Computational analysis of 20th century literature

Vivek Choksi
Stanford University
vchoksi@cs.stanford.edu

Chris Min
Stanford University
chrismin@stanford.edu

Karen Truong
Stanford University
truongk@stanford.edu

1 Abstract

We analyzed thematic and stylistic trends in a corpus of 355 popular and critically acclaimed 20th century English-language novels. First, we applied and adapted "the semantic cohort method", a vector space model meant to surface thematically similar words in a corpus; this method was originally proposed by the Stanford Literary Lab. Next, we studied trends in (1) the occurrence of words in these cohorts and (2) stylistic traits of novels, with the goal of demonstrating quantitative analysis' usefulness as a tool to enrich existing literary scholarship as well as surface new patterns in literature.

2 Introduction and the digital humanities

Our work can be considered part of "the digital humanities", a growing category of research that uses computation to study the humanities.

Digital humanities research aimed at studying literature has attracted both intrigue and criticism. Proponents, such as scholar and Stanford Literary Lab co-director Franco Moretti, are excited by the ability to research literature at unprecedented scale. Moretti's work is largely grounded on an idea that he has termed "distant reading". While humans are good at close-reading individual works of literature, there are too many books in the world for any team of humans to do a complete study. As a result, most books go unstudied. Distant reading is literary analysis at scale, performed using computation on massive amounts of text in lieu of human research.

On the other hand, critics of the digital humanities warn that computational analysis lacks the nuance of traditional literary scholarship. In the words of Yale Professor Katie Trumpeter, "[n]ovels are deeply specific, and the field has traditionally valued brilliant interpreters who create complex arguments about how that specificity works. When you treat novels as statistics, the results can be misleading, because the reality of what you might include as a novel or what constitutes a genre is more slippery than a crude numerical picture can portray" (Parry, 2010).

In light of this contentious discussion about the role of digital humanities research, we attempt to tap into the power of distant reading while being mindful of its limitations.

3 Previous work

Our project is inspired by the work of the Stanford Literary Lab, a group at Stanford that has been publishing research on "computational criticism" of literature since 2010. Since the Lab's genesis, it has published 12 "pamphlets" of exploratory digital humanities research, ranging from a study in classifying novels by genre to a semantic analysis of World Bank reports. Below, we review two of their pamphlets. The first describes a method to identify semantic trends in literature over time, and the second describes a corpus comprising "20th century canon". In our research, we adapt and expand on the Lab's semantic cohort method and apply this method to the Lab's 20th century corpus.

3.1 The semantic cohort method

In the Lab's fourth pamphlet, Ryan Heuser and Long Le-Khac describe their quantitative analysis of British novels of the nineteenth century aimed at uncovering semantic trends in British novels, particularly relating to "social restraint," "moral valuation," "partiality," and "sentiment" (Heuser and Le-Khac, 2012). They iteratively develop a methodology for constructing what they call the "semantic cohort," or "a group of words that are semantically related but also share a common trajectory through history."

They develop the Correlator, a tool to cluster words into "semantic cohorts" using the words' normalized occurrence counts. This is essentially a vector-space model for word representation, where a word vector consists of normalized frequency of that word in the corpus for each decade, and where cohorts are constructed as groups of words whose vectors correlate highly. They use the Pearson product-moment correlation coefficient as the correlation metric.

Each cohort represents a set of historically correlated but not necessarily semantically related words, and from the cohort, the authors cull semantically incoherent words by referring to the Oxford English Dictionary's Historical Thesaurus, a semantic taxonomy of every word sense in the dictionary.

The results of the methodology are promising, with an average correlation coefficient of 88% and median correlation p-value of 0.0411%. The authors retrospectively describe their methodology as "dialogic," iteratively relying upon both historical and semantic measures to build semantic cohorts of words.

3.2 Stylistic features

We also draw inspiration from work in progress at the University of Iowa, where Loren Glass leads a team of researchers extending Mark McGurl's *The Program Era* through quantitative methods (Glass, 2016) (McGurl, 2009). In order to validate and further explore McGurl's argument that we must study the development of postwar American literature alongside the rise of creative writing programs, Glass and his group hope to map the influence of graduates of the Iowa Writers' Workshop, an early powerhouse of the field. Affiliates of the Workshop include many familiar names, such as Kurt Vonnegut and Flannery O'Connor. Perhaps of even greater influence, in the context of Glass' work, are the less known names who went on to start and lead other creative writing programs at colleges and universities across the US.

