Reading Faulkner Quantitatively

An Algorithmic Criticism of the Novels of William Faulkner

CS+English Capstone Project

English 198 / CS191W

June 13, 2018

Dan McFalls
dmcfalls@cs.stanford.edu

https://github.com/dmcfalls/Faulkner

Introduction

> *"It is a curious fact that with every novel he produces William Faulkner's*
> *position in American letters becomes less easy to define."*

-Helen Neville, June 1938

There is a history of disagreement in the criticism of Faulkner's novels. Both when his novels were being released and up to the present day, there has been a great deal of difference of opinion regarding which novels have been successful, which ones have failed, which ones are crucial to his novelistic project, which ones are worth reading, which ones are not. The two late-career novels which earned him the Pulitzer Prize in fiction, *A Fable* and *The Reivers*, have been retrospectively regarded as lesser or failed works, and are largely ignored by the majority of criticism, with the former variously dubbed Faulkner at his "most turgid," "a restless residue of the Faulknerian imagination," and "an earnest and highminded mistake" (Emerson 248). By contrast, the novel that many consider Faulkner's *magnum opus* and the high point of his career, *Absalom! Absalom!*, met with mixed reception at the time of its publication, with some reviewers remarking that "this is his master work" and "a tremendous book, a novel of foremost importance", but other reviewers calling it a "disgusting book" and "a most elaborate inarticulateness" (Inge 141-163). *Pylon*, which preceded *Absalom! Absalom!* and met with almost uniformly warm reviews from critics, is today far less revered than its successor. On the whole, Faulkner has been characterized in the best and worst light throughout his career, with disagreements about his style, reservations about his subject matter, and delusions arising from his reputation causing extreme variances in opinion. With the advantage of a retrospective lens, and with such subjectivity affecting the history of Faulkner criticism, I hope to elucidate and clarify some of that criticism by taking a far-away, quantitative approach to analyzing Faulkner's novels. By investigating the basic, structural features of the novels in light of the entire trajectory

of Faulkner's career, I locate some misguided intuitions about variations in his style, provide

some commentary on why some novels are regarded as they are, and provide pieces of additional

evidence for some of the prevailing views that have come to characterize the present discussion

of Faulkner's novels.

For the primary study in this project, I have gathered several basic statistics about the

novels and charted them both over time and as pertaining to a few meaningful characterizations

of the Faulkner's novels. The basic statistics are broken into three main categories: "basic

metrics," sentence complexity metrics, and gender statistics. The "basic metrics" include all of

the data that can be extracted fairly straightforwardly from the text using simple counting,

tagging, and averages. I use two more complicated measures of sentence complexity to provide

some insight into how the novels are characterized by difficulty. The gender statistics chart the

extent to which male vs. female characters "take up space" in the novels, which, while making

no comment on the nature of their participation, gives a satisfactory high-level impression of the

extent to which female characters factor into various novels.

In addition to the high-level overviews, I use a technique called $tf\text{-}idf$ to find the words

most characteristic in a body of text, both in order to find words most particular to characters

within a given work and to find the words most particular to sections within given novels.

I will begin with a discussion of the algorithmic critical methods used, followed by a

brief discussion of the corpora. I will then give a high-level analysis of the novels, charting

quantitative changes over time and differences across meaningful categories of the works using

the basic metrics and sentence complexity metrics. I will follow with a discussion of individual

novels using a combination of the $tf\text{-}idf$ results and close reading, and I will end with some

ideas for future research and concluding remarks.

Overview of Algorithmic Methods

The high-level structural features of novels are limited in capturing information about the content of novels, but they can tell us something about the form. It is important not to assume that any of the chosen metrics have any inherent meaning about the text on their own, but rather that they are suggestive of certain hypotheses more than others. Regarding the measures of sentence complexity I have chosen, for example, nobody would agree that *The Sound and the Fury* is a simple novel to read, even though it ranks in the bottom two of all Faulkner's novels in both measures of sentence complexity—the metrics suggest complexity of a certain kind within a certain context, but are by no means holistic or conclusive. As Stephen Ramsay argues, text analysis "must endeavor to assist the critic in the unfolding of interpretive possibilities" (484). Following this mantra, I will use the computational results only to steer my analysis, always returning to the novels to understand the data, rather than the other way around.

When doing computational analysis of texts, there are several existing approaches for converting the raw text of the novel into a more useful format. Different statistics are concerned with different versions of the text, which, for my purposes, are defined as follows:

- **Text**: The plain text, split into tokens by whitespace, and converted to all lower-case.

- **Sentences**: The sentences of the novel as parsed using the Natural Language Toolkit's (NLTK) sentence parsing algorithm.

- **Stemmed text**: The text, split into tokens, with words converted to stem form, such that different forms of a word (e.g. "fly," "flies," and "flew") are considered equivalent.

- **Stemmed and filtered text**: The stemmed text with the common words in English, such as articles and prepositions, filtered out.

In the remainder of this section, I will define all of the quantitative metrics that I use in the analysis as a reference point for the reader.

Basic Metrics

- **Word count** (WC): The number of words that appear in a text.

- **Sentence count** (SC): The number of sentences that appear in a text.

- **Unique words** (UW): The number of unique character sequences in a text, ignoring case.

- **Unique words (stemmed)** [UWS]: The number of unique words in the stemmed text.

- **Average word length** (AWL): The average number of characters per word.

- **Average sentence length** (ASL): The average number of words per sentence.

- **Lexical diversity** (LD): The number of unique words in the stemmed text divided by the total number of words in the stemmed text.

- **Part of speech percentages** (Noun%, Verb%, Adjective%, Adverb%, Pronoun%): The percentage of the total words in the text of a given part of speech.

