Computational Detection of Biblical and Classical Allusions in English-Language Literature, 1771-1930

Jonathan Reeve

Introduction

How do we define literary modernity? Critics have used a variety of measures to distinguish the modern period in literature from earlier or later periods: the changing political landscape of the colonial into the post-colonial, the rise of the age of mechanical reproduction, and the development of new media like cinema, just to name a few. One of these metrics, one that appears often in discussion of periodization, is a change of attitudes regarding religion, both Christianity and other forms of religion, such as that of classical antiquity.

TODO: Eliot's "Ulysses, Order, and Myth," the influence of Frazer's *The Golden Bough*

The following describes a series of experiments in quantitative literary analysis, whereby an algorithmic approach to the detection of literary allusion is applied to a corpus of novels ranging from the late 18th century to the early 20th century. This experiment will attempt to test the mythological and religious basis of the periodization of modernity, by testing for allusions to the Christian Bible and to figures of Greek and Roman mythology.

The Experiment

Bär, Zesch, and Gurevych describe a method of detecting text reuse with a composite score aggregated from the three textual dimensions of content, structure, and style. A similar approach might be used to detect biblical allusion. However, since allusion is decidedly different than both journalistic text reuse and plagiarism, some modifications need to be made. Set-theoretical methods, text tiling, and the comparison of *hapax logomena* are best suited to comparing texts that have a good deal of similar text, but short quotations might be more difficult to identify with these methods.

The Corpus

To study the period that spans the birth of modernism, roughly in the late 19th and early 20th century, a textual corpus needed to be created.[1] Two existing corpora, Andrew Piper's txtLab450 corpus and Hendrik De Smet's Corpus of English Novels (CEN), roughly overlapped with these years, and were combined to produce the CENLab corpus used in the following experiments.

The txtLab450 corpus is a collection of 450 novels, prepared in 2016 for use at McGill University's txtLab. According to the creator, they are "drawn exclusively from full-text collections and thus should not have errors comparable to OCR'd texts" (Piper 2016). The novels are in English, French, and German, and each labeled with their languages, publication years, authors, and titles, such as `EN_1922_Joyce,James_Ulysses_Novel.txt`[2]. Only the 150 English-language novels from this corpus were used in CENLab. The dates of publications for this corpus range from 1771 to 1930. Figure 1 shows the distribution of these dates.

[1] The many difficulties of creating such a corpus—selecting, obtaining, and preparing the texts—are discussed in the eighth pamphlet from the Stanford Literary Lab (Algee-Hewitt and McGurl 2015).

[2] This feature makes it easy to extract metadata from the filename itself, using the Pandas `DataFrame.apply` method in conjunction with lambda calculus: `tdf['date'] = tdf[0].apply(lambda x: int(x.split('_')[1]))`.

Since many of the experiments to follow involve the chronological analysis of novels, this distribution is important to keep in mind, since the data are skewed somewhat by the availability of texts in a certain period. Of course, every effort was made to compute the results in ratios of word counts, but the availability of texts has an effect here nonetheless.

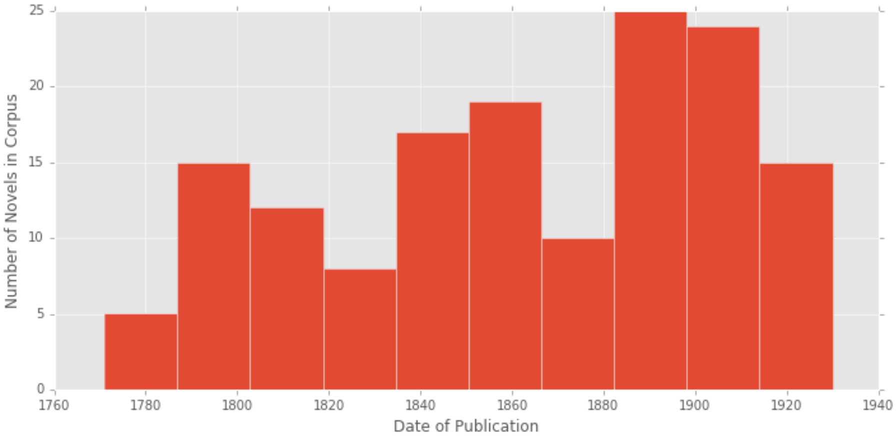

Figure 1: Histogram of Publication Dates, txtLab Corpus

The Corpus of English Novels is a similar collection of 292 novels, each published between 1881 to 1922. This corpus is distributed as part of the Corpus of Late Modern English Texts, Version 3.0. The filenames here are named similarly to those in txtLab450, although most of the metadata is distributed in an accompanying file. A histogram of the publication dates represented by this corpus is shown in Figure 2. Most of these texts were published around the turn of the century.

These two corpora were combined into the CENLab corpus that provides the basis for the experiments below. The metadata file from the CEN was merged with metadata extracted from the txtLab filenames using Python's pandas library, and this metadata was enhanced with data about the authors. A short script was written to download the Wikipedia article for each author in the corpus and convert the resulting HTML to plain text. These files were then

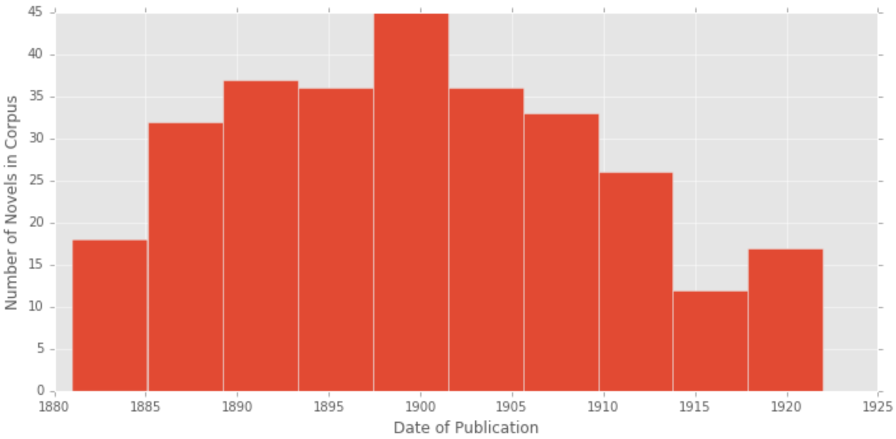

Figure 2: Histogram of Publication Dates, Corpus of English Novels (CEN)

used to populate a table with each author's gender and nationality, roughly categorized as British, American, and Canadian. From there, a short Vim macro was written to call the unix command `wc` on the txtLab texts that were missing word counts.

