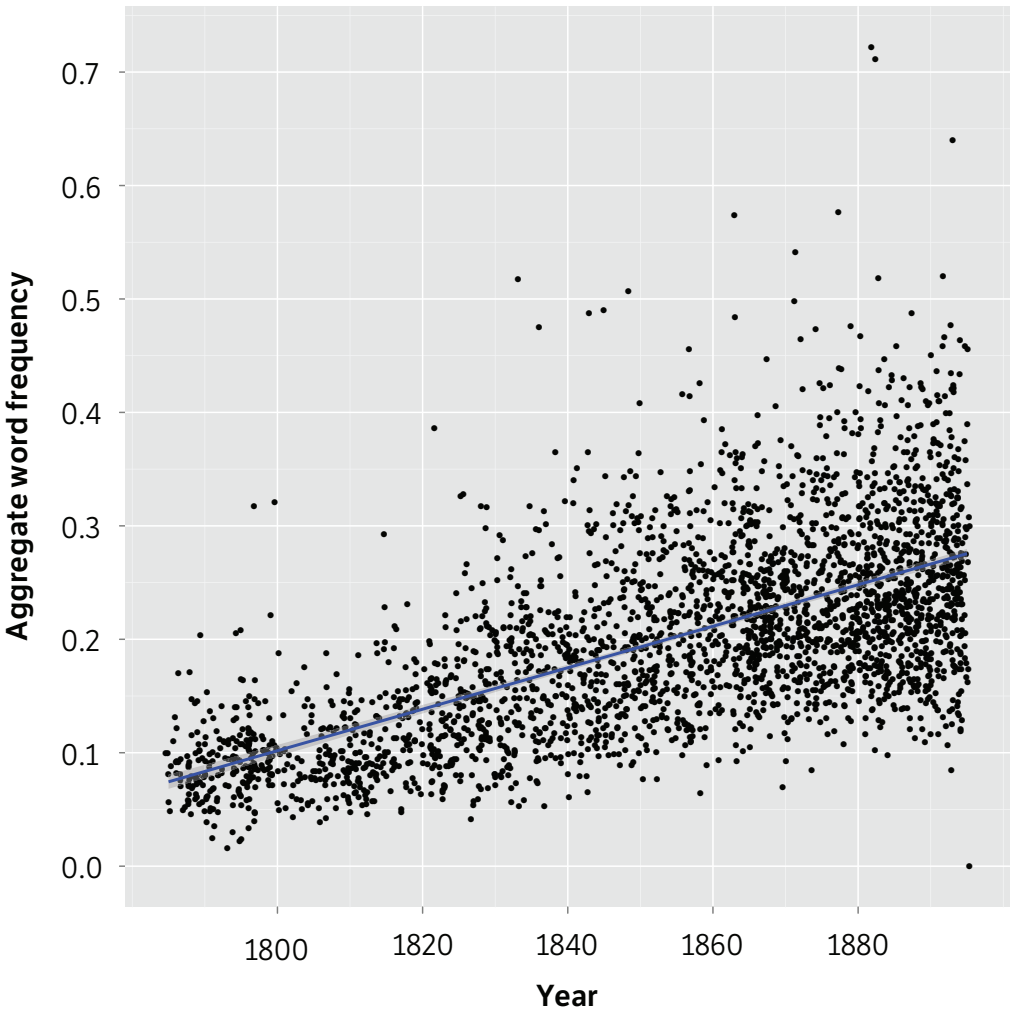

Figure 14: Aggregate term frequencies of the physical adjectives field in novels, 1785-1900.

In contrast to the values fields, the aggregate term frequency of the "hard seed" fields increases steadily across the 19th century (**Figure 15**) from 2.5% of all word tokens (~1 in every 40 words) to 5.9% of all word tokens (~1 in every 17 words), an increase in usage of over 130%.

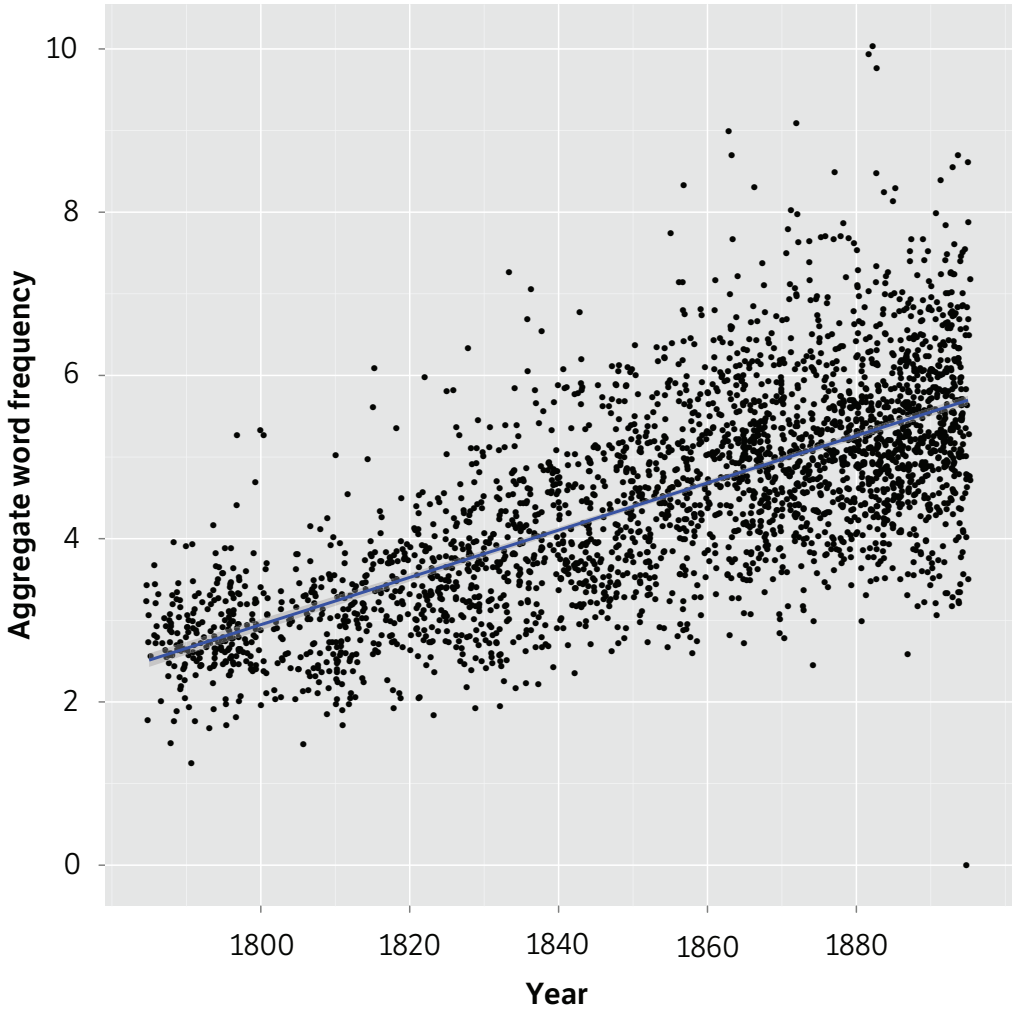

Figure 15: Aggregate term frequencies of the hard seed fields combined in novels, 1785-1900.

As we did with the abstract values fields, we looked closely at the shared characteristics of the "hard seed" words. The comparison with the abstract values words was particularly revealing. As opposed to abstractions, the "hard seed" words are concrete and physical—"wet," "stiff," "crack," "knock," "jaw," "neck," etc. They are also specific, words used to specify the particular action ("stoop," "scratch," "tilt," "crawl"...), physical orientation ("over," "under," "behind"...), physical quality ("heavy," "wooden," "crooked"...), color ("yellow," "purple", "orange," "ruddy"...), or quantity ("ten," "sixty," "hundred," "thousand"...) of an object or person. Where the abstract values words were evaluative and highly polarized, these words are non-judgmental, too rooted in the physical to refer in any direct way to abstract norms, values, and standards. And where the abstract values words were long and Latinate, these are short, often monosyllabic, and predominantly Anglo-Saxon in origin. In the context of the novel, the "hard seed" word cohort can be collectively characterized as concrete description words of a direct, everyday kind. It is these kinds of words that are rising significantly in usage over the nineteenth century.

4.3 Corroboration: Topic Modeling Data

Because these results were so striking, we wanted to make sure what we had found was in fact real. To corroborate these results, these two major trends in novelistic language, we sought another method of gathering large-scale semantic data. Topic modeling provided this complement to our semantic cohort method. A well-established procedure, topic modeling computationally groups words that tend to appear in the same context within texts; these groups can be thought of as topics or themes[14]. It offers two key differences from our methods. First, it's an unsupervised method that generates topics without subjective input from users, complementing our methods, which mix supervised and unsupervised procedures. Second, it generates topics based on co-occurrence within texts, rather than on our dual criteria of historical correlation and semantic relatedness. Thus, topic modeling gave us an entirely different lens to look at the semantic patterns in our corpus, a way to test if our results could be replicated when measured by different tools.

After generating five hundred topics of nouns, we isolated two sets of topics, those most frequent in novels published toward the beginning of the century, 1790-1830, and those most frequent in novels toward the end of the century, 1860-1900[15]. Comparing these two sets gave us a rough and ready view of historical trends in the topic modeling data. To enable the comparison of these results to our established ones, we categorized each topic into one of four types based on the characteristics of their constituent words: abstract values-type, "hard seed"-type, mixed type (topics that exhibited both abstract values and hard seed characteristics), or none of the above. The results, as shown in table 3 and figure 16, were clear:

Period	Abstract Values-type	"Hard Seed" type	Mixed-type	None of the Above
1790-1830	69%	23%	8%	0%
1860-1900	10%	64%	5%	21%

Table 3: Comparison of abstract values-type and hard seed-type words in topics from 1790-1830 and 1860-1900 showing the percentage of topics of each type among the most frequent topics of the two periods.

14 See Blei, Ng, and Jordan, the foundational paper on topic modeling for more information.

15 Due to space constraints, we cannot include the topic modeling data and procedures in this pamphlet. Please see the supplemental materials online at http://litlab.stanford.edu/semanticcohort

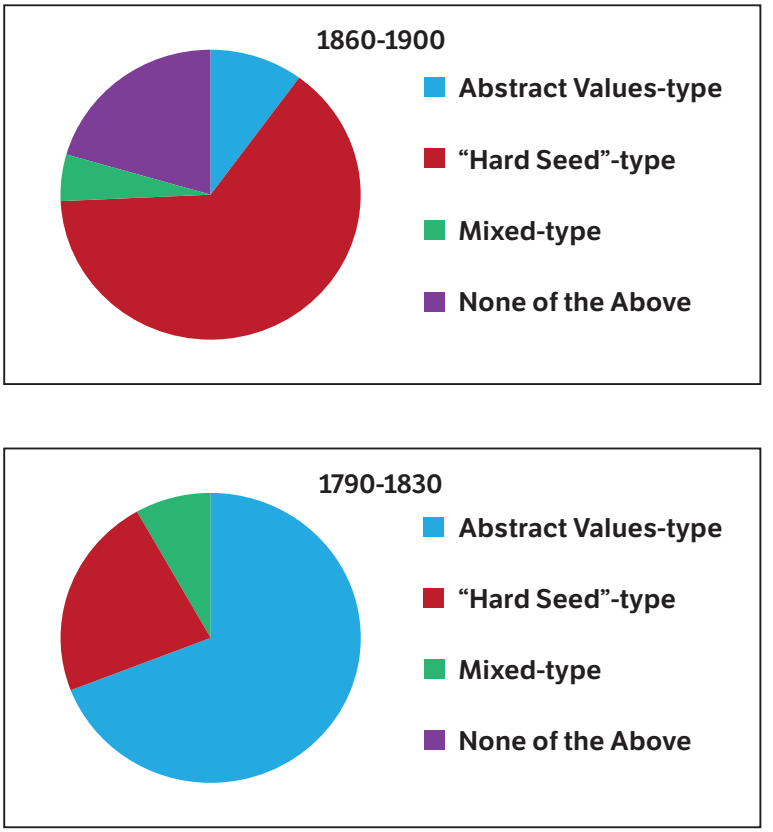

Figure 16: Pie charts of abstract values-type and hard seed-type words in topics from 1790-1830 and 1860-1900.

As in the results of our semantic cohort method, the topic modeling data confirmed opposite trends for these two kinds of novelistic language: a decline in abstract values-type words and a rise in concrete, “hard seed”-type words. What’s important here is that these same dramatic trends were found by entirely independent methods, confirming that our results are not an anomalous product of our methods but a real historical transformation in the nineteenth-century British novel.

As these trends appear real, it’s worth pausing here to emphasize their magnitudes. The abstract values fields at their height account for about 1% of all word usage in nineteenth-century British novels; the “hard seed” fields, almost 6%. These are large-scale, diffuse trends, encompassing the histories of hundreds and hundreds of words. Recognizing the scale of these changes made us all the more eager to probe into the data. What might these changes mean? What might lie behind them?

5. Discussion of Results: The Language and Social Space of the Nineteenth-Century British Novel

In any experiment, the path from data to conclusions is tricky, lined with potential pitfalls, from misrepresentation of the data to overblown readings that stretch the data beyond what they can reasonably support. The challenge of culling these massive quantities of data, analyzing them, gathering further data, and putting them into meaningful conversation brought on another bout of methodological self-consciousness. How do we interpret quantitative data on culture and literature? How do we use data to substantiate arguments about culture? Can it contribute new knowledge to literary study? These aren't just concerns of our project; they are some of the essential questions for digital humanities as a whole. The field should be ready to provide compelling and substantial answers to these questions if it is to earn mainstream credibility within the humanities. A flurry of interest and an innovative method aren't sufficient to justify an emerging research program. Thus, scholars in the humanities are right to maintain a measured skepticism until a critical mass of research projects shows how data-driven methods can make powerful and unique contributions to the field, changing the questions we can ask, the concepts we use, and the knowledge we hold. We aim to present the strides we've made toward these ends and be honest about the difficult questions that remain.

