The migration of "had"

Structural changes in the novel from 1840 to 2000

David McClure

December 6, 2016

Abstract

In recent years there has been a surge of interest in what might be thought of as a form of computational narratology – the study of the ways in which different literary phenomena are distributed across "narrative time" inside of texts. Thus far, though, these studies have largely treated narrative structure as an a-historical object. In this essay I propose a method for systematically observing the ways in which narrative structure changes over historical time. I find that one of the most significant changes in a corpus of 30,000 novels between 1850 and 2000 is a shift in what might be thought of as the temporal "boundedness" of the text. Over the course of the last 150 years, the beginnings of narratives have become increasingly marked by words that gesture backwards in time into the pre-narrative space – specifically the words "had" and "been," the past perfect (progressive) tense ("had lived in," "had been working at," "had thought about"). Meanwhile, narrative endings become marked by words that point forward into a space after the narrative -– anchored by the word "will," the future tense.

1

At the beginning of "An Introduction to the Structural Analysis of Narrative,"[1] Barthes' contribution to the 1966 special issue of *Communications* that would later seem like a crossroads in the history of narrative theory, Barthes begins by warning of the difficulties of taking an "inductive" approach to studying narrative:

> Many commentators, who admit the idea of a narrative structure, are nevertheless reluctant to cut loose literary analysis from the model used in experimental sciences: they boldly insist that one must apply a purely inductive method to the study of narrative and that the initial step must be the study of all narratives within a genre, a period, a society, if one is to set up a general model. This commonsense view is, nonetheless, a naive fallacy. Linguistics, which only has some three thousand languages to contend with, failed in the attempt; wisely, it turned deductive, and from that day on, incidentally, it found its proper footing and proceeded with giant steps, even managing to anticipate facts which had not yet been discovered. What then are we to expect in the case of the analysis of narrative, faced with millions of narrative acts? [...] It is only at the level of such conformities or discrepancies [from the general model], and equipped with a single tool of description, that the analyst can turn his attention once more to the plurality of narrative acts, to their historical, geographical, and cultural diversity. (238)

[1] Barthes, Roland, and Lionel Duisit. "An Introduction to the Structural Analysis of Narrative." New Literary History, vol. 6, no. 2, 1975, pp. 237–272. www.jstor.org/stable/468419.

It is interesting that Barthes frames this as essentially a practical problem – there are simply too many texts, too many narratives, such a large quantity of narrative "diversity" that to try to make sense of it all at once by way of some kind of bottom-up, "inductive" approach would be to drown in the sea of information. What's needed instead is a general model of narrative – a deductive theory, like Saussurean linguistics, derived from first principles. Only then, with this in hand, would it be feasible to turn to the sea of narratives in all of their "diversity" and begin to grapple with the deviations from this general model. The deductive turn – and by extension, the turn away from any kind of historically or culturally grounded understanding of narrative – is presented here as a tactical decision, a response to a problem of *scale*.

In 2016 – exactly fifty years after the publication of *Communications* 8 – we now have a set of computational tools that do, in many ways, open the door to the type of inductive approach that Barthes warned against. And, sure enough, over the course of the last five years there have been a series of projects that have started to reason empirically about the internal structure of literary texts. Most notable is Matt Jockers' work with Syuzhet, a piece of software that measures the fluctuation of "sentiment" across the text – essentially the extent to which the text is happy or sad, which Jockers argues can serve as a proxy for "plot movement" in a general sense.[2] Andrew Piper, writing in a recent issue of *New Literary History*, analyzes Augustine's *Confessions* and identifies a marker for the "conversional" narrative – one in which there is a rapid shift in the semantic register of the text around the 70% mark, the moment of conversion – and traces this signature forward in literary history to identify other conversional texts.[3] Ben Schmidt, following in Jockers' footsteps, trained topic models on a corpus of TV scripts and then looked at the distribution of different topics across the narrative interval of the scripts, and was able to tease out a snapshot of the prototypical cop drama, a crime at the beginning and a trial at the end.[4] Meanwhile, here in the Literary Lab at Stanford, the Suspense project is plotting out the movement of "suspense" and "unsuspense" across novels, and Holst Katsma has looked at changes in the level of "loudness" across Book 1 of *The Idiot*.[5] Meanwhile, in computer science, there is an older line of work that studies the sequencing of words and topics inside of documents (though with a more applied than critical bent), dating back to Marti Hearst's work on text segmentation in the 90s.[6]

These studies have made great progress in the direction of the "inductive" approach to narrative theory that Barthes rejected as impractical – instead of deducing a general model from first principles, they begin with a sea of texts and try to draw out some kind of structural insight about narratives by observing lots of individual cases. And yet, in other ways they also reproduce many of Barthes' theoretical and methodological assumptions. For one thing, they tend to be inductive only in a partial sense – they begin with large a corpus of texts and work "upwards," but they go into this inductive

---

[2]Jockers, Matt. "A Novel Method for Detecting Plot." Matthew L. Jockers. N.p., 5 Jan. 2014. Web.

[3]Piper, A. "Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel." New Literary History, vol. 46 no. 1, 2015, pp. 63-98. Project MUSE, doi:10.1353/nlh.2015.0008.

[4]Schmidt, Benjamin M. "Plot Arceology: A Vector-space Model of Narrative Structure." 2015 IEEE International Conference on Big Data (2015): n. pag. IEEE Xplore. Web.

[5]Katsma, Holst. "Loudness in the Novel." Literary Lab 7 (2014): n. pag. Web.

[6]Hearst, Marti A. "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages." Computational Linguistics 23.1 (1997): 33-64. Web.

process having already picked out high-level phenomena to put under the computational microscope (sentiment, suspense, loudness, conversionality), which come down deductively from preexisting assumptions about how narratives work. Each of these signals is fascinating in its own right, and when studied empirically sheds light on a particular set of critical questions. But if we want to study narrative in a general sense, how can we know to pick one and not the other? There are a virtual infinity of "fluctuations" across the X-axis of the text that could be studied. Why sentiment and not suspense, suspense and not loudness, loudness and not something else entirely? This question *itself* calls for a more inductive approach, a bottom-up method for selecting the signals that we use to observe narrative structure.

Second – and perhaps with more important theoretical ramifications in the present moment – these studies have largely continued to treat narrative as an almost completely a-historical object. Claims are made about historically constant characteristics of narratives – that, for example, sentiment curves can be clustered together into six archetypal shapes. But, though we have started to develop tools that can grapple with what Barthes called the "enormous task of dealing with an infinity of materials," we haven't yet *capitalized* on this, we haven't continued forward to the implicit intellectual payoff – that, once we have some kind of purchase on the question of how to make sense of narratives, we might then be able to turn our attention "once more to the plurality of narrative acts, to their historical, geographical, and cultural diversity." We haven't used our new inductive powers to start theorizing about narrative *difference*, about the ways in which these two-dimensional fluctuations across narrative time themselves change across axes of historical time, genre, nationality, or identity. Hundreds of thousands of texts spanning across centuries are lumped together, and "narrative" is studied as a unified, synchronic phenomenon.

So – we have started to wade into the question of what it might look like to study narrative inductively, from the ground up, engaging directly with the "infinity of materials." But in two important ways this new computational narratology hasn't moved beyond the assumptions of 20th century formalisms: we continue to base our inductive studies on deductive assumptions about what "marks" or "defines" narrative, and we continue to think about narrative in somewhat outdated terms as something outside of history – something sealed inside a literary system that floats above or below the historical, material, and political contexts of literature, not something that is woven inside of them.

But progress can be made on both of these fronts. At a technological level, there are certainly no longer any kind of obstacles to thinking about what a fully inductive analysis of narrative would look like – this is more a challenge of imagination than of computation. And, on the question of how to bring these narratological questions into contact with their historical and cultural contexts – modern digital library collections are now large enough that there are enough texts in any given tranche of the archive – sliced by period, genre, race, gender, nation – that it has become possible to reason empirically about how narrative structure is particular to these contexts. There is enough data, now, that we can split our corpora into lots of sub-corpora, run the same analyses on each segment, and then compare the results to tease out a "moving image" of how formal structures change across different slices of the archive.

With these two goals in mind – in this essay I will describe a very simple but very flexible method for observing the chronological organization of novels – specifically, a corpus of roughly 30,000 American novels running from 1820 to 2000. Instead of starting

with a high-level concept like sentiment or suspense, I will look at the distributions of individual words inside of the texts, modeled in a statistically flexible way that makes it possible to inductively probe through the entire dictionary and identify the words that do the most narratological *work* – the words that most strongly mark different "regions" inside the text.[7] This makes it possible, in a sense, to systematically "survey" the interior axis of literature for the first time. What exactly do beginnings, middles, and ends actually consist of, when considered across tens of thousands of novels? Do these textual regions have consistent thematic signatures, or are they essentially idiosyncratic, everything washing out in the aggregate? If they do hold together at the level of the corpus – are they what we expect? To take up a question posed recently in *Narrative Middles*,[8] a collection of essays edited by Caroline Levine and Mario Ortiz-Robles – what, precisely, can be said about the *middle* of a text? Can it only be defined negatively – not the beginning, not the end, a connective tissue in between – or is there some sort of statistical essence of middle-ness? And perhaps just as important – if these narrative regions do instantiate in computationally observable ways – how stable are they over time? Have the archetypal metaphors for narrative structure – a day, a life, a season, a courtship, a marriage – changed across literary history? Are beginnings, middles, ends, climaxes, denouements different in 1940 than in 1900, or 1860, 1820, etc.? What's the *shape* of narrative, present and past?

To return to Barthes' warning that it is a "naive fallacy" to think that narrative structure might be studied in a fully inductive way – let's take this up as a challenge to do *exactly* that, and then, if we succeed, turn our attention to what Barthes' described as the eventual outlet and application of narrative theory – a turn back to questions of structural *difference*, of the diversity of texts and narratives, of how they compare to one another, not just the ways in which they are the same.

2

So, to start – how can we inductively observe the "shape" of narratives, what Brooks called the "movement, the slidings, the errors and partial recognitions of the middle"? We have some sense that narratives have topologies – beginnings, ends, middles, climaxes, denouements, plots that rise and fall, peaks and valleys, ebbs and flows. How can we get at this computationally? Right out of the gate, the immediate difficulty is that a narrative, considered as raw information, looks more like a perfectly straight line than a surface with texture and contour – the computer just sees an ordered list of words, a flat marble surface, without any kind of inherent structure that can be hooked on to or analyzed.