Since these experiments hope to detect trends that might be revealing about the period of literary modernism, it was necessary to tag each of these texts as "modernist" or "not modernist." This was easily the most subjective and therefore the most difficult task in the compilation of this corpus. To start with, the Wikipedia article list of English-Language First- and Second- Generation Modernist Writers was used as a starting-point, to which was added novels by Willa Cather and H.G. Wells. This list of authors was then programmatically compared with the list in CENLab.

The result of this combination of corpora is a repository containing 441 novels along with a structured metadata file that describes them. This file enables many of the demographic analyses to follow, since this CSV table can easily be programmatically joined with the CSV results of the experiments, producing a unified database of text, publication dates, author information, and experimental results.

TODO: statistics about male writers, female writers; British, American, Canadian writers.

The Detection of Biblical Superlatives

The translators of the King James Bible preserved a number of the linguistic features of their source texts. One ancient Hebrew feature they retained is the superlative construction NN of NNS, a singular noun followed by "of" and a plural noun (Jones 2016, 146). A well-known example of this construction in the KJV is "king of kings," a phrase that occurs seven times—in 1 Timothy, 2 Macabees, Daniel, Ezekiel, Ezra, and twice in Revelation—a sampling which represents both Hebrew and Greek[3]. Similar constructions include "god of gods" (5 occurrences), "lord of lords" (4), "heaven of heavens" (4), "servant of servants," and "tithe of the tithes," both of which occur only once. These occur a total of 23 times in the text. The regularity of this construction makes it good candidate for a quantitative literary analysis.

To test for the presence of these constructions, a Python script was written, x-of-xs.py that scans a given text for a pattern, and logs the number of occurrences of that pattern. The pattern to be found is the regular expression `(\b(.+?)\b\sof\s\2e?s\b)`. This looks for any word, followed by "of," followed by the first word, followed by an optional "e," and ending with "s." The optional "e" is included to catch -es plurals such as fix/fixes. Granted, this pattern cannot match irregular plurals such as "mice" or "children," but a quick search of the corpus for unlikely phrases such as "mouse of mice" returns no matches.

Early tests of this program showed a few false positives, such as "it of its," as in "he emptied it of its contents." These were added to a blacklist, so that they are ignored when the program is executed. The adjusted counts are then logged

[3] These counts were generated with the command `grep -i PATTERN | wc -l`, using a segmented plain text KJV downloaded from ebible.org.

to a file, along with the filenames of the file analyzed, so that the log might be combined with other results for the final analysis.

The novel with the highest number of these superlatives is Gilbert Parker's 1918 *The World for Sale*, in which the phrase "Ry of Rys" appears 38 times. "Ry" is a Romani term which as Parker explains in his appendix "Glossary of Romany Words," means "King or ruler." In this respect, "Ry of Rys" is a perfect analog for the biblical "king of kings" (Parker 1917, 373). Other matched phrases in the novel include the analagous "queen of queens" and "Gorgio of Gorgios," both used in the superlative sense.

The novel with the second highest number of X-of-Xs superlatives is Francis Marion Crawford's 1885 *Zoroaster*, a novel set in biblical Babylon. Here, there are thirteen instances of "king of kings," as well as occurrences of the distinctly biblical phrases "song of songs," "god of gods," and two instances of "lie of lies."

The most common X-of-Xs superlative in the corpus is "heart of hearts" with 81 occurrences, a phrase that doesn't occur at all in the KJV, but has its first OED citation from Shakespeare's *Hamlet*: "Give me that man / That is not passion's slave, and I will wear him / In my heart's core, ay, in my heart of heart" (Shakespeare 2016, 141). Other common superlatives are more distinctly biblical, suggestive of the monarchical flavor of biblical theopolitics: "king of kings" with 24 occurrences, and "god of gods" with 10. Some have a celebratory tone, inherent in "wonder of wonders" (8), "song of songs," (4), and "joy of joys," (4), while others have a moralist tone, such as "crime of crimes" (4) and "sin of sins" (2). A full list of these constructions is included in the x-of-xs analysis notebook.

A chronological analysis of these superlatives, shown in Figure 3, when adjusted for the number of words in each novel, shows that the highest ratios occur between the years 1885 and 1925, overlapping with literary modernism. Among these is the famous modernist novel *Ulysses*, where the Jesuit-educated Stephen Dedalus muses somewhat biblically on the nature of soul as "in a manner all that is: the soul is the form of forms" (Joyce and Duffy 2009, 26). This phrase, "form of forms," occurs four times in the novel. Don Gifford identifies this as a classical, rather than biblical, allusion, taken from Aristotle's *On the Soul*: "As the hand is the instrument of instruments, so the mind [*nous*, soul] is the form of forms" (Gifford and Seidman 1989, 32). However, it is likely that the translators of Stephen's copy of Aristotle, whether deliberately or not, are framing Aristotle's thought in biblical terms.

