Faulkner and punctuation

Johannes Burgers

## Precursors

# Import Data

## Introduction

Since the earliest reviews, Faulkner’s style has been divisive: high art for some, bad writing for others. This friction has created a rich body of scholarship on the form of Faulkner’s writing and his use of punctuation. What has eluded critics,however, is a view of his writing as a whole. New computational techniques make it possible to sketch a more detailed portrait of his use of punctuation throughout his career.[[1]](#footnote-21) This data can be coupled with the character, location, and event data available through *Digital Yoknapatawpha*. Combined, this data can be used to understand not only how often Faulkner makes particular choices about punctuation, but also the relationship between punctuation choices and the characters present or mentioned in those particular sentences. Doing this not only confirms suspicions about specific punctuation patterns in Faulkner by earlier scholars, but it also furthers the conversation about who gets represented in Faulkner and how, and the politics of Faulkner’s style.

In the main, punctuation in Faulkner serves to stress his aesthetic of “the long sentence”. This aesthetic tries to capture a person’s past, present, and possible future in one moment in time, and thereby undermines the neat temporal distinctions that cordon off past, present, and future. Though ostensibly representative of the human condition writ large, the aggregated data tell a different story from Faulkner’s own comments on his writing. Sentence length and the experimentation with punctuation that makes it possible appears to be deeply tied to the status of the upper class White plantocracy. In Faulkner’s fiction the two most prominent threats to upper class White hegemony are the poor Whites and what was understood as miscegenation. In light of the current critical reception of Faulkner, this is not altogether surprising. To be sure, analyses of race and class are well-worn scholarly paths. What is less apparent, and indeed impossible to substantiate without computational methods, is how these themes consistently manifest themselves at the level of the sentence and his use of punctuation.

## Critical literature on Faulkner

Faulkner’s style has long been a central concern of critics and scholars. It has not always been universally well received, one critic, writing of *Absalom, Absalom!,* remarks that, “When a narrative sentence has to have as many as three parentheses identifying the reference of pronouns, it signifies mere bad writing and can be justified by no psychological or esthetic principle whatever.” (DeVoto 147) (DeVoto). While critics now uniformly agree that Faulkner’s style is more than merely “bad writing,” there is no settled view on the psychological or aesthetic principle.

107 - The long sentence, then, by Faulkner’s account as well as Aiken’s, is a technique, or “tool,” for compressing the greatest possible amount of time, or life, or motion, into the smallest possible space, in order to condense and stop it for contemplation. Adams

47 - It is as if Mr. Faulkner, in a sort of hurried despair, had decided to try to tell us everything, absolutely everything, every last origin or source or quality or qualification, and every possible future or permutation as well, in one terrifically concentrated effort: each sentence to be, as it were, a microcosm. And it must be admitted that the practice is annoying and distracting. Aiken

83 - First, there is the suspension of meaning in the long sentence, there are colons, semicolons, and dashes (sometimes there are parentheses); the sentence is a small self-contained world. Van O’Connor

Some notable perspectives included

Perhaps one of the most trenchant critics on Faulkner’s writing is Faulkner himself. Though he was notorious about self-fashioning, distorting, and, in some cases, outright lying about the genesis of particular works and his own life, the comments he makes about his style are insightful. [FOOTNOTE] Of particular interest is his idea of the “long sentence.” When Faulkner is asked about his use of long sentences at the Virginia Colleges Conference in 1957, he claims that there is no such thing as “was” because the past is part of “every man, every woman, at every moment.” This aesthetic attempts to get “his [a character in a story] past and possibly his future into the instant in which he does something” ( Faulkner at Virginia). On the surface, this has a universal appeal: men, women, and everyone is the “sum” of their past (CITE). It suggests that the moments that this past and possible future collide is in the long sentence, and, accordingly, that long sentences capture the themes to each work but also Faulkner’s writing more generally. One way of understanding Faulkner’s writing then, is to see it as an attempt to capture these long sentences, as a means of dissolving the aggregations of the past and the possible futures.

## Methodology and Editing Faulkner

Establishing a corpus of texts for computational analysis of Faulkner’s writing presents some thorny issues, one of the most significant of which is that there is not single, settled upon edition of his texts. Noel Polk famously edited “corrected” versions of Faulkner’s texts, but even he admits that his editions are far from the “definitive” (Polk 6) (POLK 6). In part, this is because Faulkner was himself inconsistent with his approach punctuation. For example, he retyped parts of the Benjy section to *The Sound and the Fury* three times, and used a different punctuation system in each iteration (Polk 14) (Polk 14). These inconsistencies are evidence of Faulkner’s constant experimentation with punctuation and his style more generally throughout his career. While it is theoretically possible to build a dynamic model of these punctuation changes, the very real practical challenge would be to digitize the typescript and manuscript pages to make them machine readable. Since no such corpus is available for the foreseable future, it makes sense to settle for the highly regarded versions created by Polk. Adhering, where possible to these editions, also allows for better alignment with the *Digital Yoknapatawpha* database, which used the Polk editions for data encoding. If these editions are imperfect, there is at least some assurance that it is the mostly likely version a modern reader will read.