Sentence Complexity Metrics

- **Average tree depth** (ATD): The average maximum depth (in distance from root) of the sentence tree produced by SpaCy's (a natural language processing tool) sentence parser (the result of which is analogous to grade-school sentence diagramming).

- **Average dependency distance** (ADD): The average distance of dependencies between words as tagged by Stanford CoreNLP's dependency parser. A parsed sentence looks like the below diagram, with each arrow representing a dependency, and the distance being

Figure 1 - The first sentence of <u>The Sound and the Fury</u>, dependency parsed by Stanford CoreNLP
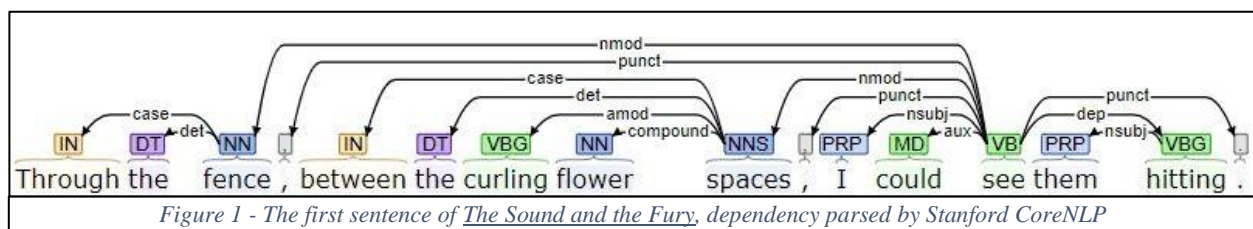
the number of words between the two words at the endpoints of the arrows, counting

inclusively. Using ADD as a measure of sentence complexity was proposed by Masanori

Oya, who proposed using the Stanford CoreNLP dependency parser and noted that

sentences with word counts between 10 and 20 words are the most effective measures of

sentence complexity (Oya 316). In order to make ADD not depend on sentence length, I

measure ADD only across the sentences of a text that are between 10 and 20 words.

Gender Metrics

- **Fem. word percentage** (Fem%): The percentage of words in the text that occur after a

  female pronoun, female name, or other female "marker word" (e.g. "woman").

- **Masc. word percentage** (Masc%): The percentage of words in the text that occur after a

  male pronoun, male name, or other male "marker word" (e.g. "father").

- **Gendered word percentages, strict** [Fem% (strict), Masc% (strict)]: The percentage of

  words in the text that occur between two female pronouns, names, or marker words

  without a male pronoun, name, or marker word between them.

- **Fem.-to-masc. weighted ratio** (FTMWR): The ratio between female and male

  pronouns/names/marker words averaged with the ratio between Fem% and Masc%.

Charts including all of the data I gathered will be publicly available, and any metrics that

appear there but not above should be self-explanatory or easily understood from the above.

A Note on the Corpora

I have limited the scope of this project to novels, and so from Faulkner's works I have

selected only those that the author referred to as novels. The texts I am using are for the most

part corrected versions of the original novels that Faulkner wrote. In line with this theme, I am

using the original text of *Flags in the Dust*, rather than the highly edited version published as

Sartoris in 1929 (for chronological analyses, I will still consider this text as having been

produced in 1929, and I will include it where one would include *Sartoris* in the ordering).

Part of the motivation for focusing on novels is that novels are typically composed of

well-formed sentences and easy for the chosen NLP tools to parse. Nevertheless, Faulkner,

originally a poet and often cited as indebted to the Romantic poets and Greek dramatists, and as a

modernist and stylistic experimentalist, blended a variety of forms, including poetry and drama,

and styles, especially stream-of-consciousness, into his works. Modern NLP suites (in the case of

this project: NLTK, SpaCy, and Stanford CoreNLP) are pretty good at handling odd inputs not

appearing as straightforward prose, but the non-standard nature of the novels' form means that

results will not always be perfect, and some of the statistics, especially those that rely strongly on

sentence-level parsing (SC, ASL, ATD, ADD), should be taken with a grain of salt.

In addition to my corpus of Faulkner novels, I am using a selection of novels from the

first half of the 20th century (and a few from the second half) that are representative of

Faulkner's contemporaries. The novelists represented are Fitzgerald, Hemingway, Hurston,

Joyce, Nabokov, O'Connor, Pynchon, Steinbeck, and Woolf). These novelists are meant as

references against whom to compare Faulkner and his work. Though not exhaustive, I have tried

to include authors whose works act as archetypes for some useful literary features (e.g. a plain

stylist in Hemingway, or a Southern author in O'Connor).

As a final reference point, I am using twenty-one randomly selected dime novels from the

early part of the 20th century as an example of "lower-brow" literature. These texts are from the

Norther Illinois University dime novel corpus, and were generated using optical character

recognition (OCR), which produces a large amount of small errors in the plain text. As a result,

some of the metrics are distorted for these texts, and a few (e.g. LD) are essentially meaningless.

The Analysis

Over the course of his career, Faulkner's novels increase in formal complexity at the

level of the sentence. *Light in August* and *As I Lay Dying* are notable for being relatively easy

novels to read as far as Faulkner's work goes, and this concords with a general phenomenon that