A quantitative study of narrative, then, has to start by picking out some kind of "signal" to look at inside the text. When push comes to shove, this generally boils down to some kind of method by which some words are selected for analysis and others are ignored.

---

[7] Of course, not even this is fully inductive. Why choose the word and not the letter, the punctuation mark, the part-of-speech, etc? And, is it necessary to think of the text as fundamentally linear? So this certainly remains deductive, as everything must, to some extent. But, my thought here is to approach this as an experiment in critical minimalism – what is the smallest and simplest sequence of critical *decisions* that result in a meaningful portrait of narrative?

[8] Levine, Caroline, and Mario Ortiz-Robles, eds. Narrative Middles: Navigating the Nineteenth-Century Novel. N.p.: Ohio State UP, 2011. Print.

For example, when Jockers looks at "sentiment," under the hood this really just a statistical model that picks out words that tend to mark "happy" and "sad" passages in hand-coded training data; when Schmidt looks at topics like "crime" or "trial," these correspond to buckets of words produced by a topic modeling algorithm; or, when the Literary Lab looks at "suspense," this ultimately maps onto a specific set of words identified by a classifier that tend to mark whether a passage is suspenseful or not. To "look at" the text quantitatively, we first need to ignore most of it – we have to choose the words that we care about, pick out one particular thread from the cloth of the text.

But, how exactly to go about this? The difficulty is that there are essentially an infinite number of threads that might be chosen, and the back-of-the-napkin math quickly stumbles into farcically large numbers. Considered mathematically – if a novel has 50,000 words, made up of, say, 5,000 unique word types, then in theory we could look at any member of what's called the "power set" of this vocabulary – all unique subsets of the 5,000 words of any size, which contains 2^n combinations. So, for our novel with 5,000 unique words, there are 2^{5000} "signals" that we might look at, an incomprehensibly large number that takes 1,506 digits to write out. Or, even if we assume that we don't care about groups larger than, say, 5 words, we'd still be left with roughly 26 trillion[9] unique combinations, which even if we could evaluate 1000 of these per second, it would still take 825 years to step through all of them.

Which of these is best, which is the most critically useful? This is where it becomes useful to start with some kind of preexisting hypothesis about what kind of "signal" or "phenomenon" is most important – suspense, sentiment, loudness – and build focused models that specifically measure those signals. This provides an a priori mechanism, essentially, for sifting through the sea of possibilities and picking a signal to look at. But, the reciprocal difficulty is that it's hard to know whether these signals are actually the ones that will tease out the most cleanly-defined portrait of narrative structure in the general, theoretical sense of the concept. Jockers implicitly assumes that sentiment just *is* plot for all practical purposes, that they are essentially one in the same, that a movement of positive or negative sentiment provides a definitive proxy for "plot" or "narrative." But is this actually true? And even if it does manage to give an image of some kind of narratological skeleton (and I think it does) – how can we know if it's the best way to do this, and not just some kind of local maximum? It's fairly easy to imagine any number of other signals that might fluctuate in interesting ways across the text. We could look for, say – some notion of "dialogism," the degree to which characters are or aren't speaking to each other; the ratio of "concrete" versus "abstract" nouns; the lengths of words or sentences; the complexity of the vocabulary; words with Latinate versus Germanic roots; and so on and so forth.

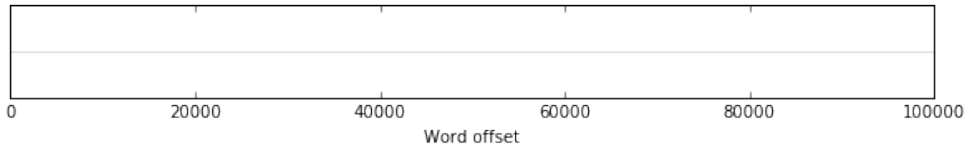
How could we go about this inductively? Though the analogy isn't perfect, this is similar to a "model selection" problem in statistics – for example, when a medical researcher is trying to figure out how to predict a certain type of cancer by looking at the genetic information from thousands of patients, where the data for each patient contains millions of individual genes that might predict the clinical result. How to figure out which of these genes are the most effective discriminators, the most powerful signals or markers for the condition? Or, for us – how might we "survey" the interior of literature in a more systematic way, as much as possible without preconceptions

[9] 26,015,651,042,712,248, precisely.

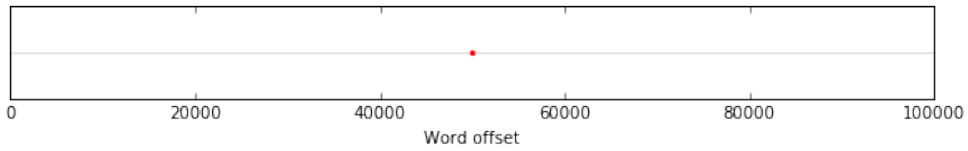
about what we might find, identify the most critically useful "filtrations" of the text?

It's not possible to probe through all 26 trillion possible combinations. What we can do, though, is exhaustively search through all five or ten thousand individual words – the complete vocabulary of the corpus – and try to find words that show the most statistically irregular or unlikely patterns of distribution across the text. Once these are in hand, we can then try to use them as signposts that point to broader literary-critical insights.

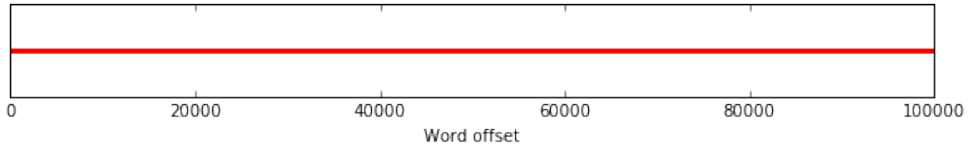
This data can be built up from information about the structure of individual texts that is extremely simple, almost to the point of being self-explanatory. We can start by representing the text as a horizontal X-axis, running from 0, the beginning of the text, to 1, the end of the text:



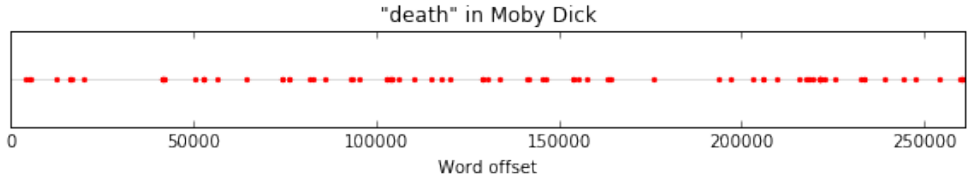
Where if, say, the text is 100,000 words long, then the 50,000th word would sit on the 50% mark on the axis:



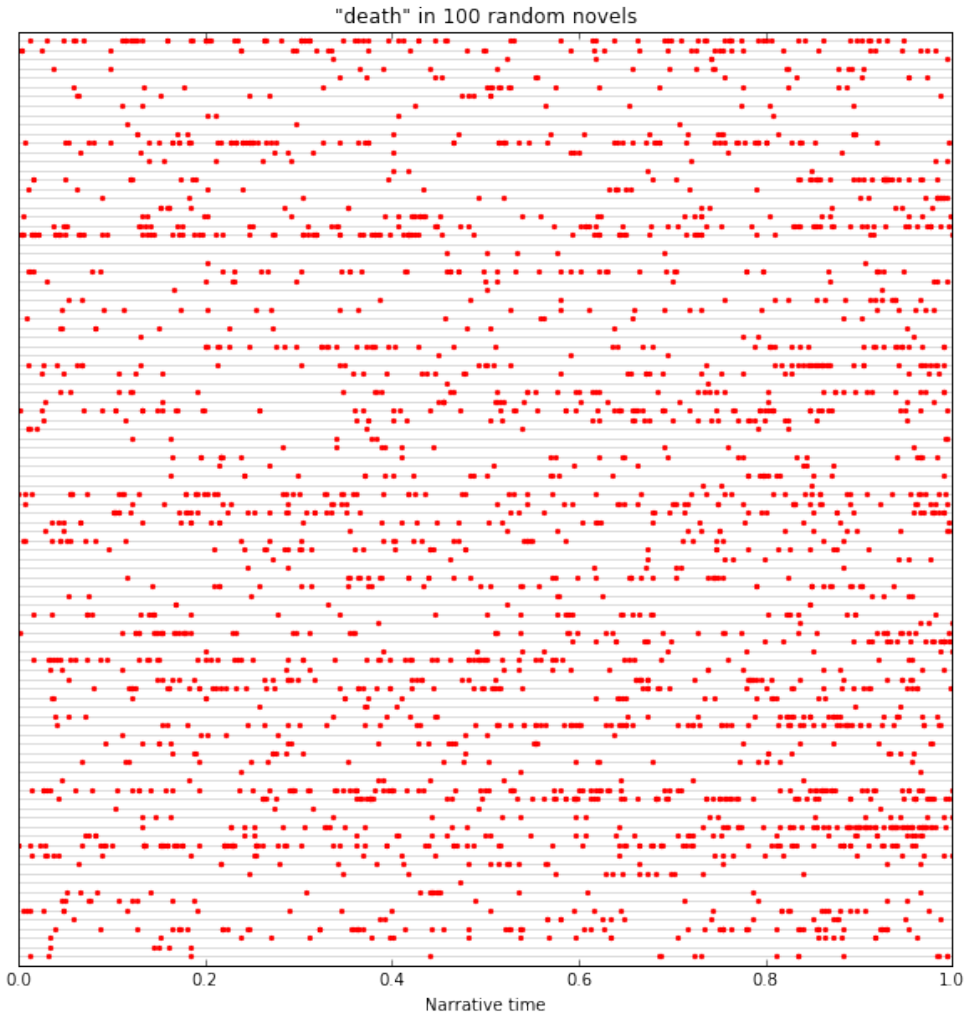
Or, all of the words at once, which blur together into a solid line:



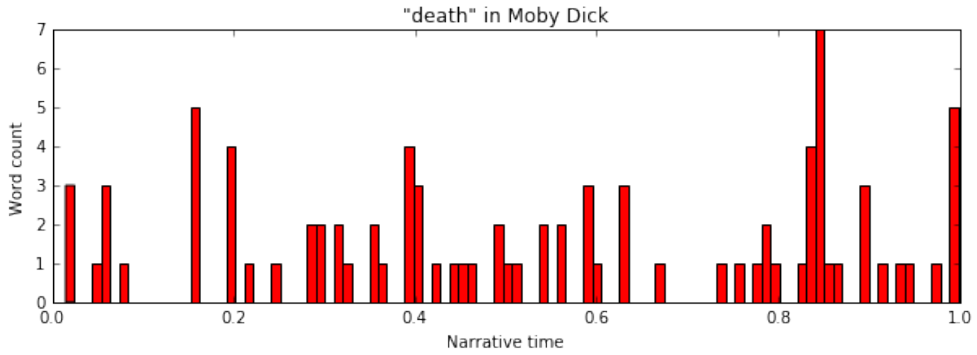
This is the raw, encountered data of the text, the straight line, the flat surface. But, instead of plotting everything, we can also just look at the positions of any individual word in the text. For example, here's the word "death" in *Moby Dick*:



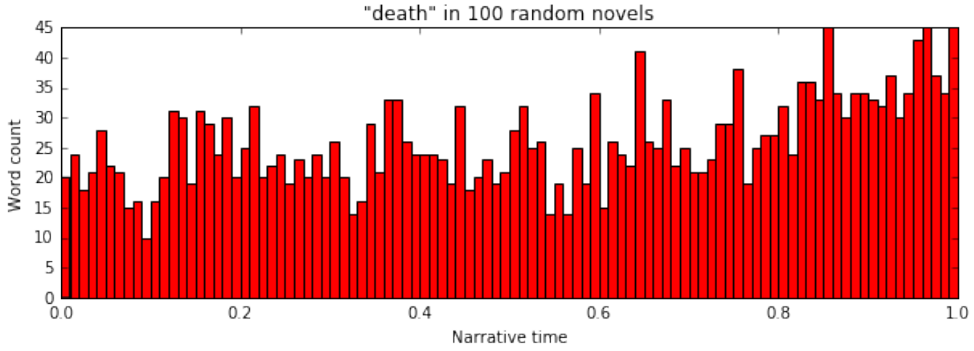
Just eyeballing this we can get a rough sense of how the word distributes inside of this individual text. But, the signal is weak, volatile, under-sampled – there isn't enough data here to say anything with confidence about the general tendency of the word. What we can do, though, is merge together this information from larger collection of texts, where each provides a set of weights across narrative time for a given word. For example, here's "death" in a set of 100 other randomly-selected novels, this time with the X-axis mapped onto a standardized 0-1 scale, which makes it possible to compare novels of different lengths:



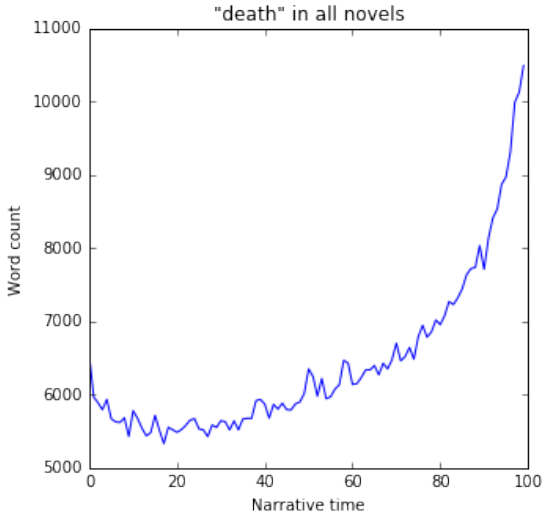
We can't just "add" these up directly, though, since different texts are almost never the exact same lengths, which means that the precise 0-1 offset values will be slightly different, even when instances of the word fall in very similar regions in the texts. To get around this, we can split up the X-axis into 100 equally sized segments, each representing one percentile of narrative time. For example, for *Moby Dick*:



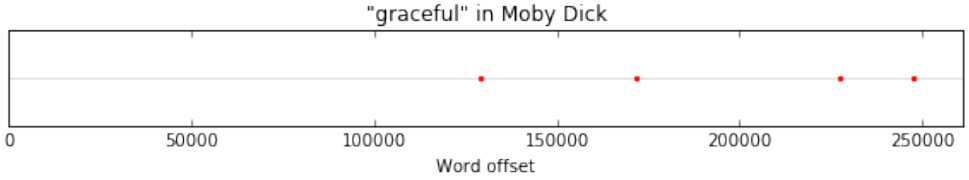
And now, with the words grouped together, we can just add up the counts of "death" in each percentile across all texts and get a combined distribution for all 100 texts:



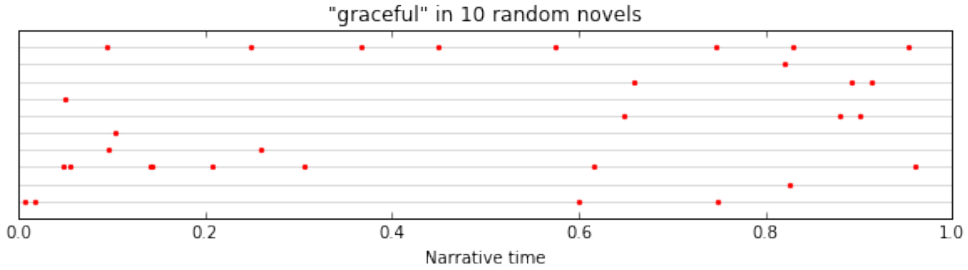
Now, we start to get a higher-fidelity version of the signal, which is what we'd expect – death tends to be an ending word. Or, beyond just 100 texts, we can do this for all of the Literary Lab's American novel corpus, ~30,000 texts from the Gale American Fiction corpus and the Chicago TextLab corpus ranging from ~1820 to 2000, this time plotted as a time-series curve across the narrative interval:



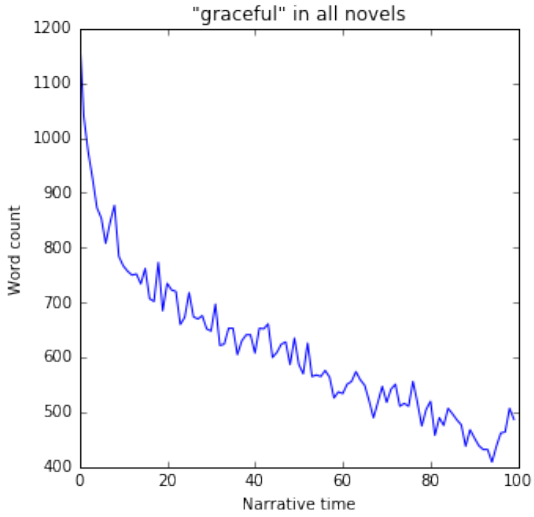
The really powerful thing about this, though, is that it makes it possible to tease out the hidden distributions of words that are so infrequent that they would rarely, if ever, show up enough times to see any kind of trend in an individual text. "Death" is a frequent enough word that we can get a pretty good sense of its distribution just by looking at data from a handful of texts, and presumably we could pick up on this if we read them directly with this question in mind – there's enough data there to start to reason anecdotally about how the word tends to distribute across the text. But for most words this isn't the case. By Zipf's Law, most words will only show up 1-2 times in any individual text, which gives nowhere close to enough statistical power to make any inferences about their general narratological structures. For example, here's "graceful" in *Moby Dick*:



And, in 10 other novels:



It's showing up just 3-4 times, if at all, and, glancing at these, looks basically random. But, when we merge together this data across the entire corpus, we can immediately see a strong pattern:

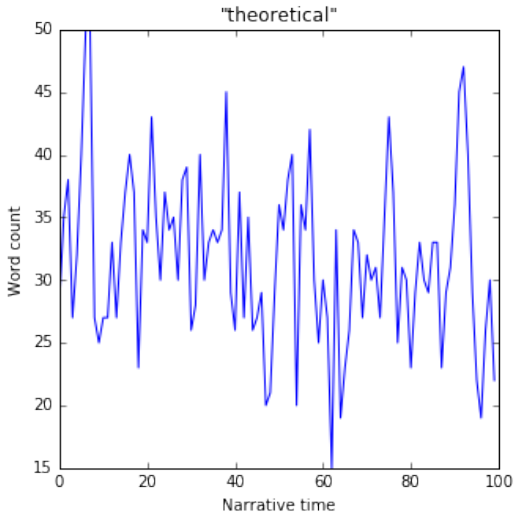


Far from being random, "graceful" concentrates strongly at the beginnings of novels. (We'll circle back shortly to why this is.) This provides an interesting epistemological power – we are using lots of texts to surface information about the general distributional tendency of the word, something that probably *doesn't actually instantiate in any individual text*. To borrow from the language of Bayesian statistics – we are teasing out information about the "prior" that governs the distribution of "graceful" across the text – there is a kind of narratological energy that makes it most likely that "graceful" will show up at the very beginning of any given text, even though there are few if any cases where either the author or reader would be aware of this as a deliberate literary decision.

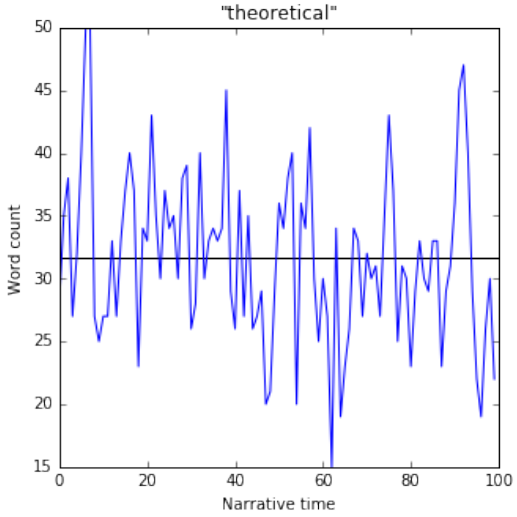
So – we can start to reason about these narratological tendencies of words, the regions of "novel time" in which they tend to concentrate when averaged out across large numbers of texts. But, where do we go from here? We can just start spot checking individual words that seem interesting, but, if we want to avoid any kind of bias or preconception about what might be most important, what we really want is some way to systematically search through all of the words in the corpus and pick out the ones that have the most irregular or anomalous distributions, the words that do the most narratological *work*.