Glass hopes to better understand the influence of these Workshop teachers and alumni on their direct and indirect literary successors. With the goal of enabling quantitative comparisons between authors and thus working towards a broad idea of similarity in writing style, Glass and his team introduce the idea of a "Style Card". [1] By considering authors' stylistic features such as type-token ratio, median sentence length, and male-female pronoun ratio, we can make "apples to apples" comparisons. While we believe stylistic features certainly cannot capture the entirety of an author or work's style, we think such measures can act as good heuristics in initial quantitative inquiries into literature.

3.3 The 20th century corpus

In a later pamphlet from the Stanford Literary Lab, Mark Algee-Hewitt and Mark McGurl set out to develop a digital corpus of 20th century novels written in English (Algee-Hewitt and McGurl, 2015). In doing so, they take on several academic, practical, and perhaps even moral challenges. After deciding to represent the century's novels through around 350 works, for practical reasons, they point out that while a personally curated list may carry too many biases and idiosyncrasies, a completely random list would be impractical to construct and likely pass over many if not all "canonical" novels. The corpus that they settle on is what we study in our analysis.

Thus having decided to construct a representative corpus with a bias for canonical works, the authors suggest an intriguing workaround to the issue of bias in any list of the best 100 novels of the 20th century. Algee-Hewitt and McGurl propose to use a "found canon," where several lists—all presumably with their own biases—of best novels in the 20th century are combined into a modularized canon. Instead of attempting to avoid bias, the authors acknowledge it, and trust future users of the corpus to conduct their research accordingly. By tagging each novel in the corpus with the "best of" list(s) to which it belongs, the authors retain all information about potential biases. Thus, the corpus becomes flexible and dynamic, easily limited, expanded, or filtered.

Algee-Hewitt and McGurl themselves point out some areas for future exploration. They foreshadow another 20th century novel corpus, populated based on quantitative demographic comparisons to ensure more equal representation. They also suggest a corpus of novels most frequently cited by literary scholars as another method of capturing the idea of "canonical." We can also see many ways to use the corpus built by the authors to explore the differences in its sources. There are

[1] We note here that Glass' work is still in progress, and the direction and details of the Style Card idea remain very much in flux. Our understanding of it stems from a presentation given by Glass and his team to members of the Stanford Literary Lab, including Franco Moretti and Mark McGurl, at Stanford on May 2, 2016.

several books that make it onto many of the lists, and others that do not–what makes these books distinctive? Can we characterize the books of specific lists?

4 Dataset

The Stanford Literary Lab granted us access to the 20th century corpus that we describe in Section 3.3. This corpus contains 355 novels written in English between the years of 1881 and 2011. These novels represent an aggregate of five "best of" lists, where each list comprises approximately 100 books from the 20th century (though a handful of these books fall outside of the years 1900-1999).

In more detail, the lists are:

1. Modern Library Boards List of 100 Best Novels of the 20th Century (ML Editors)

2. Modern Library Readers List of 100 Best Novels of the 20th Century (ML Readers)

3. Radcliffe's Rival List of the 100 Best Novels of the 20th Century (Radcliffe)

4. Larry McCaffery's List of the 100 Best Novels of the 20th Century (Experimental)

5. The yearly best-selling works of the 20th Century (Publishers Weekly)

The novels on these lists have significant overlap, as shown in Figure 1. The Literary Lab also provides metadata for each book and its author: the year the novel was published as well as the author's gender and nationality.

Lists 1, 3, and 4 consist of books selected by experts and ranked in order. List 2 consists of books selected by a group of readers and ranked in order. List 5, the Publishers Weekly list, is markedly different than the other lists in that the books were not consciously selected by experts or readers; they were selected based on sales data. Also, the books on this list are not ranked, and there is exactly one book selected from each year (i.e. the bestseller for that year).

There are too many books in the corpus to list here, so we will name just five: the top-ranked book from lists 1-4 and the most recent book from list 5.