Faulkner's earlier work is more widely read: I argue, *because* it is easier to read. Even *The Sound and the Fury*, which has a reputation of difficulty, quantitatively appears as one of the easiest works in Fau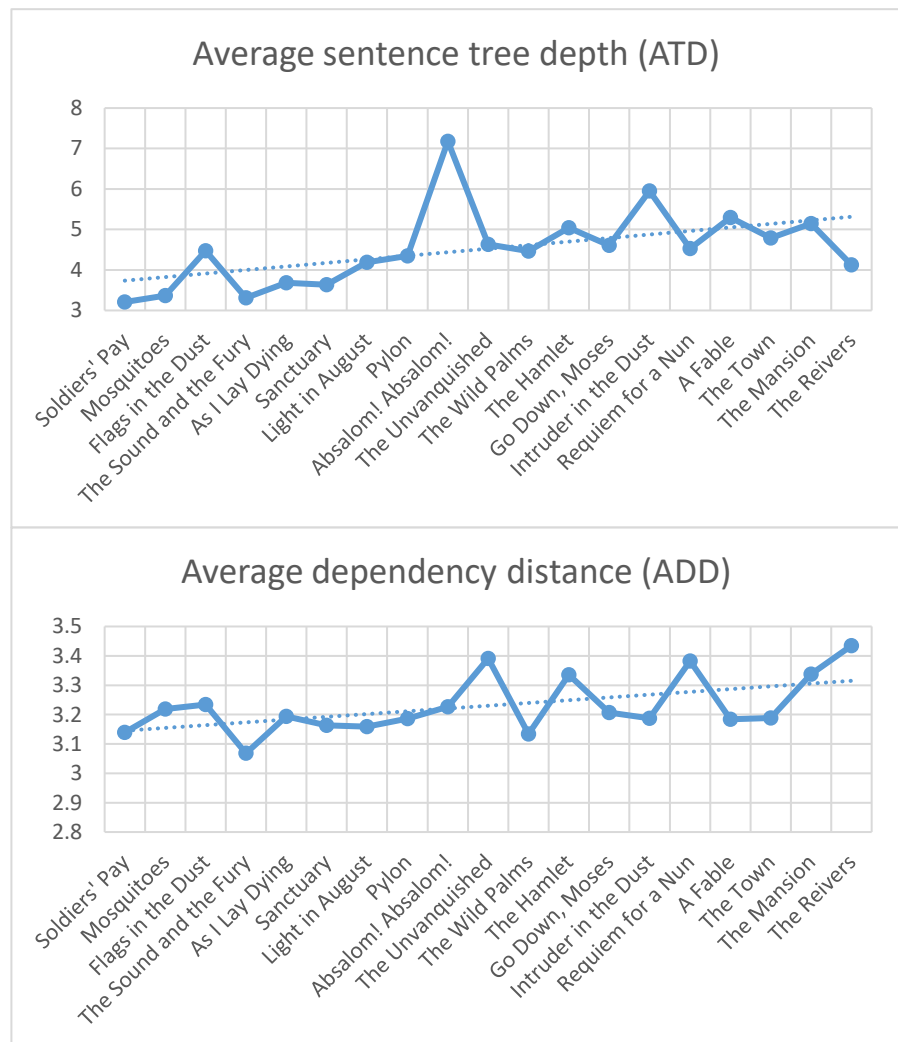lkner's oeuvre in terms of sentence complexity, which is likely attributable to the fact that both the Benjy and Quentin sections


Average sentence tree depth (ATD)

Average dependency distance (ADD)

feature many short, clear sentences whose individual meanings are clear, but which are difficult

to parse as a continuous whole. The average sentence tree depth (ATD) and average dependency distance (ADD) both increase on average over the course of Faulkner's career, but peak in different novels. The clear outlier in ATD is *Absalom! Absalom!,* which famously features a sentence often cited as one of the longest published in English. One might question whether *Absalom! Absalom!* has a high ATD solely because its contains longer sentences on average than Faulkner's other works. But if one examines the average sentence length (ASL) over the course of Faulkner's career, one sees that, although *Absalom! Absalom!* has the second highest ASL, *Intruder in the Dust* has the highest but does not exhibit a proportionally
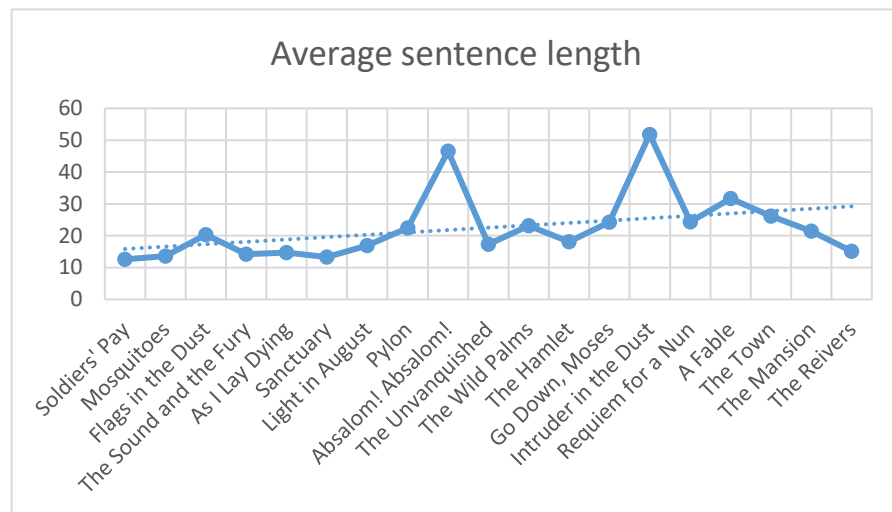
high ATD. This suggests that *Absalom! Absalom!* has sentences with are on average especially complex for their length[1]. This makes sense given the focalizing characters in the two novels: *Absalom! Absalom!* is largely from the perspective of the troubled, intellectual Quentin Compson, whereas *Intruder in the Dust* is presented from the perspective of a child, Chick, in such a way that exhibits "the objective-subjective confusion in the focus, a child in transition to an adult" (Hutchinson 39).

