Hayford’s Duplicates

*Operationalizing a Literary Theory of Herman Melville’s Moby-Dick.*

A paper by Jonathan Armoza

for a course on *Melville* by Prof. Jennifer Baker

New York University

Spring 2017

**Hayford’s Duplicates**

*Operationalizing a Literary Theory of Herman Melville’s Moby-Dick*

“Measurement anchors theories to the world they describe."

– Franco Moretti, “Literary Lab Pamphlet 6”

In 2013, Franco Moretti provided a hermeneutical treatise on how to move forward with quantification and “big data” in literary study. He entitled it, “‘Operationalizing’: or, the function of measurement in modern literary theory”.[[1]](#footnote-1) What we should be aiming for, he argues, is an “unexpected elaboration of reality” (4). It was not as if this sentiment was new for computational literary study. But with the introduction of expanded storage and heightened computational performance in twenty-first century, newer computational modeling techniques (sometimes called “machine learning”) emerged that could potentially “learn” hidden patterns from large data sets. This expectation needed to be restated as a sort of sounding of the hermeneutical depths these newer methods entailed. Even a few years later, as recently as 2016, Moretti felt the need to return to this point from a converse angle: to merely corroborate what we think we know about the texts we study but via a different means of arriving at that understanding would be an accomplishment. He bemoaned the ensuing cynicism from one of his earlier demonstrations where he showed that Hamlet was central to Shakespeare’s *Hamlet* via measures of centrality from computer science’s network theory. This was necessary, he claimed, as a means of proofing it for use over other, less structurally obvious plays:

“Corroboration, alas, is often boring to humanities scholars (and clever journalists); but it has long played a role in scientific research, and having introduced it into our field is an achievement, not a weakness of the digital humanities” (Moretti, “Literature, Measured” 5).

Humanities postulating, on the other hand, often makes broad suggestions from either *a)* the limited evidence available and/or *b)* choice selections of evidence amid a plethora of it. Despite the limits of either scenario, noteworthy insights can be derived, and this is because of the human ability/tendency to seek patterns amongst that evidence, to draw connections between events, people, and ideas when there are no immediately obvious ones. This ability is something the computational learning algorithms of today are only beginning to touch upon. Those algorithms have shown promising success in breaking through the limitations of our evidence gathering and insights. They can produce unexpected links between disparate and small evidence sets, just as they can observe patterns in amounts of evidence beyond the capacities of our senses and brains. However, even in the most unsupervised of machine learning scenarios, they are not autonomous. We set the programs in motion and, more often than not, must train them on categories of importance within our data sets. The question for humanistic study of literature has been why should we perform such mechanistic study? As Moretti’s original treatise on operationalizing asserts: we already make qualitative measurements within such study. Why not formalize our usage of it? Moretti looks to P.W. Bridgman’s *Logic of Modern Physics* for a rhetorical bridge from concepts to measurement wherein Bridgman notes that a “concept is synonymous with the corresponding set of operations” we use to derive it (5-6). Moretti dubs this the “operational approach…the process whereby concepts are transformed into a series of operations” (Moretti, “Operationalizing” 1). And how these measurements are brought about are, of course, with observations often derived via instruments.

Let’s consider the basic observation of counting for a moment to see what this might mean. The writing below will be discussing Herman Melville’s *Moby-Dick; or, The Whale*, so let’s count its components. The novelcontains three prefacing sections (“Etymology”, “Extracts (Supplied by a Sub-Sub-Librarian)”, and “Extracts”), 135 chapters, and an epilogue.[[2]](#footnote-2) Our instrument for counting these divisions of the book will be, for now, just a rudimentary Python script I wrote that looks to split words by spaces and examines words for similarity with their punctuation removed. It counts 211,352 words in Project Gutenberg’s digital edition of *Moby-Dick*. Below is a visualization of those counts across the novel’s sections.[[3]](#footnote-3)

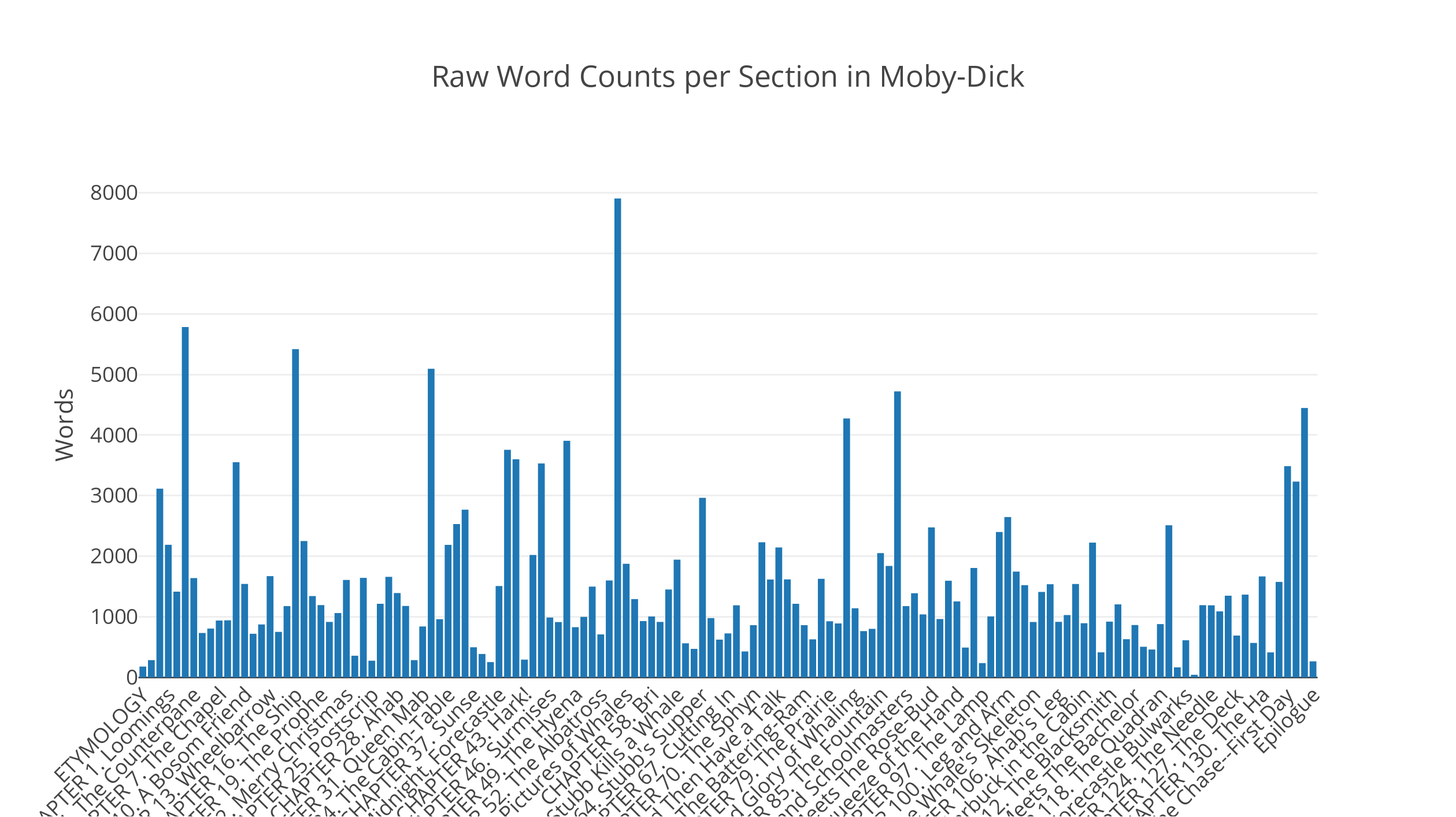


Figure 1. Counts of the words of *Moby-Dick* divided by its sections and in text-order.

From this view, it’s fairly clear there is no easily discernible pattern of section length as we move from the beginning of the book till its end. However, superlative sections are more visible. Unsurprisingly, and not much of a corroboration given the simplicity of this measurement, Chapter 54, “The Town-Ho’s Story.”, the lengthy retrospective tale told by Ishmael, is the longest of sections with 7903 words. Chapter 122, “Midnight Aloft.--Thunder and Lightning.”, the last of a small series of dialogic chapters, is the shortest at 44 words; it being merely one line spoken by harpooneer Tashtego as he lashes the Pequod’s main top sail yard in a thunder and lightning storm. Here from afar we glean Melville’s flexibility and disregard for more standardized chapter length – again, an inference we could arrive at via reading the book. This would be enough of a jumping off point to begin a closer investigation of lengthier or shorter sections and what function they serve in the novel’s composition. However, by just rearranging this data in a way more meaningful to the measurement itself, the word counts of *Moby-Dick*’s tell us something else. What if we tallied the sections into ranges of word length? Below is a visualization of such an account with ranges of 100 words.