Beyond the variance in editions, there is also the issue of the internal coherence of the textual corpus. Faulkner’s writing is quite heterogeneous, and consists of novels, short stories, poetry, film scripts, and a text that hovers between a stage play and a novel: *Requiem for a Nun*. Even within these broad categories there are further subdivisions. Speaking of the novels and short-stories more narrowly, there are the 14 novels and 54 short-stories that take place in his famous Yoknapatawpha County. Among these are “uncollected” short stories that were only ever published in magazines and then incorporated wholesale or in part into novels such as *The Unvanquished* (1938), *The Hamlet* (1940), and *Go Down, Moses* (1942). There is also a further set of stories that were unpublished in his lifetime, which are interesting in their own right or that have resonances with Faulkner’s more well-known published texts. This dense intertextuality leads to an interesting intellectual problem: how to deal with double counting. Some but not all of the stories in *The Unvanquished* are only modified slightly from their magazine versions, and including them would mean essentially double counting many of the linguistic patterns and over-representing them. Conversely, *The Saturday Evening Post* version of “The Bear” varies quite substantially from the version that appears in *Go Down, Moses*, even though the two are printed only two days apart [@corrigan2015](Corrigan%20Padgett). In this case, both texts are different enough that they should be seen as separate works. Clearly, it makes sense to remove works if they are duplicates, and keep works if the versions are very dissimilar. The problem is that it is never really the case that works are exactly the same word for word, and removing either *The Unvanquished* novel or the separate short stories means making a somewhat arbitrary distinction between a modified duplication and a revision. In short, all of the stories were kept in, but with the proviso that there is a reasonable argument to be made to exclude some texts on the basis of similarity.[[2]](#footnote-25)

Beyond these issues of similarity, there is also the issue of corpus heterogeneity.[[3]](#footnote-26) In Faulkner there are conistently different approaches to his punctuation patterns, which makes for a poor experimental design. If, for example, his changes in writing were unidirectional and gradual, as is the case with Henry James (Hoover 275), it would be possible to focus on a cluster of features and show their change over time. Instead, with Faulkner there is a high amount of variance between punctuation patterns from one text to the next, which makes it impossible to distinguish between early, middle, and late Faulkner purely on the basis of linguistic patterns. The heterogeneity of his writing militates against broad generalizations. Nowhere is this truer than when it comes to sentence length, Faulkner’s putative signature feature. True, on average Faulkner writes longer sentences than his contemporaries (17 words-per-sentence), but this number varies quite widely. A canonical work like *As I Lay Dying* uses an unremarkable 12 words per sentence, while an equally well-regarded work like *Absalom, Absalom* uses 43 words per sentence. Therefore, it makes more sense to stay away from large assumptions about how Faulkner punctuated his texts, and instead accept that what is true in some of Faulkner’s writing is not necessarily true in all of his writing. These local insights can then be stitched together to create a more global collage.

## Sentence length

The issue of corpus heterogeneity is one that complicates even one of the most basic metrics: sentence length. It does not require computational analysis to understand that some of Faulkner’s most famous works contains long sentences. For anyone who has ever taught Faulkner, it is perhaps the most common stumbling block and, indeed, complaint among students. Faulkner himself was asked about it no fewer than four times, when he lectured at at the University of Virginia (CITE). However, quantifying what readers might mean when they say Faulkner’s sentences are long is more complex.

The first issue is defining what constitutes a sentence. In terms of computation, the simplest way to define it is as any set of words where the first word is capitalized and is closed with end-sentence punctuation like a full-stop, exclamation mark, or question mark. These basic parameters are frequently used to determine sentence length in standard prose writing. This technique works well in most cases, but runs into quite a number of exceptions in Faulkner. For starters, Faulkner has a tendency to nest narratives within quoted material. For example, in *The Hamlet* (1940), much of the story is told by Ratliff, an itinerant sewing machine salesman. When he takes over as narrator, the perspective moves from third person omniscient to first person, and within this embedded narrative Ratliff quotes various characters. These quotations are themselves sentences, but also part of a longer, enframing sentence. Thus, while describing a scene between two antagonists, Ratliff says, “What will he say? What can he say except ‘All right. What do you aim to do?’” (CITE). If a sentence is defined as any set of words that starts with a capital letter and ends with end-sentence punctuation then this is four separate sentences. This seems unsatisfactory. The two sentences “All right.” and “What do you aim to do?” are embedded within the larger quote started with “What can he say except.” Not accounting for sentence embedding in *The Hamlet* actually makes the average sentence length in Faulkner appear shorter than it is. The exact opposite problem is true in *The Sound and the Fury*. Here Faulkner embeds sentences and phrases through italics. This occurs frequently in the the Quentin section where the character’s internal narration is interrupted by memories. At one point, Quentin recalls Gerald Bland’s mother talking about her son’s looks, which is interjected by a memory of Quentin fantasizing about shooting his sister’s fiance, Herbert Head: “Telling us about Gerald’s women in a *Quentin has shot Herbert he shot his voice through the floor of Caddy’s room* tone of smug approbation” (105). The main sentence is: “Telling us about Gerald’s women in a tone of smug approbation.” This is interrupted by the clause “Quentin has shot Herbert,” and possibly another sentence or a part of one, “he shot his voice through the floor of Caddy’s room.” There is no unambiguous way to determine if the two phrases should be seen as part of the same italicized sentence or if they should be read separately. There are valid interpretations for both. Similarly, while it would be computationally possible to excise the italicized text, this distorts the fact that these italics quite deliberately interrupt and extend the sentence because there is no end-sentence punctuation. It would also mean going against Faulkner’s explicit wishes to have the text presented in this way (Polk 13-14). Consequently, the sentence in *The Hamlet* (17 words) and the sentence in *The Sound and the Fury* (25 words) are roughly the same length, but in the tabulation the former is significantly shorter than the latter, only five words. This inconsistency is a feature of the entire corpus, because Faulkner varied the way in which creates long sentences throughout his career.[[4]](#footnote-28) Sentence length is therefore a rather crude and imperfect measure to understand punctuation in Faulkner, but it makes an interesting starting point for investigation.