    If one considers the ADD, the results tell a different story, namely that *Absalom! Absalom!* is not uncharacteristically complex for sentences between 10 and 20 words. This

---

[1] As a counterpoint, longer sentences certainly do seem to have an effect on the ATD; I ran the ATD algorithm on an excerpt from "The Green Coaster" in Jonathan Coe's *The Rotters' Club*, a supposedly 13804 word-long sentence, which SpaCy parsed as 11 sentences and which the algorithm gave an ATD of 17.58 (*Absalom! Absalom!* is 7.18)

suggests that *Absalom! Absalom!* achieves its complexity in the way that its sentences build—the

sentences are on average longer than in his other novels and on average those longer sentences

are relatively complex. At any rate, *Absalom! Absalom!* is a stylistic outlier among the novels

most often cited as Faulkner's greatest works[2].

Perhaps the most surprising feature of the above plots of ATD, ADD, and ASL, are that

they all increase over time. Given that the first half of Faulkner's career, beginning with *The

Sound and the Fury*, is usually retrospectively considered his most successful, this presents an

interesting possibility why. The latter half of Faulkner's career contains many works that are

either outside of Yoknapatawpha county (*The Wild Palms*, *A Fable*) or are sequels to previous

Yoknapatawpha county books (*Requiem for a Nun*, *The Town*, *The Mansion*). There is an

impression that Faulkner had a continuing impulse to express himself in increasingly complex

style but had a diminishing reservoir of material well-suited to its expression: namely, the central

stories of Yoknapatawpha. This culminated in the laboriously conceived *A Fable*, whose plot

was so complicated that the author outlined it in ink on the walls of his study. Leslie Fielder

wrote of the novel in a contemporary review: "*A Fable* finds William Faulkner at the critical

point in his career when his ideas and themes, striving desperately to open out, come into conflict

with his style which has been for several years closing in, rigidifying" (Inge 389). The discovery

of the triumphs of his early work made critics too eager to honor and praise his later work, even

if only in an attempt to give a writer overdue recognition.

Another possible explanation for the lasting success of Faulkner's earlier works is that

over time, Faulkner indisputably writes less about women. The three major metrics which chart

---

[2] If we decide that we trust the editors of Modern Library and my Narrative and Narrative Theory instructors, these would be *The Sound and the Fury*, *As I Lay Dying*, *Light in August*, and *Absalom! Absalom!*.

Fem% (strict)

Female-to-Male Weighted Ratio (FTMWR)
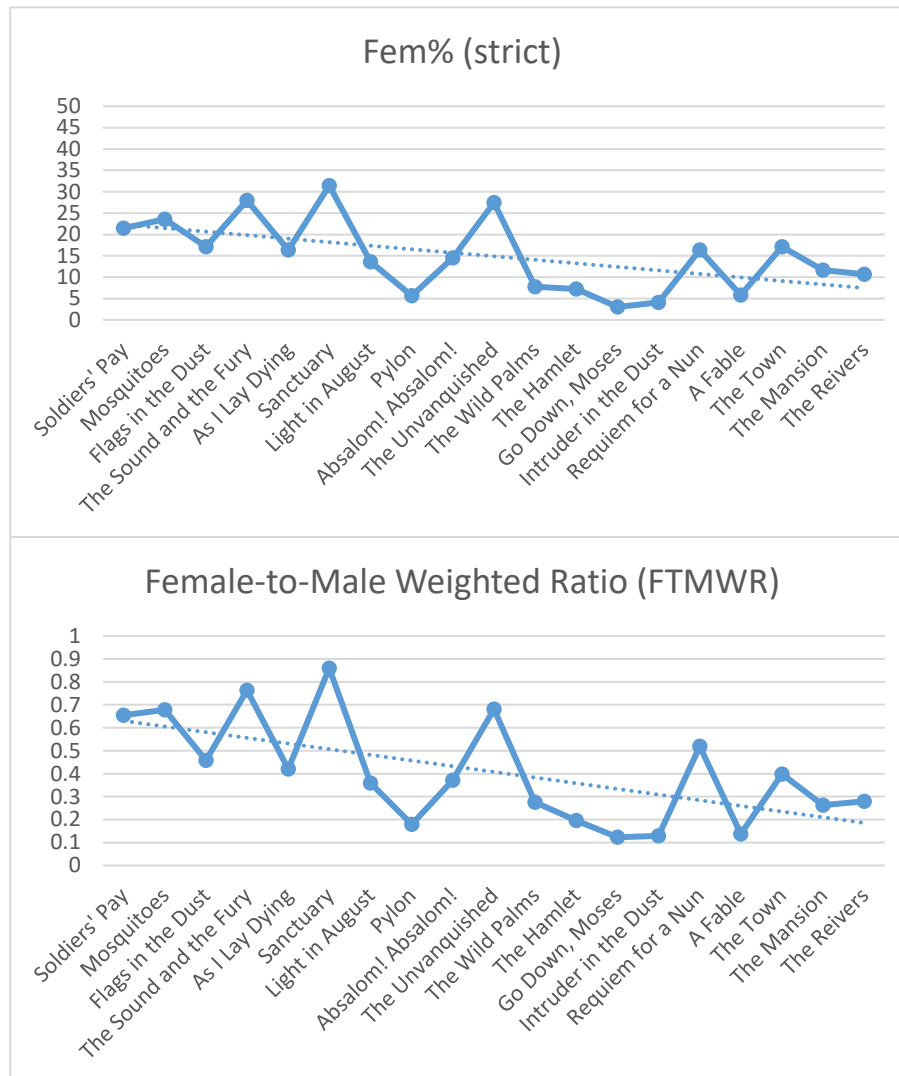
the percentage of a text inhabited by feminine presences, Fem%, Fem% (strict), and female-to-male weighted ratio (FTMWR) all produce essentially the same graph, which illustrates that, on average, the extent to which Faulkner's novels include female personae drops by about half across the span of his career. The charts showing Fem% (strict) and FTMWR, a statistic that averages the ratio of female-to-male pronouns, names, and marker words with the ratio of text enclosed by such words, show a clear and significant decline, with *Sanctuary* scoring highest and three late works, *Go Down, Moses*, *Intruder in the Dust*, and *A Fable* being notably and excessively masculine novels.