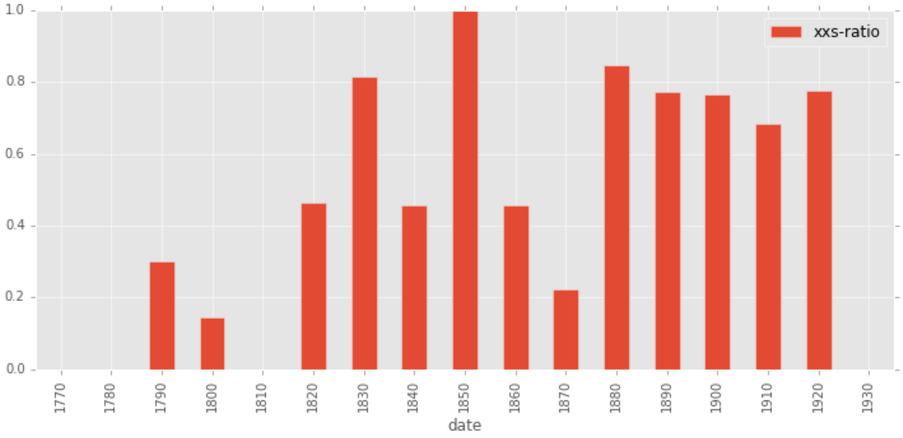

Figure 3: X-of-Xs Ratios by Decade

Detection of Biblical Possessives

Another linguistic feature of the King James Bible lies in its formation of possessives. Rather than form possessives in the typical English language fashion, with an apostrophe, such as "God's son," the KJV prefers the construction "son of God" (Jones 2016, 145). (The latter occurs 42 times, while the former occurs not even once.) As with biblical superlatives, a program was written to detect these constructions. Unlike the superlatives, however, the detection of this construction requires a more nuanced approach. First, a regular expression was used to detect the pattern `\b[A-Za-z']+\sof\s[A-Z][a-z]+`. This searches a given text for any letter or apostrophe, case-insensitive, followed by "of," followed by a capital letter, and ending with any number of lowercase letters. The capitalization requirement for the second letter enables it to better find proper names. Of course, there are noun phrases that are ordinarily constructed like this, like "plaster of Paris," but these exceptions occur infrequently in the corpus, and are few compared to distinctly biblical matches, such as "hand of God" and "curse of God." To better identify these constructions, the NLTK part-of-speech (POS) tagger was used to for a better semantic understanding of the text. The POS tagger is given about 50 characters of context for each phrase, and then is asked to guess the part of speech of the first word in the match. If the first word is not a noun, the match is not recorded. The number of these matches is then divided by the number of 's-possessives to determine the xy-ratio, that is, the ratio of "X of Y" possessives to "Y's X."

The results of this experiment vary greatly, with most novels showing an XY ratio between zero and two, but with six novels showing a ratio of one hundred or more. The novel with the highest XY score is, amazingly, George Barrow's 1843 novel *The Bible in Spain*, subtitled "The Journeys, Adventures, and Imprisonments of an Englishman, in an Attempt to circulate the Scriptures in the Peninsula." While it could be argued, based on some of the less biblical matches in this experiment, that the XY ratio represents not so much a biblical allusion as a feature of the ever-changing English language, the presence of this novel at the top of the list, and at a factor of nearly twice the XY ratio of the next highest novel, suggests that this measure is one which might correlate well with biblical allusion. Most of the biblical possessive constructions in *The Bible in Spain* are not particularly biblical in content. Some, like "port of San Lucar" and "bay of Gibraltar" could possibly be explained as literal translations from the Spanish, where this possessive structure is common. Others, however, like "word of God," "book of Christians," and "work of Scripture," are unmistakably biblical.

Viewed chronologically, XY ratios seem to peak around the mid-19th century, with secondary peaks in the late 18th century and early 20th. This trend is especially noticeable when averaging the results in groups of five years, as shown in Figure 4, where the Y scale is logarithmic. One might fairly guess that this is mostly attributable to *The Bible in Spain*, but this trend remains even when the six novels with the highest XY ratios are removed from the dataset. Is this a function of the decreasing relevance of biblical language, or is it just reflective of the changing English language?

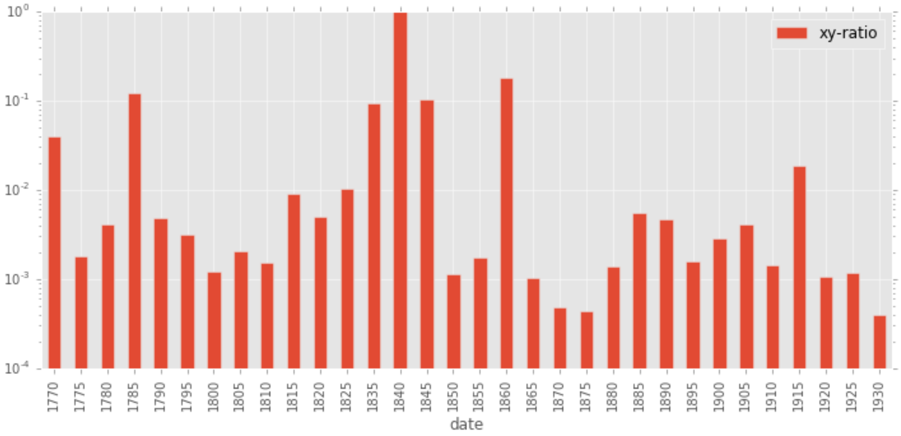

Figure 4: XY Ratios, Averages of 5 Years

Biblical Text Matching

By far the most computationally difficult task in the detection of biblical allusions is the detection of quotations themselves. Variously termed “text reuse,” “plagiarism detection,” or, more generally, “approximate string matching,” this variety of computational analysis remains an open problem in the computer science fields of digital texts forensics and natural language processing. The annual PAN contest in digital text analysis, for instance, has featured a competition in text reuse detection since 2012. A number of approaches to this problem exist, notably TODO TODO. Since this experiment aims to find allusions and short quotations, and not document- or paragraph-sized plagiarized passages, many of the document-level approaches to plagiarism detection are impractical for this task. Instead, a longest subsequence detection algorithm was used.

The text matcher program is designed to adhere to the Unix Philosophy: a program should do one thing, and do it well, and interact with the standard inputs and outputs of plain text on the command line. The program accepts either a file or a directory of files, and computes all matches between each pair of text files given. This allows a user to run the program either on, say, a single text file of the KJV, or a directory of text files where the work is broken into its constituent books, allowing the program to identify the matching biblical book of origin. It prints out its resulting matches to standard output, logging statistics about these matches to a log file.

The text matcher starts by breaking each text into tokens, using a custom tokenizing regular expression, `[a-zA-Z]\w+\'?\w*`. This differs from the NLTK's standard tokenizer, `nltk.word_tokenize()`, in that it ignores punctuation other than internal apostrophes, and keeps contractions such as "can't" intact, a feature that is useful for trigram sequences. This text is then lowercased, in order to make the text matching process case-insensitive.