Going by the imperfect measure of sentence length, Faulkner’s writing is not exceptional compared to other writing at the time. At 16.77 words per sentence, Faulkner’s sentences are only slightly longer than the average 16 words per sentence for a text in the Corpus of English Novels, which covers writing from 1881 - 1922 (Ihrmark and Nilsson 79) (Ihrmark 79) and longer still than the 13 words per sentence around the 1930s (Rudnicka) (Rudnicka). Nor is the slightly higher than average sentence length a feature of all his works. The *Sound and the Fury* is notoriously complex. Nevertheless, at only 10.202611 words per sentence, it uses fewer words on average than Hemingway’s *The Old Man and the Sea*, which uses 13 words per sentence (Ihrmark and Nilsson 82). Despite this counter intuitive disparity, Faulkner’s more verbose style could never be confused with Hemingway’s trademark concision. In part, this has to do with the distribution of sentence lengths within each text. The *The Sound and the Fury* may have a lower words-per-sentence average than *The Old Man and the Sea*, but it also contains one sentence that counts over 4,000 words. Clearly, this is less succinct than Hemingway, and probably most other authors as well. These long sentences, it should be noted, are more of a statistical exception than the rule. The normal distribution for *The Sound and the Fury*, and most of Faulkner’s other texts is heavily right-skewed. Meaning that most sentences are of average length, but a few are much, much longer. In the case of the *Sound and the Fury* only about NA% of the sentences are longer than the mean corpus sentence length. The same can be said for Faulkner’s other texts, most of these (NA%) have an average sentence length lower than the total average across the corpus. More succinctly, a small share of sentences in each work gives the impression of long sentences, and, similarly, within the corpus of Faulkner’s writing there are a minority of works that are above average in sentence length. Statistically, lengthy sentences are a local feature not a global phenomenon.

Nevertheless, lengthy sentences are seen as hallmark feature of Faulkner’s writing. In part this may be because some of Faulkner’s most canonical works feature long sentences, *Go Down, Moses* (20), *Absalom, Absalom* (43), “A Rose for Emily” (19), and “Barn Burning” (19) These works are, in a sense, more Faulknerian than a relatively unknown early text like “Elly” (9). Nevertheless, equally canonical works like “That Evening Sun” (9) , *Light in August* (15) , and *As I Lay Dying* (12) do not, on average, stand out as being particularly verbose relative to the the general corpus of texts written in English during this period. Despite these counter indications, it goes against a firmly established scholarly tradition based on Faulkner’s own statements and, indeed, common sense, to argue that lengthy sentences are not Faulkner’s particular metier. Instead, to understand why Faulkner’s sentences, for lack of a better word, *feel* long it is informative to look at their internal structure.

## Extending the Sentences: Faulkner’s Experiments in Mid-sentence puntcuation

The preceding discussion about sentence length raises an obvious question: What is a long sentence? Is it any sentence above the average sentence length of the corpus? If so, how much longer does it need to be to be consider long and not simply longer than usual? If a sentence is indeed long, is there a maximum length? At what point, does a string of words grammatically cease to be sentence? James Joyce famously wrote some of the longest sentences in English in the Penelope chapter of *Ulysses*, but these sentences are not long in the same way a sentence in Faulkner is long. While there is certainly semantic coherence underlying Joyce’s string of words, the lack of mid-sentence punctuation makes it grammatically incorrect if not necessarily syntactically incoherent (CITE flow article). This was rather point, of course, but it is an important distinction to make. Faulkner not only creates long strings of words, but also coordinates and subordinates his clauses through punctuation. Thus, whereas Joyce saw the intervention of punctuation as an infelicity to the ineluctable flow of language in the Penelope episode[[5]](#footnote-30) , the sheer profusion of punctuation in Faulkner speaks to the endless way in which meaning branches, eddies, rejoins and ends in an arbitrary terminus only to recirculate again. To use two spatial metaphors, Molly’s soliloquie is a stream, Faulkner’s writing is a watershed.