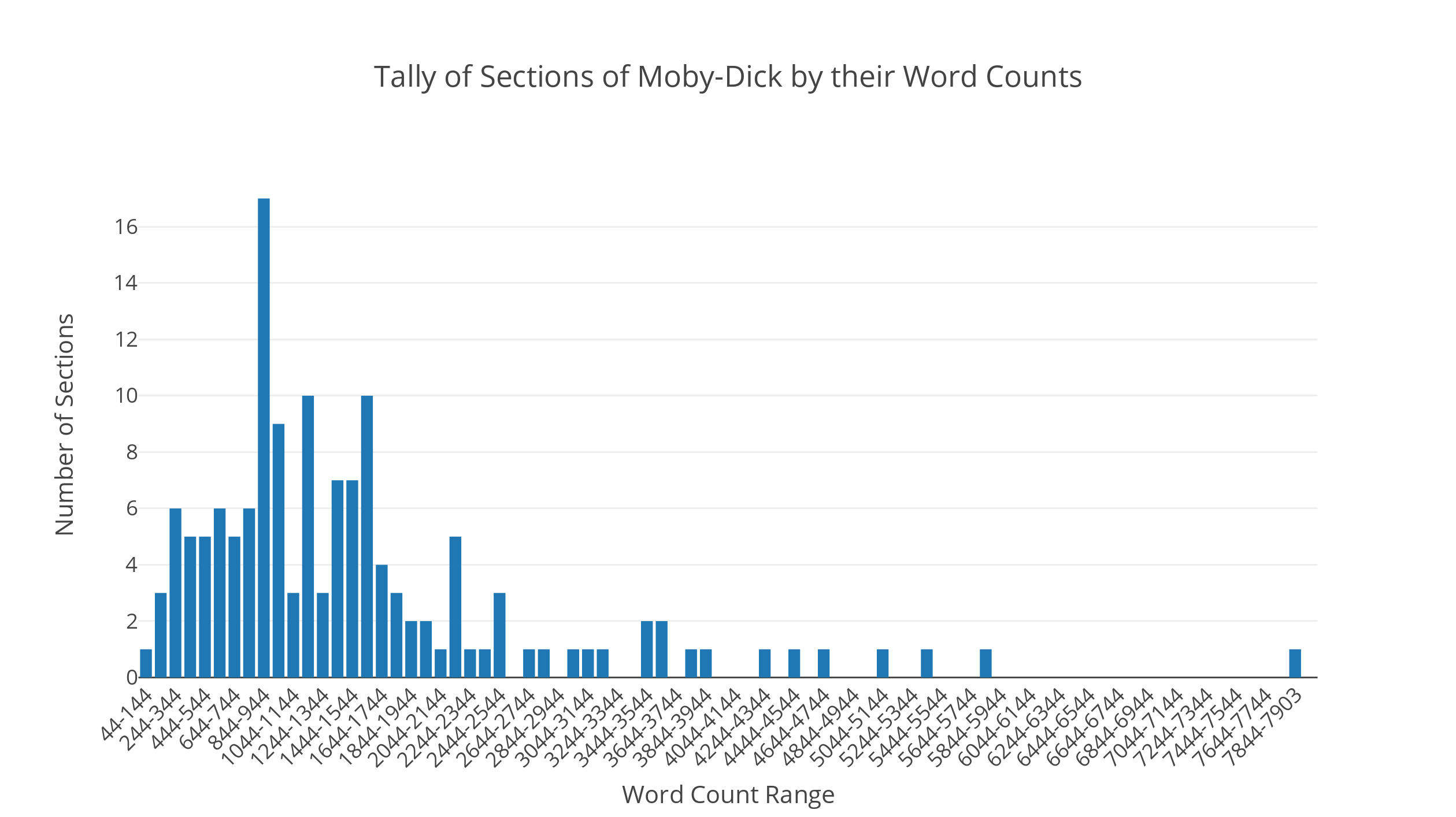
**

Figure 2. Sections of *Moby-Dick* tallied via their word counts in ranges of 100 words (Begins at 44 words since this is the smallest size section, and ends at 7903 words for the largest section.)

Here is the section length bell curve of *Moby-Dick.* Some less obvious information has now been made obvious by just taking the same measurements and reconfiguring them into a statistical fashion. “The Town-Ho’s Story.” And “Midnight Aloft.--Thunder and Lightning.” are clear outliers. While the average section length is about 1521 words, we can see that average being skewed by several lengthier sections. Instead, a more useful statistical metric would be the mode. The most commonly-sized section in *Moby-Dick* is between 844 to 944 words – there are 17 of them.[[4]](#footnote-4) What if anything makes these sections suitable for that particular common length? Do they fulfill some kind of narrative purpose? Are the boundaries – the somewhat arbitrary 100-word range – that presuppose these questions even meaningful? These and more are the kinds of questions that can be prompted via one of the most basic measurements of literature that even here forces us to “rethink the categories of literary study” (Moretti, “Operationalizing” 13). Getting just a little more complex, we can see such an analysis taking into territory quite apart from the words on the page. Let’s say we were wondering about the vocabulary of the novel. Instead of just counting words, I will extend that simple Python script to only count the unique words in each section.

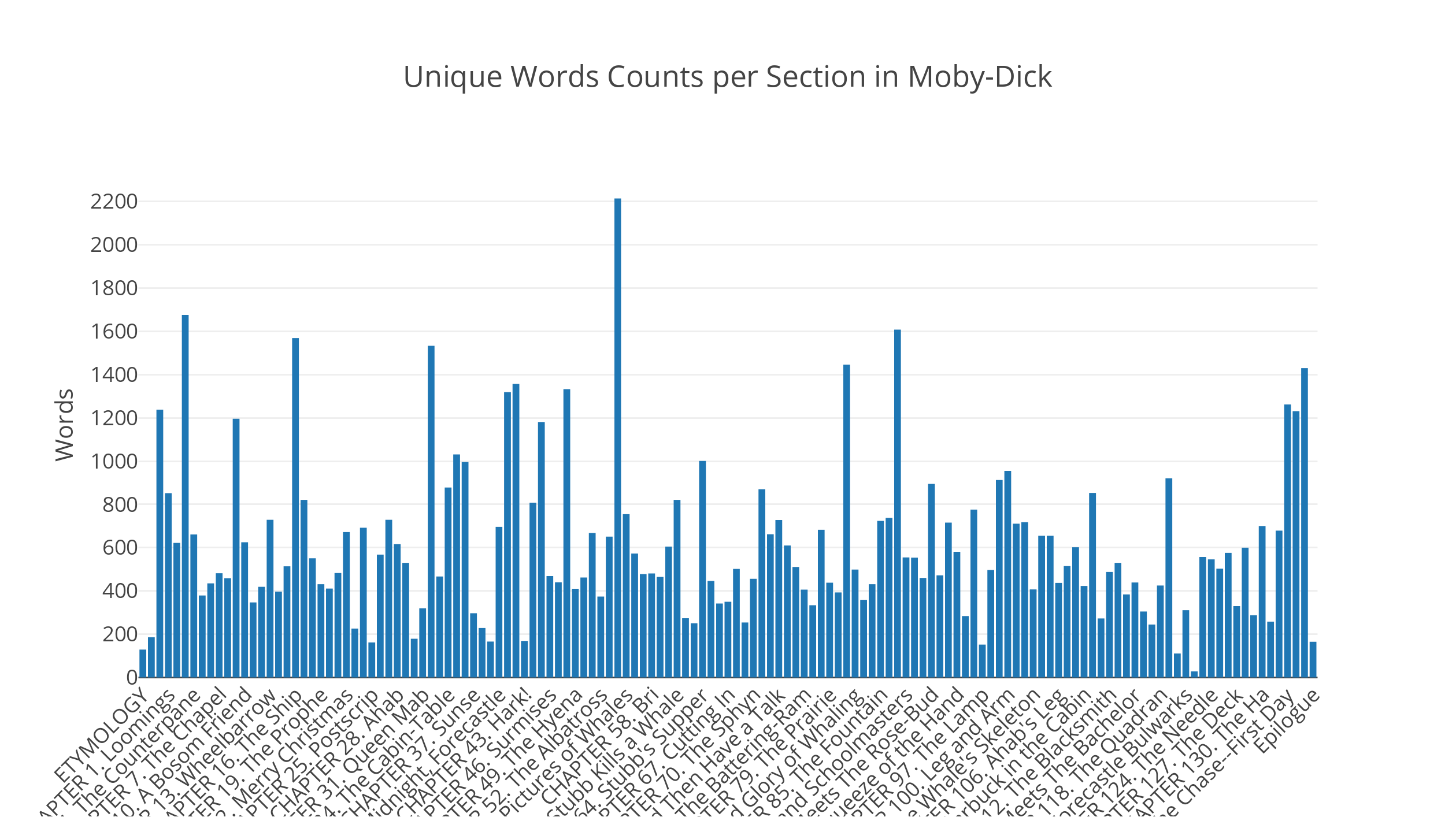


Figure 3. Counts of unique words of each section of *Moby-Dick* divided in text-order.

This looks similar to figure 1 with a few differences. In fact, at first glance it seems nearly identical, but there are notable differences in the proportions. Compared to the 212,000+ words of the novel, there are actually only about 19,800 unique words employed in it. Though the similar proportionality to the raw word count tells us there would seem to be a correlation between it and the richness of the vocabulary of each section, again a simple reconfiguring of this measurement reveals something entirely new to challenge that apprehension. Above in Figure 3 are the unique words of each section of the novel listed chronologically. What if we were to take the same chronology and account for Melville’s introduction of new words to the novel in an additive sense?