The methods we discovered in this difficult work of learning to mobilize quantitative data to make qualitative claims are as much a part of this project's payoff as the results and conclusions themselves. Thus we'll be interweaving them with the discussion of results. One tactic we adopted was to defer the moment of interpretation as long as we could, running further tests, gathering other kinds of data, visualizing our data in other ways, so we could have confidence in our results before attempting the leap to interpretation. We're attracted to the working method of the sciences in which conclusions from an experiment are part of an ongoing process of investigation; no conclusion is a final one, just the best hypothesis at the moment, always open to revision as further research is done. This is the spirit in which we hope the discussion of our results will be taken.

5.1 Initial Observations and a Spectrum of Novels

An important first step to cracking open our results was to determine the relationship between the two seemingly opposite trends we had found. So for each of the novels in our corpus we plotted its usage of the abstract values field against its usage of the "hard seed" fields. This revealed a strongly inverse relationship between the two (see **Figure 17**).

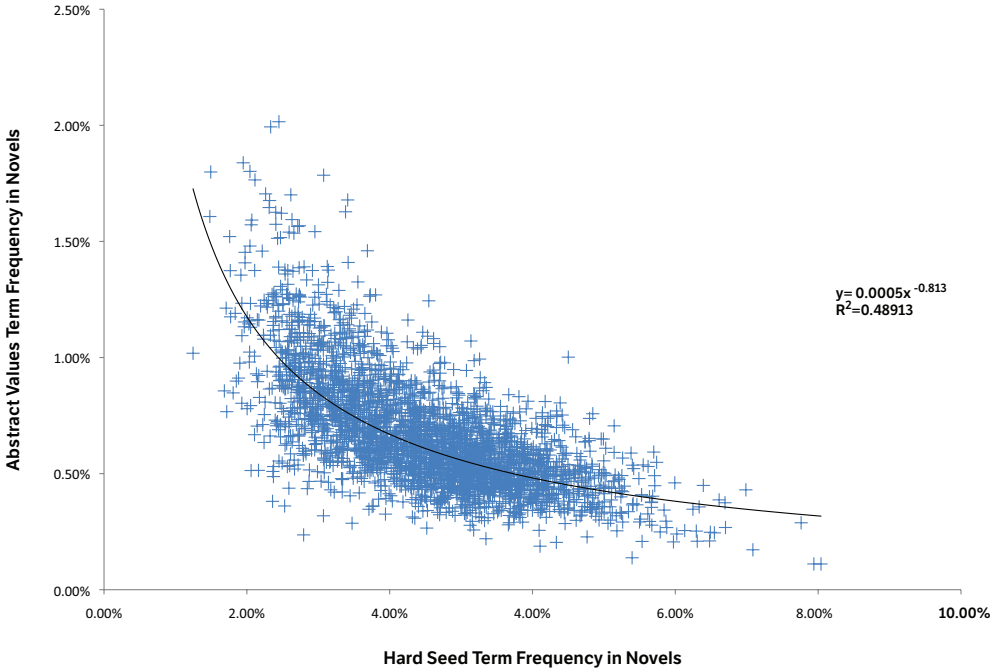

Figure 17: Inverse relationship of abstract values and hard seed word frequencies in novels. An exponential regression is superimposed.

What can be seen from this plot is the tendency for novels with high frequencies of "hard seed" words to have low frequencies of abstract values words and vice versa. The two fields' mutual exclusivity suggested it would be possible to separate out different groups of novels through their relative usage of the fields. To visualize these groups, we produced two spectra of the novels, ranked by the concentration of abstract values words or "hard seed" words. These proved enormously revealing, instrumental in moving from examining the data to interpreting them.

The spectra allowed us to see the trends through units understandable and familiar to us as readers and literary scholars, the actual novels, genres, and authors in our corpus. Instead of trying to make sense of term frequency behaviors of semantic fields, a rather abstract object, the spectra let us ask more grounded questions of the data: What kinds of novels correspond to the prevalence of one field over the other? Can we understand these trends in novelistic language more directly as changes in the kinds of novels being written? We see this process of translating data into meaningful forms as a key tactic in digital humanities work. This process follows the same kind of dialogic movement that we have pursued throughout our methods; turning to the novels helps us interpret the data in terms meaningful to literary history, while turning to the data helps us see literary history in new ways.

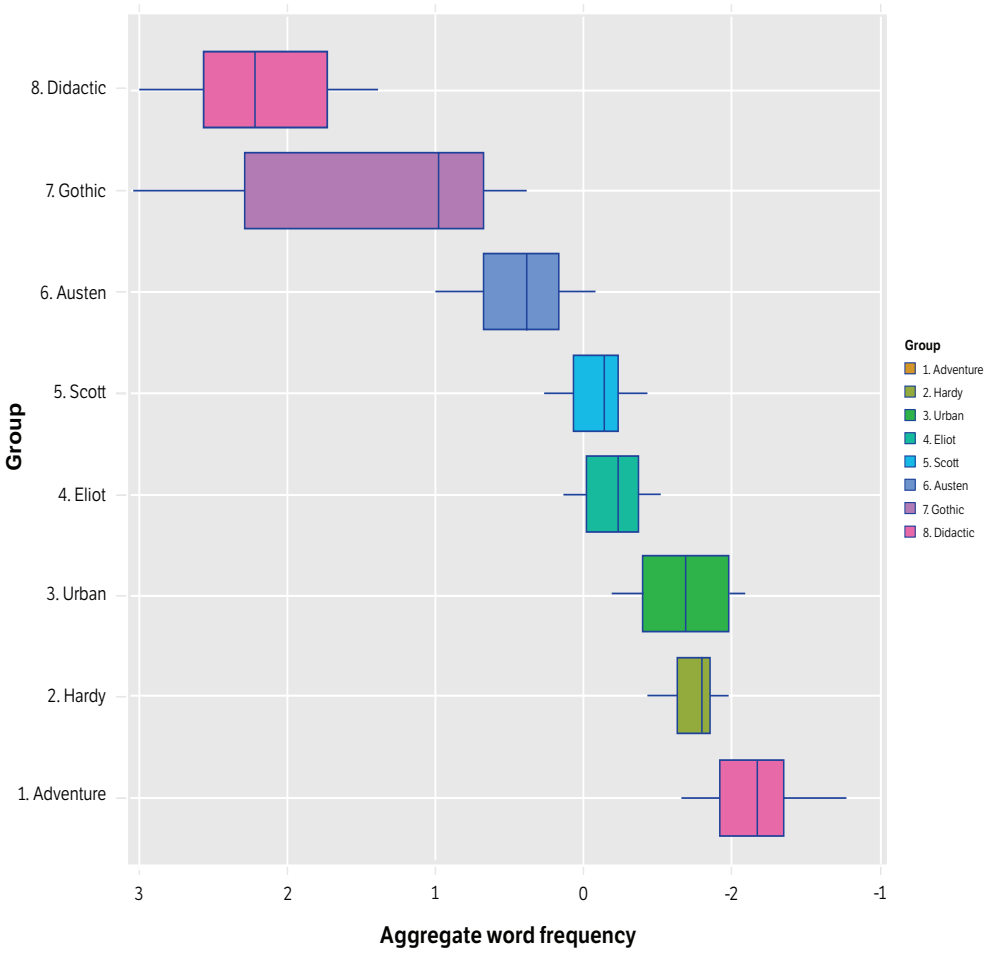

Figure 18: Spectrum of novels, authors, and genres as ranked by concentration of the abstract values fields. The x-axis shows number of standard deviations above the corpus-wide mean concentration of abstract values fields. For example, the median for the evangelical and didactic novels was around 2.25 standard deviations above the mean.

Ranking novels by their usage of the two fields indeed separates out clusters of genres and authors within the spectrum (see **Figure 18**).[16] From left to right, this shows novels with highest frequency of abstract values words to lowest (and, conversely, lowest frequency of "hard seed" words to highest). What we get is a distribution that begins at the extreme left with the evangelical novel, closely followed by the Gothic novel, then Jane Austen, Walter Scott, and George Eliot. Toward the right of the spectrum, we find the urban and industrial novel and Charles Dickens, and at the extreme right, a cluster of genres including adventure novels, fantasy, science fiction, and children's literature.[17] Given that

16 We will be focusing here on the spectrum produced by the concentrations of abstract values words. The spectrum produced by the "hard seed" words matched closely, not surprising given their strong inverse relationship.

17 We offer here some examples of the texts that make up the spectrum's genre clusters. The "Didactic" cluster includes novels such as Hannah More's *Coelebs in Search of a Wife*, Mary Brunton's *Self-Control*, William Godwin's *Things as They Are*, and Susan Ferrier's *Marriage*. The "Gothic" cluster includes novels such as Matthew Lewis's *The Monk*, Anne Radcliffe's novels, Percy Byshe Shelley's *Zastrozzi*, Kelly Isabella's *Madeline or the Castle of Montgomery*, etc. The "Urban" cluster includes the novels of Dickens, Gaskell, Gissing, Trollope, and others. The "Adventure" tag stands in for a conglomerate of genres: adventure fiction such as Robert Louis Stevenson's novels, H. Rider

it was generated quantitatively by the concentration of only two features of novelistic language, and that obviously computers have no knowledge of authors or genres, it's amazing just how suggestive the spectrum is.[18] For instance it clusters city novels together and takes Eliot's works out of chronological order and places them back a generation closer to Austen's, bringing out an affinity that many readers and critics have felt. Because of this sensitivity to genres and authors, the spectrum allowed us to see the two historical trends in novelistic language as deeper shifts in narrative mode, changes in the kinds of novels being written.[19]

One shift easily seen from the spectrum is the physical spaces of the novel expanding. The distribution moves from the tight, domestic, and village spaces of the moralistic, Gothic, and Austenian rural novel to the cities of Dickens and the exploratory expanses of the adventure, science fiction, and fantasy novels. As an initial observation, this is interesting, but we wanted to move beyond this because it didn't synthesize all the data. Thus, we worked on triangulating the movement revealed by the spectrum with the trend data we'd already found. This might help us see the larger patterns at work.

5.2 Tracing a Decline: The Waning of a Social Formation

We began by mapping the abstract values fields onto the spectrum. Recall that those fields comprise highly polarized, explicitly evaluative words related to norms of social regulation. Mapping these characteristics onto the spectrum, we can see that the change shown here goes beyond a change in physical spaces. More fundamentally, it is a change in the social space of the novel. By the term social space, we mean to point to the scale, characteristics, and force of the forms of social organization that structure the social worlds depicted in the novels. The change revealed here is an expansion from small, constrained social spaces to wider, freer ones. Think of the rigidity and tightness of social space in the evangelical, Gothic, or village novels where the character systems are limited to families

Haggard's novels, R. M. Ballantyne's *The Coral Island*; science fiction works such as H. G. Wells's *The Time Machine* and Richard Jefferies's *After London*; children's literature such as Richard Jefferies's *Bevis*, and fantasy works such as Lewis Carroll's novels and George MacDonald's *The Princess and the Goblin*.

18 This spectrum of novels produced purely through quantitative measures suggests the tantalizing possibility of categorizing genre in quantitative terms, a goal pursued to an extent already by another project at the Stanford Literary Lab conducted by Sarah Allison, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. In "Quantitative Formalism," they present results that reveal quantitative generic signals at several levels: "genres, like buildings, possess distinctive features at every possible scale of analysis: mortar, bricks, and architecture [...] the mortar, the grains of sand, of Most Frequent Words, the bricks of Docuscope's lexico-grammatical categories, and the architecture of themes and episodes that readers recognize" (8). In our project, the categorization is done not by most frequent words or lexico-grammatical categories but by semantic fields, a different kind of feature that offers certain advantages of interpretability. We have not yet followed up this intriguing possibility, but it seems the nature of this research is the continual opening up of (too) many other directions to pursue.

19 An even more intriguing possibility that emerges from this spectrum was suggested to us by Franco Moretti, who pointed out that the most canonical authors and genres of the 19th-century British novel seem to cluster in a relatively narrow range in the middle of the spectrum while the more minor genres are literally at the fringes. It's as if there were certain features of the novel, in this case, kinds of novelistic language, that can cause an author to drop out of the running for canonization simply from using it too much or too little. In other words, that there might be some kind of acceptable range for these features beyond which you are put beyond the pale. Moretti also pointed out that this seems to contradict the prevalent image of minor works as flawed derivatives of major works. In this spectrum, there seems instead to be whole ranges of form that the canonized works do not even explore, perhaps an argument for the importance of exploring the archive beyond the canon for these underexplored ranges of literary history.

or small communities. In these small social spaces, social behavior, roles, and identity are visible, monitored and tightly constrained. Moving left to right along the spectrum we see an expansion toward wider, less constrained social spaces—rapidly growing cities, London. By the time we reach the cluster of genres at the extreme right, these science-fiction, adventure, and fantasy spaces have expanded outward so far that they move beyond society entirely to exotic islands, fantasy worlds, different eras, etc.

To verify this interpretation of the spectrum as a change in social space, we built on our earlier topic modeling data, this time categorizing for types of social space.[20] Again, topic modeling provided an independent, unsupervised method of identifying patterns in language use, a parallel dataset in which we could see if social space emerged as the determining variable. Such parallel testing was particularly important in this case given the limitations of interpreting the spectrum. The powerful strategy of translating our data into readily familiar and interpretable forms comes with unavoidable costs: the drastic limitation of the sample size to the relatively canonical texts we are familiar with; and the reliance of the interpretation on our subjective conceptions of the texts. To run this parallel test, we categorized each topic into one of six types of social space:

Intimate: a private social space of intimate, often romantic, relations;

Domestic: a domestic social space of relations between those in the family and household;

Familiar: a social space of familiar relations between friends and acquaintances;

Public: the social space of the public sphere and impersonal relations;

Extra-Societal: a social space lying outside the boundary of everyday society;

Uncategorized: no indication of a social space of any kind.

This categorization produced the results shown in table 4.

Social Space	1790 - 1830	1860 - 1900
Intimate	27%	11%
Domestic	15%	16%
Familiar	15%	14%
Public	12%	35%
Extra-Societal	0%	3%
Uncategorized	31%	22%

Table 4: The most frequent topics in 1790-1830 and 1860-1900 characterized by type of social space. Percentages refer to the percentage of topics in the period characterized as indicating that type of social space.

20 See Section 4.3 for an overview of our topic modeling procedures. For a fuller account and the complete data, please see the appendix online at http://litlab.stanford.edu/semanticcohort.