To understand how Faulkner lengthens his sentences it is useful to look at mid-sentence punctuation like: colons, parenthesis, ellipses. As earlier critics observed (observation about mid-sentence) . Faulkner is not merely making sentences very long, but using punctuation to stretch out the definition of a sentence, and, in turn, calling attention to the punctuation itself. The finality of the full-stop is replaced with the indeterminacy of the unfinished sentence. What is telling about Faulkner’s writing is that he does not simply write long sentences throughout his career, but experiments with different punctuation configurations to make this possible. Arguably, any number of texts serve as good examples, but there are three salient texts that deviate from Faulkner’s own writing statistically: *Absalom, Absalom!*, *Intruder in the Dust*, and “Miss Zilphia Gant.”

To determine which of Faulkner’s texts have punctuation at variance with the corpus more generally, a statistical model of Faulkner’s short stories and novels was created.[[6]](#footnote-31) The texts were broken down into sentences, and each mid-sentence punctuation mark was counted. Intuitively, sentence length and mid-sentence punctuation are positively correlated. As sentences get longer, there is a higher likelihood of finding mid-sentence punctuation. There are some functional limits to this correlation, however. Modern punctuation tends to function as a syntactical system that regulates the number of certain punctuation marks that can be used (Schou 198) ((**Shou?**) 198). This is particularly true of colons, semi-colons, and parenthesis. For example, one line in *The Mansion* (1959) has 9 semi-colons and is over 400 words long. It pushes the limits as to how many independent clauses a reader can reasonably be expected to understand in the same ostensibly coherent thought delimited by a full-stop.

By taking the averages of the number of mid-sentence punctuation marks per sentence across all texts, it is possible to detect texts that deviate from the overall pattern. In general, phonological and syntactical phenomena tend to follow a power-law distribution ((**sun?**) and Wang 32, Newman 372) Newman; where the probability decreases exponentially with each frequency interval. Syntactically, this makes sense. In any given text, most sentences do not contain commas. If a sentence does contain a comma, it is far less likely that it would also contain a second comma, it is exponentially less likely to contain a third comma, and so forth. This rule does not apply to specific punctuation marks that are more strictly governed by the rules of grammar, such as colons, which generally do not exceed one per sentence. That said, if Faulkner did not experiment with punctuation there would be a low variance in the number of punctuation marks and a very low probability of sentences with a high number of punctuation marks. This is not borne out by the data. Plotting the distributions across all punctuation marks reveals an exponential distribution best described as chi-squared with three degrees of freedom (). Visually, this is curve that looks a bit like a slide on a playground: a steep ladder on the left-hand side with a long-curve that never quite reaches the ground on the right. This curve indicates that many texts are within the same range, but a few texts have a very high relative frequency of specific punctuation marks.

Determining when on that tail a value is outside of the expected range is a matter of some interpretation. Indeed, deciding that a data point is an outlier is an ongoing debate among mathematicians (Hawkins 9-12). As the distance between the mean and the highest value was quite far apart, the skewness of the distribution was corrected for by either taking the inverse or the log of each of the values depending on the line of best fit in a quantile-quantile plot. Using these normalized distributions, Rosner’s multiple outlier test was run on any value outside of the standard Tukey fences ().

Additionally, a secondary test was run using the Adjusted box plot method developed by Hubert and Vandervieren (Hubert and Vandervieren). The tests revealed slightly different outliers, but on the whole there is an inconsistent use of mid-sentence punctuation across the corpus. Faulkner uses some mid-sentence punctuation substantially more than in some texts versus others. This indicates some measure of experimentation with how he is creating long sentences. If he were developing only one particular technique, for example concatenating a string of independent clauses separated by a semi-colon as in *The Mansion*, the data for the other punctuation marks would likely not be as skewed, and the distribution far more normal. This is not the case, and indicates variance in punctuation techniques across his career. This variance does not appear to be strongly time dependent. There were no distinct phases to his punctuation experiments, but instead individual works stand out as being particularly experimental.