Once the text is tokenized, the tokens are grouped into trigrams, that is, n-grams where n=3. Some text reuse research, notably Caroline Lyon's, have shown success with identifying text reuse by using trigram sequence matching (Lyon, Malcolm, and Dickerson 2001). This process is computationally expensive, since for a sequence of tokens `ABCDEF`, the ngrams generated are `ABC, BCD, CDE, DEF`. For a text of length L, the text produced by all trigrams has length 3L-6, which is nearly three times the total length of the original text[4].

These trigram sets are then compared with each other using Python's included difflib module's `SequenceMatcher` class. As the module's authors describe it,

[4] Comparing all the trigrams of each text with each other, and building a list of matches by searching the texts again, is such a computationally expensive process that a Linux server was rented, in order to run these computations uninterrupted. Executing the text matching program on the full corpus, without removing stopwords from texts, could take around 12 hours to complete.

the class's algorithm ”predates, and is a little fancier than, an algorithm published in the late 1980's by Ratcliff and Obershelp under the hyperbolic name “gestalt pattern matching” (Peters 2016). The algorithm matches text approximately, automatically ignoring textual differences it considers “junk,” or differences that would be unimportant to most human readers. This, combined with the text preprocessing described above, will allow a certain amount of textual difference to consider a match between two strings: punctuation, case, and even a few letters might be different, and the match will still be found.

The novel with the highest number of matches is Harriet Beecher Stowe's 1853 novel *Uncle Tom's Cabin*, with 12 in the most conservatively configured experiment. Most of the matches come from Chapter 17, when the Quaker Simeon reads passages from the Book of Psalms in order to comfort George. Stowe does not employ this passage uncritically, however, for she notes:

> If these words had been spoken by some easy, self-indulgent exhorter, from whose mouth they might have come merely as pious and rhetorical flourish, proper to be used to people in distress, perhaps they might not have had much effect; but coming from one who daily and calmly risked fine and imprisonment for the cause of God and man, they had a weight that could not but be felt.

Another notable quotation appears at the very end of the novel. In an afterward to a later edition of the novel, Stowe asks a rhetorical question, and follows it with an unattributed quote:

> But who may abide the day of his appearing? “for that day shall burn as an oven: and he shall appear as a swift witness against those that oppress the hireling in his wages, the widow and the fatherless, and that turn aside the stranger in his right: and he shall break in pieces the oppressor.”

The quote is from Malachi 3:5, but heavily modified, for the passage in Malachi reads: “And I will come near to you to judgment; and I will be a swift witness against the sorcerers, and against the adulterers, and against false swearers, and against those that oppress the hireling in his wages, the widow, and the fatherless …” Stowe seems to be combining this passage with the later Malachai 4:1, “for, behold, the day cometh, that shall burn as an oven,” meanwhile removing the sorcerers, adulterers, and swearers.

The novel with the second-highest number of matches is Anne Brontë's *The Tenant of Wildfell Hall.* Chapter 13 of the novel features an interesting debate between Arthur and Helen, where they promote their views of Epicureanism and temperance, respectively, both views which they support through biblical quotations. In response to Helen's pleas for temperance, Arthur counters, “our friend Solomon says, ‘There is nothing better for a man than to eat and to drink, and to be merry’ ” (Brontë 1900, 210). Arthur's quote is actually a combination of Ecclesiastes 2:24 (“There is nothing better for a man, than that he should eat and drink, and that he should make his soul enjoy good in his labour.”) with Ecclesiastes 8: “a man hath no better thing under the sun, than to eat, and to drink, and to be merry.” This is a minor semantic difference, of course, but it suggests that Brontë has memorized the quote, rather than copying it out directly from the Bible.

Overall, the highest numbers of KJV text matches are concentrated in the mid-19th century. This remains true even when the top outliers are factored out. Figure 5 shows this trend, averaged by decade.

The following table shows the 10 books of the KJV that are quoted from the most, with the total number of matches found in each.

Book	Matches
Matthew	129
Luke	83
Psalms	70

Book	Matches
Job	50
Isaiah	48
Mark	46
Proverbs	39
Genesis	34
1 Corinthians	33
John	30

Composite Biblical Allusion Scoring

Biblical allusions of all types are concentrated in the mid-19th century, centered around 1840. Figure 6 shows all three biblical allusion measures: text reuse, xxs-ratios, and xy-ratios, averaged by decade. Apart from high xxs-ratios around 1918, most of the peaks occur in the mid-19th century. As with the text matches described above, this remains true even when outliers are removed. All three of these scores are added together to produce a composite biblical allusion score, the trends of which will be discussed in the "results" section below.

Detection of Classical Allusions

The detection of classical allusions is in some senses more difficult, and in others easier, than the detection of biblical allusions. Biblical semantic structure has few analogues in the Classical influence on the English language. Although some works show structural similarity to Greek myth, notably James Joyce's *Ulysses*, which is structurally and thematically modeled after Homer's *Odyssey*, these macro-level features are very difficult to detect with computational tools, at least awaiting the development of Proppean computational neo-formalism.