While the concentrations of domestic and familiar social spaces in the topics remain essentially unchanged, the major movement here is a shift in the distribution's center of gravity from intimate to public social spaces. This shift in emphasis corroborates what the spectrum suggests: a systemic expansion of social space in the novel across the century.

Thinking in terms of the abstract values, the tight social spaces in the novels at the left of the spectrum are communities where values of conduct and social norms are central. Values like those encompassed by the abstract values fields organize the social structure, influence social position, and set the standards by which individuals are known and their behavior judged. Small, constrained social spaces can be thought of as what Raymond Williams calls "knowable communities," a model of social organization typified in representations of country and village life, which offer readers "people and their relationships in essentially knowable and communicable ways" (*Country* 165).[21] The knowable community is a sphere of face-to-face contacts "within which we can find and value the real substance of personal relationships" (*Country* 165).[22] What's important in this social space is the legibility of people, their relationships, and their positions within the community. In these terms, it's easy to see how a unified system of social values and standards could undergird this legibility, providing a major scheme for making sense of people and their relationships, while shaping behavior to keep close interactions harmonious. Indeed, this general point is implicit when Williams characterizes Austen's novels as centered on "a testing and discovery of the standards which govern human behaviour" and the relation of these standards to an established social order of property and status (*Country* 113); this emphasis on conduct in Austen's work is heightened by the novels' setting within a "close social dimension" (117), in other words, a small social space. Within a small social space, the explicitly evaluative and highly polarized quality of the abstract values fields finds a natural home. Their explicitness and polarization provide clarity, clear-cut standards and categories, even binaries, for legibly representing and perceptually organizing a close community's social life. We can characterize the abstract values words as a kind of language well suited for producing social legibility, efficient engines for producing knowable communities.

---

21 The other influential term that comes to mind in describing such a social space is Ferdinand Tönnies's concept of *Gemeinschaft*, or community, in which "[u]nderstanding is based upon intimate knowledge of each other in so far as this is conditioned and advanced by direct interest of one being in the life of the other" (47).

22 It's important to note that in *The Country and the City*, Williams does not see as entirely accurate the characterization of rural communities as knowable communities. He is specifically speaking of the knowable community as a structure represented in novels, and, more broadly, as an idea. He devotes the chapter on knowable communities to complicating this model and interrogating what lies behind the point of view that would be invested in representing rural communities in this way in the nineteenth century. He points out that while within the novels Austen's communities are wholly knowable, as real communities, they are "precisely selective" (166). What is represented is not the whole social system but a network of propertied families linked by class. "Neighbors in Jane Austen are not the people actually living nearby; they are the people living a little less nearby who, in social recognition, can be visited" (166). The point is well taken and it helps us clarify that we are less interested in whether these novels accurately and comprehensively represent the social realities of community life in nineteenth-century Britain than in the social worlds as they are constructed within these novels, and the kinds of language used in that construction. Our argument suggests that the novelistic construction of "social recognition," which allows Austen to map a knowable network, may depend on the use of a kind of socially legible language. In fact, the highly polarized character of the abstract values fields would strengthen the systems of social recognition that map a knowable community by creating clear standards by which some may be excluded.

If this is how the abstract values fields are linked to a specific kind of social space, then we can make sense of their decline over the century and across the spectrum. The observed movement to wider, less constrained social spaces means opening out to more variability of values and norms. A wider social space, a rapidly growing city for instance, encompasses more competing systems of value. This, combined with the sheer density of people, contributes to the feeling of the city's unordered diversity and randomness. This multiplicity creates a messier, more ambiguous, and more complex landscape of social values, in effect, a less knowable community. Williams articulates this as a rural-urban dichotomy: "In the city kind, experience and community would be essentially opaque; in the country kind, essentially transparent [...] identity and community [in the city] become more problematic, as a matter of perception and as a matter of valuation, as the scale and complexity of the characteristic social organisation increased" (*Country* 165). Urban population growth, increasing division of labor, changing class relations—these and other factors made it more and more difficult to maintain the idea of a knowable community (*Country* 165). In such a social space, the values held by the city's multitudinous classes, communities, and subcultures overlap and conflict as much as the people making up those groups jostle, bump, and cross each other on the crowded streets. The sense of a shared set of values and standards giving cohesion and legibility to this collective dissipates. So we can understand the decline of the abstract values fields—these clear systems of social values organized into neat polarizations—as a reflection of their inadequacy and obsolescence in the face of the radically new kind of society that novels were attempting to represent. A transformation of the social space of the novel, even as urbanization, industrialization, and new stages of capitalism were drastically reshaping the actual social spaces of Britain. The decline we see in this kind of language is a trace of the waning of an entire form of social organization, an entire way of life, from the world of the novel.

The change is not a comfortable one. Alienation, disconnection, dissolution—all are common reactions to the new experience of the city. Wordsworth precisely articulates this in his description of London in the 1805 *Prelude*, seventh book:

> How often, in the overflowing streets,
>
> Have I gone forwards with the crowd, and said
>
> Unto myself, 'The face of every one
>
> That passes by me is a mystery.'
>
> ...
>
> And all the ballast of familiar life—
>
> The present, and the past, hope, fear, all stays,
>
> All laws of acting, thinking, speaking man—
>
> Went from me, neither knowing me, nor known. (258, 260)

Wordsworth brings out the experience of the city, the wide social space, as the experience of close proximity to an anonymous diversity of people, a seemingly endless stream of strangers. What happens to novels as they try to capture this experience? The effect

on character would be particularly strong. The protagonist or focalizer within this overwhelming social space finds himself in much the same position as Wordsworth: most everyone in the "overflowing" crowd is a stranger. With the absence of the knowable community's face-to-face relationships ("neither knowing me, nor known") and the dissolution of shared social values ("All laws of acting, thinking, speaking man / Went from me"), he has neither the knowledge nor the stable schema to place these strangers and in turn make sense of his position and relationship to them. The perceptual disorientation of the city corresponds to this breakdown in social legibility. Alongside this is a feeling of loss, the loss of the human connections that ground not only identity but the sense of mutual responsibility at the core of ethics and conduct. Seen another way, the anonymity of wider social spaces dissolves the social accountability and visibility that makes for the regulation of a tightly knit community. In capturing this experience, shifting its language and represented social space, the novel touches deeply on the historical and sociological changes in Britain: the shift from community (*Gemeinschaft*) to society (*Gesellschaft*), to use Ferdinand Tönnies's terms (33). By the mid-nineteenth century, Britain had become the first place in the history of world to have more people living in cities than in the country (Williams, *Country* 217). In the context of such transformations, it would be surprising not to see profound changes in the novel, an art form at the height of its powers and cultural importance.

5.3 Tracing a Rise: The Hard Seed Fields in Action and Setting

Having understood the change manifested in the decline of the abstract values fields, a question remains: why is there a correlation between that trend, an expanding social space, and the rise of the "hard seed" field? We present several possibilities. Given the sheer magnitude of the "hard seed" field, and the fact that it's far more semantically and conceptually diffuse than the abstract values fields, it shouldn't be surprising to find multiple factors at work.

We can begin by considering the experience of setting and character within urban and wider social spaces. To keep this grounded, let's look under the hood of our data at a few sample passages. For example, a passage from Great Expectations, which as a whole exhibits the highest concentration of the hard seed fields among the canonical city novels in our corpus.[23] With Pip leaving the marsh country of Kent to pursue his expectations in London, few novels represent the contrast between rural and urban social spaces so memorably. The interface between these two, Pip's first day in London, provides a stark encounter with the city, its spaces and people, in their concrete reality. Words from the hard seed fields are in bold, but note also the preponderance of other concrete words:

> Of course I had no experience of a London summer day, and my spirits may have been oppressed by the **hot** exhausted air, and by the dust and grit that lay thi**ck** on everything. But I **sat** wondering and **waiting** in Mr. Jaggers's close room, until I really could not bear the **two** casts on the shelf **above** Mr. Jaggers's chair, and got **up** and **went out**.

23 It's probable that among the close to 3,000 novels in our corpus, there are lesser known city novels exhibiting even more extreme concentrations of the hard seed fields, but, as we mentioned in Section 5.2, the usefulness of translating data into familiar forms always comes at this cost of leaving out the less familiar.

> When I told the clerk that I would take a **turn** in the air while I **waited**, he advised me to **go round** the corner and I should **come** into Smithfield. So, I **came** into Smithfield; and the shameful place, being all asmear with filth and fat and **blood** and foam, seemed to stick to me. So, I rubbed it **off** with all possible speed by **turning** into a street where I **saw** the great **black** dome of Saint Paul's bulging at me from **behind** a grim stone building which a bystander said was Newgate Prison. Following the wall of the jail, I found the roadway covered with straw to deaden the noise of passing vehicles; and from this, and from the quantity of people **standing** about, **smelling** strongly of spirits and beer, I inferred that the trials were on.
>
> While I **looked** about me here, an exceedingly dirty and partially drunk minister of justice asked me if I would like to step in and hear a trial [...] As I declined the proposal on the plea of an appointment, he was so good as to take me into a yard and **show** me where the gallows was **kept**, and also where people were publicly whipped, and then he **showed** me the Debtors' Door, **out** of which culprits **came** to be **hanged**: heightening the interest of that dreadful portal by giving me to understand that, "**four** on' em" would **come out** at that door the day after to-morrow at **eight** in the morning, to be killed in a row. This was horrible, and gave me a sickening idea of London: the more so as the Lord Chief Justice's proprietor wore (from his hat **down** to his boots and **up** again to his pocket-handkerchief inclusive) mildewed clothes, which had evidently not belonged to him originally, and which, I took it into my **head**, he had bought cheap of the executioner. **Under** these circumstances I thought myself well rid of him for a shilling. (165-66)

Of the 399 words in this passage, 41 are hard seed words. They account for 10.3% of the passage, a rate over two times the average in 1860-61 when *Great Expectations* was published. That percentage would be even higher if we included all the other concrete description words in the passage[24]. The hard seed words are truly integral to the linguistic fabric of this passage. So what are they doing?

We see three major uses: constructing setting, narrating actions, and characterization. From "the **hot** exhausted air" and "the dust and grit that **lay thick** on everything" to the "filth and fat and **blood** and foam" and "the Debtors' Door, **out** of which culprits **came** to be **hanged**," this language is instrumental in rendering the physical spaces and settings of this world. They help construct the city's diverse spaces—Jaggers's office, the cattle market at Smithfield, the roadway by Newgate prison, the facilities of the jail—place them in spatial relation, and bring them to life down to the particular ritual, spatial arrangement, and timing of the debtors' executions. Reading Dickens's description, we can almost map out these settings, following Pip's trajectory through the city. The spatial character and concreteness of the hard seed fields make them a language suited for constructing settings that are imaginable as physical spaces.

Within these spaces, of course, there are characters acting. So it's not surprising that another prevalent use of the hard seed words is for narrating actions and movements: "I sat wondering and **waiting**," "**got up and went out**," etc. Some proportion of the us-

24 Please see appendix C for a discussion of the limits we encountered in filling out some of the hard seed fields, hence the fact that our fields do not account for all the concrete description language in these passages.

age of hard seed words, particularly the action verbs and locative prepositions, belongs to this expected baseline of narrative rather than descriptive function. But in light of the construction of setting as imaginable spaces, this function takes on another significance. These action verbs and locative prepositions are more often than not spatial, interdependent with setting as they describe characters' movements within rendered physical spaces. For example, in the passage describing Pip's walk, almost all the hard seed words serving narrative functions are spatial movements: "When I told the clerk that I would take a *turn* in the air while I **waited**, he advised me to go *round* the corner and I should *come* into Smithfield. So, I came into Smithfield; and the shameful place, being all asmear with filth and fat and **blood** and foam, seemed to stick to me. So, I rubbed it **off** with all possible speed by *turning* into a street…" (165, emphases added). These actions don't merely take place within a setting; their spatial character actively contributes to the sense of the setting as an imaginable space. The integral role of the hard seed words in establishing this kind of setting suggests that part of their rising trend is tied to an increasing spatialization of setting in the nineteenth-century British novel. This spatialization is a growing treatment of setting as more than a functional backdrop, aspect of mood, or even a historically and socially specific place, but as a material space with physical dimensions, orientation, and constrictions within which characters act. This mode of setting holds that narrative unfolds within spaces. Description, in this mode, draws heavily on hard seed type words and continually makes the reader conscious of space. From the tightness of Jaggers's office to the single view of church and prison to the roadway directed by the walls of the prison, the specific spatial juxtapositions and layout of this part of London (as Dickens imagines it) are decisive elements in the passage; everywhere Pip turns, he deals with the brute materiality of the city as a space.

The physicality of setting. We may seem to have drifted a long way from our discussion of changing social spaces, but actually the two are closely related. That the rise in the hard seed fields comes in part from an increasing spatialization of setting is just one part of the story. The other part of the story is profoundly social. To see it, we have to consider that how setting is rendered may be related to what settings are being rendered. We observed earlier an expansion of physical spaces across the spectrum (see Section 5.1). This is another factor in the rise of the hard seed fields. As the physical spaces depicted within novels widen and become more varied, it makes sense that more concrete description language would be needed to render those worlds. The novelty, variety, and specialization of physical spaces being represented demands this language. But when we look at why and how this works, we can see that this change in physical spaces stems from the same transformations in social organization we've been tracing. These changes generated new, unfamiliar spaces calling for description, the drastically expanded London for instance. While the space of the city isn't new, urbanization reshaped it and brought it a new predominance. The consequent division of labor and differentiation of society led to more and more specialized spaces that needed to be rendered with particularity: factories, urban slums, railroad stations, police stations, professional offices, etc. But as Pip's walk through four distinct spaces in the span of three paragraphs reveals, it wasn't just the proliferation of spaces that mattered but their dense concentration and heterogeneous juxtaposition. The jarring contrasts created by urban density bring forward the particularity of each space, adding to this need for specific description. Another factor explaining the rise of hard seed words, then, are the tangible effects on setting of an expanding and diversifying social space.