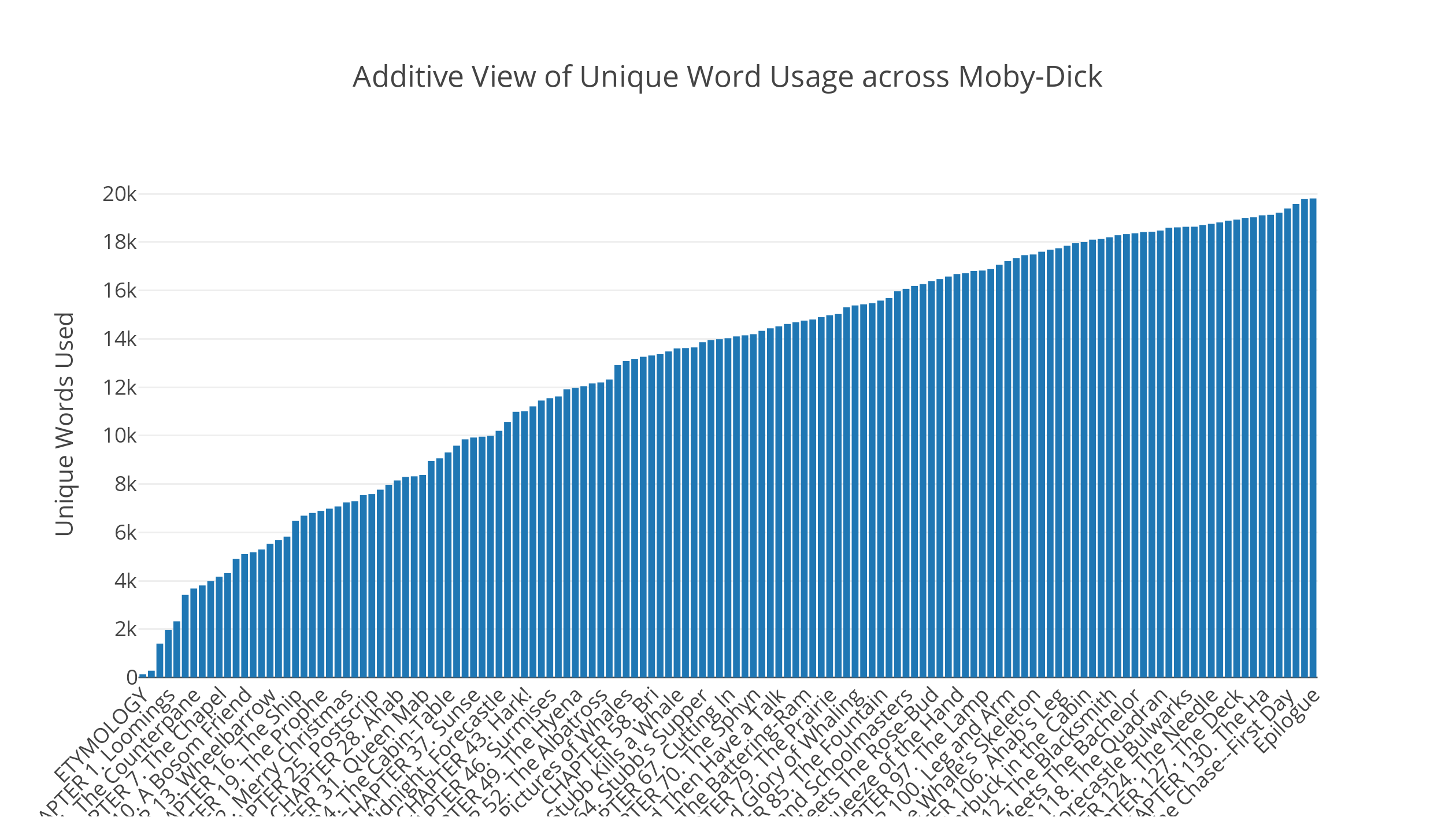


Figure 4. A view of unique words being introduced as the vocabulary of *Moby-Dick* grows over its sections, moving from unique word 1 till its 19,800th unique word.

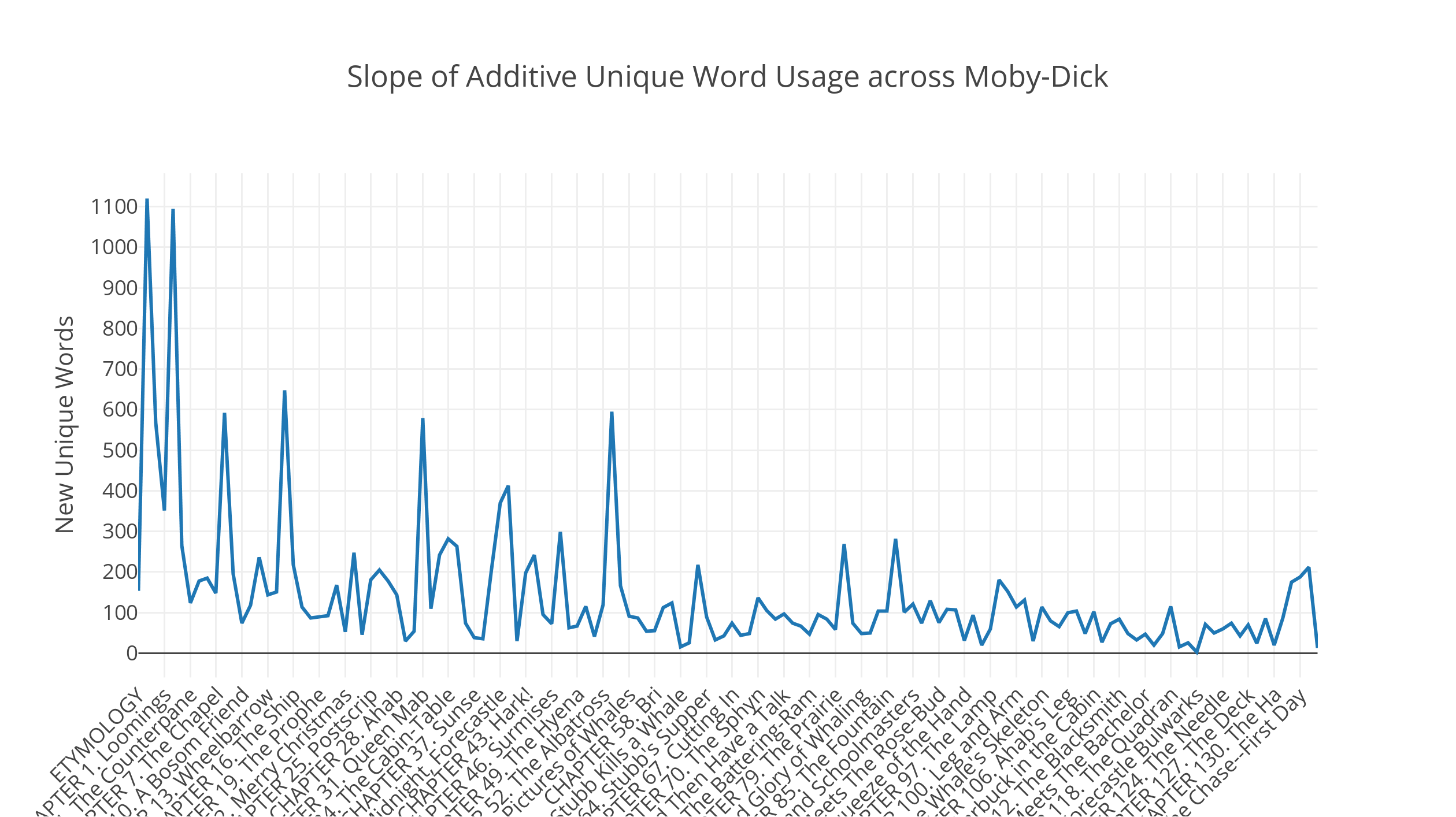


Figure 5. The slope of word introduction from Figure 4. A shift upward represents more words being introduced than the previous chapter did, and a shift downward, the opposite.

Here we are forced to confront change over a time in another sense: the rate at which new words are used in *Moby-Dick*. By doing so, we are asked to consider something else about that vocabulary. Does an increase or decrease of vocabulary introduction across sections tell us something about the function or quality of those sections? In the second figure, we this slope of change a bit more directly. As expected, many words are introduced in the beginning and this tapers as the novel progresses, but there are also significant peaks and troughs in the development of the novel’s vocabulary in its later sections that warrant investigation. These findings also move us beyond the page. The measurement being made is the counting of the novel’s components: its linguistic parts and how they are stratified across its authorial divisions. But might this rate of vocabulary introduction also act as Melvillian signature? One that changes over the course of his life? On the basis of this measurement – we can check!

But I must apologize for a bit of subterfuge here. In order to discuss the idea of quantitative measurement for literary studies I purposely left out a more complex concept or research question that would prompt such measurement. In my examples above, the concept broken out into operations is, “What is the length of this book? Chapter? Etc.” Moretti states outright that, “Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts” (“Operationalizing” 1). The idea here is that such operations put into heuristic form provide a literary theory with both “points of contact” (4) and “precision” (5) to that which the theory attempts to conceptualize. This is for the sake of added detail, verification, falsification and/or – if we are lucky – that “unexpected elaboration of reality” (4). I would not want to leave the impression that measurement is necessarily a stable quantity though. It might be more stable than say, a Marxist spatial theory of the lawyer’s office in Melville’s “Bartleby”, but measurement depends on the instrument used and the context in which it is being made. This instability is formalized by ideas like “degrees of confidence” or “probability”, and even the simplest of measures have room for flux. For instance, an account by “Voyant Tools” – a popular web application for text mining by Stéfan Sinclair – of the same digital edition of *Moby-Dick* indicates the novel contains 212,998 words with 17,341 of them being unique.[[5]](#footnote-5) That’s not an enormous difference from the accounting made by my Python script, but it does give a sense of the contingency of the measurements being made, and indeed the forms being considered. This points us to the stability of form. What is it we are counting? Why is *Moby-Dick* considered large? One reason is its high number of pages which are constructed of many sections which are constructed of even more paragraphs which are constructed of thousands of sentences which are constructed of hundreds of thousands of words. These divisions we consider and their successive amalgamation are the operations that help us to develop concepts like the physical “largeness” of *Moby-Dick*. Even if they are implicitly made, as is the case here, they denote a basic operational approach to texts that we are already making. And these are just some forms we implicitly consider. Moretti’s suggestions include morphology, genre, register, system, and style. Since they are the “repeatable element[s] of literature”, this is what we seek to count, measure, and perhaps most importantly, what we seek *to question* (Moretti, “Literature, Measured” 6).

**Melville; or, The Grumbling Carpenter**

“I don’t like this cobbling sort of business―I don’t like it at all; it’s undignified; it’s not my place…I like to take in hand none but clean, virgin, fair-and-square mathematical *jobs*, something that regularly begins at the beginning, and is at the middle when midway, and comes to an end at the conclusion; not a cobbler’s job, that’s at an end in the middle, and at the beginning at the end.”