Faulkner's ability to write about women, as with most aspects of Faulkner criticism, has been widely disagreed upon. And not unreasonably so—it's unclear in what light to view Temple Drake's decision to condemn an innocent man essentially to his death, or Lena Grove's relentless, naïve pursuit of the father of her child, or another future. One view says that these are

simplistic, two-dimensional characters modeled on stereotypes about womanhood; another says

that they are original, perplexing characters who excise themselves from the constraints placed

upon them in acts and manifestos of defiant definition. O.B. Emerson gives this especially

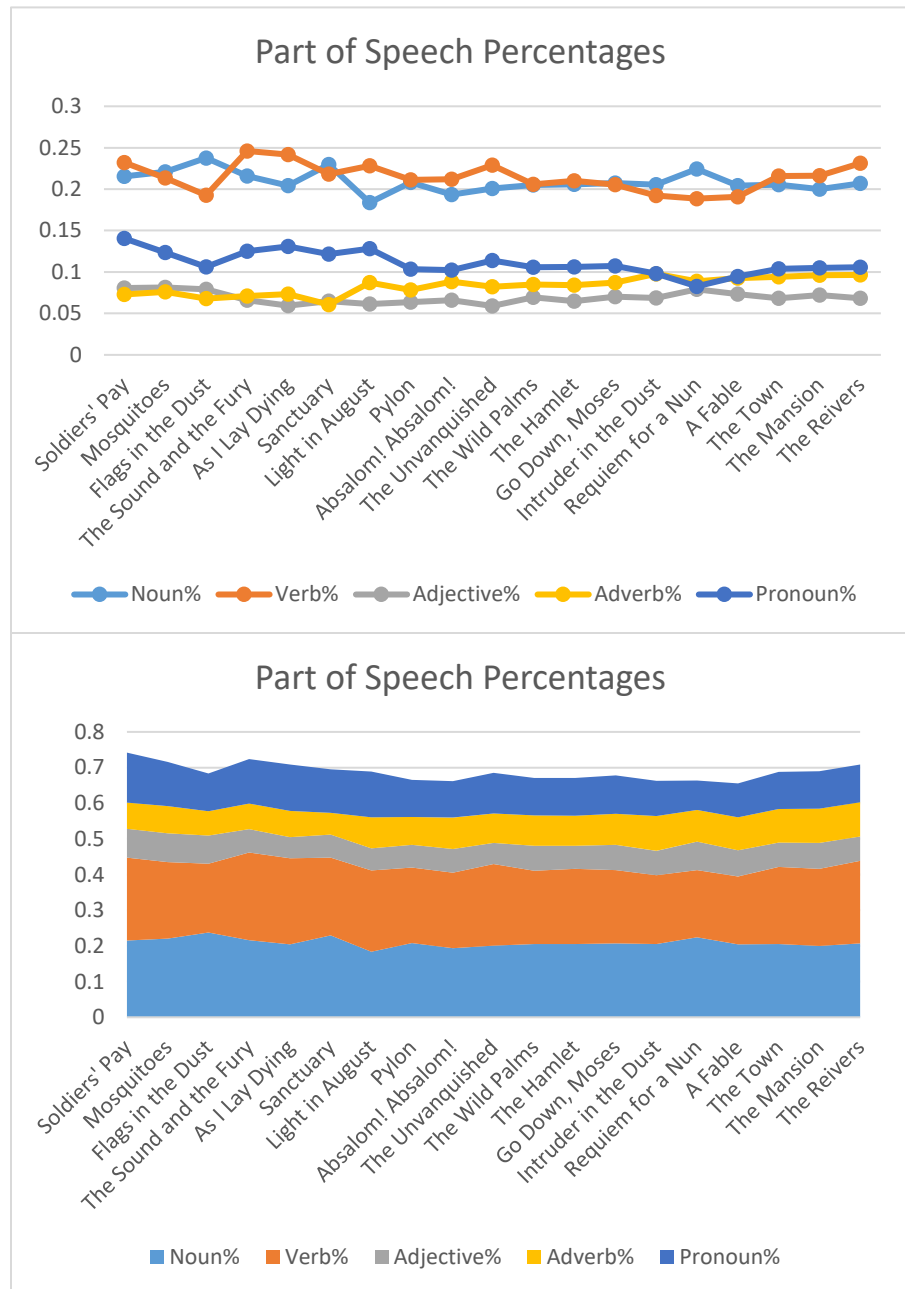amusing juxtaposition about the extent to which critics disagree on the matter:

> "In 1929 writing on *Sartoris*, J. Dana Tasker observed that Faulkner understands women better than he
>
> understands men. Twenty-eight years later Irving Malin observed that Faulkner's women are superficial,
>
> and Faulkner's fiction suffers from the author's inability to portray women because he knows so little about
>
> them." (Emerson 12).

What changed over twenty-eight years? I would argue that part of it has to do with Faulkner's

novels including less female presence over time. That *Requiem for a Nun* scores about half of

what its predecessor *Sanctuary* does on all the gender metrics, despite being its sequel, is telling.

Novels from the late career like *Go Down, Moses*, which are still relatively highly regarded

critically, may suffer from reaching too much toward examples of "Great American Novels" in

search of attaining status as the definitive one—*Go Down, Moses* in which "The Bear"

dominates has a Fem% score of 9.84 and its major inspirational source, *Moby-Dick*, scores

9.74—thus these late works suffer from the chronic deficiency that these "great American

novels" excessively tend toward a masculine flavor (Wallach). The data and criticism suggest

together that Faulkner is at his best in the novels that have more robust feminine presences[3].