5.4 Tracing a Rise: A Social Transformation in Character

The question we posed at the beginning of the last section still looms. The final piece that has been missing from our discussion is how all this relates to the decline of the abstract values fields. These pieces come together in the third function of hard seed words as a language of characterization. Characterization is the nexus where the abstract values enter the picture because they are a language profoundly invested in character, invested in describing, evaluating, and organizing character. By comparing the abstract values and hard seed fields as languages of characterization, we can see most clearly what's at stake in the shift from one to the other. The experience of social relation and the representation of character necessarily change as social space expands. Wordsworth captured the alienating effect of facing streams of strangers on the crowded city streets, but in the fully urban novels of Dickens we can see the effect this unfamiliarity and randomness have on the representation of people. Pip's introduction to London presents the city's streets and spaces as an encounter with unordered sensory details. People become just another one of these details. Pip sees a "quantity of people **standing** about, **smelling** strongly of spirits and beer" (165), an undifferentiated mass that, alongside the noise of vehicles, simply reinforces the passage's overall feeling of sordidness. Even when people are individuated, like the proprietor, their characterization reveals a change. The horror of London seems to manifest for Pip in the proprietor's character, specifically in his attire: "(from his hat **down** to his boots and **up** again to his pocket-handkerchief inclusive) mildewed clothes, which had evidently not belonged to him originally, and which, I took it into my **head**, he had bought cheap of the executioner" (166). Character here inheres almost exclusively in appearance. This mode of characterization works through surface and physical detail, fitting for a social space of strangers, random encounters, and sheer variety of people. With the breakdown of social legibility and human connection in the city and wider social spaces, one encounters people not just as strangers but as appearances, attire, anonymous bodies. Recall then the behavior of one of the "hard seed" fields, body parts words (**Figure 19**).

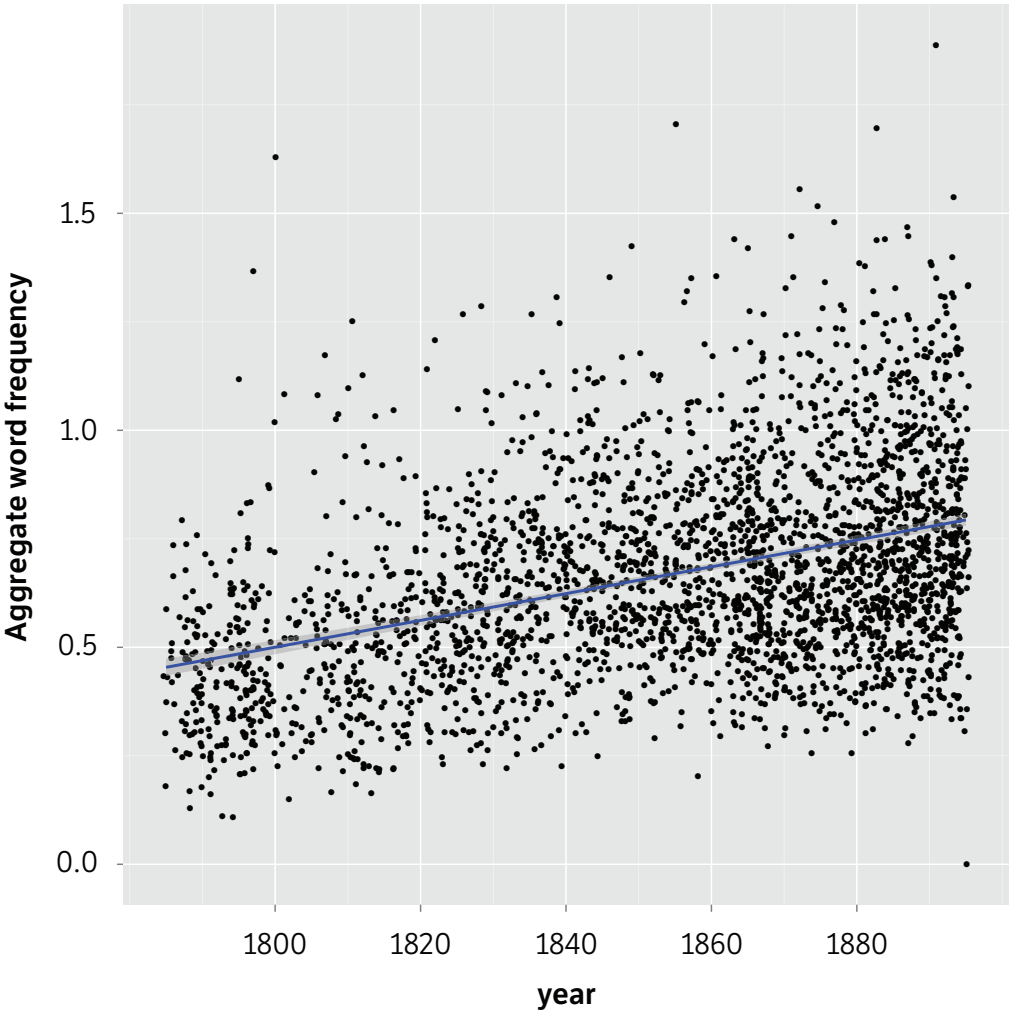

Figure 19: Aggregate term frequencies of the body parts field in novels, 1785-1900.

Its rise corresponds tightly with the shift to wider social spaces within novels. In the disorientation and unknowability of these communities, the attempt to place people within stable social schema comes up against surfaces of opaque physicality.

This comes through very clearly in another passage from *Great Expectations*, the first time Pip gets a good look at Mr. Wemmick. The introduction of a character is a highly saturated moment of characterization, so it's revealing to see the particular mode of description at work here. Again hard seed words are in bold, and note the prevalence of other concrete description words:

> Casting my eyes on Mr. Wemmick as we **went along**, to **see** what he was like in the light of day, I found him to be a **dry** man, rather short in stature, with a square **wooden face**, whose expression seemed to have been imperfectly chipped out with a dull-edged chisel. There were some marks in it that might have been **dimples**, if the material had been softer and the instrument finer, but which, as it was, were only dints. The chisel had made **three** or **four** of these attempts

> at embellishment over his nose, but had given them **up** without an effort to smooth them **off**. I judged him to be a bachelor from the frayed condition of his linen, and he appeared to have sustained a good many bereavements; for, he wore at least **four** mourning rings, besides a brooch representing a lady and a weeping willow at a tomb with an urn on it. I noticed, too, that several rings and seals **hung** at his watch-chain, as if he were quite laden with remembrances of departed friends. He had glittering **eyes** - small, keen, and **black** - and **thin wide** mottled **lips**. He had had them, to the best of my belief, from **forty** to **fifty** years. (171)

Two points jump out from this passage. First, the density of physical description. A lavish amount of attention is paid just to the topography of Wemmick's face, as well as to visual details such as the fraying of his linen and the number of mourning rings he wears. The hard seed words (which make up 11.4% of the passage) and other concrete nouns and adjectives are integral to this level of description. Second, notice how character emerges from this tableau of physical details. At no point does the passage directly state anything about Wemmick's character or personality. What we learn about Wemmick is implicit; it must be inferred from the physical details. This points to a general pattern where characters in the urban novel encounter strangers and must read them initially as bodies and appearances, and from there, may work to infer more about identity, character, and position. As inferences, these impressions are never entirely certain. Accordingly, Pip's language here is consistently couched in the subjective act of perception, interpretation, and inference: "I *found* him to be a dry man"; "whose expression *seemed*"; "I *judged* him to be"; "he *appeared*"; "I *noticed*"; "*as if* he *were*"; "to the best of *my belief*."[25] The tangible surfaces

---

25To counter the possible objection that this characterization through perception rather than definition is solely a consequence of the homodiegetic narration of this passage, consider the language in these descriptions from another hard seed-rich novel, Trollope's Phineas Finn, which is narrated from a heterodiegetic point of view (hard seed words are in bold):

"Her **eyes**, which were large and bright, and very **clear**, never seemed to quail, never rose and sunk or **showed** themselves to be afraid of their own power. Indeed, Lady Laura Standish had nothing of fear about her. Her **nose** was perfectly cut, but was rather large, having the slightest possible tendency to be aquiline. Her **mouth** also was large, but was full of expression, and her **teeth** were perfect. Her complexion was very bright, but in spite of its brightness she never blushed. The shades of her complexion were set and steady. *Those who knew her said* that her heart was so fully **under** command that nothing could stir her **blood** to any sudden motion. As to that *accusation* of straggling which had been made against her, *it had sprung from ill natured observation* of her modes of **sitting**. She never straggled when she **stood** or **walked**; but she would **lean** forward when **sitting**, as a man does, and would use her **arms** in talking, and would **put** her **hand** over her **face**, and pass her **fingers through** her hair—after the fashion of men rather than of women—and she seemed to despise that soft quiescence of her sex in which are generally found so many charms. Her hands and feet were large—as was her whole frame." (31, emphases added)

"Her **eyes** were large, of a dark **blue** colour, and very bright,—and she used them in a manner which is as yet hardly common with Englishwomen. She seemed to intend that you should know that she employed them to conquer you, **looking** as a knight *may have **looked*** in olden days who entered a chamber with his sword drawn from the scabbard and in his **hand**. Her **forehead** was broad and somewhat **low**. Her nose was not classically beautiful, being broader at the **nostrils** than beauty required, and, moreover, not perfectly **straight** in its line. Her **lips** were **thin**. Her **teeth**, which she endeavoured to **show** as little as possible, were perfect in form and colour. They who criticised her severely said, however, that they were too large. Her **chin** was well formed, and divided by a **dimple** which gave to her **face** a softness of grace which would otherwise have been much missed. But perhaps her great beauty was in the brilliant clearness of her dark complexion. You *might almost fancy* that you could ***see*** *into it so as to read the different lines **beneath** the **skin**.*" (303, emphases added)

of body, attire, manner are evocative, revealing, even symbolic, but they aren't definitive.

To see the distinctiveness of this kind of characterization, let's compare Wemmick's description to a similar character introduction in *Pride and Prejudice*, one of the canonical novels in our corpus that is lowest in concentration of hard seed words (2.96% vs. *Great Expectations*'s 7.17%) and highest in abstract values words (1.10% vs. *Great Expectations*'s 0.44%). In the first sustained description of Mr. Collins, Austen draws heavily on the abstract values words and a much more direct presentation of character, which we can see in the choice of verbs (abstract values words are in bold):

> Mr. Collins *was* not a **sensible** man, and the deficiency of nature *had been* but little assisted by education or society; the greatest part of his life *having been* spent under the guidance of an illiterate and miserly father; and though he belonged to one of the universities, he had merely kept the necessary terms, without forming at it any useful acquaintance. The subjection in which his father had brought him up, *had given* him originally great **humility** of manner, but it was now a good deal counteracted by the self-conceit of a weak head, living in retirement, and the consequential feelings of early and unexpected prosperity. A fortunate chance had recommended him to Lady Catherine de Bourgh when the living of Hunsford was vacant; and the **respect** which he felt for her high rank, and his veneration for her as his patroness, mingling with a very good opinion of himself, of his authority as a clergyman, and his rights as a rector, *made* him altogether a mixture of **pride** and obsequiousness, self-importance and **humility**. (104, emphases added)

There are no physical details to speak of and they aren't necessary because there's no need for perception or inference. The narrator tells us directly about Mr. Collins's character and identity, what he is rather than what he appears or seems to be. The description places him in the social schema of the abstract values words, which make up 2.9% of the passage. Mr. Collins is insensible, self-conceited, obsequious, and self-important, and by the standards of this social world, the valuation of those qualities is absolutely clear. This is characterization not as perception and interpretation but definition.

The contrast between the two descriptions is extreme. From Collins's description to Wemmick's, there is a complete disappearance of abstract values words and an increase in the frequency of hard seed words of almost 600%. Setting the two side by side lets us see our data trends more tangibly on the page as a change in the very linguistic texture of these novels. But it also lets us see clearly that this is more than a change in word choice, it's a change in representation. Where characterization in *Pride and Prejudice* is definitive, direct, and evaluative, in *Great Expectations* it is ambiguous and inferential; not only are the character traits implicit behind surfaces of physical detail, but the valuation of those traits is at a further remove from clarity. The passage from *Great Expectations*, the canonical city novel richest in hard seed words, then exemplifies a mode of characterization that presumes no direct access to character, a mode characteristic of the less knowable community of urban social spaces. If these samples are representative,[26] a major part of what

26 A serious "if" for sure. As always in this kind of research there is more work that can and should be done. The decision to cut off a research project to write up its results is always to some extent arbitrary. The change in characterization is suggested by the trends we found, coupled with the spectrum's revelation of expanding social spaces and the corroborating topic modeling data; however, one of the major directions we have not yet had time to follow up is to take a rigorous and representative sampling of passages, read them, and code them to further test the arguments we

we are seeing in the "hard seed" trend is a shift in characterization away from explicit comment, judgment, and placement within a value system to a more indirect mode of presenting bodies, appearances, details. A shift from direct to indirect characterization. With the decline of the abstract values fields, we saw a dissolution of the stable social schema organizing relations as social spaces became wider and more complex. In the aftermath, we find a mode of characterization that reflects a mode of social relation made radically new through its sudden ubiquity: the everyday encounter with hundreds of strangers. The shift from the direct characterization underwritten by the abstract values fields to the indirect characterization underwritten by the hard seed fields is a major change in representation. But behind this, a more fundamental change in perceptions of social formation, from one that presumes the epistemological luxury of transparent social knowledge, or what may amount to the same thing, a transparent social order, to one that no longer has this luxury.

We argued earlier that the hard seed fields' usage for setting revealed a change not just in *how* setting is rendered but *what* settings are rendered. As a consequence of the same underlying social transformations, characterization undergoes a similar process. We've established the change in how characters are represented but there remains the question of change in what kinds of characters are represented. The challenge presented by the characters in a wider, more complex social space is finding a language adequate for capturing these messy, conflicting, multitudinous social landscapes. Recall how the abstract values fields were revealed to be inadequate for doing this. The concrete language of the hard seed fields, however, can represent this greater range and openness of identity, position, and character. They can render ambiguity and variability very well. As we saw in the contrast between Wemmick's and Collins's descriptions, the language of materiality isn't explicitly valued, so it isn't organized into rigid categories and binaries like the axes of abstract values; correspondingly, it doesn't categorize its objects along these rigid axes. Instead, it sets out an enormous, almost infinite range of non-hierarchical nuances and differences. Rather than "moral" and "immoral," we have "hard," "rough," "liquid," "sharp," "stiff," "crooked," and so on. If we keep in mind how physical detail can imply qualities, character, and identity, we can see how such language offers more range and specificity of human characteristics than a polarized, uniform field of social norms. The concrete fields then are a kind of language that can render a larger, more variegated character system, exactly what a novelist would need to express the character of more complex social spaces.