1. Ulysses (James Joyce)

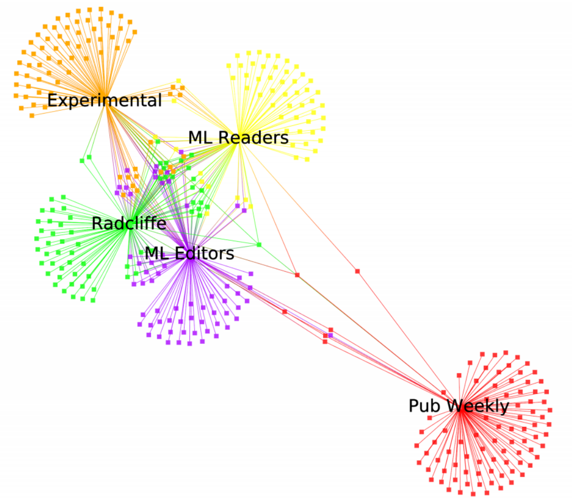

Figure 1: Force-directed graph showing the overlap of novels in the corpus's five lists, as depicted in The Stanford Literary Lab's 8th pamphlet. (Algee-Hewitt and McGurl, 2015)

2. Atlas Shrugged (Ayn Rand)

3. The Great Gatsby (F. Scott Fitzgerald)

4. Pale Fire (Vladimir Nabokov)

5. The Testament (John Grisham)

5 Methodology

We take an exploratory approach to our analysis rather than focusing on a single task, and our explorations broadly involve two distinct tools and both unsupervised and supervised approaches.

In our investigation, we attempt to emulate the careful approach that The Stanford Literary Lab takes. In particular, the Lab's work tends to successfully (1) start from hypotheses that are grounded in traditional literary analysis, (2) focus on specific corpora, and (3) avoid drawing conclusions that overestimate the interpretive power of the methodology and corpus. We believe that the Lab's research is valuable in part because it is sensitive to criticisms of the digital humanities, and we hope to frame our methodology with this same sensitivity.

5.1 Tools

5.1.1 Semantic cohorts

In our initial approach, we apply the semantic cohort technique to our dataset. Following from the methodology of Heuser and Le-Khac (Heuser and Le-Khac, 2012), we first count occurrences of stemmed words in every book. We group the books both over time by decades of the 20th century as well as by their membership on various best-of lists. This allows us to represent each word as a vector, where each element of the vector is the normalized occurrence frequency of that word in a particular decade or on a particular list. Next, we construct a “Correlator” tool to calculate Pearson correlation coefficients between each pair of word vectors. The Pearson correlation between vectors x and y is calculated as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

The method yields groups of words (semantic cohorts) that trend in the same pattern over time.

5.1.2 Stylistic features

The original semantic cohorts method was meant to identify groups of words and surface trends in the thematic content of a corpus over time. However, English-language literature in the 20th century saw changes not only in thematic content but also in writing style. By measuring writing style according to features such as noun-to-adjective ratio, we would theoretically be able to confirm and/or surface long-term stylistic trends in the 20th-century “canon”. To this end, we analyze each book according to the following features.

- Noun-adjective ratio
- Sentence length
- Male/female pronoun ratio
- Vocabulary size
- Type-token ratio
- Various part-of-speech ratios
- Flesch-Kincaid grade level

The type-token ratio measures the richness of a text’s vocabulary, and is the ratio between the number of different words in a text (the number of ”types”) and the number of total words (the number of ”tokens”). The Flesch-Kincaid grade level derives from the Flesch-Kincaid readability score and measures the readability of a text. The grade level is normalized to indicate the difficulty of a text as a US school grade level. A text with a Flesch-Kincaid grade level of 10 would be appropriate material for a 10th grade student, while a text with a level of 15 would be suitable for a college junior. The grade level can be calculated as below:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+1.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

We identify part-of-speech related features using NLTK’s built-in POS tagger, while deriving syllable counts for Flesch-Kincaid grade levels from phonemes using the Carnegie Mellon University Pronouncing Dictionary, which is also included in the NLTK package.

5.2 Approaches

As the Literary Lab and Glass before us, we apply our aforementioned tools to test hypotheses proposed and widely accepted previously, but perhaps yet unverified at scale. The realities of close reading often limit individual scholars and even entire academic disciplines to studying only a select corpus of texts. We aim to validate our methods (and, more generally, the digital humanities) as tools to confirm and further extend the theories built upon the rigorous analysis of smaller corpora.

We also attempt to use computational methods to surface new insights and trends. By inverting the order of our process and studying the corpus with the goal of finding quantitatively interesting patterns, we hope to develop new hypotheses to explore either through further, more nuanced computational analysis or close reading. As such, digital humanities may serve as a double-edged tool, both confirming existing hypotheses and surfacing new ones.