Before we can do this, though – before we can compare this notion of irregularity across words – we first need to be more specific about what it means to say that any given word's distribution is "irregular." We can glance at the plot for death and notice

that it seems clearly lopsided, but how do we formalize this? This is an important question to ask, because, in the same way that most words in an individual text only appear a handful of times, a great many words will only appear a small number of times in the entire corpus, and, even when we stack up 30,000 texts, the combined distributions will still be extremely under-sampled. For instance, take a word like "theoretical", which only shows up 3,166 times in the entire corpus:

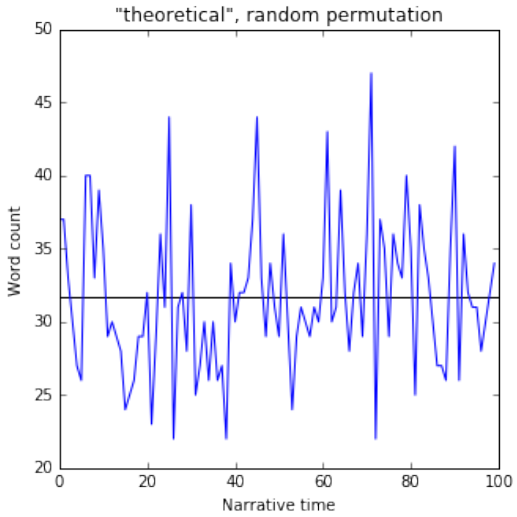


If you squint at this, there might be some kind of downward trend? But, how do we know if this is significant, or just noise in the data? We need to be able to say whether the observed trend for a word differs significantly from what we'd expect to see if there were no effect. In this case, this null hypothesis is simply that there is *no relationship between the frequency of a word and the position in the text.* So, in other words, if a word appears a total of 1000 times in the text, we would expect that it would show up exactly 10 times in each of the 100 percentiles – that the line would be perfectly flat across the text. For "theoretical," we'd expect about 32 times per percentile:

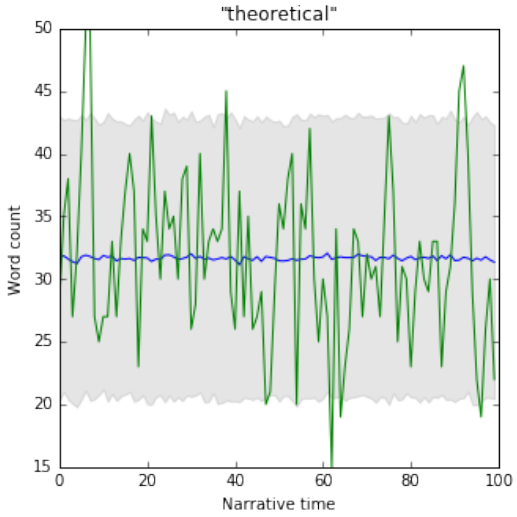


Where the flat black is the expected value, if there were no effect, and the blue line is the observed data. So, at some points, the actual counts are higher than expected, and at other points lower. But how much above or below does the observed series need to be in order for it to be significant? There's a closed-form formula that we could

use to calculate this, but it's also possible to directly simulate the expected error, which in many ways is more intuitive. We can treat the flat line – the expected distribution – as a multinomial probability distribution, and then take random samples from this distribution, which will give us examples of the kind of real-world data we would expect to see if there were no significant trend. For example, here's one random sample:

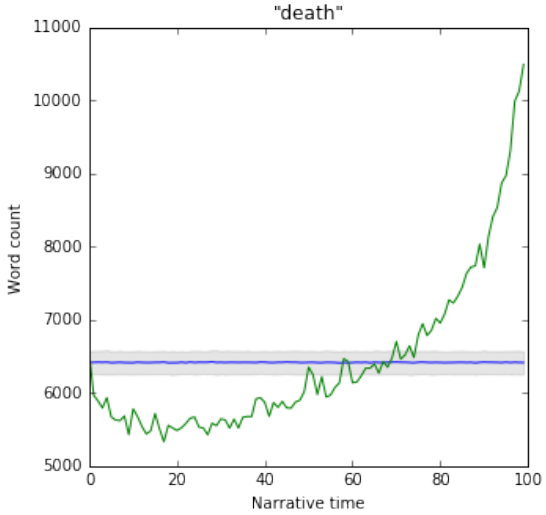


Already, we can see that this doesn't look too different from the observed values in the last plot. To formalize this, we can "permute" the baseline distribution – draw from it, say, 1000 times. For each of the 100 percentiles, this gives us 1000 values that were produced from the random sampling. From these, we can then compute a simple standard deviation, and then use this to define a "band" around the expected flat distribution that marks the boundary beyond which an observed point would be considered statistically significant. For example, here's the two standard deviation band, which means that if an observed point falls above or below this, there's a ~95% chance that it's not just a random fluctuation – that the word is significantly over or underrepresented at that position in the text:



So, in this case, with the low-frequency word, we can see that the actual curve almost always falls inside of the 2 standard deviation band, meaning that we can't say with much confidence that there's any general relationship between the position in the text

and the frequency of the word – either the word doesn't have any kind of chronological anchoring in the text, or we just don't have enough data to be able to say. But, if we turn back to "death," and do the same thing:



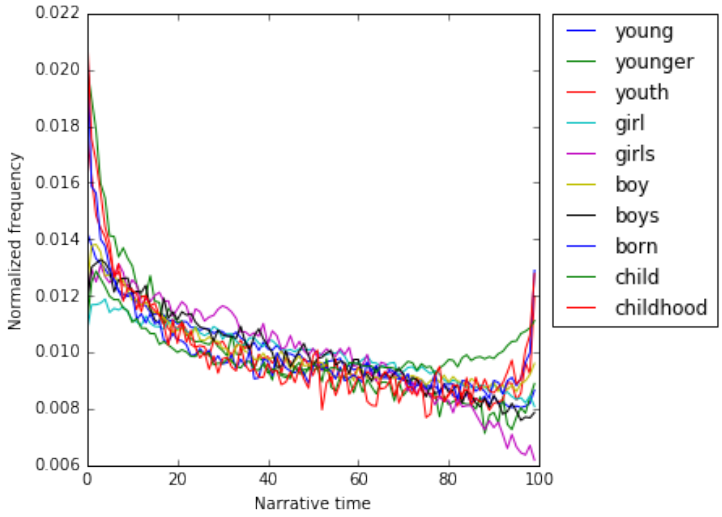
We can see that "death" is far above the expected value at the end of the text – precisely, 51.6 standard deviations – meaning we can say with enormous statistical confidence that death is concentrating at the ends of novels. Now, with this in hand, the last step is to find some way of converting this into a single score that captures the degree to which a word is irregular, and in a way that can be compared across words. There are a couple ways to do this, but one simple way is to just take the average number of standard deviations away from the expected value for each of the 100 percentiles – the average of the zscores in each bucket. This will give high scores for words that have the most lopsided "trends" across the text. For example, for "death," the average zscore is 10.03, whereas for "theoretical" it's just 0.95.

With this in hand, we can now turn back to the question of how to inductively pick out the words that are the most irregular – words that provide a kind of narratological scaffolding across thousands of novels. To do this, we can just step through each word in the dictionary, compute this metric, and then sort on the values. Here are the 500 most irregular words, with the most irregular first:

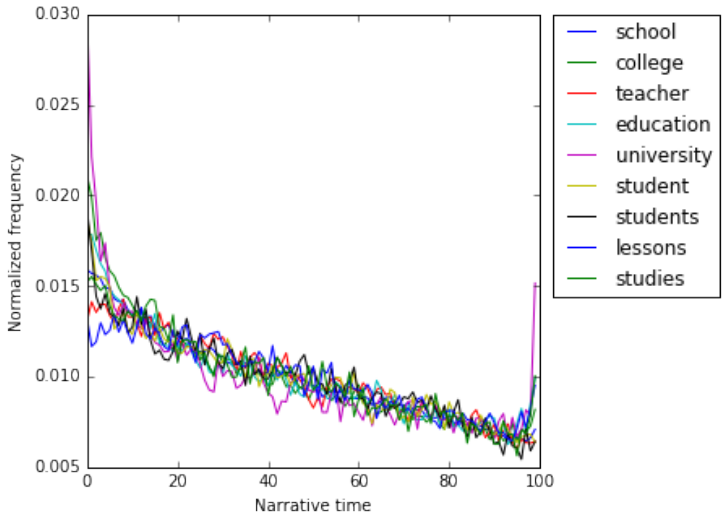
> a, you, i, the, young, of, years, me, him, school, father, iii, mother, love, to, will, ii, death, t, he, said, hair, now, that, god, she, iv, it, what, we, boy, have, do, age, old, year, tall, and, dead, not, girl, girls, vii, an, blue, again, viii, miss, family, if, college, letter, ix, brown, don, heart, vi, small, life, told, large, kill, his, always, chapter, can, like, think, know, joy, or, tell, happiness, die, books, its, prisoner, killed, forgive, little, could, boys, has, your, loved, black, gun, would, had, handsome, xi, xii, new, in, last, arms, trial, back, older, pretty, stranger, xiii, about, aunt, good, tears, done, then, class, big, jury, xx, pain, go, liked, happy, beauty, shall, must, xix, be, her, broad, dying, children, prison, saw, mrs, rather, wedding, well, younger, did, xiv, youth, high, gone, green, cried, nose, is, witness, table, teacher, xvi, still, project, which, child, xviii, rich, us, very, lived, usually, evidence, xvii, leave, dinner, no, mr, uncle, often, most, died, xv, murder, was, come, let, with, education, some, hope, arrest, fine, parents, pistol, longer, enemy, son, six, all, asked, wounded, my, hands, twenty, hand, wore, cannot, white, confession, away, born, truth, english, book, grandfather, author, red, beautiful, gray, murderer, escape, built, against, complexion, return, whispered, came, five, sometimes, daughter, see, m, summer, dress, prisoners, terrible, why, world, nice, wife, university, town, from, slender, any, conversation, want, news, going, get, through, eyes, strength, ladies, believe, much, guilty, save, ve, village, agony, knew, talk, been, quite, houses, sort, how, bride, tried, testimony, forgiveness, are, country, laughed, attack, re, battle, went, message, lady, thin, study, left, dark, three, social, should, end, land, cry, suffering, features, anything, forever, peace, graceful, thought, sun, shot, house, tea, hat, student, yes, mean, v, they, manners, sea, their, sorrow, fell, century, students, night, story, defendant, hospital, fiction, store, street, one, once, plan, gentleman, fat, published, face, such, x, justice, supper, papa, lad, cell, attractive, ll, mamma, name, stories, wait, four, am, cousin, dear,

alive, generally, court, shop, two, married, kitchen, many, novels, troops, somewhat, thirty, home, farm, taste, bright, final, yellow, revolver, man, soul, society, say, dance, whose, replied, eat, this, kind, trembling, height, moment, west, england, darkness, soft, so, mercy, blond, prosecution, business, despair, wronged, as, bullet, skin, cook, soldiers, eldest, thick, help, long, novel, type, grandmother, leather, guard, interest, fashioned, dog, heaven, elder, answer, horror, half, meal, favorite, style, york, color, coffee, speak, fired, early, revenge, pleasant, fashion, fish, square, sent, safe, ain, grief, grave, earth, master, companion, fair, ask, marry, women, curls, acquaintance, desperate, than, work, past, arm, guards, play, sure, words, anguish, sheriff, windows, lessons, attorney, modern, childhood, car, rare, witnesses, got, kissed, shoes, history, reader, smile, breakfast, lost, golden, flames, studies, found, stood, french, charming, tunnel, chairs, alone, body, fate, music, haired, guilt, bread, destroy, middle, brick, really, jail, seven, might, pink, firing, interested, right, wrote, habit, mama, weapon, release, wide, amused, wooden, chin, tobacco, crime, subject, neighbors, youngest, fear, silver, voice, delicate, but, manner, inherited, dining, sacrifice