In thinking of character in wider social spaces, we should return to a range of the spectrum we've been neglecting in our discussion: the cluster of fantasy, adventure, science fiction, and children's novels at the extreme right. In many ways, the extra-societal spaces of these genres offer encounters with an even broader and wilder range of character types than the urban novel: cannibals, pirates, talking animals, princesses, goblins, citizens of future and past societies; the range of characters not contained by the stable social schema of the abstract values is nearly limitless. (It's pretty amusing, though, to imagine

---

are outlining here. This would be a time-consuming procedure, but a valuable one.

hordes of chaste cannibals, reserved goblins, and deferential pirates.) By attending to this extreme end of the spectrum, we see that the transformations we've been tracing do not end at the urban novel, important as it is in diagnosing some of the key changes; instead they place the urban novel within a larger pattern encompassing these other genres. Indeed, many of the arguments we've put forward are applicable to an even greater degree in these outlier genres. The abstract values fields would have little purchase in their exotic extra-societal settings. In "savage" adventure settings, the lack of social norms or laws is often a key source of conflict; in science fiction or fantasy settings, the protagonist is commonly dropped into entirely unfamiliar social worlds. The perceptual disorientation that the concrete field captures in urban novels is also at work in these unfamiliar adventure settings, which can be even more aggressively estranging than the modern city. The perspective of curiosity at entirely novel experiences and spaces in these genres would naturally lead to a heavy use of concrete language as these new settings are perceived and described.

Taking account of these other genre clusters on the spectrum is crucial. To theorize from our data, we have been emphasizing two paradigmatic social spaces in this discussion—tight rural spaces and wide urban spaces—but of course the spectrum is a spectrum not a binary, and both the abstract values and hard seed trends are long shifts not sudden transformations. The data shows a whole range in the use of these kinds of language, a range of linguistic positions that nevertheless follows a clear direction. This suggests a range of social spaces depicted in the novel even as the dominant movement is one of expansion. This spectrum of rural, urban, mixed, intermediary, and extra-societal novelistic spaces echoes the actual complexity of the transformations in British society. For the sake of clarity and space, we haven't emphasized some of the intermediary positions on the spectrum. Our arguments have picked out a few important points to represent the range of a larger movement. Keeping these complications in mind, remaining alert to nuances and potential outliers,[27] we have aimed to draw out the larger patterns in our data. After all, that is a big promise of quantitative literary history.

5.5 Conclusion: From Telling to Showing

So to conclude this discussion we should tease out one more macroscopic pattern that several threads laid out here have implied. The growing inadequacy of explicitly evaluative language, the change in characterization to an indirect mode of presenting concrete detail, the inversely related trends of the abstract values and "hard seed" fields—all of these point to an overarching shift in the novel's narration and style: a shift from telling to showing. Given the range of concrete description words in the "hard seed" cohort, many of which are not necessarily anthropocentric, we can see a broad change in the general mode of perceiving and representing people, objects, spaces, and actions in the novel. This change from abstract, evaluative language to concrete, non-evaluative language doesn't necessarily indicate the disappearance of evaluation. Given the patterns we've seen, it would be more accurate to say that the modes of evaluation and characterization

27 The case of Hardy's position on the spectrum was one of these outliers, though not a large one considering his novels account for 11 of our almost 3,000 texts. His relatively high usage of the hard seed fields despite predominantly rural settings affirms our suggestion that multiple variables are at work in the hard seed trend. Perhaps most crucial for explaining Hardy's case is the shift from telling to showing that we present in the conclusion of this section.

changed, moving from explicit to implicit narration, from conspicuous commentary to the dramatization of abstractions, qualities, and values through physical detail. As we saw with characterization in the urban novel, though, this change is not a direct translation of categories of abstract values into ones of concrete detail. The change in mode is a change in quality, toward a finer-grained, more variegated and complex range of characteristics. Simultaneously, the indirectness of the mode, in which there is no clear one-to-one correspondence of concrete detail to abstract quality, suits the ambiguity and flux of widening and changing social spaces.

An attempt to sum up then: a pervasive expansion of social space in the nineteenth-century British novel in reaction to parallel changes in the actual social spaces of Britain; a concomitant concretization of novelistic language that constructs, reflects, and critically responds to this change in social experience; a spatialization of setting; a move from direct to indirect characterization; a fundamental shift in narration from telling to showing. A complex system of changes where, in varying degrees, each shift simultaneously drives and is driven by the others. In the end, the complexity of the range of mechanisms and forces we've been tracing cannot be separated from just how large these patterns are. Spanning nearly 3,000 novels and encompassing about 5% of their language use, this data could have revealed little more than noise and random variation. But what emerges from the data is a system, a history of the novel with a definite shape. And that may be the most striking discovery of all.

Postscript: A Method Coming to Self-Consciousness

If we've stressed throughout this paper the necessity of methodological reflection, it's because we believe that such self-consciousness is the only way for an emerging field of research to refine itself and reach maturity. On a more individual level, it's also because self-consciousness (in the more anxiety-ridden sense) has defined the experience of this whole project for us as we struggled to figure out exactly what we were doing and if it made any sense. While we haven't yet dispelled that sense of self-consciousness, this is perhaps a good thing, as that feeling and the thinking it drove led us to figure out some things about how to do this kind of research. The purpose of this postscript then is to zoom out from our project and sum up the methodological lessons learned along the way that we feel pertain to the larger enterprises of digital humanities and quantitative cultural study. We know full well that we don't have all the answers. We only hope that these suggestions make some contribution toward the goal of a healthy disciplinary self-consciousness (in both senses of the word).

The general methodological problem of the digital humanities can be bluntly stated: How do we get from numbers to meaning? The objects being tracked, the evidence collected, the ways they're analyzed—all of these are quantitative. How to move from this kind of evidence and object to qualitative arguments and insights about humanistic subjects—culture, literature, art, etc.—is not clear. In our research we've found it useful to think about this problem through two central terms: *signal* and *concept*. We define a signal as the behavior of the feature actually being tracked and analyzed. The signal could be any number of things that are readily tracked computationally: the term frequencies of the 50 most

frequent words in a corpus, the average lengths of words in a text, the density of a network of dialogue exchanges, etc. A concept, on the other hand, is the phenomenon that we take a signal to stand for, or the phenomenon we take the signal to reveal.[28] It's always the concept that really matters to us. When we make arguments, we make arguments about concepts not signals. Few indeed would be interested in a key, overlooked difference in the term frequencies of the 50 most frequent words between two authors, but if, instead, we found a key, overlooked difference in authorial style, more ears would probably perk up. Authorial style is one of the tantalizing concepts that the term frequencies of most frequent words could potentially give us access to. The point here is that in the digital humanities, the interest and impact of our arguments rely on concepts, but what we can tangibly grasp and point to are merely signals. Thus, the problem of moving from numbers to meaning can be formulated more precisely as the problem of bridging the always-existing distance between the signals we have and the concepts we want them to represent.

There are two directions from which to attack this problem. We can work to bring the signals closer to the concepts, or, from the other direction, work to fit the concepts more closely to the signals. These two directions correspond to the two major methodological processes where we face challenges in the quantitative study of culture (around which we have organized this paper): 1. Identifying the signals—the process of experimental design and data collection, including defining one's object of study, choosing features to track, designing and running tests, and gathering data. 2. Building the concepts—the process of data analysis, interpretation, and argument.

With respect to the first direction of bringing signals closer to the concepts of interest, several strategies emerged in the course of our project. First, it's crucial to be as precise and conservative as possible when thinking about the signal in an experiment; in other words, not to make a signal more than it really is. Signals are always smaller and more modest in scale than the big concepts we want them to stand for. Such desire can make it tempting to over-read a signal, making it carry more meaning than it can reasonably support. A good example of this can be found in Aiden and Michel's Culturomics article, in which term frequency data showing the changing usage of year names—"1940," "1972," etc.—in a large corpus are taken to reveal a change in cultural memory (178-179). This is an example of building a big, weighty concept on the foundation of quite a modest signal. In this case, the distance between signal and concept is just too large a leap to make. If the goal is to reduce the distance between the signal and concept so that it can be bridged effectively, then first having an accurate and reasonable evaluation of the signal's scope is necessary.

Given that the modesty of a signal increases its distance from the concepts of interest, the other key strategy is to design experiments so that the signals produced are as clean and robust as possible, that is, signals that are as reliable and comprehensive as possible. This is certainly not an original point, but it's important enough to remind ourselves of it. In many ways this is the true test of experimental design, how close one can get the signal to match the concepts of interest. Cleanliness in text-based quantitative research can mean everything from considerations of OCR quality to identifying sources of data error and more. The comprehensiveness of the signal is where the scale of data really makes a difference. We can make very different kinds of arguments from signals based

28 Of course there are many ways to express these two ideas and clearly these terms are a little fuzzy, but their point is pragmatic, to let us think usefully about our methods.

on thousands of novels than from signals based on only a handful. In our project, this was why we needed a large corpus that approaches the magnitude of a comprehensive set of nineteenth-century British novels. But this push for magnitude was not just in our corpus-building but also in defining our objects and features. We spent so much time figuring out how to construct massive semantic cohorts because these larger, cleaner signals allowed us to investigate much larger concepts.

As for bringing concepts closer to signals, the process of interpreting data, in some ways this is even thornier. As a discipline, humanists are in the profession of interpretation, but in empirical data we are faced with an entirely different kind of text to read. There's a temptation when working with empirical data to jump to conclusions, possibly because it projects a certain truth-bearing aura. If anything, though, we found that there are important continuities from our familiar practices of interpretation, even with this entirely different kind of text. The same careful attention to nuance and complexity that humanists have developed in close reading texts pays dividends when close reading data.

The first lesson we learned in reading data may go against the familiar interpretive practices in the humanities, but is crucial nonetheless. This is to collaborate, especially with those who have had extensive training in working with data. There's no reason humanists need to reinvent the wheel when it comes to the protocols of statistics and data analysis when other fields have already figured out many of these issues. The second lesson is that it can help to translate and visualize the data in forms that are more immediately interpretable to us as scholars of literature and culture. For instance, the spectra of novels, genres, and authors that we constructed from our data were instrumental in helping us see the data in familiar terms and see our familiar categories in new ways.

The third lesson may be the most powerful. In moving from signal to concept, from numbers to meaning, what may be needed in fact is more numbers. More numbers and different kinds of numbers. As humanists, we may fear that gathering more quantitative data only moves us farther from the qualitative meaning we seek, but we're suggesting that having more kinds of data actually moves us closer to finding meaning. To understand this, we should first see how there's a way to be anecdotal in one's evidence even when working with massive corpora of millions of texts; the problem there is not where one is looking for a signal, but rather what kind of signal one is isolating from that massive archive. If we were to run an experiment on millions of texts but isolate the term frequency data for a single word, we can see the anecdotal limitations of that data for making any kind of larger argument. A single dimension of data is tricky to interpret because it can be explained in any number of ways. For example, when we first isolated the trend for the abstract values fields, we had a hunch that the decline we were seeing was a shift in value systems from late-eighteenth-century values to Victorian values. But when we contextualized that data with the inversely related trend of concrete description words, a very different picture emerged that diminished this initial hunch as a plausible interpretation. Going on to set these two sets of data against the generated spectra of novels and the topic modeling data led to further refinements of interpretation.

The experience of having to revise our interpretations taught us one of the common pitfalls of interpreting data: the tendency toward validation. Because it's difficult to bridge the distance between signal and concept, we tend to read data in terms of the concepts we already have at hand, in our case, a historical narrative that draws a clear line between Victorian and late eighteenth-century British society. Another clear example of the problem of validation can be seen in Dan Cohen's research as presented in "Searching for

the Victorians." In that piece, Cohen presents frequency plots of several words related to faith and concludes that his data confirms the Victorian crisis of faith (and thus seems reliable). Even a visual inspection of the plots, though, reveals that these words do not share any single, declining trend. Many of the words presented have distinct and complex trends, some rising in the Victorian era before declining. In such cases, a familiar concept is applied too hastily to the data, thus flattening the data's nuances and complexities. A troubling corollary to this is a tendency to throw away data that do not fit our established concepts. When Cohen discards a striking correlation between "belief," "atheism," and "Aristotle" as an accident of the data, he does just this. Whether or not the correlation is accidental should be decided by statistical analysis rather than the feeling that it doesn't make sense. If we required all data to make sense—that is, fit our established concepts—quantitative methods would never produce new knowledge. If the digital humanities are to be more than simply an efficient tool for confirming what we already know, if they are to be a method for making new discoveries and exploring the unexplored, then we need to check this tendency to seek validation.

In contrast to this validation model of working with data, what we are suggesting as our third lesson is not to throw away data but to aim for more data. We learned that the open-ended process of interpretation can be made more rigorous when it has to triangulate multiple dimensions of data and account for a wide set of related observations. This approach can be summarized as a hypothesis-testing mode of interpretation. Engaged in a constant dialogue between evidence and interpretation, hypothesis testing seeks to eliminate potential theories by testing them against multiple forms of data, resulting in a stronger argument. The focus shifts from whittling away data that do not fit our theories to whittling away theories that do not fit the data. We believe this model, when combined with robust, carefully collected data, can provide a new and powerful form of investigation for humanities scholarship.

This whole section has been focused on methodological pitfalls and ways to be more rigorous in our interpretations, but given all this, we still feel it important that the digital humanities not back away from asking the big questions and making the big claims. If the field is to be taken seriously as an integral part of the discipline, it will need to continue refining its methodologies, but more importantly, it will need to ask questions that are important to the field and make original arguments that push forward our current concepts and theories. For all we have learned from the admirable restraint the sciences show toward making conclusions, we find great value in the humanistic modes of argument that put forward possibilities and powerful ideas, which may not yet be conclusive or certain, but which drive further study and force us to look at what we thought we knew in new ways.