Three texts provide useful insights. *Absalom, Absalom*, and *Intruder in the Dust*, and “Miss Zilphia Gant” were all marked as outliers. As one of Faulkner’s most challenging texts, it is unsurprising that the punctuation patterns in *Absalom, Absalom!* diverge from the corpus. Averaging around 43 words per sentence, the text is marked by a heavy use of parentheses. Thus, a sentence will be interrupted by parenthetical statements, which in themselves are not usually sentences, leading to sentences of extraordinary length. In one sentence stretching over 1,000 words, he uses nine parenthetical statements as a running commentary on the main sentence, which runs across three pages (Faulkner, *Absalom, Absalom!* 148–50). Some of these parentheticals are themselves several lines. All of this is complicated by the fact that this 1,000 word sentence appears in Chapter 6. This entire chapter is surrounded by parentheses. As a result, the parentheses in this sentence are actually nested within surrounding parentheses. More plainly, the sentence is an aside to an aside. Beyond this particular sentence, the nesting continues throughout the Chapter 6 and at some point the the word Quentin is nested 4 levels deep (Faulkner, *Absalom, Absalom!* 168). That is, as an aside to an aside to an aside. Not only, then, are Faulkner’s sentences very long in *Absalom, Absalom!* they are also, in a sense, deep. Each parenthetical statement adds more commentary and context, which itself requires more commentary and context. It is fitting that peeling away all of these parentheses like layers of an onion reveals the absent center: Quentin. The tale is ultimately about the teller.

In a slightly different fashion, Faulkner instrumentalizes colons in *Intruder in the Dust* to create sentence depth. To give an extreme example, one sentence in *Intruder in the Dust* is over 1,600 words long and features eighteen colons. The sentence deploys colons as a clausal adjunct. Nunberg labels this phenomenon colon-expansion, whereby the content following the colon expands or elaborates on the preceding clause (Nunberg 30). Importantly, he points out that there are only two constraints to a colon-expansion. First, there is a semantic limit to the extent to which something can be elaborated, and, two, colon-expansions cannot themselves contain other colon-expansions (Nunberg 30-31). Faulkner violates the former by dint of the latter. Each sentence dilates in meaning as he adds on more clausal adjuncts and elaborations. While it is impractical to quote Faulkner’s multi-page sentence in its entirety, the opening sequence captures this effect: “Because he was free: in bed: in the cool familiar room…(34)”. With each colon more details about the narrative situation are revealed. The choice for concatenating clauses with a colon and not simply writing new sentences, appears to be similar to the depth effect created through parentheticals in *Absalom, Absalom!*. With each new colon the reader is tasked with processing a new piece of information that falls within the scope of one sentence. The narrative does not move forward but instead moves deeper.

It does not require inferential statistics to understand that the punctuation in *Absalom, Absalom!* and *Intruder in the Dust* are unusual, even for Faulkner. One surprising text that is a statistical outlier is “Miss Zilphia Gant.” This is a relatively obscure Faulkner story about a woman who… It sits in the middle of a cluster of stories from 1927-1934 that average a higher number of ellipses per sentence than other Faulkner texts. To mark this as a general feature of Faulkner’s style during this period would be a mistake, however; as canonical works like *The Sound and the Fury*, *As I Lay Dying*, and *Light in August*, all written around this time, do not feature nearly as many ellipses. Instead, this may have been a punctuation mark he was experimenting with in some specific contexts. “Miss Zilphia Gant” is perhaps the most extreme example.

Piece on Zilphia gant

The different techniques Faulkner uses to achieve his long sentences show a sustained commitment to an artistic vision he expressed throughout his career. Far from limiting experimentation to a particular phase or type of writing, his punctuation speaks to a constant desire to capture the long sentence. To give every character their due as experiencing their past present and future simultaneously. While Faulkner was certainly aspiring to capture a universal human condition, when these long sentences appear is not as universal as it might first seem. As a reader, it would be impossible to keep track of which characters appear in which long sentences. It is possible to capture this computationally, doing so reveals a stark fact: the long sentence appears to be reserved for some people in Yoknapatawpha county. Nor is it that every past is part of the “sum of the past” only those pasts about race are interesting.

# The Demographics of Punctuation

Understand how the long sentence may or may not apply to certain characters requires knowing what type of characters are in what specific texts. Fortunately, the *Digital Yoknapatawpha* project has at least some of this data available. Created with the mission to encode every character, location, and event in Faulkner’s Yoknapatawpha fiction, *DY* represents a rich database of every character in every event in every text that takes place in Yoknapatawpha. Started in 2012, the database has been meticulously built by over thirty Faulkner experts through a strict peer-review process that has seen each of the thousands of records receive multiple editorial passes. All data entry for the database was manually performed, which leaves room for human error and a variance in interpretations, but the continuous cycles of peer-review have kept variances to a minimum. The database focuses exclusively on the 14 novels and 54 short stories that takes place in Yoknapatawpha, and excludes roughly 30%% of the Faulkner’s writing in novels and short stories. This caveat aside, the database still represents the majority of Faulkner’s writing. In *DY* each event is any time a character or group of characters is present or mentioned at a specific location for a limited duration of time. Each character, in turn, is identified by different characteristics, race, class, gender, rank and other important attributes. By calculating what types of characters occur in each event, it is possible to build a demographic frequency model on how often and when characters occur in the corpus.