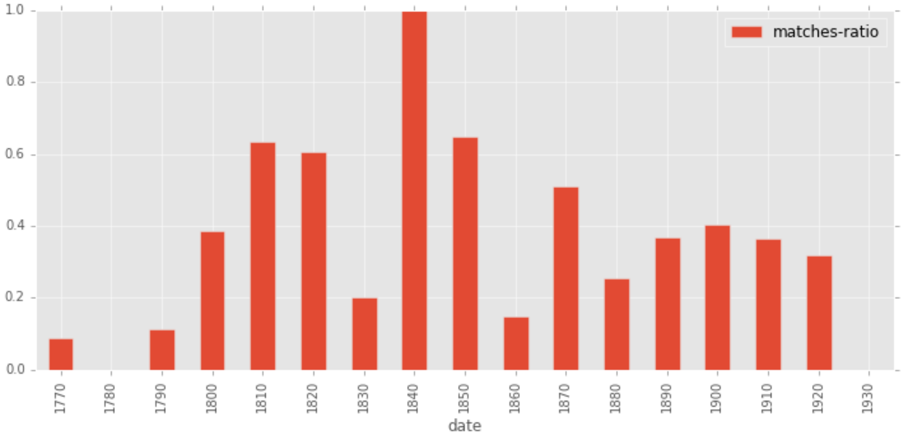

Figure 5: KJV Text Matches by Decade

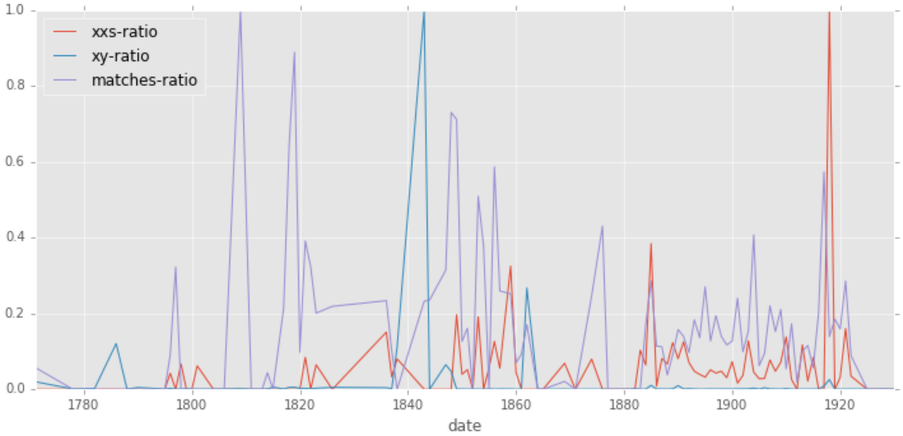

Figure 6: Biblical Allusions by Decade: Matches Ratios, XXS Ratios, and XY Ratios

On the other hand, classical names are more easily recognizable as allusions. "Zeus," for instance, is a much more uncommon English name than "Matthew" or "Mark," meaning that an instance of the word "Zeus" in a text is much more likely to be a classical allusion than an instance of the word "Matthew" is a biblical allusion. To this end, much simpler tools than the previously described text matcher are available.

One such tool is GNU grep, one of the core utilities distributed with GNU/Linux. Grep simply searches a given text for a pattern. Given a file of patterns, it will search the text for every pattern in the file. Thus, simply by creating a file containing a list of names from classical antiquity, it might be possible to construct a crude quantitative analysis, looking for instances of Greco-Roman mythological figures in the corpus. To begin, the Wikipedia article List of Greek Mythological Figures was converted to plain text, and parsed into a text file where every line is a single mythological figure. This file contains 698 mythological figures, with both their Greek and Roman names, and was deduplicated by running the sequence of Linux commands `cat mythological-figures.txt | sort | uniq > mythological-figures-sorted.txt`. The resulting text file was then fed into Grep, and run against the texts in CENLab with the command `grep -wf mythological-figures.txt cenlab/texts/* > classical-matches.txt`. In order to retain the counts of all the matches, a similar command was run, `grep -cwf mythological-figures.txt cenlab/texts/*`. The results, which are colon-delimited, are then converted to comma-separated values with GNU `sed`, and appended to the prior biblical results using Python's pandas module.

The novel with the highest number of classical references (classical names, that is), with a total of 410, is Lydia Child's 1836 *Philothea*, subtitled, perhaps unsurprisingly, *A Grecian Romance*. A sample of the text will indicate just how many Greek names are present: "It was the last market hour of Athens, when Anaxagoras, Philothea, and Eudora, accompanied by Geta, the favourite slave of Phidias, stepped forth into the street, on their way to Aspasia's residence" (Child 2016). The novel even goes so far as to contain an appendix listing all of the Greek gods in both their Greek and Roman forms.

The work with the second highest number of classical references, at 375, is Edith Nesbit's 1904 *The Phoenix and the Carpet*, a somewhat mythological children's novel. Most of its references, it turns out, are the word "Phoenix." The novel with the third highest number of classical references is Charles Kingsley's 1853 novel *Hypatia*. Like *Philothea*, this is another historical novel, set in 5th-century Alexandria, that "exalts the Greek Neoplatonic philosopher Hypatia who was torn to pieces in ad 415 by a mob of infuriated Christians" (Birch 2009).

Notably, the novel with the sixth highest number of classical references is James Joyce's *Ulysses*. While the majority of these matches is instances of the word "Myles," the ancient king of Laconia and the first name of Joyce's character Myles Crawford, many more obvious classical allusions are present here. The presence of "Mercury," "Venus," "Mars," "Jupiter," and "Neptune" is not only a nod to Joyce's astrophysical themes in the novel, but to a particularly Roman flavor of Greek mythology, since these are all, like "Ulysses" is to "Odysseus," Roman names for Greek mythological figures.

Ulysses aside, the greater chronological trend here is one of decline. Figure **??** shows decade averages for the number of classical names in a novel, divided by its number of words, and plotted here on a logarithmic scale. Since the range of results here was so dramatic, with some novels featuring hundreds of classical names and others just one or two, another test was performed, discarding the top three historical and mythological novels discussed above. The results of that test also confirm the early 19th century as the time of greatest number of classical allusions.