In a sense we want it both ways. The dual proposition to be rigorous in our experimental design, methods, and interpretations while pushing to ask the big questions and make the big claims may seem to be in tension, and they are. But this is the tension that is native to the digital humanities as they straddle two disciplinary models that have often been seen as antithetical to each other. So we've returned in the end to a major theme of this paper, the dialogic mindset and method that oscillate between the two disciplines. If there's one takeaway from the methodological journey of this process it is that. The digital humanities and those looking on as this emergent research develops can see this central tension as a problem. And too often this problem revolves around ugly issues of disciplinary turf

and encroachment. But to us it seems far better to try to get past such issues and, in the modest spirit of a field still figuring things out, take criticisms from both scientists and humanists as legitimate concerns, opportunities to learn to do what we do better. To do so, to strive to integrate the rich resources of both worlds is to explore the ways in which this tension, more than anything, can be productive and full of possibility.

References

Allison, Sarah, Ryan Heuser, Matthew Jockers, Franco Moretti, and Michael Witmore. *Quantitative Formalism: An Experiment.* Stanford: Stanford Literary Lab, 2011. Print. Pamphlets of the Stanford Literary Lab 1.

Austen, Jane. *Pride and Prejudice.* Ed. Robert Irvine. Peterborough: Broadview, 2002. Print.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research 3* (2003): 993-1022. Print.

Borgman, Christine L. "Scholarship in the Digital Age: Blurring the Boundaries between the Sciences and the Arts." *Proceedings of Digital Humanities 2009: The 21st Joint International Conference of the Association for Literary and Linguistic Computing, and the Association for Computers and the Humanities*. 22-25 June 2009, University of Maryland. Ed. Kate Singer. College Park, MD: Maryland Institute for Technology in the Humanities, 2009. xvi. Print.

Cohen, Dan. "Searching for the Victorians." *Dan Cohen's Digital Humanities Blog*, 4 Oct. 2010. Web. 24 Jan. 2011.

Cohen, Patricia. "Analyzing Literature by Words and Numbers." *The New York Times* 4 Dec. 2010: C1. Print.

Crystal, David. "Semantics." *A Dictionary of Linguistics & Phonetics*. Ed. David Crystal. Oxford: Blackwell, 2003. 410-11. Print.

Dickens, Charles. *Great Expectations*. Ed. Charlotte Mitchell. London: Penguin, 1996. Print.

Manovich, Lev. "Activating the Archive, or: Data Dandy Meets Data Mining." *Proceedings of Digital Humanities 2009: The 21st Joint International Conference of the Association for Literary and Linguistic Computing, and the Association for Computers and the Humanities*. 22-25 June 2009, University of Maryland. Ed. Kate Singer. College Park, MD: Maryland Institute for Technology in the Humanities, 2009. xv. Print.

Michel, Jean-Baptiste, et al. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* 14 Jan. 2011: 176-82. Print.

Tönnies, Ferdinand. *Community & Society*. Trans. and Ed. Charles P. Loomis. New York: Harper, 1963.

Trollope, Anthony. *Phineas Finn*. Ed. Simon Dentith. Oxford: Oxford UP, 2011. Print.

Williams, Raymond. *The Country and the City*. New York: Oxford UP, 1973. Print.

Williams, Raymond. *Culture and Society: 1780-1950*. New York: Columbia UP, 1958. Print.

Wordsworth, William. *The Prelude* 1799, 1805, 1850. Ed. Jonathan Wordsworth, M. H. Abrams, and Stephen Gill. New York: Norton, 1979. Print.

Appendix A. Table of Data about Each Field

The following table indicates the magnitude, number of words, and correlation values for the semantic fields in our study. Column A indicates the percentage of the words in our corpus belonging to the respective field. Column B shows the number of words in the field after the initial word cohort was developed with semantic taxonomies, in other words, after stage 3 of our process. Column C shows the number of words remaining in the field after the statistical filtering of stage 4, which represents the final version of the field and is the basis for all further results. Column D indicates the average correlation coefficient for these words, while Column E indicates their median correlation p-value.

Field	[A] Percent of words in corpus	[B] Number of words after OED (stage 3)	[C] Number of words after filtering (stage 4)	[D] Average correlation coefficient	[E] Median correlation p-value
Social Restraint	0.19%	155	136	91%	0.002%
Moral Valuation	0.24%	124	118	92%	0.002%
Sentiment	0.17%	116	52	77%	0.157%
Partiality	0.01%	34	20	92%	0.002%
Abstract Values	*0.61%*	*429*	*326*	*88%*	*0.041%*
Action Verbs	1.99%	257	248	73%	0.742%
Body Parts	0.65%	147	111	71%	0.777%
Colors	0.13%	96	46	57%	6.160%
Locative Prepositions	1.09%	28	27	74%	0.499%3
Numbers	0.37%	46	44	73%	0.679%
Physical Adjectives	0.20%	32	32	79%	0.227%
Hard Seed	*4.43%*	*606*	*508*	*71%*	*1.510%*

Appendix B: Full List of Words in Each Semantic Cohort

The following appendices list the words included in each semantic cohort. Words are ordered from most to least frequent. A percentage following a word indicates its share of the aggregate frequency of the field; the absence of this percentage indicates that its share of the aggregate is less than one percent.

B1: Moral Valuation

character [8.7%], honour [8.0%], conduct [5.4%], respect [5.0%], worthy [4.3%], temper [3.2%], innocent [3.0%], shame [3.0%], admiration [2.9%], manners [2.9%], dignity [2.4%], guilty [2.3%], ashamed [2.2%], virtue [2.2%], sin [2.1%], moral [2.1%], wicked [1.8%], contempt [1.7%], respectable [1.5%], goodness [1.5%], admired [1.4%], principle [1.4%], reproach [1.3%], innocence [1.3%], disgrace [1.3%], admire [1.3%], reputation [1.3%], guilt [1.2%], vice [1.2%], merit [1.2%], esteem [1.2%], unworthy [1.1%], virtuous, foul, dignified, respected, despise, despised, malice, nobility, excellence, esteemed, wickedness, morality, integrity, malicious, valour, infamous, reproof, disgraceful, disgraced, sinful, malignant, sordid, notorious, shameful, contemptible, etiquette, infamy, righteous, iniquity, corruption, rectitude, baseness, perverse, corrupt, faultless, chaste, laudable, outrageous, pernicious, scandalous, despicable, guiltless, unpardonable, depravity, villainy, disdained, magnanimous, depraved, malignity, magnanimity, reprobate, misconduct, degenerate, tainted, ignominious, licentious, corrupted, heinous, ignominy, righteousness, flagrant, immoral, unwholesome, irreproachable, indecent, iniquitous, reprove, debased, chastity, reputable, inexcusable, debauchery, licentiousness, perverseness, untainted, debauch, badness, turpitude, incorruptible, ribaldry, corrupting, debauched, uncorrupted, undefiled, lewd, corruptible

B2. Partiality

correct [24.3%], prejudice [15.3%], partial [11.4%], disinterested [8.6%], partiality [7.8%], prejudiced [5.0%], detached [4.3%], bias [3.3%], impartial [3.2%], inveterate [3.2%], detachment [3.1%], bigotry [2.0%], disinterestedness [1.8%], prepossessed [1.3%], impartiality [1.3%], prepossession [1.1%], prejudicial [1.1%], bigot, prepossess, sober-minded

B3. Sentiment

heart [47.2%], feeling [15.3%], passion [7.2%], bosom [5.2%], emotion [4.2%], sentiment [2.2%], ardent [1.5%], coldly [1.5%], pang [1.2%], blushing [1.2%], affect [1.1%], passionately [1.1%], ardour, pathetic, sensibility, heartless, sentimental, fervour, vehemently, fervent, fervently, vehement, impassioned, zealous, unfeeling, ecstasy, ardently, insensibility, callous, fervid, hard-hearted, pathetically, feelingly, cold-blooded, tearless, ebullition, heartlessness, cold-hearted, unblushing, sentimentality, dispassionate, dispassionately, fervency, ardor, passioned, sentimentally, callousness, mawkish, unimpassioned, frigidity, unsentimental, sentimentalist, master-passion

B4. Social Restraint

gentle [6.9%], pride [6.0%], proud [5.7%], proper [4.3%], agreeable [3.5%], humble [2.8%], becoming [2.8%], sensible [2.4%], vanity [2.3%], gracious [1.7%], elegant [1.7%], vulgar [1.6%], delicacy [1.6%], reserve [1.5%], subdued [1.4%], mild [1.3%], reserved [1.3%], simplicity [1.3%], reasonable [1.3%], haughty [1.2%], caution [1.2%], courtesy [1.2%],

polite [1.2%], modest [1.2%], prudent [1.2%], coarse [1.2%], indulge [1.1%], sober [1.1%], propriety [1.1%], indulgence [1.1%], refined, decent, rash, politeness, cautious, boast, superiority, moderate, restraint, discretion, humility, courteous, restrain, excess, modesty, civility, gentleness, restrained, insolent, extravagant, deference, thoughtless, appropriate, impertinent, insolence, softness, excessive, refinement, judicious, impetuous, extravagance, considerate, gross, genteel, affectation, presumption, impertinence, discreet, condescension, improper, scrupulous, impudence, rudeness, conceit, impudent, decorum, conceited, orderly, pompous, presumptuous, profligate, decency, wanton, affront, impropriety, prodigal, condescending, moderation, ungracious, demure, uncouth, unrestrained, unbecoming, solicitous, indiscreet, arrogance, haughtiness, decorous, unseemly, temperate, petulance, arrogant, supercilious, inconsiderate, petulant, mildness, nicety, overbearing, ostentation, sobriety, urbanity, conformity, disorderly, intemperance, unassuming, indecent, indelicate, intemperate, unrelenting, seemly, circumspection, indecorous, prodigality, circumspect, wantonness, immoderate, incivility, indelicacy, boastful, rusticity, grossness, bragging, indecency, superciliousness, immodest, ungenteel

B5. **Action Verbs**

see [6.2%], come [5.2%], go [4.5%], came [3.9%], look [3.5%], let [3.2%], looked [3.1%], went [3.0%], saw [2.7%], put [2.5%], going [2.3%], get [2.2%], seen [2.2%], turned [2.0%], stood [1.8%], got [1.8%], looking [1.7%], work [1.7%], gone [1.7%], keep [1.4%], open [1.4%], sat [1.3%], coming [1.2%], lay [1.2%], turn [1.1%], kept, close, opened, ground, walked, stand, walk, seeing, lie, show, standing, turning, looks, broken, comes, run, wait, sit, sitting, ran, caught, waiting, getting, moved, watch, grew, closed, break., broke, showed, touch, watched, bent, touched, lying, works, putting, hung, walking, goes, waited, move, keeping, watching, dropped, opening, running, grown, hurt, eat, shown, catch, growing, drop, lifted, grow, leaning, lies, working, breaking, moving, worked, flung, pulled, touching, showing, hanging, rolled, hang, stands, swept, sees, knocked, turns, gets, fetch, knock, picked, leaned, walks, crept, letting, pull, bending, glimpse, slipped, slip, pick, closing, jumped, catching, crushed, lift, eating, keeps, shows, dropping, rolling, drops, roll, ate, stooped, lean, lifting, quivering, bend, pulling, puts, smell, runs, sweep, sits, jump, grows, hanged, eaten, leap, knocking, creeping, sweeping, gotten, lain, strolled, crush, leaped, strode, fling, touches, stooping, quivered, shivering, breaks, crack, kick, trip, creep, shiver, picking, swung, stoop, stroll, opens, tap, shivered, flinging, kicked, quiver, swinging, fetched, jumping, watches, hangs, cracked, leaping, crushing, swing, moves, lets, slipping, tapped, waits, leapt, kicking, glimpses, strides, rolls, scratch, strolling, tripped, tapping, slice, grinding, stride, scratched, crawled, crawl, catches, hurts, dropt, cramped, sweeps, crawling, lifts, slips, knocks, cracking, closes, tripping, eats, smelling, bends, scratching, grind, smells, pulls, hurting, leans, fetching, leaps, kicks, tilt, creeps, flings, jumps, slices, cracks, cramp, tilted, picks, trips, stoops, scratches, shivers, taps, tilting, crushes, swings, strolls, quivers, crawls, fetches, sliced, grinds, smelled, tilts

B6. **Body Parts**

eyes [12.8%], hand [12.5%], face [11.0%], head [8.6%], hands [6.5%], eye [3.4%], arms [3.2%], lips [3.0%], arm [2.9%], feet [2.6%], hair [2.4%], blood [2.2%], foot [1.7%], ear [1.5%], mouth [1.4%], ears [1.4%], cheek [1.3%], breast [1.2%], neck [1.1%], cheeks, brow, tongue, fingers, shoulder, shoulders, knees, teeth, forehead, nose, legs, throat, finger, lip, limbs,

flesh, knee, skin, nerves, beard, chest, leg, waist, chin, bones, lap, veins, heels, elbow, limb, nerve, palm, bone, eyelids, heel, fist, stomach, wrist, thumb, vein, elbows, skeleton, lungs, nostrils, muscles, skull, muscle, palms, toe, toes, jaw, liver, ribs, tooth, wrists, forefinger, ankle, hip, spleen, gorge, ankles, knuckles, bowels, marrow, thigh, sinews, dimples, thumbs, loins, belly, rib, nostril, eyeballs, shins, dimple, eyebrow, scalp, shin, eyelid, womb, sinew, thighs, gut, nape, kidney, jowl, loin, artery, instep, knuckle, gland, eyelash, intestine, pus, tendons

B7. Colors

white [17.7%], black [15.0%], red [8.6%], blue [8.4%], green [8.0%], gold [7.3%], grey [6.6%], brown [5.7%], silver [4.4%], yellow [2.7%], gray [2.3%], crimson [2.1%], scarlet [1.5%], purple [1.5%], pink [1.4%], sanguine, orange, ruddy, dun, whiteness, sable, blackness, azure, verdure, tawny, buff, tan, russet, saffron, sapphire, bice, fallow, redness, mignonette, celeste, indigo, chamois, carmine, topaz, castor, ciel, sorrel, nankeen, burnet, aqua, putty, teal, chartreuse, cerulean, puce, vermeil

B8. Locative Prepositions

out [16.2%], up [16.0%], over [10.1%], down [8.7%], away [6.7%], back [6.4%], through [6.0%], under [5.2%], off [4.9%], between [3.7%], within [2.3%], behind [1.8%], above [1.7%], along [1.7%], beyond [1.5%], around [1.3%], across [1.3%], beside, beneath, front, outside, amid, throughout, inside, toward, alongside, underneath