5.2.1 Supervised

We broadly categorize any analysis motivated by a hypothesis or literary question as supervised. When using semantic cohorts, after identifying a particular trend we want to verify or explore, we establish a list of words we believe most closely associate with the given trend. After retrieving the semantic cohorts associated with each word in our list, we cull the list by hand, removing words unrelated to our topic. We attempted to use the Oxford English Dictionary’s Historical Thesaurus as did the Literary Lab, but found the thesaurus largely focuses on semantics before our era of study. Most entries describe historical meanings stretching back into the early- and mid-sixteenth century, while we considered the semantics of words in the 20th century exclusively. Given the relatively recent timeframe we consider and the lack of similar, academically vetted resources addressing more modern history, we decide to proceed with the Oxford English Dictionary (which itself does contain some historical definitons) without a supplementary historical thesaurus. With the resulting culled semantic cohort, we proceed to analyze its change across different groupings such as time and best-of lists.

We approach the supervised use of stylistic features similarly, identifying hypotheses to investigate before selecting the appropriate features to analyze.

5.2.2 Unsupervised

Given a set of seed words, the semantic cohort method can surface groups of related words and report trends in their occurrence counts in the corpus over time. However, in order to generate the most interesting results, the method relies on the researcher to input the most interesting seed words. We hypothesize that we could find some of the most interesting seed words in an unsupervised way, by automatically ranking words and cohorts by some metric that approximates "interestingness". We consider each word or cohort's frequency, normalized by total word count, using the following metrics, among others, to measure their fluctuation across variables such as time and best-of lists.

- Standard deviation
- Sum of (absolute) change
- Sum of positive change
- Sum of negative change

The motivation for studying fluctuation across time derives from hoping to find salient chronological trends. When considering differences across best-of lists, we make comparisons across the composition and nature of the lists themselves.

Unsupervised approaches with stylistic features again runs in a similar vein. We consider each feature and its correlation with time and best-of lists, as well as the authors' gender and nationality. We attempt to discover trends or outliers based on their quantitative remarkability, as opposed to preconceived notions of importance.

6 Results and analysis

6.1 Unsupervised results

The unsupervised approaches with semantic cohorts generally prove quite noisy, and require much selective interpretation. As an example, the cohort with greatest absolute fluctuation decade by decade was that of the word `saucepan`. The most correlated (stemmed) members of the cohort itself are, in order of decreasing correlation, `illustri, enter, handicraft, felt, conquer, revolt, penit, his, grave, ahoy`. While we see hints of ideas relating to conflict, we cannot draw strong semantic links between any member of the cohort and `saucepan`.

We see some more promising results when considering standard deviation across time. Within the top ten cohorts with greatest standard deviation, we find the cohort for `she: her, waitingroom, blubber, uninjur, indescrib, mettl, rosi, toddl, acquiesc, volley, cling, furnitur, obes, met, convict, incongru, brokenheart, abund, intox, dinner`. Although we find some noise, many of the terms relate to 20th century ideas of femininity: `waiting room, acquiesce, furniture` and `dinner` seem to reference the docile domestic woman, while terms like `uninjured, indescribable, cling` and `brokenhearted` conjure images of damsels in distress. The fact that the cohort for a term as common as `she` includes these words seems a strong indicator of their semantic relevance.

We see the cohort's frequency trend downward over time, which aligns with our understanding that such tropes declined in both usage and popular acceptance throughout the course of the century. We see a curious rise of the cohort from the 70's to 90's, which is difficult to explain without reading the books that cause the cohort's increase.

We turn our focus to unsupervised use of stylistic features and find that the most interesting trends had not the greatest but least change over time. We compare the the books on the ML Editors list to those on the ML Readers list and find that, perhaps contrary to cynics' suspicions, the books do not vary significantly in either their Flesch-Kincaid grade or type-token ratio.