    Two other noteworthy statistics, the part of speech metrics and lexical diversity, are less

provocative but help clarify some otherwise tricky questions regarding Faulkner's work. I have

---

[3] A note on the numbers for comparison's sake: Faulkner's highest Fem% score is 44.79 in Sanctuary and his lowest is 9.74 in *Go Down, Moses*. Among some of his contemporaries, *The Great Gatsby* scores 34.23, *The Old Man and the Sea* scores 1.70, *To the Lighthouse* scores 59.47, and O'Connor's *A Good Man is Hard to Find* scores 49.86. The highest score in the modernist corpus is Nabokov's *Lolita* scoring 69.85. Which demonstrates that these metrics only show how much space female presences take up in a novel, but they say nothing about the capacity in which they figure in the texts.

presented the five part of speech metrics in two ways: from the line graph one can discern which kinds of words are characteristic in given works, and the area graph shows that the percentages stay relatively consistent over time. The line graph shows a few notable outliers, beginning with *Light in August*, which has a relatively high pronoun to noun ratio. This accords with one distinctive stylistic feature of *Light in August*, characters not listening to each other and being described repetitiously in their steadfastness, as in the following description of Lena:

> "But she is not listening apparently…She is not thinking about this at all. She is thinking about the coins knotted in the bundle beneath her hands. She is remembering breakfast, thinking how she can enter the store this moment and buy cheese and crackers and even sardines if she likes" (Faulkner 26).

The diminished amount of verbs in *Requiem for a Nun* is something of a red herring—given that a great deal of the novel is constructed as a drama, and that the words "said" and "would" are at the top of the list of most frequent (stemmed/filtered) words in nearly all Faulkner's novels, it's unsurprising to see fewer nouns in that work. *Flags in the Dust* is the only other novel in which a name ("Bayard," in this case) appears more frequently than the word "said." A contemporary review suggested that from the novel, "above all else one will remember the suffering mind and imagination of young Bayard Sartoris" (Inge 29). The obsession of the book with a single man and his thoughts comes across now instead as a weakness—in the subsequent four books that he would publish, Faulkner would find lasting power in his ability to create dynamic casts of characters, rather than singular, gravitational personages.

It is surprising that *Absalom! Absalom!* does not have a significantly high percentage of adjectives relative to the other texts. The novel famously opens upon Miss Coldfield's office on a "long still hot weary dead September afternoon," and the novel seems so distinctively stuffed with adjectives that one contemporary reviewer criticized the novel for having sentences "freighted with adjectives" (Faulkner 3, Emerson 37). Yet the text as a whole contains no abnormal amount of descriptors; Faulkner's style stays fairly consistent throughout his career in this regard. My guess is that reviewers and readers alike are prone to remembering distinctive passages and incorrectly extrapolate those passages' features to the rest of the text, distorting their impressions of the novels' structural qualities due to a few striking exposures.

A curious fact about Faulkner's novels is that his most critically revered novels tend to also be those that are the least lexically diverse. Shorter novels tend to have disproportionately high lexical diversity, since they have fewer chances to recycle words, but *The Sound and the Fury* and *As I Lay Dying*, two of Faulkner's shortest novels, and *Light in August* and *Absalom!*

Absalom!, two of his longer ones, all fall below the average LD for his oeuvre. Also curious is the fact that the average LD for all novels set in Yoknapatawpha

Lexical diversity (LD)
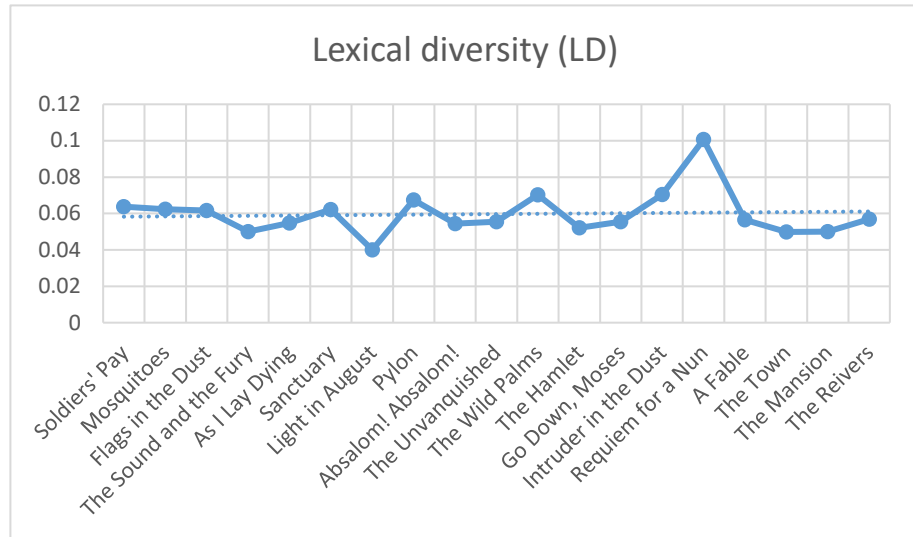
country is less than the average LD for all novels set elsewhere. As for why this might be, I am

inclined to think that monotonous and repetitive effects that the past has upon the present for the

characters in Yoknapatawpha manifest themselves in the language. As words ("honeysuckle,"

"adze," "wisteria") accrue meaning for the characters, they gain an inertia in the text.