– The Pequod’s carpenter, Chapter 126 “The Life Buoy” in *Moby-Dick; or, The Whale*

In a lecture-turned-essay entitled “Unnecessary Duplicates”, long-time Herman Melville editor and critic, Harrison Hayford, takes this grumbling artisan above as Melvillian representative aboard the Pequod. Hayford writes on the many strange instances of duplicate (and sometimes vestigial) settings, events, and characters in *Moby-Dick* – this, one of the “greatest” of American novels.[[6]](#footnote-6) Though these duplications are sometimes lost to even the experienced *Moby-Dick* reader amid the enormity and reputation of the novel, what is at stake for Hayford is an explanation as to the how these duplicates serve as some of *Moby-Dick*’smost glaring imperfections andwonders. The seeming “unnecessary” nature of many of these duplicates prompted Hayford, after years of reading, to offer his own set of compositional hypotheses as to how they might have come about. Central to Hayford’s idea of the novel’s composition are *three* primary stages of a draft. And as we have no direct manuscript from which to verify claims about Melville’s drafts, Hayford relies on incidental evidence and inference as well as the occasional comment Melville provided in personal letters discussing the novel’s progress. Computational study of writing style has been employed in authorship attribution studies, and often proceeds with evidence gathered at similar oblique angles of observation. This subject of cobbling versus carpentry is an interesting one to experiment with in that light. If computational analysis of writing can show a degree of confidence in authorship based on unique or even untoward instances of language use, could it be used similarly to detect differences in writing drafts or writing for different effect/purpose? Below I will attempt one exploration of that general question using Hayford’s ideas on the drafting of *Moby-Dick*. First, I offer his hypotheses as chronologically as I can, as Hayford’s explanation of these potential drafts and his evidence in the essay is (perhaps beneficially) circuitous. For the sake of space, I will cover some of the broader strokes of these hypotheses and fill in their details where appropriate.

Stage 1 presents a more plain narrative in two sections, one in which Ishmael potentially: (1) sets out from New York to one whaling port, New Bedford, and then (2) leaves out to sea on a whaling ship. Absent from this narrative are three central characters Queequeg, Ahab, and Bulkington, and in their place are Peleg[[7]](#footnote-7) and Bildad, with Peleg as a newer captain of the ship (not necessarily yet dubbed the “Pequod”) and Bildad as the owner/former captain/pilot. Characters like Queequeg, Tashtego, and other “savages” may have been aboard, but they were likely not central characters nor harpooneers. We will return to why this possibility is likely in a moment, but what is important to note about this first stage is that Hayford suggests that much of the sea narrative (Chapters 22-Epilogue) was written before some of the more celebrated passages we now know in the finalized shore narrative (Chapters 1-21). With so little definitive evidence as to the shape of this stage of the draft, Hayford produces an outline via a subtractive logic that cuts away from Stage 2 and 3 of the drafting process. It should also be said that Hayford suggests that there may have been other stages and other smaller substages that he does not account for in this writing. He lists many questions he feels that the plot holes and duplicates of *Moby-Dick* (which we are about to see) elicit, including some that remain beyond the ambit of his essay.

Stage 2 introduces what Hayford calls “roles” for characters in the story involving the protagonist and voyage (Hayford 52). Within the third chapter in the Spouter Inn in New Bedford, Hayford displays how obvious it is that the introduction and prompt exit of Bulkington is a graft into the narrative – a mere four paragraphs where the Grampus crew and this character is introduced, soon to be Ishmael’s “shipmate” (Melville “Chapter 3. The Spouter-Inn.”). From Bulkington’s description we get a sense of the potential roles that Melville had in mind for the character:

“His face was deeply brown and burnt, making his white teeth dazzling by the contrast; while in the deep shadows of his eyes floated some reminiscences that did not seem to give him much joy” (“Chapter 3. The Spouter-Inn.”).

Hayford proposes that Bulkington was to take on two different roles in the novel, one of a seasoned whaleman that is teacher and “comrade” to Ishmael, and also that of a “truth-seeker” harboring the deep resentment against nature (i.e. the whale) (Hayford 47).[[8]](#footnote-8) However, as we know, Bulkington is not to take on either of these roles in the more final draft(s) of *Moby-Dick.* The now famously vestigial Bulkington is given a short passage of literary grandiosity – the entirety of chapter 23, “The Lee Shore” – and then summarily dismissed from the rest of the novel without explanation. Hayford ventures that it may have been that Melville could not bear to remove this good bit of writing, going back to the aforementioned in chapter 3 and lateradding to Ishmael’s claim of Bulkington being his future shipmate, the parenthetical “though but a sleeping-partner one, so far as this narrative is concerned” (Melville “Chapter 3. The Spouter-Inn.”).

Stage 3 splits the roles of Bulkington between Queequeg, a new or possibly re-used member of the whaling ship’s crew as seasoned mentor and comrade to Ishmael, and Ahab, a brand new captain that takes on this role of troubled “truth-seeker.” Peleg and Bildad both become vestigial duplicate owners/mates/pilots and Hayford makes a persuasive case given their actions and dialogue at the signing of Ishmael on the Pequod, the piloting of the ship out of the harbor, and their calling of the men aft – an act duplicated chapters later in chapter 36, “The Quarter-Deck” by Ahab, who like many other characters who seem to have been subsequently written in by Melville (e.g. Fedallah and his men, Queequeg) has been “hiding out” since their Bulkington-like grafts in earlier chapters (Hayford 51). The naming of the ship as “The Pequod”, the centralizing of Queequeg, and the humanization and promotion of non-white characters like Tashtego and Daggoo to the role of harpooneer are all deemed as additions of this draft. And while the chapters featuring Queequeg and his befriending of Ishmael in the shore narrative are rich, Queequeg’s comradeship *and* centrality seems to almost entirely drop away once the ship is underway in Chapter 22. Why is this? Hayford ventures that this is because the rough sea narrative of Stage 1 was completed before Stage 3, at which point Melville went back and inserted Queequeg’s fleshed out character and, subsequently, a spate of seemingly unnecessary duplicates. There are two ports, two inns, two landlords, two sleeps (on the planed bench at the Spouter Inn and then in the bed with Queequeg), two to three captains (Peleg, Bildad, and Ahab), and two signing scenes (one for Ishmael while Queequeg is “hiding out” on his 24-hour “Ramadan”, and one for Queequeg). Hayford has more, but these will be enough to mention here.

Painstakingly examining the prose and dialogue surrounding Queequeg and Ishmael in later chapters, Hayford also convincingly shows that there is no longer any comradery between the two – either in the mat-making or coffin scenes, for instance. In the former, Queequeg and Ishmael don’t even regard each other, and in the latter Queequeg does not call upon Ishmael for help in his deathly state to retrieve Yojo from his bag but “one” random sailor (Melville “Chapter 110. Queequeg in His Coffin.”). In the famed monkey-rope scene where Ishmael and Queequeg are tied together, there is similar lack of comradeship – Ishmael merely refers to Queequeg as a “savage” (Hayford 57). He notes that a few prefixed additions of words in these scenes like “my” or “poor” before “Queequeg” is all the evidence that brings us a sense that Ishmael regards Queequeg as anything other than an unusual crewman, similar to language used to describe Tashtego (57). And the virtual absence of this comradery also suggests the now-invisible nature of the Stage 2 drafting, where it is likely that Bulkington’s dialogue as harpooneer and “truth-seeker” was reassigned to Ahab. The assignment of the title of harpooneer to Queequeg, Tashtego, and Daggoo also becomes a possibly haphazard late-stage addition. Hayford points out that they are mentioned as resting in the forecastle with the other men, away from their specified quarters in Chapter 33 “The Specksynder” – where the harpooneers of the Pequod are said to “take their meals in the captain's cabin, and sleep in a place indirectly communicating with it” (Melville). None the less, the humorous, touching, and lengthening prose of Queequeg’s addition to the shore narrative versus his smaller role in the larger sea narrative warrant some attention. Is there a way we could test Hayford’s hypothesis, some way to measure the difference between Queequeg as comrade versus Queequeg as harpooneer?

**Probable Cobbling**

“Seat thyself sultanically among the moons of Saturn, and take high

abstracted man alone; and he seems a wonder, a grandeur, and a woe. But

from the same point, take mankind in mass, and for the most part, they

seem a mob of unnecessary duplicates, both contemporary and hereditary.”

– Ishmael(?), Ch. 107, “The Carpenter” in *Moby-Dick; or, The Whale*

There is likely more than one way of testing Hayford’s hypotheses, but what follows in the subsequent pages is an experiment as to how we might do so. One of the challenges of working with textual data is the sheer scale of pattern recognition and investigation it represents – and thus entire fields of study that attempt to do so. When I refer to scale though, I’m also referring to the mathematics of measuring such information. Recall that *Moby-Dick* is over 210,000 words long with 17,000 – 20,000 “unique” words, all of which, have been arranged in repetitive, systematic fashion to tell the tale of Ishmael, Queequeg, Ahab, the white whale, and really more, if scholarly interpretation is to believed. It’s a complex book, so I would like to propose a complex set of operations to measure it. Hayford’s premises are admittedly incomplete, though they do offer rough boundaries and points of contact from which to base a quantifiable study. He also tells us what drafting scenarios are more likely than others. Here Hayford has supplied us a bevy of concepts, one of which are these roles that Melville might have had in mind for the characters of the novel. Broadly, that concept states that the novel can be divided spatially in a way that reflects the chronologies of its development. Since there is no remaining manuscript to contrast with, the (really “a”) finalized version of the novel is split between a shore narrative and a sea narrative. The versions of Queequeg that remain are split across those two parts. This division can be complexed in a number of ways given unclear narrative perspectives or sometimes it very clearly being an act of retrospective (see “Chapter 54. The Town-Ho’s Story.”), but Hayford’s premise and evidence suggest this shore vs sea dichotomy suffices to reflect the novel’s drafts.

Now that we have the concept, the question is how do we operationalize it. What do we measure and how? Earlier, there was the example of counting words, but clearly it’s the type of words and how they are employed that allow for Hayford’s reading. One possibility is to look to the categories of linguistics, parts of speech, to see what kinds of words are being employed to that effect. The categories themselves may be anachronistic to the work but they function as implicit and achronological. A fair case can be made that they are a mix of conscious and subconscious language patterning.[[9]](#footnote-9) Hayford makes this case when he asserts that the understanding of Ishmael and Queequeg’s close friendship is oversold and the trick of the later-written/edited shore chapters and the small insertions of affectionate phrases. Here, for instance, is Hayford disputing the general claim that Chapter 72 “The Monkey-Rope.” presents evidence of the “bosom friendship” between Ishmael and Queequeg:

“[Queequeg’s] special relationship to Ishmael is specified by epithets at two points: the first reference is ‘*my particular friend Queequeg*, whose duty it was, as harpooneer…’; the second is ‘*my dear comrade and twin-brother*, thought I.’…In the light of what follows, I argue that Melville later inserted ‘my particular friend’ and ‘my dear comrade,’…Nothing else in that scene of the chapter is written in a way that presumes or requires the pair to be comrades already…[I]t suggests the likelihood that it was his writing of this scene and this metaphor that opened to Melville the possibility of making the pair bosom friends in the shore sequence when he removed Bulkington from the comrade role” (Hayford 57-58, italics mine).