Classical Text Matching

An attempt was made to find classical allusions by comparing texts against Thomas Bulfinch's popular and influential *Mythology*, a compilation of his works

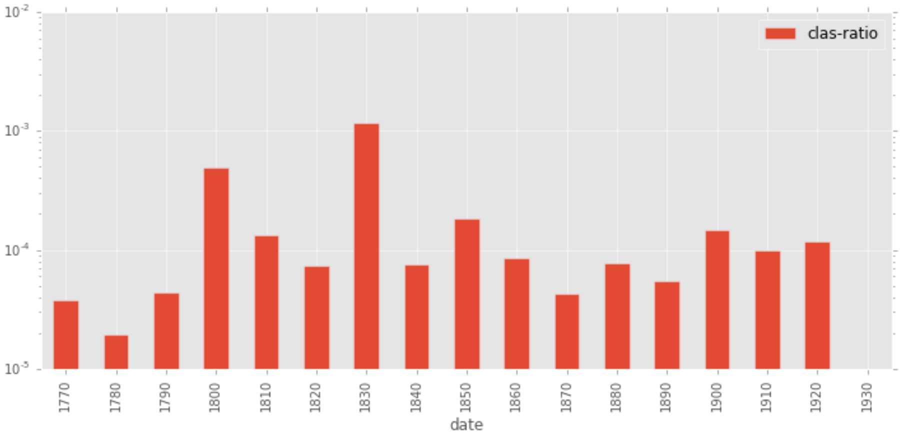

Figure 7: Classical Ratios, by Decade, on a Logarithmic Scale

The Age of Fable, *The Age of Chivalry*, and *Legends of Charlemagne*. Since the latter two works, and part of the first, deal with periods beyond that of classical antiquity, these parts were cut out, and the remainder matched against the CENLab corpus. The results were disappointing. Rather than finding instances of CENLab authors quoting Bulfinch, what was found instead was Bulfinch quoting CENLab authors, mostly Walter Scott. Additionally, there were a few instances of Bulfinch and an author, Hannah More, for instance, both quoting the same passage from Milton or Wordsworth. There were no detectable instances of CENLab authors quoting Bulfinch. A similar attempt was made with Ovid's Metamorphoses, in particular the popular 1717 English edition translated by John Dryden, Alexander Pope, and others. The text matcher did not find any matches with this text, either. Until a more suitable source text can be found (Hesiod's Theogony?), we might assume that, despite the presence of many classical allusions, there are very few quotations from English texts in classical mythology. In this respect, we might say that biblical allusions are more inherently *textual* than classical ones, more reliant on syntax and semantics.

Results

Classical and Christian mythologies in literature are competing and complementary. Figure 8 shows classical and biblical allusions, averaged by decade. There appears to be a general decline in both these varieties of allusion from about 1850 until the present, but with a very minor resurgence around 1880-1920, coincident with the period of literary modernism. Furthermore, biblical and classical allusions seem to happen in answer to each other. After classical allusions peak in the early 19th century, there is another peak of biblical allusions. This trend is most noticeable when outliers are removed, as shown in Figure 9.

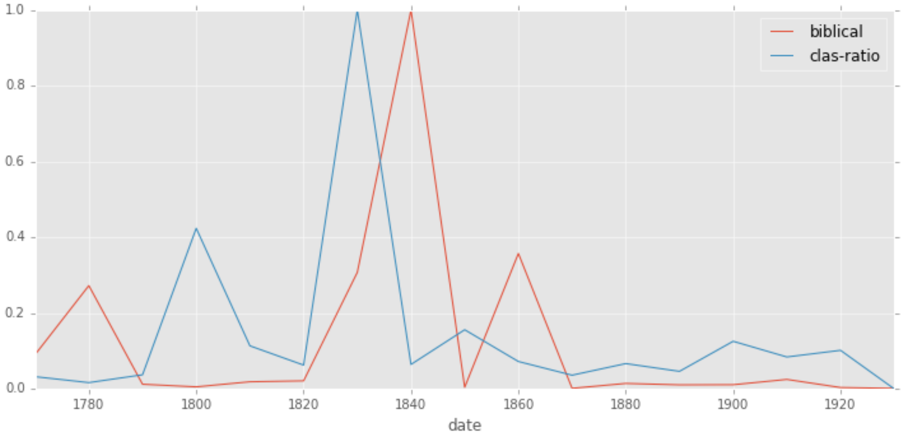

Figure 8: Biblical and Classical Allusions by Decade

Could this be a cyclical trend? Might classical and Christian themes be shown to compete, alternating in decades? Perhaps these cycles are compatible in some way with Franco Moretti's assertion, in "Graphs, Maps, and Trees" that literary genres usually cluster in 30-year groups, or Colin Martindale's theories, in *The Clockwork Muse*, of the periodicities of artistic movements.

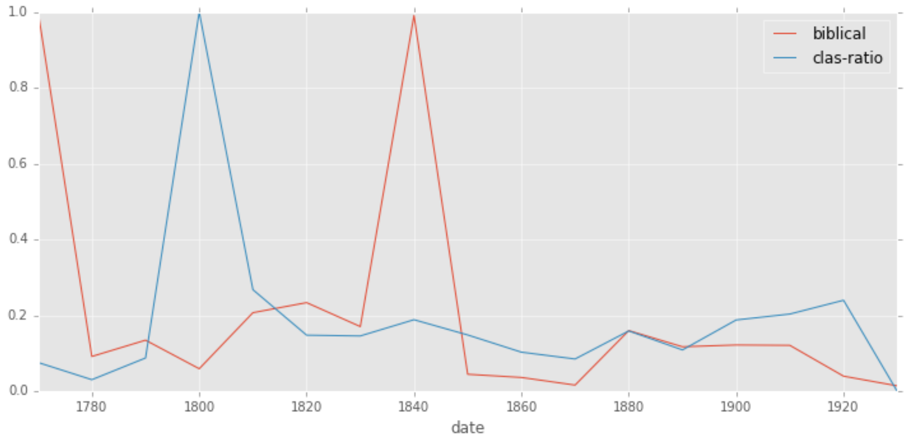

Figure 9: Biblical and Classical Allusions by Decade, Outliers Remoed

Demographics

An analysis of biblical allusions, separated by the authors's genders, shows that, on the whole, male writers are more allusory than women. Figure 10 shows that male writers have almost all the biblical allusions until the beginning of the 19th century, when allusions are shared by both men and women. When outliers are removed, however, this picture changes, and men only have high biblical allusion scores at the end of the 18th century, while women have the highest scores throughout the 19th.

TODO: averages for men and women; averages by nationality.

This trend is almost exactly the opposite for classical allusions. Figure **??** shows that women writers have the highest numbers of classical allusions, even when outliers such as historical novels have been removed.