Based on the demographic frequency model, the world of Yoknapatawpha is heavily skewed towards Whites (79%) and men (80%). The distributions are also heavily left-skewed, meaning that the overwhelming majority of stories feature Whites and men, and in only very few stories do they not represent the majority of characters. Given that Faulkner was himself a White male, this is not entirely surprising. These measures give a rough indication as to the composition of demographics within the corpus. Race, class, and gender can also be linked to one another to get a better sense of the finer distinctions between characters. For example, upper class White males constitute around (24%) of the characters who appear, whereas for enslaved black women this number is a paltry (1%). Clearly, there is a stark disparity in terms of who gets represented.

Using these two data models, that of all the punctuation, and that of all the characters, it is possible to get an understanding of the relationships between characters and punctuation. As with all such inferential models, the result are very much dependent on the underlying assumptions about the data. Suffice to say that each statistical analysis sheds light on possible relationships, but there is not one single test that provides a tidy answer. Given this, relationships between characters and punctuation can be imagined at three different scales: book, character type, and individual characters. In the first, the total character composition of the book is taken and compared to sentence length, in the second the character types in each event is compared to sentence length, and in the third each individual character is compared to sentence length. As the tests get more granular, there are more data points, and generally, there is a higher confidence that the pattern is not random.

### Texts

These demographic distributions per work can be matched to the punctuation distributions. By combining these data models, it is possible to see if there is a correlation between sentence length and the types of characters that occur. To do this, a Pearson correlation was run between a race, class, gender composite of all the characters.[[7]](#footnote-34) The magnitude of a Pearson correlation runs from -1 to 1, where a negative correlation indicates an inverse relationship and a positive correlation indicates a direct relationship. Generally, a correlation above .5 is considered moderate and above .6 is considered strong (CITE THIS). There was a moderate positive correlation between sentence length and four types of characters: Enslaved women who have both Black and White ancestry (0.535), upper class men who have both Black and White ancestry (0.513), Men who have both Black and White ancestry and whose class is indeterminable (0.53), and upper class women who have both White and Black ancestry (0.334). Needless to say, in Faulkner’s highly stratified social world, these characters are anomalous. If short stories are removed from the data set, this trend becomes even more pronounced. Indicating that this is more of a feature of the novels than the corpus as a whole. By and large, this correlation highlights those novels that deal with the transgression of the South’s strict race and class boundaries. Indeed, the three novels that prominently feature the disruptive effects of black and white ancestral entanglement all feature longer than average sentences: *Absalom, Absalom* (43), *Intruder in the Dust* (39), and *Go Down, Moses* (20). The only notable exception is *Light in August* (15), which has Joe Christmas’s racial ambiguity as its main theme, but does not, on the whole complement this stylistically with long sentences as expected. Another confounding feature is that the other set of texts with above average sentence lengths is the Snopes trilogy (*The Hamlet*, *The Town*, *The Mansion*), which is actually notable for being almost entirely devoid of non-White characters (theresa). This discrepancy also underscores the limitations of this analysis. There may be some relationship between race and sentence length, but without knowing anything more about the actual sentences this is at best a tenuous link. After all, just because a text features more non-White characters than other texts does not mean that they are central to the action, or appear in those long sentences.

### Character Types

In the Digital Yoknapatawpha database, the events data table stores the coincidence of characters and locations, where each event represent one action by a single or group of characters in one location. As the definition of an event is semantic and not syntactic, it effectively means that events are not coextensive to sentences. Occasionally, an event can run for several pages capturing multiple sentences or it can be only a few words in the middle of a sentence. A digital version of the text can be delimited by the beginning and ending of events, because the database stores the first 8-10 words of each event. This process is involved, and, on occasion requires, manually aligning the beginning and ending of events. The result is a data table that shows each sentence and each event that falls within that sentence if it is a short event, and, conversely, each sentence that falls within an event for longer events. Consequently, each sentence has attached to it event data that includes what characters are present and mentioned. This gives an approximation of the relationship between characters and sentences. Notably, there is no way of knowing what role each character plays in the event: that data simply does not exist. It could be that a character is simply in the background and is not the main part of the action. The baseline assumption is that a frequent recurrence of a character indicates some salience even if they are not the focus of every event.