What’s interesting about this particular hypothetical edit is that it visits upon several central aspects of Hayford’s ideas on the drafting and seems like a point of genesis for them. The first is that Melville went back during a later stage and inserted phrases that would seem to superficially bolster the relationship between the two characters. The second is that this physical tie-turned-metaphor – the “monkey-rope” – is also a possible genesis (Hayford uses the term “likelihood”) for Melville’s own rewritten shore narrative where Queequeg becomes a richer character and takes on a larger role. You will note that this hypothesizing all turns on a set of phrases. The two I have highlighted in the above excerpt have a similar construction:

***<possessive pronoun> <adjective(s)> <noun(s) or proper noun>***

One path for measurement here could be to look for such sequences of words with those parts of speech in the sentences of *Moby-Dick*. It’s true that their sequence informs their function, but another way to think of this is merely to regard their presence. And at this low level of the text, the question following that measurement would become, “Can the presence of particular amounts of parts of speech (POS) in sentences reflect style?” The trouble with style detection that relies on just the measurement of counting is that at the analytical stage once those counts are made, the method for analysis relies on visual inspections or formulae (e.g. the Burrows’ Delta that detects style via most frequent words used) that often can miss fine- and small-scale analyses based on that larger patterning of large swaths of text. In other words, once we derived frequency tallies and the like, there is no easy way of going back to the relationships between words in the individual sentences where those words were deployed. This is why computational text analysis has turned to more complex modeling techniques that can attempt to find those hidden patterns for things like thematics. Given the possibility that, as Hayford’s hypothesizing suggests, pattern formation reflects a multi-stage process and is therefore inconsistently distributed or unstable, the suitability for probabilistic modeling as one possible operation becomes apparent.

The first measurement we will be making is to count the POS in each sentence of the novel. I will use the recently developed and well-supported POS tagger program called “spaCy.”[[10]](#footnote-10) It should be said here that the act of part of speech tagging itself is always a probabilistic venture. Even manual tagging is prone to error or different interpretation. In the case of modern POS tagging programs like “spaCy”, a model of language use is created and “trained” via a machine learning algorithms over a set of small in-memory networks meant to mimic biological nervous systems. These are referred to as neural networks. The more data that flows into the network, the more possibility it has to learn how to correctly assign POS to words. “spaCy” is trained on a corpus of over 1 gigabyte of digital English language texts. I want to take care to describe here the instrument with which we’ll be relying to take measurements and then offer analyses on the language of *Moby-Dick*. It is not a small point to be overlooked. For each measurement there is a probability of error and one of success. The creators of “spaCy” claim it renders 93% tagging accuracy (Honnibal “Citation Information #272”). Once the tagging is done for each sentence, we will rely on a set of counts of the number of instances of POS types in that sentence. The produced measure is then an ordered series of numbers.

In linear algebra, a sequence of related numbers (often tied to a coordinate system) is referred to as a vector. So for each sentence, we have a POS vector and thus a model or representation of our text data. The next operation for our inquiry will then be to use a means of establishing relationships between those vectors. The suggestion here is not so bold in a literary context. It is also inherently decompositional. Sentences are components of systems that are paragraphs that are chapters that are novels. What we would like to understand is if sentences and their POS serve particular roles in those systems. One recently popular set of methods for identifying clusters of vector data is known as collaborative filtering.[[11]](#footnote-11) These methods are used in several common contexts such as recommendation systems. Netflix, for instance, sponsored a collaborative filtering contest in 2006 that challenged researchers to make improvements above their algorithms’ accuracy for inferring user ratings of movies. The general premise is that between its users and its movies there are attributes of either that draw one to the other and elicit a response. What are the basic questions that they would be interested in answering via model of their data? (1) What about a user makes them watch a movie and/or give it a particular rating? (2) What about a movie makes users watch and/or give those ratings? It is proposed that the answers to those questions may be characterized or approximated by recovering those latent factors. Note that the factors are not the answers, but they are suggestive of them. This is the power and weakness of presenting a system in this way. The model encodes our assumptions about a problem set and proceeds to approximate answers to questions we might have, given what we can observe. If done correctly, once the result is determined, an assessment is made over that result and the model is revised according to external information (other data or our intuition).

One of the successful methods of collaborative filtering is – in the very same sense – decompositional: matrix factorization. Here it will be enough to lightly describe the mathematics involved in this process. Each POS vector (recall, a sequence of related numbers) represents a linear equation with a set of known coefficients (the POS category counts). A vector with three coefficients could be written as such: . Given a set of vectors we stack them up into a table. In linear algebra this stack of vectors is referred to as a matrix. In formal notation we erase the variable names for clarity. Let’s say we had three POS vectors, representing three texts. A matrix would look like a row by column table of those coefficients. In the Netflix examples these counts are replaced by ratings, each vector representing a user. And there may be large holes (zeroes) in this matrix due to missing values. One task could be to infer what those missing ratings might be based on the ratings of other users who have watched similar movies to a user of interest. Another task could be simply to group users for some recommendation purpose. The latter is what matrix factorization is more typically used for today.[[12]](#footnote-12) In either case, we would like to know how those latent factors contributed to the values of this matrix and to do so we factorize it, attributing some portion of each of its values to latent user factors and the remaining portion to latent movie factors. This produces two matrices, one for users and the other for movies, the coefficients of which will inform the proportional contributions to our known data: the stack of user rating vectors.[[13]](#footnote-13) The idea is that these two matrices can be multiplied together to produce our observed data – thus they are factors. The values of these matrices are determined over a number of iterations of the factorization method until an approximated error value is deemed stable enough (i.e. factorization results consistently produce an error value of *x*%). For the purposes of this writing, we will look to a Bayesian probabilistic approach for this method called Probabilistic Matrix Factorization (PMF). Matrix factorization, in general, solves the problem of high dimensional data. One implicit but important property of this decomposition is that in order for matrices to be multiplied they must share a dimension. If one has a height of 3 users, the other has to have a width of 3 ratings, and so forth. This has the effect of allowing us to take a matrix of 50 million users and 50,000 movies and greatly reduce one dimension in each of the resultant factor matrices. The suggestion of this lower dimension value (it could be 1,000 or 2 or whatever we feel proper) is given to the method and this is the height of one matrix it approximates and the width of the other matrix it approximates. The lower the suggested shared dimension, the more detail is lost. The hope is that some details are not helpful or even distracting to the pattern detection problem we are interested in solving.

The Netflix example helps our understanding of this method, but users and items are only one metaphor for this kind of model. Another intuitive metaphor more appropriate to our POS-tagged data set is one proposed by biologists: patients and genes. In this modeling metaphor, we try to identify latent factors that are “hidden patients” and “hidden genes.”[[14]](#footnote-14) The task is then to identify clusters of patients with similar genetic traits, or the inverse, genetic traits with similar patients. Overlaying this metaphor to the text to POS problem can thus produce “hidden sentences” or “hidden POS.” In this case, we look to identify those hidden sentences around which we can group actual novel sentences as being somehow linked by their similar POS usage.[[15]](#footnote-15) We run the factorization algorithm several times to produce a consensus of the coefficients of those matrices, determine a hierarchy of those consensus results and then flatten the hierarchy to produce a cluster assignment for each sentence. These results tell us much more than a mere cluster assignment ID though. In fact, they present an association of each sentence with one of those hidden sentences. Each hidden sentence is proportionally representative of the features of each of the sentences assigned to it. In this case the features are POS counts. The hidden sentences are also thus far more representative of the collection of objects as a whole – a model of a model. We will refer to these representations as PMF produced POS profiles, or PMF-POS profiles for short. PMF-POS profile vectors (or “hidden sentences”) are averages of the POS vectors of those sentences grouped together in the PMF model. As we will see, profile vectors enable understandings of a collection of objects that are simultaneously more comprehensive and individually representative.

To turn back to the question of whether there is a way of stylistically determining the difference between Queequeg as comrade and Queequeg as harpooneer, this different understanding is exactly the hope. As we have seen, Hayford’s understanding of the draft stages of *Moby-Dick* are both locally and globally informed. His years of editing, writing, and teaching about Melville and *Moby-Dick* allowed him this more comprehensive perspective. But not everyone is Harrison Hayford. Here instead we allow the model to tune itself through suggested parameters (shared matrix dimension, statistical prior distribution parameters) and through adjustment via error correction across multiple runs. For this experiment, the PMF model and such accompanying suggestive and corrective modeling functions are supplied by a matrix factorization tool designed by computational biologists called “Nimfa” (an acronym for the most basic descriptor of this matrix factorization modeling technique: “nonnegative” matrix factorization). “Nimfa” identifies 393 semi-distinct POS sentence patterns in this model of the sections of *Moby-Dick*. If we just look at explicit mentions of Queequeg in these sentences,[[16]](#footnote-16) framing them under Hayford’s major divisions of drafts between shore and sea narrative, and plot the instances of sentences identified by the POS sentence groupings, we get a graph that looks like the one below.