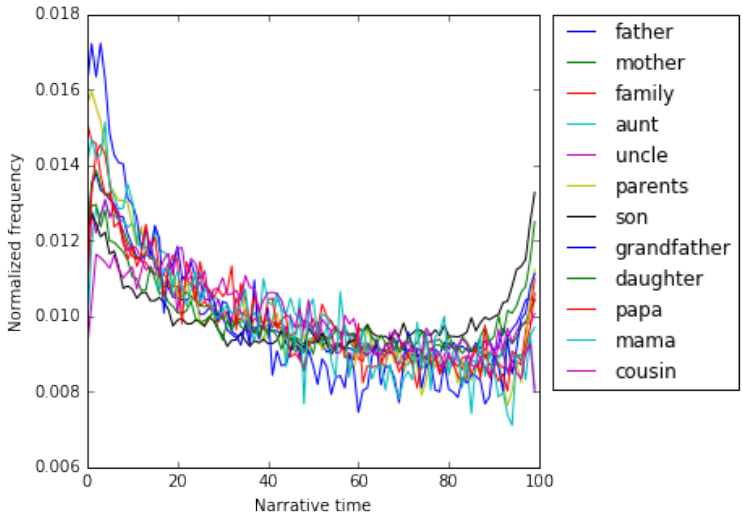
There's lots to look at here, but, as a gut check on the method, we can immediately see groups of words that map onto intuitive notions about various types of narrative unities – narrative as a life, a day, a courtship, etc. For example, words about youth and childhood peak at the very beginning, and then fall off across the rest of the text:



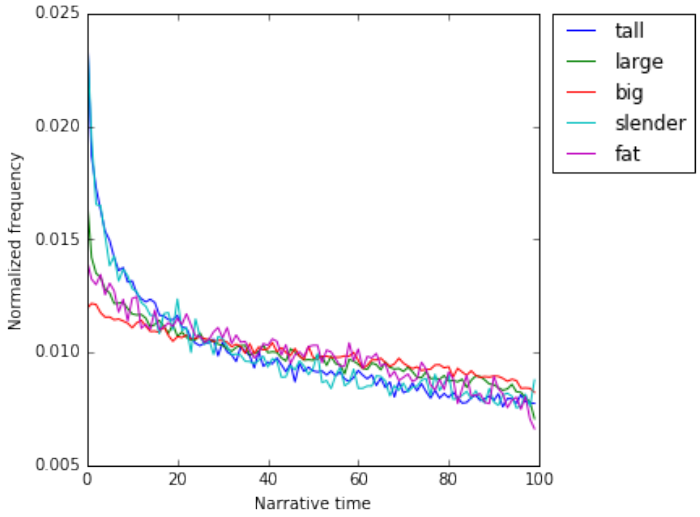
Likewise with words about education:



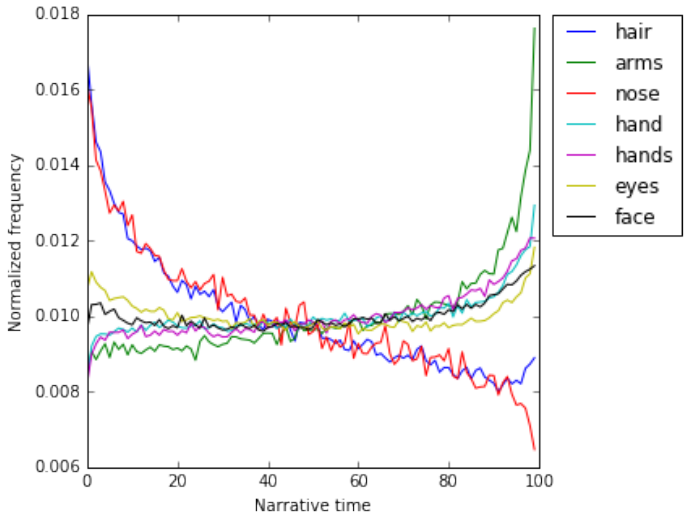
And, in the same vein, words about family life:



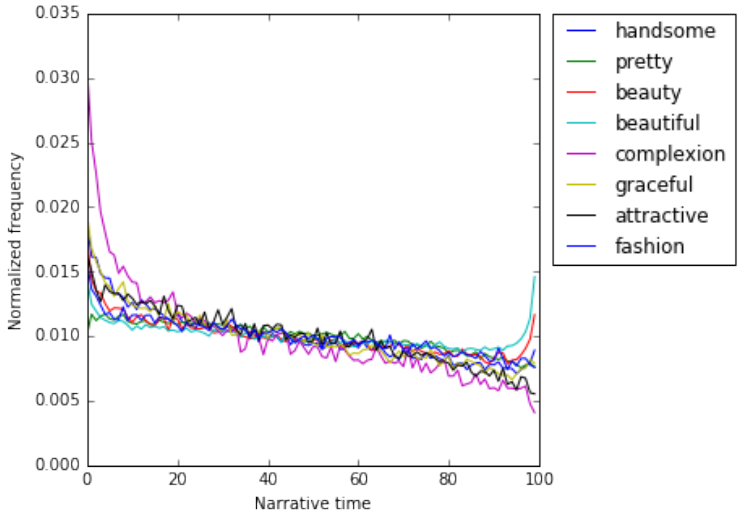
We can also see a concentration of words used to *describe people* at the beginning, which also makes sense – characters have to be introduced at the start:



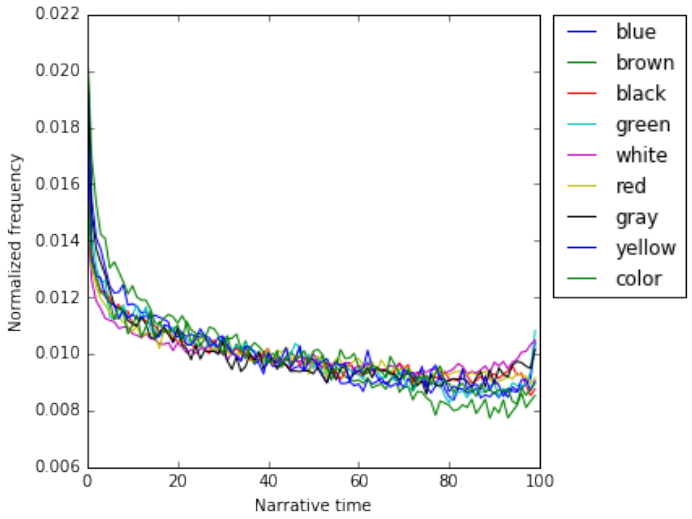
Body parts, meanwhile, are interestingly spit between beginnings and ends. We can make out an interesting distinction between body parts that are used to describe physical appearance ("hair" and "nose"), which spike at the beginning, and body parts that are associated with displays of emotion, which peak at the end ("arms" getting thrown around people, the taking of "hands"). And "eyes" and "face," which seem to be associated both with description and emotion:



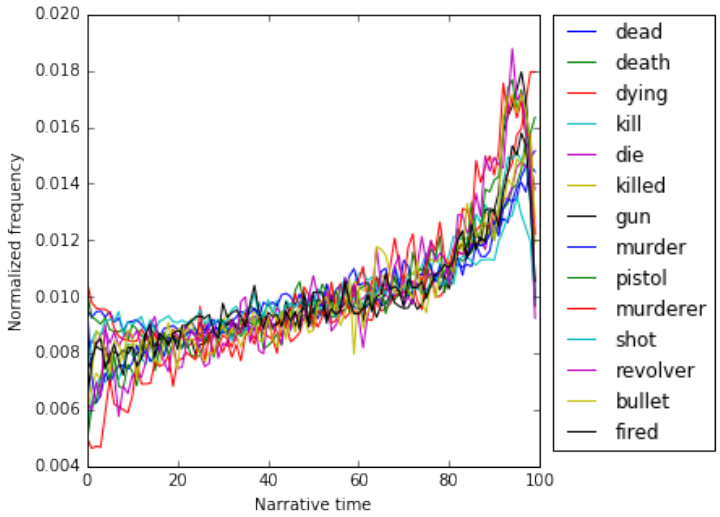
Or, along the same lines, a set of words related to fashion and (mostly good) appearance:



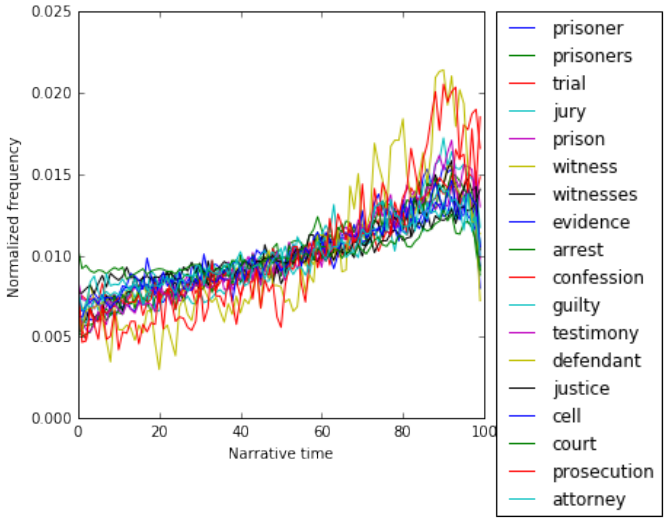
In a more general sense – *colors* show up at the beginning, also used in obvious ways in the process of painting the fictive scene into existence:



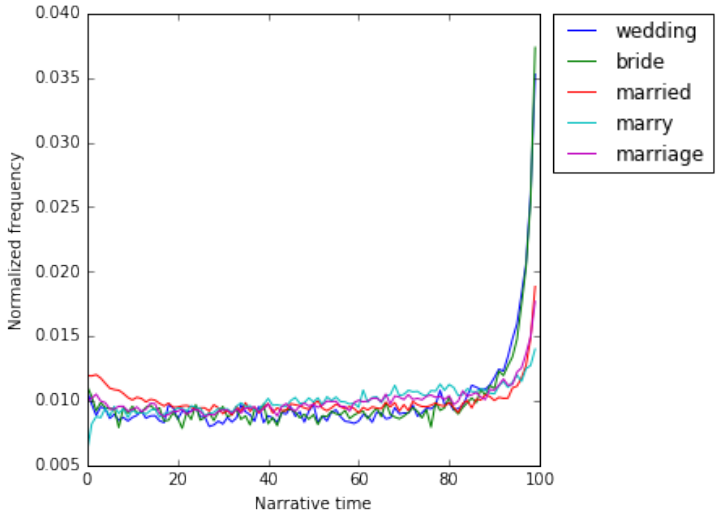
And, at the end? First and foremost, we see words related to death, particularly involving guns, many of which seem to peak just before the end at what looks like a "climax" of some sort, right around 95%:



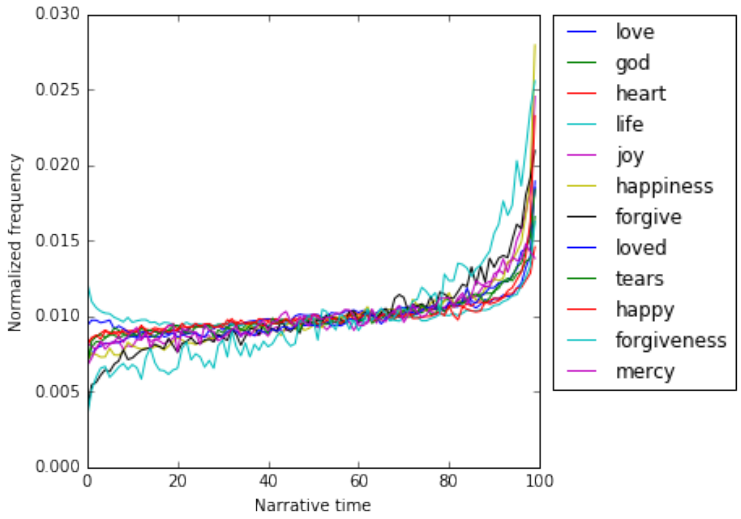
Related, words about criminal justice – the trial, broadly writ – not unlike the topics that Ben Schmidt identified in the TV cop dramas:



Meanwhile, perhaps the flip side of ending-as-death, the marriage plot:

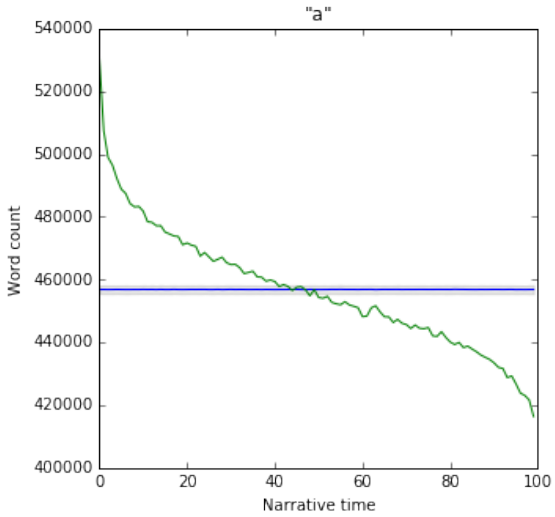


And last, along these lines but in a more general sense, a group of words associated with *resolutions* – in which enemies are reconciled, lovers united and reunited, sins forgiven, loose ends tied up, narrative tension discharged, the plot moved into its final pose:



So, this all looks about right – beginnings tend to be about childhood, family, education, and the introduction of people, places, and things; endings are about death, trials, weddings, reconciliation, forgiveness, various flavors of living happily ever after. Not many of these signals about narrative content are surprising – but this is useful as a confirmation that the method is effective, a gut check that we're picking up on literary signals that jibe with real reading experiences.

Much more interesting on this list, though, is the presence of a collection of very high frequency words that, it turns out, also have extremely skewed distributions across narrative time. For example, take the word "a" – the indefinite article – which comes out as *the* most statistically irregular word in the entire dictionary. The distribution of "a" has a highly symmetrical, almost mathematical structure that looks sort of like an inverse "logit" function – it begins high, falls off very quickly, declines more gradually through the middle of the text, and then falls off sharply at the end, mirroring the drop at the beginning:



What to make of this? This is an interesting case where it's not actually very useful to check our intuitions by dipping back down into individual passages in the corpus, since a word like "a" is so common that it will show up in a huge range of contexts, which makes it hard to reason about in the abstract. But, just considered grammatically, "a"

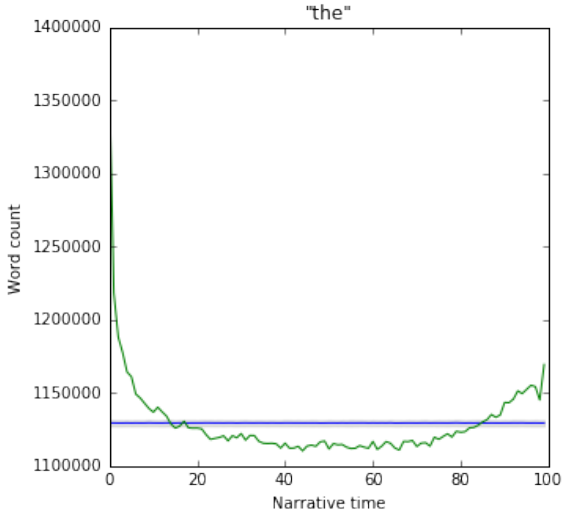
is generally used when a noun is introduced for the first time, when an object makes its first appearance in some narrative context. For example, you might say – "a man was walking down the street" – but then, after that, you would switch to the definite article, "the" – "the man walked into a shop," etc. So, with this in mind – and with the caveat that this drifts somewhat into speculation, and that it would take more work to back this up empirically – the obvious interpretation is to think of this graph as a kind of statistical x-ray of the *rate at which newness is introduced into the text*. At the beginning, this is very high, as the fictive world is painted into existence, people and places are introduced, the pieces of the narrative are put into place. It then falls off rapidly after the beginning once the stage is set and the plot starts to spool out, as the fictive matter introduced at the beginning shifts into the register of the definite "the." Then, over the course of the middle of the text, the plot continues to move into new fictional space – new people, new places, new objects – but more slowly than at the beginning, which had to bootstrap the entire fictional world into existence from the ground up. And finally, at the end, this rate of movement falls off quickly – as if the text arrives at its ending position a bit early, rapidly stops moving into new fictive territory around the 95% mark. "A," in other words, gives an empirical signal for the "speed" of the novel, in one sense of the idea – the degree to which it is moving into fictive contexts that have to be introduced for the first time, as opposed to staying put inside of contexts that have already been blocked in by the preceding narrative.

But, beyond the interpretation of any one word – what's really most interesting here from a theoretical standpoint is that there's any effect, that "a" shows any kind of narratological tendency across the text when considered across tens of thousands of novels. "A" is purely a function word, the 6th most common word in the English language, and bears no semantic content. And again, remember that the null hypothesis here is that we would just see a flat line, that there would be no relationship between the frequency of the word and the position in the narrative. But, not only is there a relationship, but there's a spectacularly powerful one, in fact the most significant of any word in the dictionary – in this corpus of 30,000 novels, in the first percentile of the text, "a" shows up almost 80,000 times more than we would expect under the null hypothesis, which corresponds to a zscore of 112, where anything above 2 is considered significant.

This becomes much more interesting. Unlike the more semantically focused signals – youth, death, marriage – this starts to feel like a keyhole view into some sort of sub-semantic narratological architecture. But, it's not clear exactly what we should make of this. Does it make sense to think of this as just a lower-level version of the types of constraints enforced by, say, genre conventions like the marriage plot, which push "marriage" to the end? Or does that overestimate the importance of literary convention, at this level – is there some sense in which this trend is *inevitable*, that this must be the case, that it would be nearly impossible to write a narrative that doesn't show this pattern? As far as "a" is concerned – is a novel like a ball placed at the top of a steep, smooth incline, where it would take some kind of huge structural exertion not to roll down the slope, not to show this symmetrical falloff in the frequency of the word? (This raises the tantalizing empirical question – are there any texts that invert this, in which "a" *increases* monotonically across the text – texts that roll uphill?)

Or, to take another example. If "a," the indefinite article, shows this monotonically decreasing pattern, then we might expect the definite article, "the," to do the opposite, to consistently increase across the text – as more and more fictive material is sketched

into existence, perhaps "the" becomes increasingly necessary to denote all the objects that have already been introduced? In fact, though, "the" – the fourth most irregular word on the list, and the most common word in the language – looks like this:

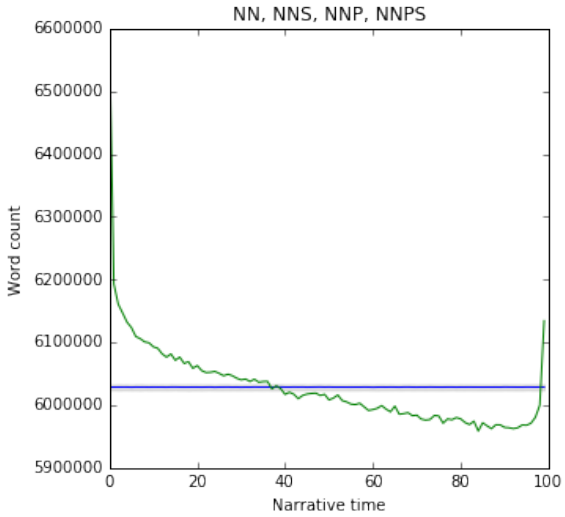


A huge spike at the beginning that falls off very quickly in the first 10% of the text, stays low through the middle, and then spikes back up at the end, though less dramatically than at the beginning. So, "a" and "the" – flip sides of the same coin, grammatically speaking – seem to do conceptually different work at a narratological level. Both seem to mark beginnings and ends, but in different ways. "A" shows something about how they are different – the beginning is a deluge of newness as the fictional world is introduced, a fast start out of the gates, and the end is slow, stationary, in the sense that it has stopped moving into new territory and occupies the space already created across the beginning and middle. But, "the" spikes at both the beginning and the end, and seems to show something about how they are similar, a way in which the end is some kind of return to the beginning.