Since there are roughly ninety-thousand sentences in the entire corpus ranging from sentences one word to six thousand words long there is a really broad range of sentence types in which characters can appear. A Pearson correlation was used to establish the relationship between a character being present and sentence length. This is a very large sample size, and the -value, the odds that the correlation is random, tends to be quite low. Based on this data, the results are at odds with those of the text-level correlation above. Whereas the presence of characters with multiple ancestries tended to be predictive of sentence length at the level of the work, at the level of the sentence there is a strong correlation between sentence length and whether an upper class White male is present (0.86) or mentioned (0.742). Given upper class White males tend to be the protagonists of these novels. This is not all that shocking. It makes sense that the longest sentences appear when the main character is part of the action. What is quite telling is who is less likely to be present. It is true for both upper class White females and lower class White males that their correlation coeffiecent is higher when they are mentioned (0.625 and 0.497 respectively) than when they are present (0.478 and 0.311 respectively). The difference between these correlations is notable, as in each case the correlation coefficient drops quite a bit. That said, this drop does not mean that White women and lower class White men are being spoken about. Mentioned in the database is a catch-all term for any character who is mentioned in an event but does not actually appear. It could be that the event only references a character. Furthermore, correlation is not causation. If one of these characters is present it does not mean that the sentence length will get shorter, only that they are less likely to be present in longer sentences. All those caveats aside, scenes in Faulkner in which two White men, usually upper class, are discussing a white Woman who is absent, through highly wrought and experimental prose that collapses past, present and future is are not rare. Indeed, Quentin Compson discussing his sister Caddy with his father and Quentin and Shreve imagining Judith Bon, are well-mined critical veins. Likewise, both *The Town* and *The Mansion* feature two upper class White men who recount and critique the actions of the lower class white men in the Hamlet of Frenchman’s bend and the town of Jefferson. Thus, while some caution is necessary when interpreting the value of these correlations, they do comport with the social picture of Yoknapatawpha that other critics have also painted. What falls away from this picture is the correlation between sentence length and any troubling of the racial binary of the Jim Crow South. There is only a weak correlation between mixed ancestry characters and sentence length. A plausible explanation for this is that the correlation between sentence length and character type was run across the entire corpus. Since, characters of mixed ancestry do not appear in all texts, their sentence length is effectively 0 for quite a number of rows. Their individual presence in long sentences gets lost in the general mass of writing.

It is possible to tease out this relationship by performing a more fine-grained analysis and looking at character type and sentence length correlations within each work. The pattern that occurs here is quite telling. Essentially, upper class white men are consistently have a moderate to strong correlation with longer sentences, but there are also other character types whose correlation suddenly becomes stronger in relation to sentence length. Thus, in the short story ‘Lion’ both Upper Class White men and the lone lower class male with Native American and White ancestry, Boon Hogganbeck, strongly correlate with longer sentences (0.98, 0.745). However, this is likely due to the fact that Boon happens to appear frequently in scenes with an upper class white male, Quentin Compson. Indeed, the average sentence length when an upper class white male is alone is 55 words-per-sentence, which drops to round(lion\_alonegroup$alone,0) when there are other characters present.[[8]](#footnote-36)

One thing to consider with all these charactertypes is that their distribution across the corpus.

They are the relational objects that constitute the protagonist subject. Since we have the correlations for each event we also have the events specific characters appear in. Thus, we can run a correlation on which characters tend to appear in longer sentences. Here again, the top ranking is unsurprising: Quentin Compson. When Quentin appears the sentences get longer. The same is true to a lesser extent when Caddy Compson is present. Importantly, only when Jason Compson is mentioned does sentence length go up. A similar dynamic is at play, albeit very weakly, between Judith, Charles Bon, and Henry Sutpen. Judith has a stronger correlation to sentence length than both men, but only when she is Mentioned.

Finally, a different way to think about correlation is by work. One of the things that doing a correlation across the corpus masks is that not every type of character is present in every type of text. Mixed ancestry characters only appear in 13 of the 68 texts. Of these texts, only four characters would be considered major characters. So do do a correlation across an entire corpus hides these characters in the mass of data. Instead, each text has its own correlation environment. Thus, each character should be correlated with sentence length by text to get a more representative picture. Doing this shifts correlations from being very uncommon to being very ubiquitous. In part this is because major characters tend to correlate with longer sentences. Thus the correlation is a proxy for character centrality. That said, there is an interesting phenomenon in this data. Some characters have a higher correlation rate when they are mentioned versus when they are present. Generally speaking, when an event mentions, a middle class white male, a poor white male, or a lower class white female sentence length goes up more often than when they are present. This again, speaks to the power of upper class white men, who have a positive correlation with sentence length in over half the texts, to speak about others. There is no definite way to know who is doing the actual mentioning, but a correlation between characters present and mentioned reveals that: This makes intuitive sense because throughout the Town and The Mansion Gavin Stevens and Ratliff talk about Flem Snopes. This is actually born out by looking at the correlations by name, where Flem Snopes, even though he is putatively the main character in three novels has a more numerous correlation to sentences where is is mentioned versus those where he is present.

### Notes

Consider the long sentence and the flow of history. Who gets to be part of that history? How do we suture folks back into their histories. Who is embedded in these long sentences? Is it Quentin? Is it Stevens?

I can correlate sentence length to character demographics per text weighted frequency of character demographics. That is to say, does the sentence length increase when there are more upper class white characters?

I have a pretty decently cleaned up version of the numbers. Should go through the sentences again to see if there’s any major regex errors.

There do not appear to be any strong correlations between sentence length and the racial composition of the character. This effect might appear stronger when combined.

No correlations with race class gender either

“Afternoon of a cow” was published in 1947 nearly 40 years after Faulkner wrote it. Some of the texts are not always published when written.

Compson appendix, name for a city,

Adams, Richard P. *Faulkner: myth and motion*. Princeton University press, 1968.