However, we see that the two lists vary much

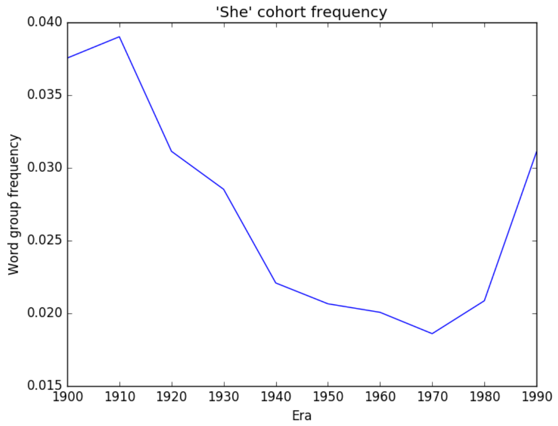

Figure 2: Normalized frequency of the semantic cohort seeded by the word ’she’.

List	Flesch-Kincaid Grade		Type-Token Ratio	
	Average	Std. Dev.	Average	Std. Dev.
ML Readers	5.011	2.30	0.0915	0.027
ML Editors	5.85	1.955	0.0955	0.035

Figure 3: Comparing ”readability” across the ML Editors and Readers lists.

more across time. The ML Editors list draws mostly from the earlier half of the century, while the ML Readers list covers more of the latter. We find evidence against the cynical intuition that the everyday reader would prefer “easier” novels. The two lists’ significant overlap in the middle of the century further supports the hypothesis that readers and editors share similar literary tastes more often than we may initially assume. We speculate that the editors’ focus on older books may come from a tendency for editors, like academics, to consider only works which have endured a certain period of critical evaluation to be eligible to join the canon. On the other hand, readers may be more likely to have a shorter memory and favor more recent novels.

We also note that Figures 4 and 5 exclude one outlier each. William Faulkner’s *Absalom, Absalom!*, from the ML Readers List, has a staggering Flesch-Kincaid grade of 17.601, while James Joyce’s *Finnegan’s Wake*, from the ML Editors List, has an impressive type-token ratio of 0.2657. The two outliers are perhaps unsurprising, given both Faulkner and Joyce’s reputations as difficult authors. However, it is reassuring to see our methodology naturally surface and verify such results.

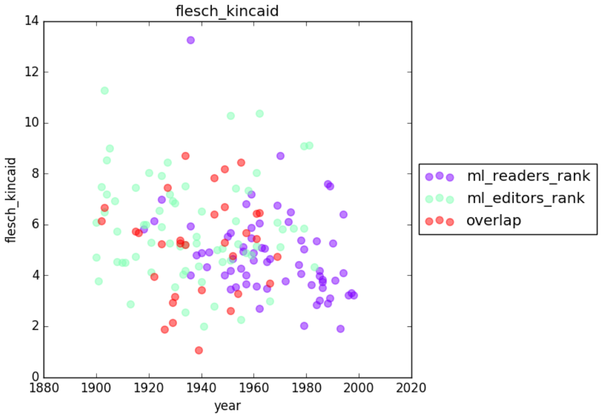

Figure 4: Flesch-Kincaid Grade in the ML Editors and Readers lists.

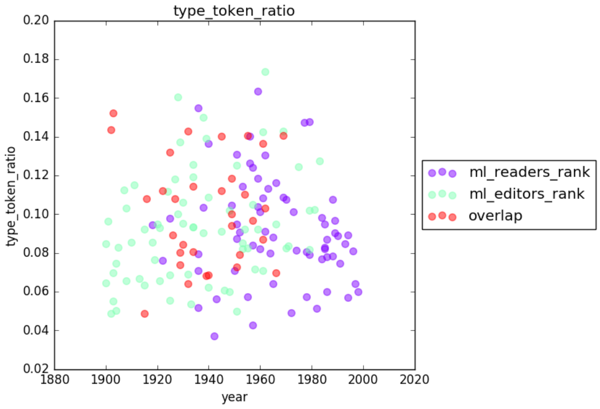

Figure 5: Type-token ratio in the ML Editors and Readers lists.

6.2 Supervised results

In our supervised task, we seek to further explore paths discovered in our unsupervised exploration. We also look for semantic cohorts and stylistic features that reflect historical events or movements of the 20th century, both from literary and broader historical perspectives.

6.2.1 Interpretations by gender

We further explore the thread begun with the `she` cohort by analyzing stylistic features by author gender. More specifically, we find that male vs. female pronoun usage by author varies significantly based on gender, as shown in Figure 6.