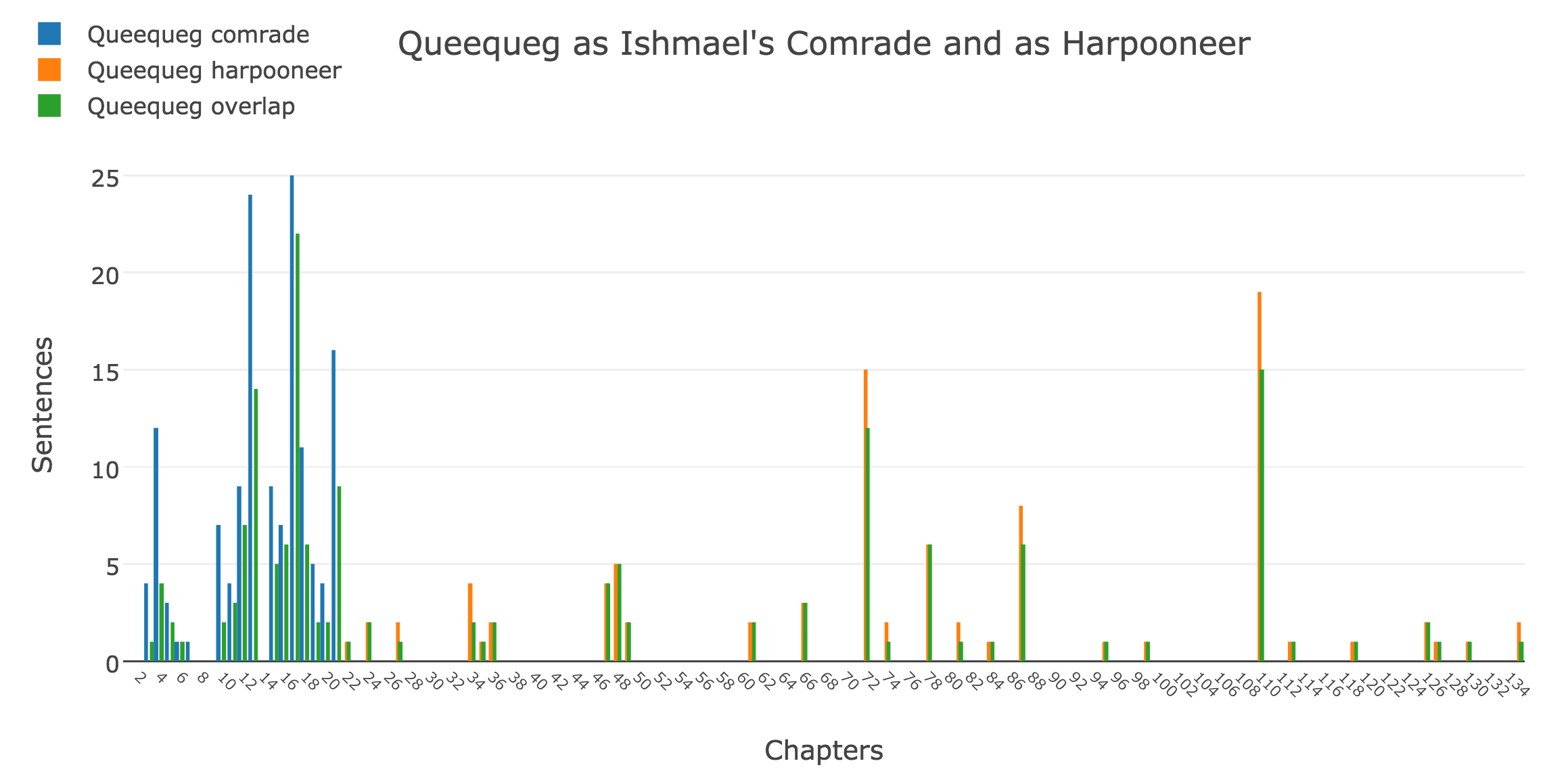


Figure 6. Sentences mentioning Queequeg split across Hayford’s division of *Moby-Dick*’s shore and sea narratives. Here blue represents Queequeg on shore as Ishmael’s “comrade”, and orange represents Queequeg at sea as the Pequod’s “harpooneer.” The green shows overlaps in the POS pattern groups as identified by a probabilistic matrix factorization (PMF) model of the novel.

From the outset, this distant view makes clear the sheer sparsity and comparative dearth of Queequeg’s presence in the novel after the Pequod leaves Nantucket. All sentences on the shore are marked in blue and the on the sea in orange. The green indicates when those sentences share a sentence group ID in the PMF model. At times that overlap is almost complete, but in other notable chapters there are significant, uniquely grouped sentences. Those examples include both much critiqued and some lesser critiqued Queequeg chapters including “Chapter 3. The Spouter-Inn.”, “Chapter 12. Biographical.”, “Chapter 16. The Ship.”, “Chapter 20. All Astir.”, “Chapter 72. The Monkey-Rope”, and “Chapter 110. Queequeg in His Coffin.” The question before us is how to meaningfully access the model’s underlying data and when to take the step or operation that will be a comparison with the inductive proposals of Hayford’s hypotheses. At first glance, taking a look at the unique sentences across that shore-sea division seem an obvious step. If Queequeg is indeed written differently across that division, albeit with emendations once the bosom friendship of the shore narrative had been conceived in Stage 3, then there should be some sort of stylistic or linguistic difference that conveys a richer characterization. The statistical leap here though is to question why we should move away from considering the mode, or most common sentence types. Just like with the count data it would seem that consistent styling would be more worthy of consideration than inconsistent styling. Humanities-style postulating also frequently likes to consider the functional/aesthetic value of outliers – something that contradicts the statistical notion employed here: central tendency. Disregarding sentences with shared PMF groups however is precisely the outcome of the measurement we would like here. Or, what (possibly) makes these writings different? This is akin to throwing out that central bell of Figure 2. As we’ll see in a moment, looking at sentences in the most highly used POS patterns (most highly identified PMF groups) with a superlative lens is not very helpful.

However, we don’t have to entirely disregard the most common POS patterns. We can look at them in a sort of statistical relief – what their profundity in the whole of the identified sentences denotes. It turns out that the sentences of Queequeg on shore versus Queequeg on the sea tell us that Melville was also writing in a more diverse manner in the former. On shore, writing mentioning Queequeg consists 142 sentences and are assigned to 73 PMF groups while on sea writing mentioning Queequeg consists of just 91 sentences across 47 PMF groups. But shore Queequeg has 39 unique groups across 56 sentences and sea Queequeg has only 13 unique groups across 15 sentences. Recall that the shore narrative includes 21 chapters and the entire novel has another 114 plus an Epilogue. Where the plain sentence to chapter ratio denotes a more obvious Queequeg-density issue, the PMF groupings tell us of another. Combining those two measures of density – length and POS-usage – we see that not only is Queequeg comparatively absent in mention, the way he is being written about once on the sea is about 17% less diverse. So how can we talk about that diversity in a more meaningful?

When we look at the uniquely grouped sentences of both sections, we also find that they are uniformly distributed. Each sentence was different enough in its POS usage to be assigned to a different group in the PMF model. While we could consider each of those sentences’ POS counts, remember this fine line of operationalization and compositional-theoretical conjecture we are attempting to traverse. To stick with the PMF model – the reasoning behind why sentences were identified as unique – we can instead consider the PMF-POS profile surrogates and attempt to determine the combined POS tendency of those “hidden” sentences. Let’s look at the average of the POS profiles of those sentences identified as having POS patterns unique to the shore (Queequeg as “comrade”) and sea (Queequeg [mostly] as “harpooneer”) narratives.

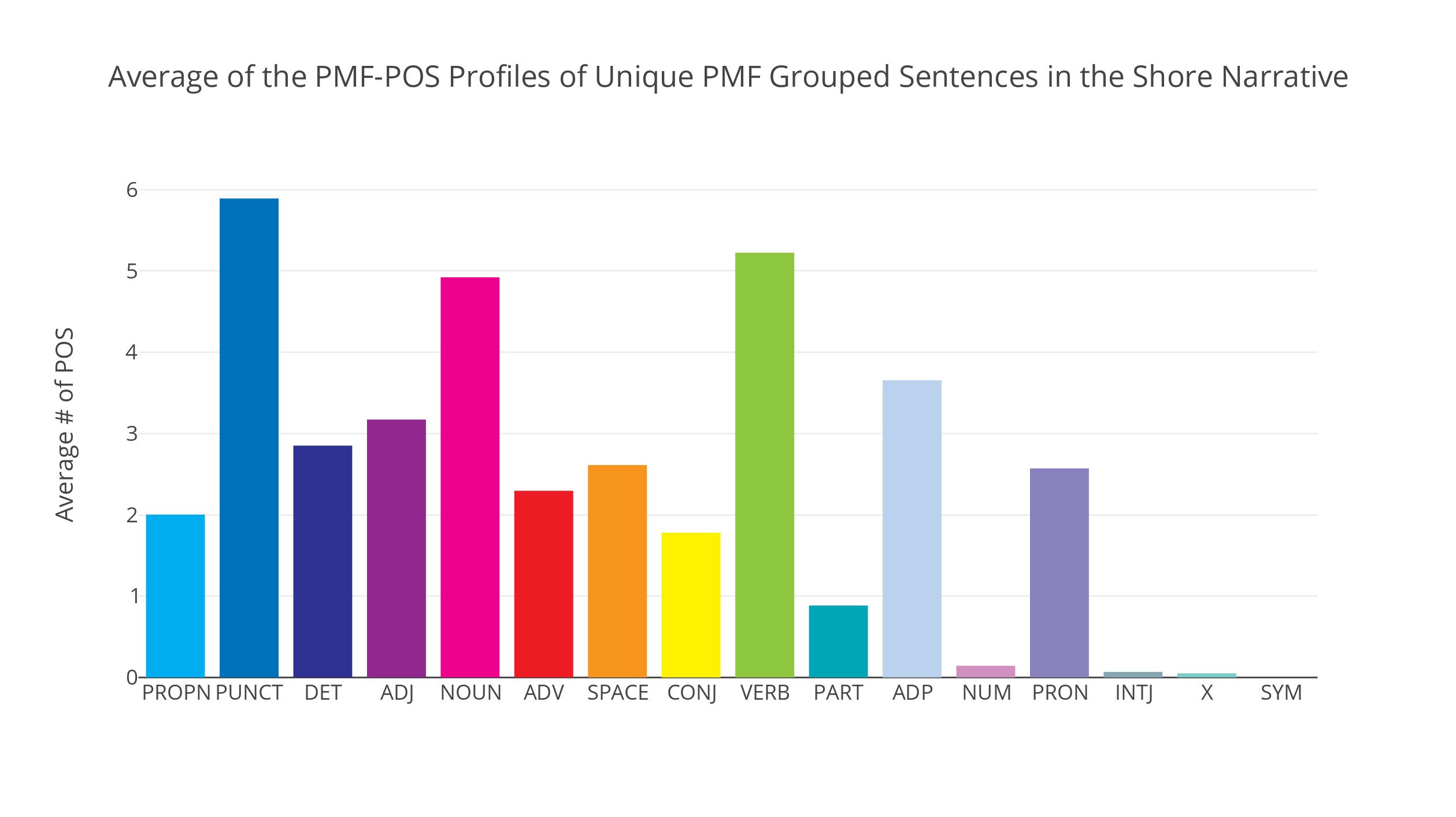


Figure 7. PMF-POS profiles of sentences with POS patterns unique to the shore narrative are averaged together to produce this POS sentence count dynamic.