Grouped by nationality, it seems that British writers are the most biblically and classically allusive, although the majority of their allusions happen in the late 18th and early 19th centuries. Figure 12 shows biblical allusions by nationality, again with outliers removed. This suggests that British writers are more likely

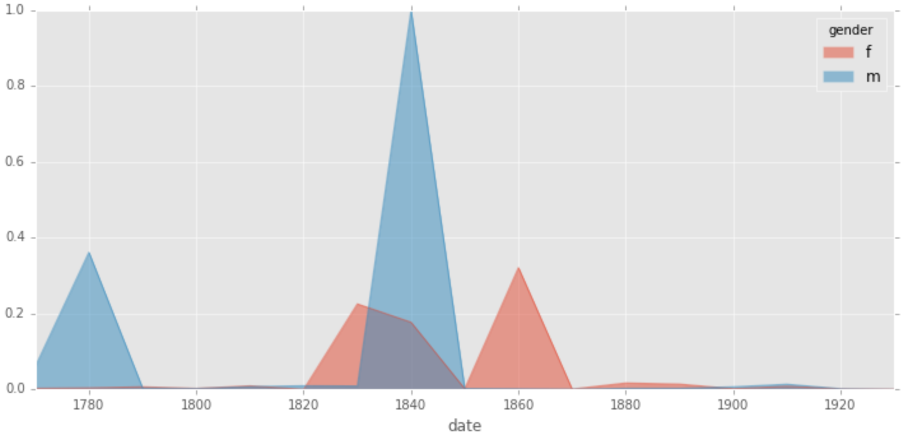

Figure 10: Biblical Allusions by Gender and Decade

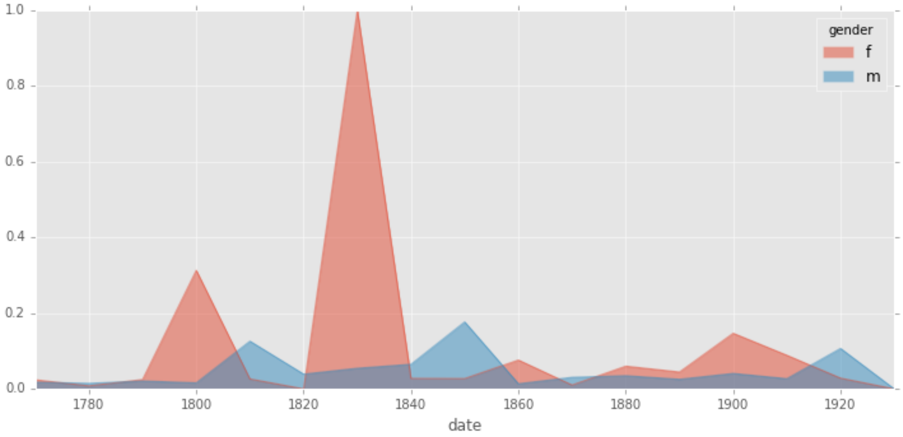

Figure 11: Classical Allusions by Gender and Decade, Outliers Removed

to be classically allusive at the end of the 18th century, while Canadian writers are more likely to be so at the end of the 19th. However, this is almost certainly attributable, at least in part, to the demographics of the corpus—CENLab doesn't feature as many 18th century American novels as British, and there are almost no older Canadian novels, either.

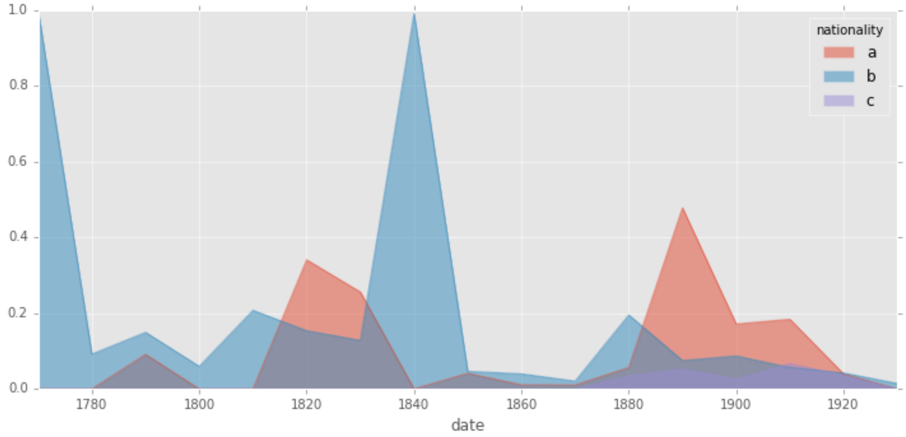

Figure 12: Biblical Allusions by Nationality and Decade, Outliers Removed

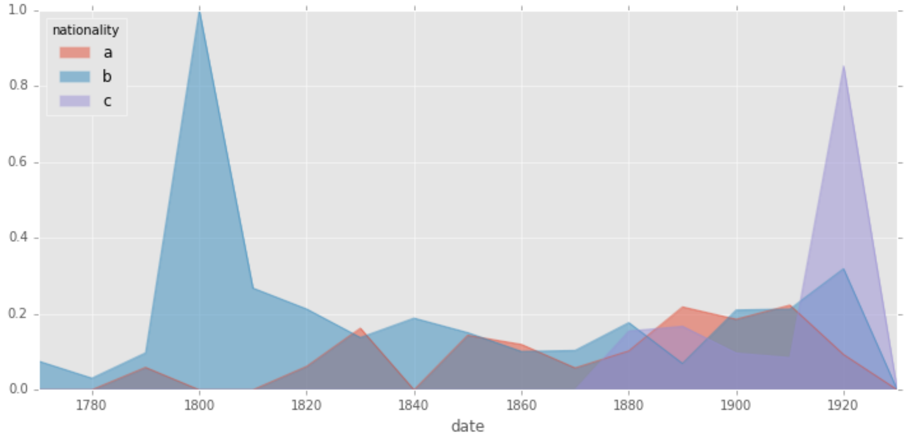

Figure 13: Classical Allusions by Nationality and Decade, Outliers Removed

When these data are grouped by whether the authors are considered to be modernist writers, it seems that modernist writers are not as allusory as their