Aiken, Conrad. *William Faulkner: The Novel as Form*. Edited by Dean Morgan Schmitter, Mac Graw-Hill Book Company, 1973, pp. 45–52.

DeVoto, Bernard. *"Witchcraft in Mississippi." Saturday Review of Literature, 15 (October 31, 1936), 3-4, 14*. Edited by M. Thomas Inge, Cambridge University Press, 1995, pp. 144–49.

Faulkner, William. *Absalom, Absalom!* Knopf Doubleday Publishing Group, 2011.

---. *Absalom, Absalom!* Knopf Doubleday Publishing Group, 2011.

Hlavsa, Virginia V. James. *Faulkner and the Thoroughly Modern Novel*. University Press of Virginia, 1991.

Hoover, David L. “Modes of Composition in Henry James: Dictation, Style, and What Maisie Knew.” *The Henry James Review*, vol. 35, no. 3, 2014, pp. 257–77, <https://doi.org/10.1353/hjr.2014.0024>.

Hubert, Mia, and Ellen Vandervieren. “An Adjusted Boxplot for Skewed Distributions.” *Computational Statistics & Data Analysis*, vol. 52, no. 12, 2008, p. 51865201, <https://doi.org/10.1016/j.csda.2007.11.008>.

Ihrmark, Daniel, and Johan Nilsson. “A Corpus Stylistic Analysis of Development in Hemingway’s Literary Production.” *The Hemingway Review*, vol. 40, no. 2, 2021, pp. 71–93, <https://doi.org/10.1353/hem.2021.0005>.

Newman, Mej. “Power Laws, Pareto Distributions and Zipf’s Law.” *Contemporary Physics*, vol. 46, no. 5, Sept. 2005, pp. 323–51, <https://doi.org/10.1080/00107510500052444>.

Polk, Noel. *Children of the Dark House: Text and Context in Faulkner*. Univ. Press of Mississippi, 1998.

Rudnicka, Karolina. *Variation of Sentence Length Across Time and Genre: Influence on Syntactic Usage in English*. Edited by Richard J. Whitt, John Benjamins Publishing Company, 2018.

Schou, Karsten. “The Syntactic Status of English Punctuation.” *English Studies*, vol. 88, no. 2, Apr. 2007, pp. 195–216, <https://doi.org/10.1080/00138380601042790>.

Sun, Kun, and Rong Wang. “Frequency Distributions of Punctuation Marks in English: Evidence from Large-Scale Corpora.” *English Today*, Dec. 2018, <https://doi.org/10.1017/S0266078418000512>.

Van O’Connor, William. “Faulkner.” *The Georgia Review*, vol. 12, no. 1, 1958, pp. 83–86, <https://www.jstor.org/stable/41395494>.

1. Before it became possible to do computational analysis of texts, there have been any number of precursors that anticipate the methodology. Notably the concordances (CITE), but also the work of Virginia Hlavsa in 1991 uses word counts to make its argument (51-53) Hlavsa [↑](#footnote-ref-21)
2. Overall, the revised texts represent about 5% of the corpus. They range from being near duplicates to differing substantially from the version of the story that appears in the novel. The total duplicated texts from these works is about .4% of the entire corpus, which is substantial enough to acknowledge, but not significant enough to skew the data. [↑](#footnote-ref-25)
3. To simplify matters, his poetry and movie scripts can be left out of consideration for different reasons. The volume of poetry is so low that computational analysis is not a fruitful methodology. Movie scripts have multiple authors and the punctuation patterns will likely to vary too much from script to script. The only major Faulkner text that has been excluded from analysis is *Requiem for a Nun.* It hovers between novel and play, and also, as it happens contains one of Faulkner’s longest sentences. Yet, since much of the text is written in the form of a play, it was excluded from analysis. [↑](#footnote-ref-26)
4. Indeed, in an earlier piece on punctuation in *Absalom, Absalom!* sentence length was calculated by accounting for embedded quotation. This created sentences of, on average, 49 words per sentence, which is substantially higher than the 43 words per sentence arrived at in this paper that counts embedded quotes as individual sentences. [↑](#footnote-ref-28)
5. There is an extensive critical literature on the “flow” of Molly’s soliloquy exhaustively documented by Derek Attridge in “Molly’s Flow: The Writing of”Penelope” and the Question of Women’s Language”. [↑](#footnote-ref-30)
6. As previously indicated, *Requiem for a Nun* was removed from this data set, because it was too much at variance with the other texts. [↑](#footnote-ref-31)
7. Unless otherwise indicated the threshold for significance was set at <.05. More plainly, only correlations where there was a less than 5% chance of the pattern being random were included. [↑](#footnote-ref-34)
8. It is not possible to run a correlation test on only those sentences where a character is alone, because there are too few data points. Generally, though it is true that a character alone, irrespective of race, class, or gender tends to be embedded in longer sentences than in a group. That said, the plurality of events where a character was alone featured upper class white males (33% [↑](#footnote-ref-36)