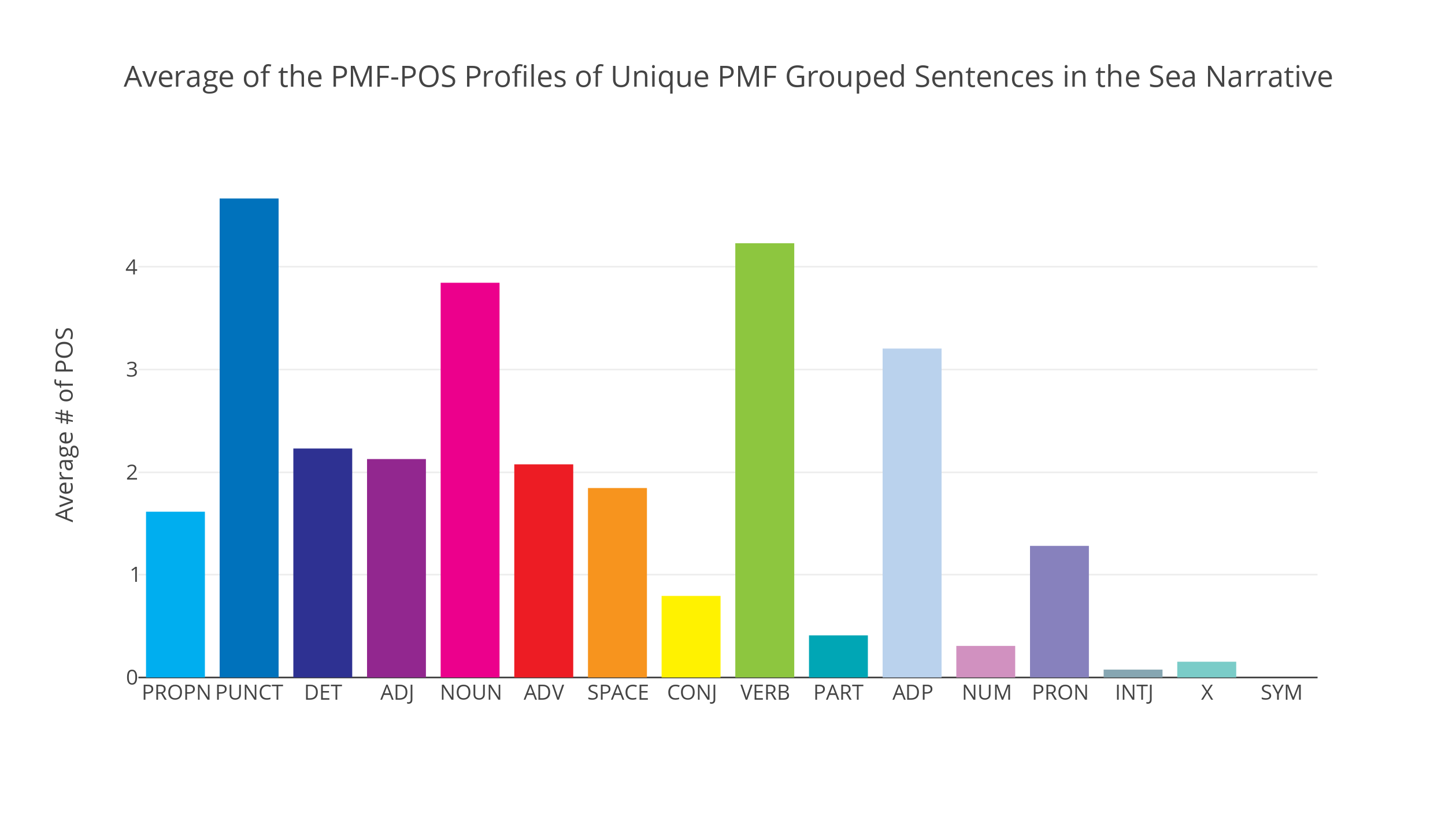


Figure 8. PMF-POS profiles of sentences with POS patterns unique to the sea narrative are averaged together to produce this POS sentence count dynamic.

Each graph above gives a sense of the dynamics of what, on average, distinguish these sentences that were identified as unique from the other shared POS-patterns of the Queequeg sentences. Immediately we see a disparity in length. In the shore narrative these Queequeg sentences have a full 9-10 more words than their sea narrative counterparts (38.136 average words to 28.872 average words). If we account for that difference by normalizing the counts (assessing their proportionality by summing the POS category counts of each profile separately and then dividing each separate count by that sum), we can see the percent differences between the PMF-POS profiles of these Queequeg sentences. The hope is that the operation of taking their average does not obliterate too much of the outlying POS features that make these sentences unique. Instead of concentrating on the exact values, the idea is to use the changes in POS between to guide an investigation of them.

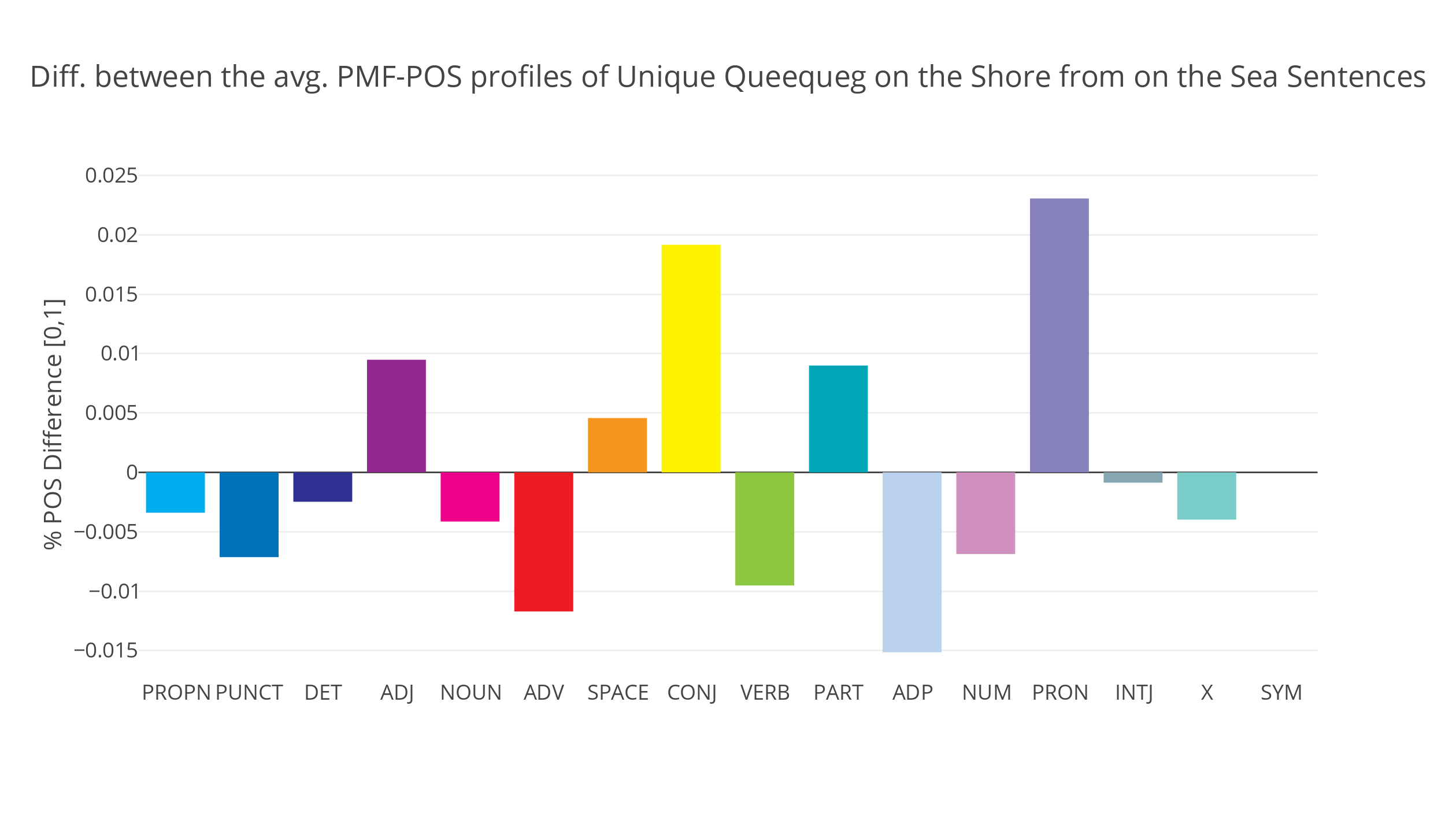


Figure 9. The percent difference in POS counts between Figures 7 and 8. Here we see difference from Stage 1 to Stage 3. When we flip the order, the percent dynamics are reversed.

Above, we can see that biggest swings (again, albeit seeming small shifts that have been attenuated by the process of averaging) are an increase of conjunctions, adpositions, and pronouns moving from the proposed earlier Stage 1 sea narrative and the later Stage 3 shore narrative. The larger sentence size of the shore sentences ostensibly explain the increased conjunctions, and maybe even the adpositions. The largest shift in POS seems to be in the increased number of pronouns. An optimistic and educated guess here is that the increased number of pronouns reference Queequeg, possible “he” or “his”, lead to a personalization rather than references to him using more objectified language as Hayford notes (e.g. “the savage”).