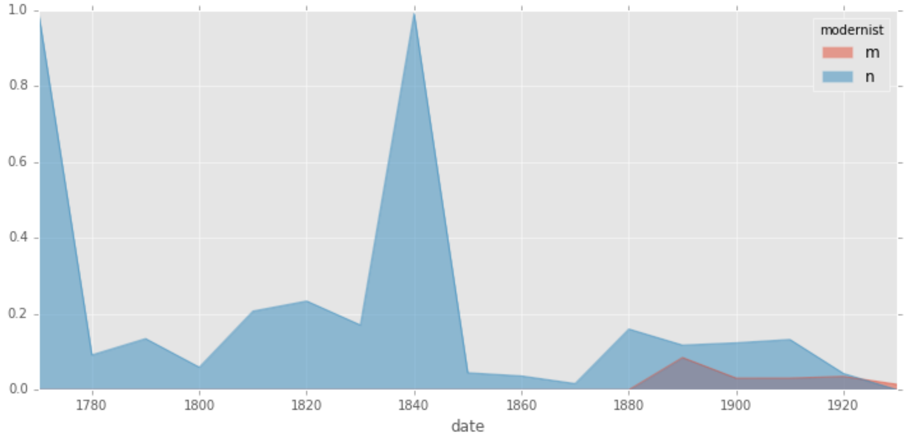

Figure 14: Biblical Allusions by Modernist and Non-Modernist Writers, Outliers Removed

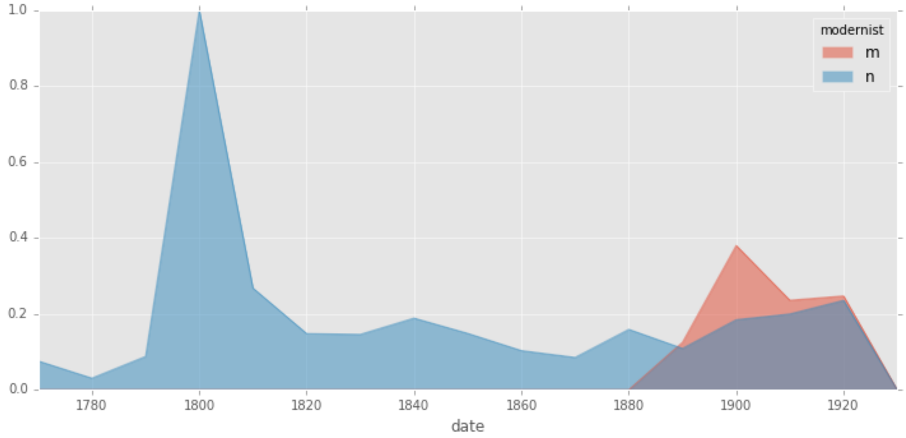

Figure 15: Classical Allusions by Modernist and Non-Modernist Writers, Outliers Removed

older counterparts. Figure 14 shows biblical allusions by non-modernist (n) and modernist (m) writers, with outliers removed, grouped by decade. With classical allusions, however, modernist writers out-allude non-modernist writers in the 20th century, as shown in Figure 15.

Conclusions

Although these results are highly suggestive of an overall trend in literary history, more data are needed in order to make even provisional pronouncements. The uneven shape of the corpus, with most of its novels concentrated in the early 20th century, while it may indeed reflect the numbers of novels published during these periods, could skew even the most balanced and normalized ratios.

Works Cited

Algee-Hewitt, Mark, and Mark McGurl. 2015. "Between Canon and Corpus: Six Perspectives on 20th-Century Novels." Stanford Literary Lab. http://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Bär, Daniel, Torsten Zesch, and Iryna Gurevych. 2012. "Text Reuse Detection Using a Composition of Text Similarity Measures." *Proceedings of COLING 2012* 1: 167–84. https://www.cdc.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/publikationen/2012/COLING_2012_DaB_published.pdf.

Birch, Dinah. 2009. "Kingsley, Charles." *The Oxford Companion to English Literature.* Oxford, UK: Oxford University Press. http://www.oxfordreference.com/view/10.1093/acref/9780192806871.001.0001/

acref-9780192806871-e-4203.

Brontë, Anne. 1900. *The Tenant of Wildfell Hall .: With an Introduction by Mrs. Humphry Ward . (the Haworth Ed.).* Harper & bros. [c 1900].

Child, Lydia. 2016. "Philothea: A Grecian Romance. Internet Archive." https://archive.org/stream/philotheagrecian00chil#page/22/mode/2up.

Gifford, Don, and Robert J. Seidman. 1989. *Ulysses Annotated: Notes for James Joyce's Ulysses.* University of California Press.

Jones, Norman W. 2016. *The Bible and Literature : The Basics.* New York: Routledge.

Joyce, James, and Enda Duffy. 2009. *Ulysses.* Mineola, NY: Dover.

Lyon, Caroline, James Malcolm, and Bob Dickerson. 2001. "Detecting Short Passages of Similar Text in Large Document Collections." In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, 118–25. http://homepages.stca.herts.ac.uk/~comrcml/plagiarism_01.pdf.

Martindale, Colin. 1990. *The Clockwork Muse : The Predictability of Artistic Change.* New York: BasicBooks.

Moretti, Franco. 2003. "Graphs, Maps, Trees." *New Left Review*, no. 24 (November): 67.

Parker, Gilbert. 1917. *The Works of Gilbert Parker.* C. Scribner's Sons.

Peters, Tim. 2016. *Difflib* (version 3.5). Python Software Foundation. https://docs.python.org/3.5/_sources/library/difflib.txt.

Piper, Andrew. 2016. "txtLAB450. a Multilingual Data Set of Novels for Teaching and Research. .txtLAB @ Mcgill." January 18. http://txtlab.org/?p=601.

Shakespeare, William. 2016. "The Tragedy of Hamlet, Prince of Denmark.

Folger Digital Texts.” http://www.folgerdigitaltexts.org/html/Ham.html.

Stowe, Harriet Beecher. 1853. “Uncle Tom’s Cabin. Project Gutenberg.” http://www.gutenberg.org/files/203/203-h/203-h.htm.