B9. Numbers

two [23.7%], 1 [12.1%], three [9.9%], 2 [4.7%], ten [4.1%], thousand [3.9%], four [3.8%], five [3.7%], hundred [3.5%], six [2.8%], 3 [2.5%], twenty [2.3%], 5 [1.7%], pair [1.6%], 4 [1.5%], seven [1.5%], couple [1.5%], eight [1.3%], fifty [1.3%], twelve [1.3%], dozen [1.1%], nine [1.1%], 6, thirty, forty, fifteen, 11, eleven, eighteen, 7, 0, 8, sixteen, sixty, 9, seventeen, fourteen, 10, seventy, nineteen, million, thirteen, eighty, ninety

B10. Physical Adjectives

round [20.7%], hard [10.4%], low [9.8%], clear [7.4%], heavy [5.7%], hot [3.4%], straight [3.4%], wide [3.3%], sharp [3.3%], big [3.0%], thick [2.6%], rough [2.5%], slow [2.4%], thin [2.4%], empty [2.3%], apart [2.3%], dry [2.1%], bare [1.9%], wet [1.8%], clean [1.6%], loose [1.5%], flat [1.1%], wooden [1.1%], stiff, tight, dusky, backward, transparent, liquid, ripe, crooked, bushy

Appendix C. Semantic Categorization with the OED's Historical Thesaurus

To explain our procedure in using the OED's historical thesaurus requires a brief overview of the thesaurus's structure. The historical thesaurus is a semantic taxonomy of all the word senses in the OED. As a taxonomy, it is organized in a tree-like structure starting with three root categories: the external world, the mind, and society. Each of these root categories is divided into smaller and smaller "branches" until it breaks down to individual word senses. For example, the sense of "integrity" that means the lack of moral corruption is categorized as society > morality > virtue > absence of moral flaw. In the structure of the historical thesaurus, word senses that are closely related in meaning cluster together in the same branch or nearby branches. For instance, "rectitude," in the sense of conforming to standards of morality, lies on the next branch over from "integrity." Both word senses fall under the overarching category of virtue.

Our general procedure in using the OED's historical thesaurus was as follows. We would take the word cohorts generated by Correlator and look up those words in the historical thesaurus, noting their categorizations. These categorizations helped us more precisely identify the semantic content in these proto-semantic fields. To add to the fields, we would then draw words from those categories and from nearby "branches" that shared the same overarching categories. In selecting words, we would filter out ones that would be anachronistic for the period and ones too obscure to have substantial frequencies. In practice, this meant filtering out a substantial number of words; the OED is so exhaustive that many of the words have negligible frequencies. Remember as well that this step of filling out fields with the historical thesaurus was followed by a final step of statistical filtering, so some of the words added in the OED stage did not end up in the final semantic cohorts.

Below we record the semantic categories used in identifying and constructing each of our semantic fields. They offer a textured view of the semantic and conceptual content of our semantic fields.

C1. Semantic Categories in the Abstract Values Fields

C1.1 Moral Valuation

society > morality > [noun] > moral qualities or endowments >
society > morality > moral evil > [adjective] > immoral or unethical
society > morality > moral evil > moral or spiritual degeneration
society > morality > moral evil > wickedness >
society > morality > moral evil > wrong conduct
society > morality > moral fitness or propriety > moral impropriety >
society > morality > virtue > [noun]
society > morality > virtue > righteousness or rectitude
society > morality > virtue > morally elevated quality
society > morality > virtue > absence of moral flaw
society > morality > virtue > purity
society > the community > society in relation to customs, values, or beliefs > customs, values, or beliefs of a society or group >
society > religion > faith > spirituality > sin
the external world > abstract properties > action or operation >behaviour or conduct >
the mind > emotion or feeling > humility > feeling of shame >
the mind > mental capacity > contempt
the mind > mental capacity > esteem

C1.2. Partiality

the mind > mental capacity > mental acceptance, belief > expressed belief, opinion >bias, prejudice
the mind > mental capacity > faculty of knowing > conformity with what is known, truth > freedom from error, correctness >
the mind > emotion or feeling > absence of emotion > without emotion [adjective] > emotionally detached
society > morality > rightness or justice > wrong or injustice >wrong or unjust [adjective] > partial or biased
society > morality > rightness or justice > [noun] > impartiality
society > morality > virtue > morally elevated quality >unselfishness > unselfish [adjective] > free from personal interest

C1.3. Sentiment

the mind > emotion or feeling > [noun]
the mind > emotion or feeling > seat of the emotions
the mind > emotion or feeling > relating to the emotions [adjective]
the mind > emotion or feeling > in relation to/connected with the emotions [adverb]
the mind > emotion or feeling > sentimentality
the mind > emotion or feeling > absence of emotion
the mind > emotion or feeling > strong feeling or passion
the mind > emotion or feeling > zeal
the mind > emotion or feeling > capacity for emotional perception> sensitiveness or tenderness
the mind > emotion or feeling > manifestation of emotion >

C1.4. Social Restraint

the external world > abstract properties > action or operation >behaviour or conduct > good behaviour
the external world > abstract properties > action or operation >behaviour or conduct > good behaviour > restrained or moderate behaviour
the external world > abstract properties > action or operation >behaviour or conduct > good behaviour > [noun] > seemly behaviour or propriety
the external world > abstract properties > action or operation >behaviour or conduct > bad behaviour
the external world > abstract properties > action or operation >behaviour or conduct > bad behaviour > lack of moderation or restraint
the external world > abstract properties > action or operation >behaviour or conduct > a standard of conduct > [noun] > acting according to some standard, fashion, etc.
the external world > abstract properties > action or operation >manner of action or operation > lack of violence, severity, or intensity
the external world > abstract properties > action or operation >manner of action or operation > care, carefulness, or attention >caution
the external world > abstract properties > action or operation >manner of action or operation > carelessness
the external world > relative properties > order, orderliness >agreement, harmony, or congruity > suitability or appropriateness
the external world > sensation > physical sensibility >moderation in sensuous gratification
the mind > emotion or feeling > pride
the mind > emotion or feeling > humility
the mind > emotion or feeling > composure or calmness
the mind > mental capacity > understanding, intellect > wisdom, sagacity > prudence, discretion

the mind > aesthetics > good taste > pleasing fitness
the mind > aesthetics > good taste > refinement
the mind > aesthetics > bad taste > lack of refinement
society > morality > virtue > purity > chastity > modesty or decency
society > morality > moral evil > evil nature or character > lack of magnanimity or noble-mindedness > self-interest > [noun] >selfishness
society > morality > moral evil > licentiousness > profligacy, dissoluteness, or debauchery

C2. Semantic Categories in the Hard Seed Fields

A few of the hard seed fields presented particular challenges for the process of semantic categorization as fields like the action verbs and physical adjectives are linked less by meaning than by grammatical function. In those cases, filling out the fields with the historical thesaurus was nearly impossible as those words were distributed so diffusely across the taxonomy's branches. For these fields, we generally stuck with the words in the Correlator-generated word cohorts. Below we present the semantic categories used in identifying and constructing the fields where semantic categorization was feasible.

C2.1. Action Verbs

Not categorized.

C2.2. Body Parts

the external world > the living world > body > sense organ
the external world > the living world > body > part of body
the external world > the living world > body > external parts of body
the external world > the living world > body > skin
the external world > the living world > body > nail
the external world > the living world > body > hair
the external world > the living world > body > structural parts
the external world > the living world > body > speech organs

C2.3. Colors

the external world > matter > colour > named colours

C2.4. Locative Prepositions

the external world > abstract properties > space or extent >relative position
the external world > abstract properties > space or extent >distance or amount of distance
the external world > abstract properties > space or extent >extension in space
the external world > abstract properties > space or extent >direction

C2.5. Numbers

the external world > relative properties > number > specific numbers

C2.6. Physical Adjectives

Not categorized

Appendix D. Ranked List of Novels

The below data represent a ranked list of the novels in our corpus for their frequency of usage of the abstract values words. Although only the more canonical novels provided by Chadwyck-Healey are shown here due to space constraints, the entire corpus of 2,958 novels was first segmented into equal-sized groups called quartiles, each of which contains 25% of the novels. For a fuller listing of this data, please see http://litlab.stanford.edu/semanticcohort.