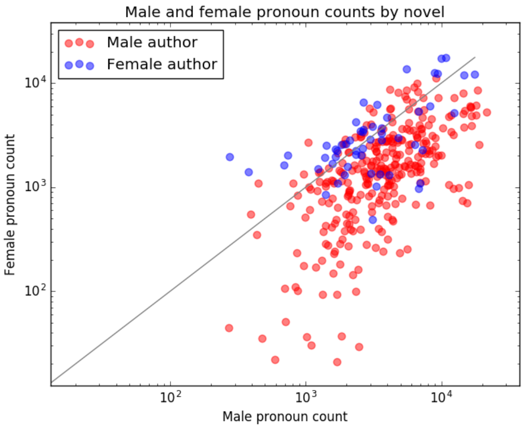

Figure 6: Male pronoun usage plotted against female pronoun usage on a log-log scale. We also plot a line of slope 1 (representing equal usage of male and female pronouns) for reference. Each data point is a book.

While the male authors' leaning towards using more male pronouns does not raise eyebrows, the clustering of female authors close to equal usage of male and female pronouns surprised us. We imagine the usage of male pronouns could be inflated because of the increasingly less-fashionable convention of using male pronouns in the generic case. Still, our findings both reflect suspicions that male authors write about men while raising the question of why we do not see a parallel in female authors. We believe these results could provide a quantitative springboard for those interested in gender representation in 20th century literature.

6.2.2 Interpretations by year

Given that the 1900s were rampant with international conflict, one theme we expect to see is the rise of militarism in literature. Using the semantic cohorts with frequency by decade, we examine the cohort seeded by the word weapon. The first 15 members of this cohort, in order of decreasing correlation, are: tank, binocular, plane, coordin, jack, armor, data, target, blast, trigger, branch, alert, fuel, tactic, survivor. Qualitatively, most of these terms are semantically related to weapon. Physical objects, like tank, binocular, and armor, can be considered hyponyms of weapon. Others seem to associate with actions one might take with weapons. Figure 7 illustrates an upward trend in the use of words in the 'weapon'-seeded cohort.

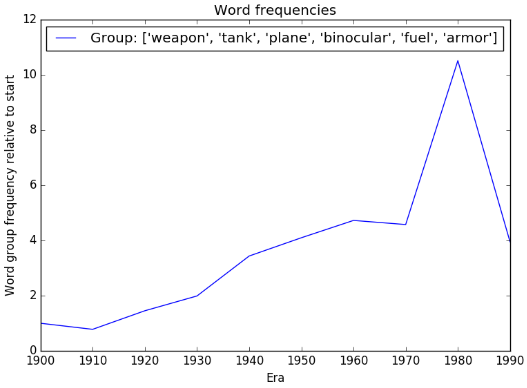

Figure 7: Normalized frequency of the semantic cohort seeded by the word 'weapon'.

The winding path that leads us to study the exquisite cohort and its subtleties perhaps best illustrates the potential for quantitative analysis as a tool when leveraged by human users. We begin by experimenting with the social restraint cohort as presented by Heuser and Le-Khac (Heuser and Le-Khac, 2012), wondering if we can find trends that continued from the 19th into the 20th century. While exploring the cohort's member's cohorts in a recursive fashion, we stumble upon the exquisite cohort, comprised mostly of "extreme" words such as absurd, dreary, and extraordinary. We are particularly struck by how the cohort contains words of both significantly positive and negative connotation, which leads us to believe something beyond pure semantics drives the co-occurrence of the words.

We split the cohort into two—one for positive words and another for negative words, while discarding ambiguous or neutral words—and consider their usage patterns separately. The positive cohort (exquisit, amiabl, extraordinarili, refin, inspir, philanthropi, extraordinari, eloqu, scrupul, generous, goodhumour, heroic, splendid, jocular, beauti, buoyant, magnific, chivalri, delight, genial, genius) has an average Pearson correlation coefficient of 0.9478 while the negative cohort (ridicul, absurd, condescend, dreari, prig, obstinaci, tiresom, abomin, sham, conceit, inferior, ignobl, roguish, abhorr, vulgar, shabbi, sordid, insens, inarticul, indol, obstin) has a Pearson correlation coefficient of 0.9507. We see that both cohorts, when considered over time, trend downwards. Although our hypothesis requires support from further analysis and close reading of the texts, we believe our observations may reflect the rise of minimal and realist literature in the postwar period and leading into the 21th century. We map the cohorts in Figures 8 and 9.