Character Analysis with $tf\text{-}idf$

Departing from high-level analyses of Faulkner's novelistic career, I will now briefly

focus on a technique for identifying the words most distinctive to a given body of text within a

corpus of texts, and how it can be applied to identify words most distinctive to given characters

within *The Sound and the Fury* and *As I Lay Dying*. $Tf\text{-}idf$, or "term frequency"-"inverse

document frequency," is a measurement of how often a word appears in a particular text relative

to how infrequently it appears across an entire corpus. Stephen Ramsay originally proposed

using $tf\text{-}idf$ to obtain the words most characteristic to the six narrator-characters in Woolf's *The

Waves*, exploiting the experimental structure of the novel. The entire text of the novel (except for

the cinematic, naturalistic introductions to the chapters) has the form of the speech of one of the

six characters, which of whom is indicated explicitly by the phrase "said _____". Ramsay treated

the collection of words spoken by each character as its own text, and the novel as the corpus, and

applied $tf\text{-}idf$ accordingly (Ramsay 484).

Using Ramsay's article as an inspiration, I have used a similar application of $tf\text{-}idf$ to

compute the words most distinctive to characters in *As I Lay Dying*, in which each chapter is

composed of a character's narration, and *The Sound and the Fury*, which contains four sections

each focalizing one of four major characters.

I use a simple modification of $tf\text{-}idf$ that uses logarithms to smooth the terms. Defining

tf as the number of times a term appears in a section, df as the number of sections in which the

word appears, and N as the number of sections, my formula is as follows:

$$tf\text{-}idf = (1 + \log(tf)) \cdot \log\left(1 + \frac{N}{df}\right)$$

Consider the twenty highest-weight terms from each section of *The Sound and the Fury*:

Benjy Section		Quentin Section		Jason Section		Dilsey Section	
Words:	Weights:	Words:	Weights:	Words:	Weights:	Words:	Weights:
versh	6.38533	shreve	8.225634	dam	8.116477	negro	6.261311
charlie	6.16931	spoade	7.463906	earl	6.853142	sheriff	6.169318
said	5.41357	gerald	6.790018	check	5.856836	breddren	5.737564
the	5.37669	anse	6.430883	i	5.750165	the	5.653430
bellering	5.31530	bland	6.348329	to	5.472870	ben	5.480959
and	5.21493	dalton	6.169318	says	5.337740	and	5.411281
you	5.21391	squire	6.071747	the	5.330024	hears	5.145734
bowl	5.14573	ames	6.071747	and	5.269712	de	5.081425
to	5.11785	julio	5.967876	you	5.261195	to	4.998007
i	5.09282	the	5.887118	telegraph	5.145734	luster	4.997227
luster	5.09279	loaf	5.737564	a	5.113157	he	4.968143
vershs	4.74125	i	5.513735	it	5.009053	seed	4.956169
branch	4.71945	and	5.484695	busy	4.956169	presently	4.956169
it	4.71904	twentyfive	5.468701	she	4.924667	a	4.932360
he	4.66943	shreves	5.315305	that	4.741571	of	4.901061
caddy	4.59888	a	5.241886	jews	4.741259	said	4.846113
on	4.51196	you	5.195322	headache	4.741259	vanished	4.741259
parlor	4.49316	to	5.175137	ford	4.741259	somethin	4.741259
overshoes	4.49316	vest	5.145734	alley	4.741259	you	4.735500
froze	4.49316	trout	5.145734	account	4.741259	in	4.702189

Some of the words are not surprising to see—it makes sense that Shreve would be associated with Quentin's section, since he doesn't appear in any others. Some of the results are more illuminating, though: a quick glance reveals Benjy's obsession with his sister Caddy, and Quentin's obsession with Dalton Ames, her lover. It's notable that the word "presently" is so particular to Dilsey's section: it suggests from her focal point an importance of the present moment and a kind of interaction with time in which the past is less intertwined with the present (it most certainly is for Benjy and Quentin). Knowing that "presently" appears only and distinctly in Dilsey's section and the epilogue lends a new impression to lines like these:

> "Mrs Compson returned to her room. As she got into bed again she could hear Dilsey yet descending the stairs with a sort of painful and terrific slowness that would have become maddening had it not presently ceased beyond the flapping diminishment of the pantry door." (Faulkner 268)

Thinking further about how time seems subject to Dilsey inversely to the way that it commands Benjy and Quentin, Dilsey's strange, prophetic lines take on a special weight: "I've seed de first en de last…Never you mind me…Never you mind,…I seed de beginning, en now I sees de endin" (297). In a mire of the past, Dilsey in some sense predicts the future.

The words captured by $tf\text{-}idf$ are merely suggestive, but they provide further evidence for the pathways of inquiry we deem worth pursing in light of them. The results from *As I Lay Dying* are perhaps more intriguing, as the cast of characters for which there is analysis is more populous. I've provided a few examples below. The top words for Darl, "lantern," "motion," "glare," "moment," all point to the incident where he inexplicably sets the barn his family is staying in on fire. Many of Dewey Dell's highest-weight words are associated with her unwanted pregnancy: "daughter," "guts," "naked." Addie's most distinctive vocabulary—"violated," "tricked," "terrible"—makes plain the horror of her speech from the grave and her status as a victim, bringing the texture of *As I Lay Dying* as a Southern Gothic novel to the forefront.