Which sentences from either narrative are most similar in POS proportion, not just certain POS category count differences? As if to answer for Hayford’s claims about the kinds of characterization that Melville deployed across draft stages, the model replies. The most “comrade” or shore-stage-like Queequeg is found in the fitting Chapter 12, “Biographical.”:

“In vain the captain threatened to throw him overboard; suspended a

cutlass over his naked wrists; Queequeg was the son of a King, and

Queequeg budged not” (Melville).

The story here is retrospective but it gives us a scene that depicts a Queequeg with some of the most narrative agency he has in the entire novel. Here, a young Queequeg flees his homeland for adventure, climbing aboard an American ship and defying the captain’s order to disembark. We see the increased pronoun usage as expected and the sense of agency imbued in the character. A sentence later the captain relents, observing, “*his* desperate dauntlessness, and *his* wild desire to visit Christendom.” Similarly, when looking to what sentence in the sea narrative closely fits the difference in POS proportion, it corroborates Hayford’s premise of Queequeg as secondary, patronized, as a less than equal crew member. Stubb speaks sarcastically,

“[F]irewood?--lucifer matches?--tinder?--gunpowder?--what the

devil is ginger, I say, that you offer this cup to our poor Queequeg

here” (Melville).

Here we see the limits of a POS tagger given unclear punctuation and capitalization – “spaCy” is unsure of the bounds of the sentence – but as if to also temper our skepticism on the counting of categories like POS which hide the rich qualities of the words beneath them, we see that the PMF model at least correlates with Hayford’s suspicion about the imbalance in the qualities of Melville’s prose. In the last phrase, the possessive pronoun “our”, matched with “I” and “you”, sets of an us versus Queequeg as other, and we again see that familiar sequence of *possessive pronoun*, *adjective*, and *proper noun*, “Queequeg.” It seems as if the model has connected enough dots to point us to yet another example of Stage 3 meddling. Is this another insertion, a case of Melville’s “cobbling” as Hayford proposed? Or is this mere coincidence? The probabilistic weighting of the iterative pattern-matching operations involved in the modeling give us more *confidence* that it is not. And that degree of belief is itself the threshold between Hayford’s concept and our complex measurement.

We should take caution here, lest these results seem too conclusive. The question I have asked the model by parsing through and investigating the dynamics of its underlying data (the POS counts) harbors the initial presumptions of Hayford’s hypotheses on the drafting of *Moby-Dick*. These – perhaps unnecessary – duplicates that the PMF model has found can only be teased apart by an initial concept and set of operations meant to put that concept into practice. In order to properly test Hayford’s premises, a number of suggestions counter to those divisions would be made and then tested as to their comparative likelihood. Never the less, it would seem that – for now – the PMF model of POS has uncovered some probable and promising evidence that Hayford is on to something with regard to the inequitably-styled characterization of Queequeg in *Moby-Dick*.

**Appendix**

**A.** *List of sections in Moby-Dick; or, The Whale with 844-944 words*

CHAPTER 7. The Chapel

CHAPTER 8. The Pulpit

CHAPTER 12. Biographical

CHAPTER 20. All Astir

CHAPTER 47. The Mat-Maker

CHAPTER 57. Of Whales in Paint; in Teeth; in Wood; in Sheet-Iron; in

CHAPTER 59. Squid

CHAPTER 70. The Sphynx

CHAPTER 76. The Battering-Ram

CHAPTER 79. The Prairie

CHAPTER 80. The Nut

CHAPTER 103. Measurement of The Whale's Skeleton

CHAPTER 106. Ahab's Leg

CHAPTER 109. Ahab and Starbuck in the Cabin

CHAPTER 112. The Blacksmith

CHAPTER 115. The Pequod Meets The Bachelor

CHAPTER 118. The Quadrant

**Bibliography**

Armoza, Jonathan. “Part-of-Speech Profiling and Stylistic Textual Assessment Using Probabilistic Matrix Factorization,” 2016.

Bridgman, P.W. “The Logic of Modern Physics,” 5–6. New York, NY: Macmillan, 1927.

Buell, Lawrence. *The Dream of the Great American Novel*.

Cambridge, MA: Harvard University Press, 2016.

Burrows, John. “Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship.” *Literary and Linguistic Computing* 17, no. 3 (2002): 267–87.

Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. “On Clustering Validation Techniques.” *Journal of Intelligent Information Systems* 17, no. 2/3

(2001): 107–45.

Hayford, Harrison. “Unnecessary Duplicates.” In *Melville’s Prisoners*, edited by Hershel Parker, 39–68. Evanston, Illinois: Northwestern University Press, 2003.

Honnibal, Matthew. “Citation Information #272,” February 22, 2016. <https://github.com/explosion/spaCy/issues/272>.

———. *spaCy*, 2016. spacy.io.

Koren, Yehuda, Robert Bell, and Chris Volinsky. “Matrix Factorization Techniques for Recommender Systems.” *IEEE Computer Society*, August 2009, 42–49.

Melville, Herman. *Moby-Dick; or The Whale*. Edited by Daniel Lazarus, Jonesey, and David Widger. The Project Gutenberg Ebook Edition., 2016. <https://www.gutenberg.org/files/2701/2701-0.txt>.

Moretti, Franco. “Literature, Measured.” *Pamphlets of the Stanford Literary Lab* Pamphlet 12 (April 2016). <https://litlab.stanford.edu/LiteraryLabPamphlet12.pdf>.

———. “‘Operationalizing’: Or, the Function of Measurement in Modern Literary Theory” Pamphlet 6 (December 2013). <https://litlab.stanford.edu/LiteraryLabPamphlet6.pdf>.

Salakhutdinov, Ruslan, and Andriy Mnih. “Probabilistic Matrix Factorization.” *University of Toronto*, 2008.

Sinclair, Stéfan. *Voyant Tools*, 2017. <http://voyant-tools.org/>.

Vala, Hardik, David Jurgens, Andrew Piper, and Derek Ruths. “Mr. Bennet, His Coachman, and the Archbishop Walk into a Bar but Only One of Them Gets Recognized: On The Difficulty of Detecting Characters in Literary Texts.” *Conference on Empirical Methods in Natural Language Processing*, 2015.

Zitnik, Marinka, and Blaz Zupan. “Nimfa: A Python Library for Nonnegative Matrix Factorization.” *Journal of Machine Learning Research* 13 (2012): 849–53.

———. *Nimfa.* 2012. <http://nimfa.biolab.si/>.

1. The title format will be familiar to readers of Melville, albeit with the substitution of a colon in place of the semicolon. [↑](#footnote-ref-1)
2. “Extracts” is left out of all of the following analysis, as it is explicitly quotation from external sources and not Melville’s styled prose. [↑](#footnote-ref-2)
3. Since space and interactivity are often limiting factors in reading graphs such as these, more interactive web-browser accessible versions of each figure are included with this paper on its GitHub project. Direct links to each figure are offered in the Appendix. [↑](#footnote-ref-3)
4. To satisfy any curiosity, a list of these sections can be found in the Appendix. [↑](#footnote-ref-4)
5. My counts were 211,352 total words and 19,800 unique ones. The discrepancies are likely accounted for in different ways of normalizing and counting variant spellings, contractions, hyphenated words, etc. [↑](#footnote-ref-5)
6. *Moby-Dick* is seen as partially responsible for the idea of the “Great American Novel.*”* See Buell, *The Dream of the Great American Novel* (2016). [↑](#footnote-ref-6)
7. Hayford guesses that Melville originally dubbed him the clichéd ‘Pegleg’ in an early, rougher draft (Hayford 54). [↑](#footnote-ref-7)
8. Here is where Hayford admits his own premise conflates the possible roles of Peleg and Bulkington, and thus also the possibility of other unmentioned drafting stages. [↑](#footnote-ref-8)
9. This is the case for style-detection via function-word usage in authorship attribution studies. [↑](#footnote-ref-9)
10. This use of “spaCy” accounts for proper nouns, punctuation, determinants, adjectives, nouns, adverbs, spaces, conjugations, verbs, participles, adpositions, numbers, pronouns, interjections, and symbolic characters. All remaining untaggable words are accounted for in a category, X. [↑](#footnote-ref-10)
11. The following prose that explains the probabilistic matrix factorization modeling method has been taken and adapted from my previous exploration of the method used over Emily Dickinson’s fascicle manuscript books, “Model as Archive: Computational Perspectives on Emily Dickinson”. A copy of this writing can be found alongside the code and writings for this project on GitHub. [↑](#footnote-ref-11)
12. Inferring missing values has been found to produce the unfortunate result of making computation of these problems increasingly difficult. [↑](#footnote-ref-12)
13. Acceptable coefficients of these matrices are determined after a factorization has been calculated by the minimization of a related, error-based objective function. In the case of the method used in this writing, “Probabilistic Matrix Factorization”, it is the minimization of an error-based objective function that is informed by prior statistical distributions representing the data and the two matrices resulting from factorization. The mathematics behind these methods is described in a separate paper, my “Part-of-Speech Profiling and Stylistic Textual Assessment Using Probabilistic Matrix Factorization” (listed in the bibliography and available in the “Hayford’s Duplicates” GitHub project) and also in Salakhutdinov and Mnih’s article, “Probabilistic Matrix Factorization.” [↑](#footnote-ref-13)
14. This is the metaphor employed by a set of researchers using the matrix factorization code library, “Nimfa,” used for the PMF results below. See bibliography and <http://nimfa.biolab.si/>. [↑](#footnote-ref-14)
15. The prior statistical distributions (mentioned in footnote 11) given to the method as assumptions for this modeling demonstration are nothing but standard bell-curves (a.k.a. normal or Gaussian distributions). Though such prior given statistical information given to a modeling method tells it how it should expect incoming data to roughly be distributed, this expectation is also quickly overtaken as more data is added to the model. [↑](#footnote-ref-15)
16. Programmatic identification of characters is a notoriously challenging problem, and one with which even human readers can, at times, produce inconsistent results. See, for instance, Piper et al.’s “On the Difficult of Detecting Characters in Literary Texts.” The best way of doing so for a more rigorous study of the novel would be to read around each explicit mention, or other surrogate terms (e.g. “savage”) to determine if a sentence is referring to the character in question. [↑](#footnote-ref-16)