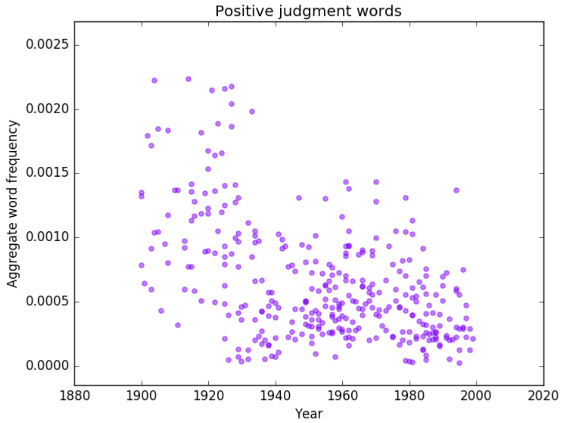

Figure 8: Positive words from the exquisite cohort. Each data point is a book.

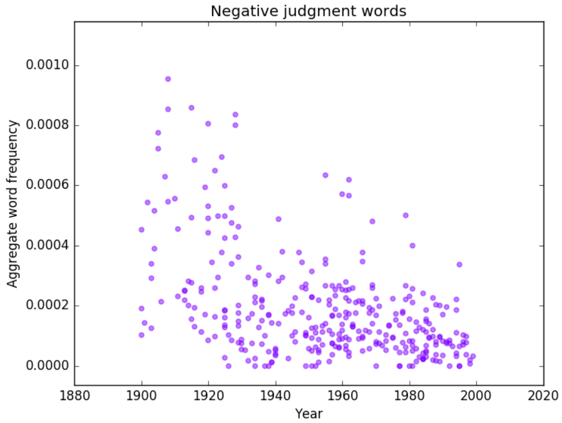

Figure 9: Negative words from the exquisite cohort. Each data point is a book.

7 Conclusion

Our discovery of both positive and negative "extreme" words trending downwards over time gives rise to many questions and opportunities, both within our realm and that of our predecessors. We imagine similar results of analysis on a larger data set—such as the corpus of books written by Iowa Writers' Workshop authors, as being compiled by Glass—would support McGurl's theories on the Workshop's influence on postwar American literature. These possibilities reflect the youth of digital humanities as a field of study. Larger, structured, open source digitized corpora are difficult to come by, especially compared to the relative ubiquity and massive size of brick-and-mortar libraries. Such data sets would allow researchers to ask broader questions with more general answers. Similarly, while we leverage tools from the broader field of natural language processing, we find programmatic access to literary analysis resources relatively limited.

The nascency of the digital humanities also means that opportunities abound. We see that quantitative tools like semantic cohorts and stylistic features can help both support previous literary hypotheses and surface new ideas. We take a broad approach because our goal is not to supplant or mimic traditional literary analysis by proving a point—old or new—about 20th century literature. We instead hope to demonstrate the power of digital humanities as a companion to reading and reasoning about literature and its nuances. As the digital humanities mature, we expect to see old hypotheses both strengthened and challenged; unexpected, novel ideas and patterns brought to the fore; and more books read.

References

M Algee-Hewitt, M McGurl. 2015. *Between Canon and Corpus: Six Perspectives on 20th-Century Novels.* Stanford Literary Lab.

V Ganjigunte, A Song, and F Y Choi. 2013. *Success with Style: Using Writing Style to Predict the Success of Novels.* Association for Computational Linguistics.

L Glass. 2016 *The Program Era Project.* https://www.lib.uiowa.edu/studio/project/program-era-project/ Web. 6 June 2016.

A P Hackett. 1967. *70 Years of Best Sellers: 1895-1965.* R. R. Bowker Company.

R Heuser, L Le-Khac. 2012. *A quantitative literary history of 2,958 nineteenth-century British novels: The semantic cohort method.* Stanford Literary Lab.

M Korda. 2001. *Making the List: A Cultural History of the American Bestseller 1900-1999.* Barnes & Noble, Inc.

A Liu. 2012. *Where Is Cultural Criticism in the Digital Humanities?.* Debates in the Digital Humanities.

M McGurl. 2009. *The Program Era: Postwar Fiction and the Rise of Creative Writing*. Harvard University Press.

M Parry. 2010. *The Humanities Go Google*. The Chronicle of Higher Education.

A van Cranenburgh and C Koolen. 2015. *Identifying Literary Texts with Bigrams*. Association for Computational Linguistics.