DARL		DEWEY DELL		ADDIE	
Words:	Weights:	Words:	Weights:	Words:	Weights:
lantern	10.6279240877	tub	7.74040081959	violated	5.81858876392
motion	9.66221287526	lafe	7.27172070042	tricked	5.81858876392
glare	9.42096611271	daughter	7.2348981274	switch	5.81858876392
moment	9.15671018307	guts	6.84227213506	deeds	5.81858876392
rigid	8.86458880559	thief	6.61621283359	terrible	5.58436978214
mack	8.86458880559	sneak	6.61621283359	shape	5.51761854224
expression	8.86458880559	row	6.61621283359	school	5.10682781837
pas	8.53802488926	moaning	6.59020864581	fear	5.10682781837
chalkline	8.53802488926	nuzzles	5.81858876392	sin	4.94206919193
carefully	8.53802488926	mi	5.58436978214	anse	4.8311709949
alert	8.53802488926	3	5.58436978214	words	4.83110643763
moves	8.20333017508	cow	5.51761854224	worst	4.69440077791
returns	8.167797256	sack	5.27836240506	worlds	4.69440077791
reins	8.167797256	shake	5.10682781837	vessel	4.69440077791
raincoat	8.167797256	pale	5.10682781837	revenge	4.69440077791
clear	8.167797256	stint	4.69440077791	refused	4.69440077791
pale	7.93547276792	neighbors	4.69440077791	necessary	4.69440077791
lifts	7.78782345813	lafes	4.69440077791	invented	4.69440077791
stoops	7.74040081959	eats	4.69440077791	geese	4.69440077791
profound	7.74040081959	dish	4.69440077791	flow	4.69440077791
gums	7.74040081959	cupboard	4.69440077791	ending	4.69440077791
erect	7.74040081959	clumps	4.69440077791	echo	4.69440077791
coat	7.74040081959	blotched	4.69440077791	boiling	4.69440077791
slowly	7.62922749448	blast	4.69440077791	aloneness	4.69440077791
current	7.45793080415	darkness	4.67548508897	strange	4.49116914928
surface	7.27172070042	slope	4.49116914928	cora	4.2756655179
yards	7.2348981274	rushing	4.49116914928	blood	4.25675736859
woodenbacked	7.2348981274	naked	4.49116914928	i	4.06706496117
stooping	7.2348981274	kilt	4.49116914928	father	4.06588163939
snuff	7.2348981274	died	4.49116914928	children	4.06588163939

Naturally in Vardaman's list (not shown) the word "fish" appears, but only beneath the more distinctive words "track," "moon," and "train," all of which suggest a somewhat Romantic characterization of the family's youngest son. Cash's section has the words "build" and "stress" that one would expect, but also "music" at the top of the list, which is not something one might intuitively associate with the linearly-minded carpenter. There are numerous possibilities for inquiry; I leave it to the reader to peruse the lists at their leisure and decide what resonates.

Conclusion

Faulkner has made a particularly good case study because his career has clear early, middle, and late periods, and because he experimented with several styles and subject matters

that lent his work noticeable variability. Other authors with more limited corpora might not have work that changes as much structurally over time. Still, all the methods used in this study are easily exported to any corpora of literature. Even the $tf\text{-}idf$ analysis, which demands some kind of structure to assign words to texts, could be generalized in an approximated fashion by using a method that makes assignments statistically or using the closest relevant collocation as an indicator of the section to which to assign a given word.

For future research, I would be interested in investigating more nuanced ways of analyzing gender in texts computationally. Generalizing $tf\text{-}idf$ successfully for novels and stories would certainly be a worthwhile goal. On the more ambitious end of things, my project has shown that even the state-of-the-art text and sentence parsers make frequent mistakes when presented with experimental literature—developing better ways of dealing with non-standard word order, association, punctuation, etc. would be a step forward for computational critics.

This study has by no means been exhaustive; rather, it has only scratched the surface of what algorithmic criticism can uncover when applied to large collections of literature. I acknowledge that there is yet more to do even with Faulkner, whose short stories I have not even touched. Despite my limited scope, I have hopefully been able at least to show that Faulkner's increasingly complicated style without enough corresponding substance was related to his declining long-term success, and that similarly his failure to continue to write about women in the same capacity as in his early works contributed to what we might perceive as a stagnation of his career. If only slightly, I hope that this study and future ones like it will provide strong evidence and generative material that helps Faulkner critics (and literary critics, for that matter), come one step closer to agreeing about anything.

Works Cited

Emerson, O.B. *Faulkner's Early Literary Reputation in America*. Ann Arbor, Michigan, UMI

  Research Press, 1984. Print.

Faulkner, William. *The Sound and the Fury*. New York: Vintage Books, 1984. Print.

Faulkner, William. *Absalom! Absalom!*. New York: Vintage Books, 1972. Print.

Hutchinson, D. "The Style of Faulkner's Intruder in the Dust." *Theoria: A Journal of Social and*

  Political Theory, No. 39, October 1972, pp. 33-47.

Inge, Thomas. *William Faulkner: The Contemporary Reviews*. Cambridge, Cambridge

  University Press, 1995. Print.

Kirk, Robert, and Marvin Klotz. *Faulkner's people, a complete guide and index to characters in*

  the fiction of William Faulkner. Berkeley, University of California Press, 1963. Print.

Northern Illinois University Libraries. Twenty-one 20th century dime novels. *Nickels and*

  Dimes: A Digital Archive of Dime Novels. Northern Illinois University, 1425 W. Lincoln

  Hwy., DeKalb, IL 60115. Web. 05 June 2018.

Oya, Masanori. "Syntactic Dependency Distance as Sentence Complexity Measure."

  Proceedings of the 16th Conference of Pan-Pacific Association of Applied Linguistics.

  Mejiro University, 2011.

Ramsay, Stephen. "Algorithmic Criticism." *A Companion to Digital Literary Studies*. Eds. Ray

  Siemans and Susan Schreibman, pp. 479-491. 2013.

Wallach, Rick. "Moby Bear: Thematic and Structural Concordances Between William

  Faulkner's "The Bear" and Herman Melville's Moby Dick." *The Southern Literary*

  Journal, Vol. 30, No. 1, Fall 1997, pp. 34-54.

<u>Appendix</u>

All of the code, data, and corpora for this project are publicly available on the GitHub

page for the project, which can be found at the following address:

https://github.com/dmcfalls/Faulkner