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
Fourth quartile (Q4): Top 25% novels by frequency of Abstract Values				
1799	Hays, Mary	The Victim of Prejudice	2.02%	2.47%
1796	Hays, Mary	Memoirs of Emma Courtney	1.99%	2.34%
1806	Dacre, Charlotte	Zofloya or The Moor	1.79%	3.09%
1809	More, Hannah	Coelebs in Search of a Wife	1.62%	2.49%
1796	Inchbald, Mrs.	Nature and Art	1.56%	2.71%
1797	Robinson, Mary	Walsingham or The Pupil of Nature	1.54%	2.97%
1798	Shelley, Mary Wollstonecraft	The Wrongs of Woman Or Maria	1.41%	3.42%
1805	Godwin, William	Fleetwood Or The New Man Of Feeling	1.37%	2.92%
1792	Bage, Robert	Man As He Is	1.36%	2.54%
1810	Shelley, Percy Bysshe	Zastrozzi	1.33%	3.56%
1799	Godwin, William	St Leon A Tale of the Sixteenth Century	1.28%	2.38%
1796	Bage, Robert	Hermsprong or Man As He Is Not	1.27%	2.56%
1837	Shelley, Mary Wollstonecraft	Falkner	1.27%	3.81%
1810	Brunton, Mary	Self Control	1.26%	2.75%
1811	Shelley, Percy Bysshe	St Irvyne or The Rosicrucian	1.26%	2.96%
1792	Holcroft, Thomas	Anna St Ives	1.26%	2.60%
1794	Godwin, William	Things As They Are	1.25%	2.65%
1805	Opie, Amelia Alderson	Adeline Mowbray Or the Mother And Daughter	1.23%	3.17%
1835	Shelley, Mary Wollstonecraft	Lodore	1.18%	3.44%
1788	Shelley, Mary Wollstonecraft	Mary	1.17%	3.31%
1794	Holcroft, Thomas	The Adventures of Hugh Trevor	1.16%	2.74%
1790	Radcliffe, Ann Ward	A Sicilian Romance	1.14%	2.63%
1791	Inchbald, Mrs.	A Simple Story	1.14%	3.09%
1800	Moore, John	Mordaunt	1.13%	1.97%
1796	Lewis, M. G.	The Monk	1.10%	3.35%
1794	Austen, Jane	Lady Susan	1.08%	2.24%
1788	Smith, Charlotte Turner	Emmeline the Orphan of the Castle	1.07%	3.08%
1840	Thackeray, William Makepeace	A Shabby Genteel Story	1.05%	4.45%
1796	Burney, Fanny	Camilla or A Picture of Youth	1.03%	3.64%
1818	Ferrier, Susan	Marriage	1.02%	3.21%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
1813	Austen, Jane	Pride and Prejudice	1.00%	2.96%
1895	Allen, Grant	The Woman Who Did	0.98%	4.74%
1832	Lytton, Edward	Eugene Aram	0.98%	4.24%
1795	Fenwick, E.	Secresy Or The Ruin On The Rock	0.96%	3.24%
1814	Burney, Fanny	The Wanderer or Female Difficulties	0.96%	3.38%
1828	Lytton, Edward	Pelham Or The Adventures Of A Gentleman	0.95%	3.80%
1791	Radcliffe, Ann Ward	The Romance of the Forest	0.93%	2.84%
1830	Shelley, Mary Wollstonecraft	The Fortunes of Perkin Warbeck	0.92%	3.64%
1810	Porter, Jane	The Scottish Chiefs	0.89%	4.03%
1849	Froude, James Anthony	The Nemesis of Faith	0.89%	4.14%
1797	Radcliffe, Ann Ward	The Italian	0.88%	2.63%
1823	Shelley, Mary Wollstonecraft	Valperga	0.88%	3.65%
1811	Austen, Jane	Sense and Sensibility	0.87%	3.14%
1885	Meredith, George	Diana of the Crossways	0.86%	4.14%
1844	Disraeli, Benjamin, Earl of Beaconsfield	Coningsby or The New Generation	0.86%	2.71%
1826	Shelley, Mary Wollstonecraft	The Last Man	0.86%	3.76%
1848	Bronte, Anne	The Tenant of Wildfell Hall	0.85%	4.48%
1837	Trollope, Frances Milton	The Vicar of Wrexhill	0.85%	3.70%
1814	Barrett, Eaton Stannard	The Heroine Or Adventures Of Cherubina	0.84%	4.99%
1891	Meredith, George	One of Our Conquerors	0.84%	4.35%
1831	Shelley, Mary Wollstonecraft	Frankenstein Or The Modern Prometheus [3rd ed.]	0.83%	2.94%
1818	Austen, Jane	Northanger Abbey and Persuasion	0.83%	3.51%
1857	Bronte, Charlotte	The Professor	0.82%	5.23%
1872	Linton, E. Lynn (Elizabeth Lynn)	The True History of Joshua Davidson	0.82%	4.68%
1836	Gore, Mrs. (Catherine Grace Frances)	Mrs Armytage or Female Domination	0.81%	2.84%
1821	Galt, John	The Ayrshire Legatees or The Pringle Family	0.81%	3.19%
1818	Austen, Jane	Northanger Abbey and Persuasion	0.81%	3.85%
1818	Shelley, Mary Wollstonecraft	Frankenstein or The Modern Prometheus	0.81%	2.83%
1858	Farrar, F. W. (Frederic William)	Eric or Little by Little	0.80%	6.41%
1814	Scott, Walter, Sir	Waverley or Tis Sixty Years Since	0.80%	3.15%
1814	Austen, Jane	Mansfield Park	0.80%	3.47%
1823	Scott, Walter, Sir	Quentin Durward in the Waverley Novels	0.80%	3.19%
1879	Meredith, George	The Egoist	0.80%	4.09%
1819	Scott, Walter, Sir	The Bride of Lammermoor	0.79%	3.24%
1823	Galt, John	The Entail or The Lairds Of Grippy	0.79%	3.54%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
Third quartile (Q3): Top-middle 25% novels by frequency of Abstract Values				
1847	Bronte, Anne	Agnes Grey	0.78%	4.37%
1818	Scott, Walter, Sir	The Heart of Mid Lothian	0.78%	3.51%
1794	Radcliffe, Ann Ward	The Mysteries of Udolpho	0.77%	3.37%
1861	Thackeray, William Makepeace	Lovel the Widower	0.77%	5.35%
1848	Thackeray, William Makepeace	Vanity Fair	0.77%	5.03%
1824	Hogg, James	The Private Memoirs And Confessions Of A Justified Sinner	0.77%	4.50%
1853	Gaskell, Elizabeth Cleghorn	Ruth	0.76%	5.74%
1816	Austen, Jane	Emma	0.75%	3.22%
1859	Meredith, George	The Ordeal of Richard Feverel	0.75%	5.28%
1867	Ouida	Under Two Flags a Story of the Household and the Desert	0.75%	6.26%
1862	Thackeray, William Makepeace	The Adventures of Philip on His Way Through the World	0.75%	4.76%
1819	Scott, Walter, Sir	Ivanhoe A Romance in the Waverley Novels	0.74%	3.60%
1823	Scott, Walter, Sir	St Ronan's Well in the Waverley Novels	0.74%	3.32%
1863	Oliphant, Mrs. (Margaret)	Salem Chapel Chronicles of Carlingford	0.74%	6.22%
1849	Bronte, Charlotte	Shirley	0.74%	5.13%
1826	Disraeli, Benjamin, Earl of Beaconsfield	Vivian Grey	0.74%	3.76%
1822	Galt, John	The Provost	0.74%	3.92%
1849	Thackeray, William Makepeace	The History of Pendennis	0.73%	4.85%
1861	Meredith, George	Evan Harrington	0.73%	5.06%
1821	Scott, Walter, Sir	Kenilworth in the Waverley Novels	0.73%	3.81%
1850	Bell, Robert	The Ladder of Gold	0.71%	4.82%
1817	Scott, Walter, Sir	Rob Roy in the Waverley Novels	0.71%	3.38%
1793	Smith, Charlotte Turner	The Old Manor House	0.70%	3.30%
1876	Meredith, George	Beauchamp's Career	0.70%	4.32%
1805	Austen, Jane	The Watsons	0.70%	3.81%
1858	Thackeray, William Makepeace	The Virginians	0.69%	4.56%
1819	Scott, Walter, Sir	A Legend of Montrose	0.69%	3.37%
1850	Thackeray, William Makepeace	Rebecca and Rowena	0.68%	4.56%
1821	Galt, John	Annals Of The Parish	0.68%	4.39%
1824	Scott, Walter, Sir	Redgauntlet in the Waverley Novels	0.68%	3.78%
1848	Dickens, Charles	Dombey and Son	0.68%	6.07%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
1854	Thackeray, William Makepeace	The Newcomes: Memoirs of a Most Respectable Family	0.68%	4.71%
1816	Scott, Walter, Sir	Old Mortality in the Waverley Novels	0.67%	3.73%
1872	Eliot, George	Middlemarch: A Study of Provincial Life	0.67%	4.67%
1856	Thackeray, William Makepeace	The Memoirs of Barry Lyndon	0.66%	4.23%
1858	Eliot, George	Scenes of Clerical Life	0.66%	5.70%
1864	Charles, Elizabeth Rundle	Chronicles Of The Schonberg Cotta Family	0.66%	3.76%
1863	Reade, Charles	Hard Cash: A Matter of Fact Romance	0.66%	6.61%
1876	Eliot, George	Daniel Deronda	0.66%	4.95%
1831	Peacock, Thomas Love	Crotchet Castle	0.66%	3.95%
1816	Scott, Walter, Sir	The Black Dwarf in the Waverley Novels	0.66%	3.72%
1857	Trollope, Anthony	Barchester Towers	0.65%	4.10%
1853	Bronte, Charlotte	Villette	0.65%	5.39%
1840	Thackeray, William Makepeace	Catherine	0.65%	5.08%
1815	Scott, Walter, Sir	Guy Mannering Or The Astrologer in the Waverley Novels	0.65%	3.97%
1888	Ward, Humphry, Mrs.	Robert Elsmere	0.65%	6.31%
1884	Besant, Walter	Dorothy Forster	0.64%	4.36%
1847	Disraeli, Benjamin, Earl of Beaconsfield	Tancred or The New Crusade	0.64%	3.19%
1893	Gissing, George	The Odd Women	0.64%	4.86%
1889	Stevenson, Robert Louis	The Master of Ballantrae	0.63%	5.21%
1887	Barry, William Francis	The New Antigone	0.63%	5.34%
1871	Lytton, Edward	The Coming Race	0.63%	3.20%
1845	Disraeli, Benjamin, Earl of Beaconsfield	Sybil or The Two Nations	0.63%	3.69%
1866	Eliot, George	Felix Holt The Radical	0.63%	4.99%
1817	Austen, Jane	Sanditon	0.62%	3.76%
1852	Thackeray, William Makepeace	The History of Henry Esmond Esq	0.61%	4.92%
1860	Eliot, George	The Mill on the Floss	0.61%	5.79%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
Second quartile (Q2): Bottom-middle 25% novels by frequency of Abstract Values				
1786	Beckford, William	Vathek Translated from the original French	0.61%	3.70%
1840	Trollope, Frances Milton	The Life and Adventures of Michael Armstrong the Factory Boy	0.61%	4.69%
1870	Disraeli, Benjamin, Earl of Beaconsfield	Lothair	0.61%	2.79%
1850	Kingsley, Charles	Alton Locke, Tailor and Poet	0.61%	5.40%
1815	Scott, Walter, Sir	The Antiquary in the Waverley Novels	0.61%	3.82%
1878	Payn, James	By Proxy	0.60%	3.97%
1849	Gaskell, Elizabeth Cleghorn	Mary Barton: A Tale of Manchester Life	0.60%	6.11%
1861	Reade, Charles	The Cloister and the Hearth: A Tale of the Middle Ages	0.60%	6.81%
1818	Peacock, Thomas Love	Nightmare Abbey	0.60%	3.55%
1844	Dickens, Charles	The Life and Adventures of Martin Chuzzlewit	0.59%	5.35%
1858	Trollope, Anthony	The Three Clerks	0.59%	4.65%
1861	Wood, Henry, Mrs.	East Lynne	0.58%	6.10%
1847	Bronte, Charlotte	Jane Eyre	0.58%	5.73%
1855	Gaskell, Elizabeth Cleghorn	North And South	0.58%	5.94%
1863	Eliot, George	Romola	0.58%	5.78%
1857	Dickens, Charles	Little Dorrit	0.57%	5.83%
1891	Gissing, George	New Grub Street	0.57%	4.86%
1853	Yonge, Charlotte Mary	The Heir of Redclyffe	0.57%	4.93%
1839	Dickens, Charles	The Life and Adventures of Nicholas Nickleby	0.57%	5.33%
1862	Sala, George Augustus	The Seven Sons of Mammon	0.56%	5.00%
1889	Gissing, George	The Nether World	0.56%	5.57%
1866	Gaskell, Elizabeth Cleghorn	Wives and Daughters	0.56%	5.26%
1854	Dickens, Charles	Hard Times	0.55%	6.24%
1858	Trollope, Anthony	Doctor Thorne	0.55%	4.30%
1855	Trollope, Anthony	The Warden	0.55%	4.49%
1841	Dickens, Charles	Barnaby Rudge	0.55%	6.09%
1847	Bronte, Emily	Wuthering Heights	0.54%	6.12%
1859	Collins, Wilkie	The Woman in White	0.53%	5.41%
1881	White, William Hale	The Autobiography of Mark Rutherfor Dissenting Minister	0.53%	4.29%
1874	Hardy, Thomas	Far from the Madding Crowd	0.53%	6.31%
1862	Collins, Wilkie	No Name	0.52%	5.53%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
1861	Trollope, Anthony	Framley Parsonage	0.52%	4.31%
1868	Collins, Wilkie	The Moonstone	0.51%	5.65%
1863	Gaskell, Elizabeth Cleghorn	Sylvia's Lovers	0.51%	6.58%
1874	Trollope, Anthony	Phineas Redux	0.50%	3.73%
1856	Yonge, Charlotte Mary	The Daisy Chain or Aspirations: A Family Chronicle	0.50%	5.18%
1859	Eliot, George	Adam Bede	0.50%	6.82%
1864	Thackeray, William Makepeace	Denis Duval	0.50%	5.23%
1880	Trollope, Anthony	The Duke's Children	0.50%	4.05%
1886	Stevenson, Robert Louis	Strange Case of Dr Jekyll and Mr Hyde	0.50%	5.55%
1841	Dickens, Charles	The Old Curiosity Shop	0.50%	5.86%

First quartile (Q2): Bottom 25% novels by frequency of Abstract Values

1850	Dickens, Charles	The Personal History of David Copperfield	0.50%	6.05%
1869	Blackmore, Richard Doddridge	Lorna Doone	0.49%	6.23%
1886	Burnett, Frances Hodgson	Little Lord Fauntleroy	0.49%	6.18%
1867	Trollope, Anthony	The Last Chronicle of Barset	0.49%	4.33%
1837	Dickens, Charles	The Posthumous Papers of the Pickwick Club	0.49%	5.67%
1870	Collins, Wilkie	Man and Wife	0.49%	5.80%
1865	Dickens, Charles	Our Mutual Friend	0.49%	6.33%
1861	Eliot, George	Silas Marner the Weaver of Raveloe	0.48%	6.23%
1866	Collins, Wilkie	Armadale	0.48%	5.71%
1870	Dickens, Charles	The Mystery of Edwin Drood	0.48%	5.93%
1891	Hardy, Thomas	Tess of the D'Urbervilles	0.47%	6.05%
1851	Borrow, George Henry	Lavengro the Scholar the Gypsy the Priest	0.47%	5.06%
1848	Newman, John Henry	Loss and Gain	0.47%	4.07%
1873	Hardy, Thomas	A Pair of Blue Eyes	0.47%	5.95%
1838	Dickens, Charles	Oliver Twist	0.47%	6.55%
1853	Dickens, Charles	Bleak House	0.47%	5.80%
1871	Black, William	A Daughter of Heth	0.46%	6.70%
1862	Braddon, Mary Elizabeth	Lady Audley's Secret	0.46%	6.04%
1886	Hardy, Thomas	The Mayor of Casterbridge	0.46%	6.08%
1876	Trollope, Anthony	The Prime Minister	0.46%	3.79%
1883	Broughton, Rhoda	Belinda	0.46%	6.32%
1873	Trollope, Anthony	The Eustace Diamonds	0.45%	4.15%
1859	Dickens, Charles	A Tale of Two Cities	0.45%	6.76%
1872	Butler, Samuel	Erewhon or Over the Range	0.44%	4.30%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
1887	Hardy, Thomas	The Woodlanders	0.44%	6.23%
1866	Kingsley, Charles	Hereward the Wake Last of the English	0.44%	6.18%
1864	Trollope, Anthony	The Small House at Allington	0.44%	4.66%
1872	Hardy, Thomas	Under the Greenwood Tree	0.44%	7.00%
1839	Ainsworth, William	Jack Sheppard	0.44%	4.86%
1875	Trollope, Anthony	The Way We Live Now	0.44%	4.34%
1892	Baring, Gould Sabine	In The Roar Of The Sea	0.43%	6.65%
1869	Trollope, Anthony	Phineas Finn	0.43%	4.22%
1861	Dickens, Charles	Great Expectations	0.42%	7.17%
1892	Grossmith, George	The Diary of a Nobody	0.41%	5.77%
1892	Hardy, Thomas	The Pursuit of the Well Beloved	0.41%	5.81%
1865	Trollope, Anthony	Can You Forgive Her?	0.41%	5.04%
1839	Taylor, Meadows	Confessions Of A Thug	0.39%	4.73%
1895	Ward, Humphry, Mrs.	The Story of Bessie Costrell	0.39%	8.62%
1854	Surtees, Robert Smith	Handley Cross or Mr Jorrocks's Hunt	0.38%	6.21%
1843	Dickens, Charles	A Christmas Carol	0.37%	6.78%
1878	Hardy, Thomas	The Return of the Native	0.37%	6.37%
1886	Stevenson, Robert Louis	Kidnapped	0.36%	6.93%
1880	Hardy, Thomas	The Trumpet Major	0.36%	7.10%
1857	Hughes, Thomas	Tom Brown's School Days	0.36%	8.34%
1834	Marryat, Frederick	Peter Simple	0.34%	5.85%
1891	Morris, William	News from Nowhere	0.34%	5.73%
1885	Jefferies, Richard	After London or Wild England	0.33%	6.11%
1887	Haggard, H. Rider (Henry Rider)	She: A History of Adventure	0.32%	6.89%
1890	Doyle, Arthur Conan, Sir	The Sign of Four	0.30%	7.09%
1878	Russell, William Clark	The Wreck of The Grosvenor	0.30%	7.88%
1871	Chesney, George Tomkyns	The Battle Of Dorking	0.29%	7.50%
1883	Schreiner, Olive	The Story of an African Farm	0.29%	9.79%
1858	Ballantyne, (R. M.)	The Coral Island	0.28%	6.47%
1863	Kingsley, Charles	The Water Babies	0.27%	8.71%
1895	Wells, H. G. (Herbert George)	The Time Machine	0.27%	7.88%
1885	Haggard, H. Rider (Henry Rider)	King Solomon's Mines	0.25%	8.14%
1877	Sewell, Anna	Black Beauty	0.25%	8.50%
1893	Stevenson, Robert Louis	Island Nights Entertainments	0.24%	8.55%
1871	MacDonald, George	At the Back of the North Wind	0.24%	8.03%

Date	Author	Title	Frequency of Abstract Values	Frequency of Hard Seed
1883	Stevenson, Robert Louis	Treasure Island	0.21%	8.48%
1866	Carroll, Lewis	Alice's Adventures in Wonderland	0.21%	8.32%
1872	MacDonald. George	The Princess And The Goblin	0.21%	7.98%
1872	Carroll, Lewis	Through the Looking Glass and What Alice Found There	0.17%	9.09%
1882	Jefferies, Richard	Bevis The Story of a Boy	0.11%	10.05%