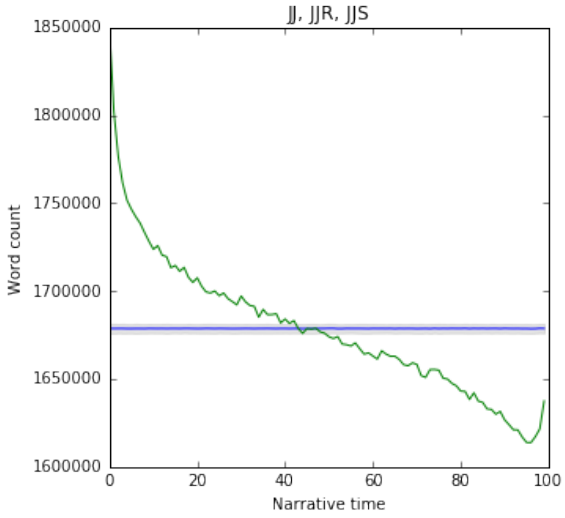
But, in what sense? What does "the" mark, at a literary register? Again, any one of these questions could be taken up as a project in its own right, but, it's easy to speculate – perhaps one way to think about beginnings and endings is that they both have to concern themselves with specific, concrete things? This fits with the idea that the beginning has to essentially hammer together the "stage" of the narrative – the room, the house, the table, the chair, the street, the train, the material platform on which the plot plays out. And, the end? There's an intuitive sense in which the end seems to return to this "surface" of the fictive world, though it's a bit harder to put a finger on what the gist of this is. One way to think about it, I guess, is just to say that things tend to happen at the end – pistols are fired, people are killed, lovers are reunited, the bride walks down the aisle, etc. The action of the narrative crescendos, the pieces of the plot are moved into their final positions – Robert Jordan blows the bridge; Ralph dies and Isabel goes back to England, only then to return to Osmond in Italy; Lucy and George marry and go back to Florence. We can imagine, if a novel is turned into a screenplay, that the beginnings and ends have the largest amount of "description" in the text of the script, the prose narration that the screenwriter threads in between the lines, the rough sketch of what would appear on the screen but not in the dialogue – the layout of the

set, the physical movements of the characters, the broad strokes of the photography.[10] Beginnings and ends are preoccupied with narrative *stagecraft*, which pulls the narrative out into a register of description, physicality, specificity, concreteness, *particularity* – the domain of "the."

Indeed, though it would take lots of analytical and interpretive work to really tease out the mechanics of this in any detail, we can start to triangulate onto this sort of account by looking at looking at the combined trends of much larger cohorts of words in the dictionary. For example, to hook onto the extent to which the narrative is concerned with objects or particulars – we can merge together all words that get tagged as nouns by a part-of-speech tagger. As with "the," we see a large spike at the beginning, and a smaller (though still statistically enormous) rise at the end:



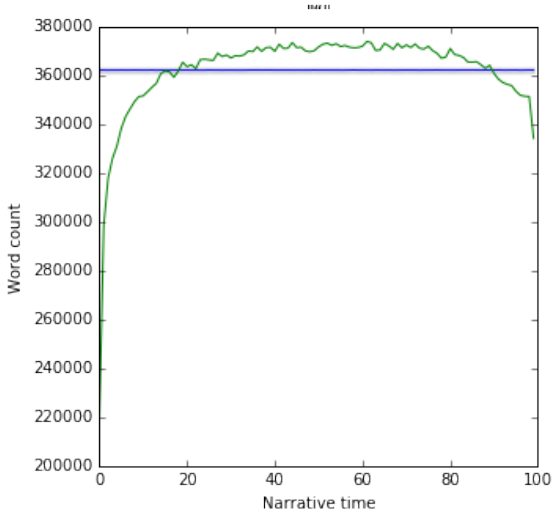
And, something similar for adjectives, which we can assume provide some kind of simple proxy for "descriptiveness":



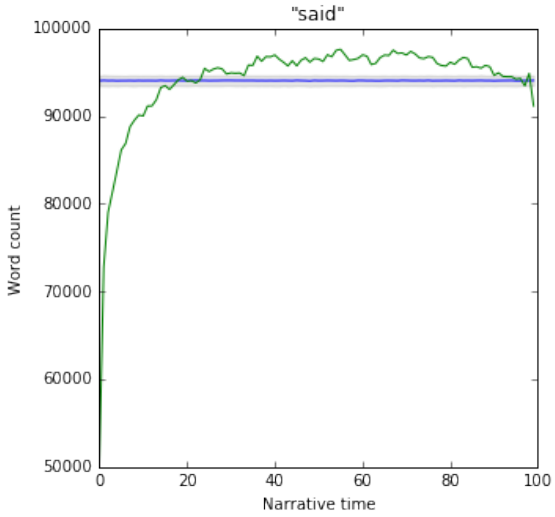
So, beginnings and ends are filled with objects and descriptions, particulars, marked by "the" – but "a" is only needed at the beginning, when the fictive matter of the text is first being narrated into existence. But, this really just scratches the surface –

[10]This also could be empirically checked, if we had a corpus of film scripts.

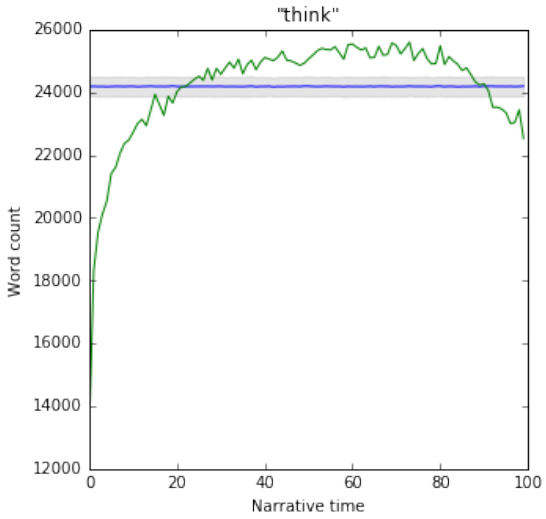
there are many more questions than answers here. If narratives begin and end on the physical "surface" of things, the description of objects – then what sits "below" this surface, outside the beginning and the end, in the middle? To pick up a variation of the question posed by Levine and Ortiz-Robles in *Narrative Middles* – if we can say that beginnings and ends are concrete, physical, or external, then how do we interpret the *absence* of this signal across the middle of the text? A narrative middle that's non-concrete, non-physical, non-material, somehow internal and abstract? What is this, exactly? Instead of just saying what this isn't, what *is* this middle, in a positive sense? Maybe – some sort of psychological interiority? Or, to pick up with the screenwriting analogy – maybe dialogue, interactions between characters? Indeed, the quotation mark – a good if not completely foolproof proxy for dialogism, the degree to which characters are speaking to each other – is an almost perfect inverse of "the": very low at the beginning, high through the middle, and then down again at the end, though not as strongly as at the beginning:



Likewise with "said":

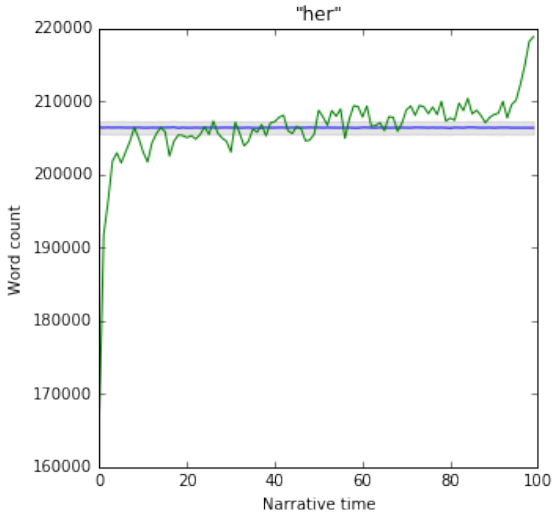


And, perhaps the flip side of "said" – "think":

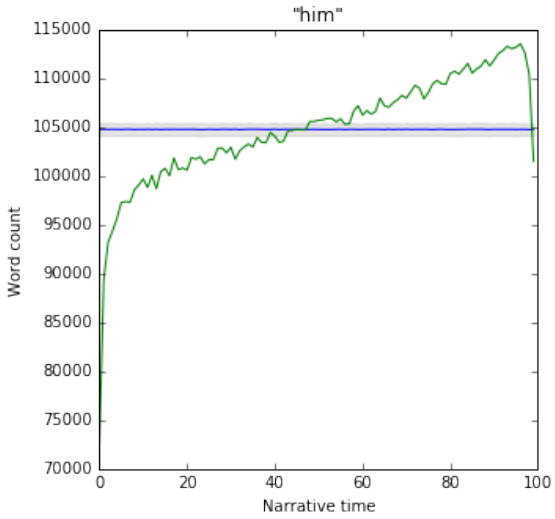


This looks about right, really, though it's interesting to see it shake out of the data so organically. In the most schematic sense, then, can we say that – novels begin with description, pass through a middle filled with speaking and thinking, and then circle back to description at the end? The center of the novel is human discourse – inter-personal (speech) and intra-personal (thought) – bookended by physical description and action?

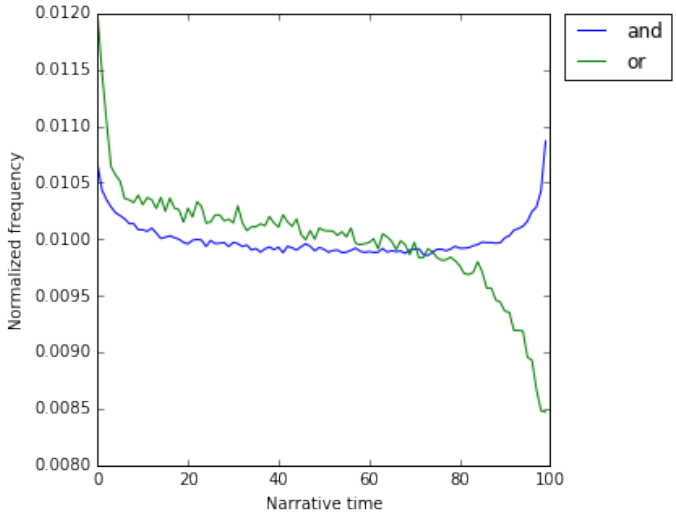
Other things are more mysterious. For example, it appears that gender has distributional patterns across narrative time. What should be made, for instance, of the fact that "her" is basically flat across the novel, but then spikes suddenly at the end:



Whereas "him" rises in a remarkably linear way across the text, and then craters in the last 5%?



Or that "and" and "or" should behave so differently at the end?



Or for that matter, what's happening with any of the other high-frequency stopwords that show irregular patterns – "I," "you," "of," "me," "to," "will," "we," "what," "do," etc.? Each of these feels like a loose thread that might be tugged at, one little strand in a larger matrix of energies that clearly act on narratives in very powerful and fundamental ways. But, the full shape of this, and how it all fits together – are questions for another time.

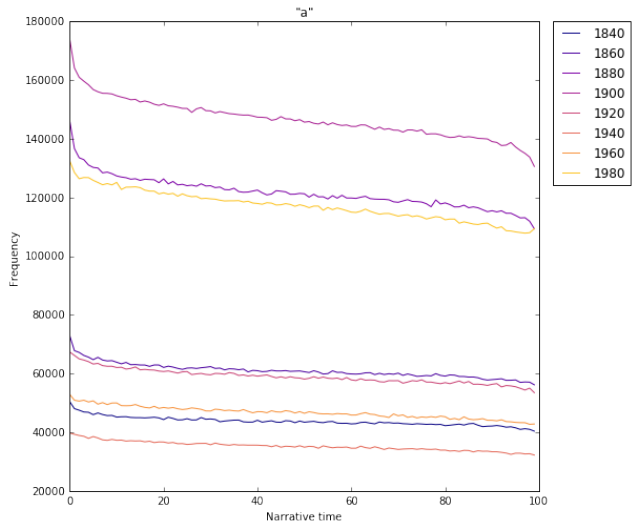
3

So – a way to "image" the narratological structure of individual words and then probe through the dictionary and pick out the words that provide the most narratological structure, that deviate most significantly from the expected flat line. We can inductively identify the most narratologically salient threads in the fabric of the text, which seem to come in two varieties. First, a layer of subject matter words that more or less map onto expectations about genre conventions and traditional notions of narrative unity – youth, school, family at the beginning; death, trials, reconciliations, and weddings at the end. And second, alongside this, a more sub-structural layer that plays

out at the level of function words and parts of speech, which seems provide a glimpse onto what might be thought of as the underlying "physics" of narrative chronology, operating below the layer of plot of subject matter, like the steel frame of a building – always there, always doing work, but invisible from the inside, hidden behind the walls.

So far, though, we've followed in the footsteps of previous studies in treating the corpus as completely synchronic information. To build up the composite distributions for each word, we have to merge together the distributions extracted from each individual text. And, so far, we've been looking at al 30,000 texts at once, which has the effect of collapsing any kind of difference across historical time or cultural context. To turn back to our original theoretical objectives – how can we "three-dimensionalize" this data, cast it along a third axis of historical context, turn our attention "once more to the plurality of narrative acts, to their historical, geographical, and cultural diversity"?

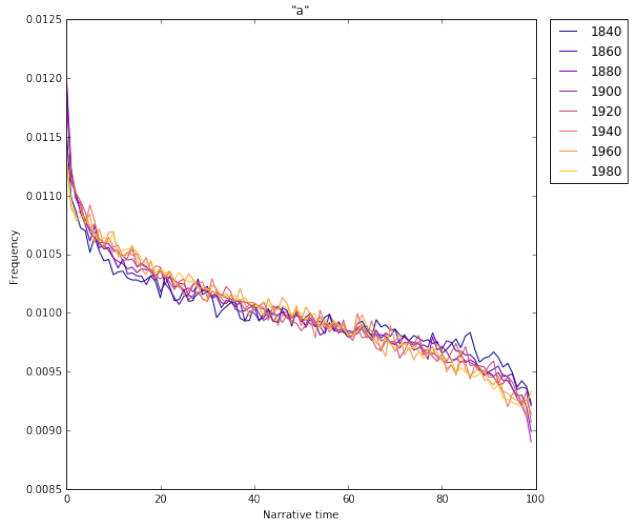
We can start to reason about this in a simple, anecdotal way just by splitting up the texts into a series of sub-corpora, broken out by the year of publication, and then constructing these trends for words across narrative time in each of these time-sliced segments. For example, to pick back up on "a" – if we split the corpus into a set of eight 20-year segments, starting in 1840 and ending in 2000, we can build up eight different time series for "a," one for each of these sub-corpora, and then plot them on top of one another:



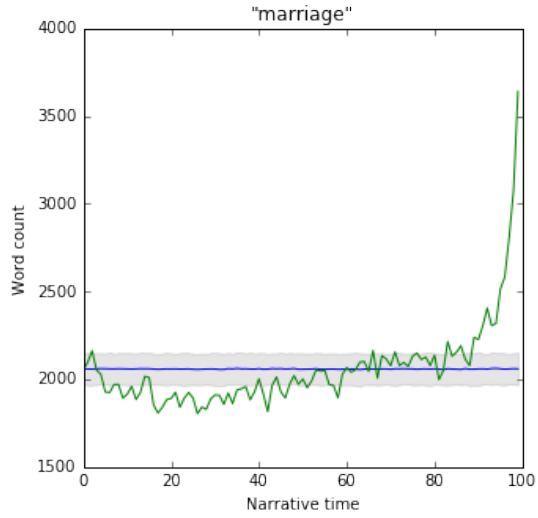
The curves don't line up vertically on the Y-axis because "a" appears a different number of total times in each of the sub-corpora, since each has a different number of novels (eg., we have more from 1900 than 2000) and because the relative frequency of any given word can shift over time – the kind of diachronic changes in overall volume that get surfaced by something like the Google n-gram viewer.

But, in our case, we don't care about changes in total volume – we just care about changes in the trends of the words across novel time, the "shapes" of the distributions along the X-axis. To get a better sense of this, we can flatten out the differences in total volume by relativizing the curves – dividing the count in each percentile by the total number of times that the word appears in the time slice, which produces a density function, meaning that the counts will all add up to exactly 1. This has the effect of moving everything onto the same scale on the Y-axis, which makes it easier to look

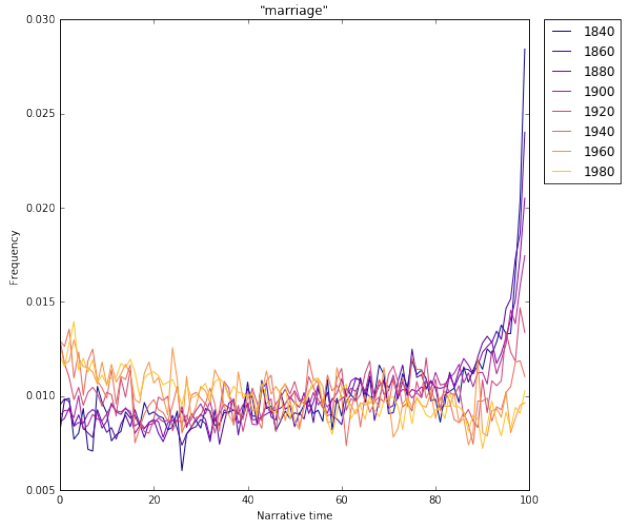
at the changes in the trend. For "a," we can see that the distributions stay more or less the same over time, perhaps becoming a bit "steeper" over the course of the 20th century:



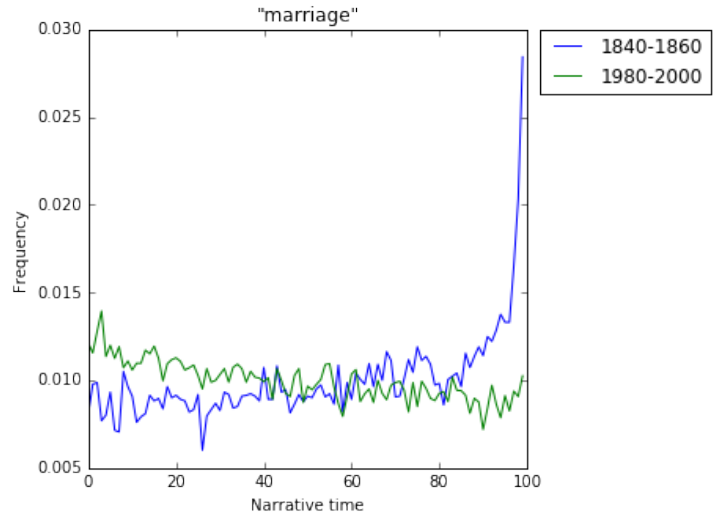
But, in other cases, this very much isn't the case. For example, take a word like "marriage" – here's the distribution across narrative time for all 30,000 novels:



A big spike at the end, and also a smaller peak at the beginning of the text. But, when we split this out into the 20-year slices, we can see that this concentration at the beginning of the narrative is coming almost exclusively from the novels that were published in the second half of the 20th century, and that the spike at the end is coming mainly from novels published in the 19th century:



To cut through some of the noise, let's just look at the earliest and latest novels. Here's 1840-1860 in blue, and 1980-2000 in green:



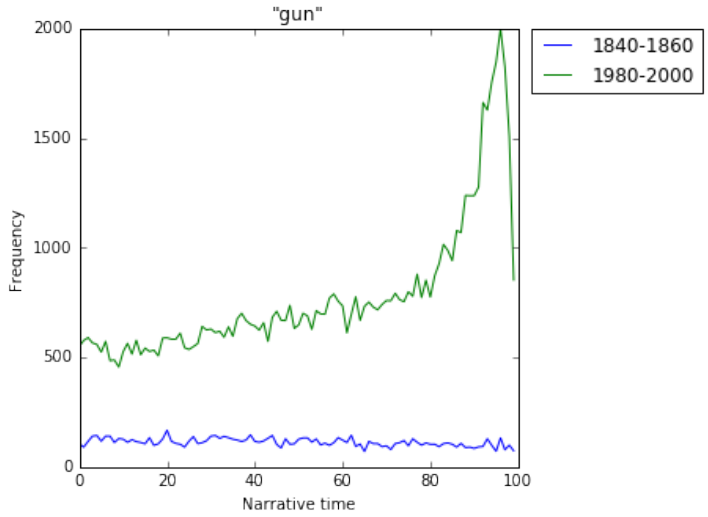
So, far from concentrating in both the beginning and the end, marriage essentially switches narratological sides. In the middle of the 19th century marriage is relatively infrequent at the start of the text, rises slowly across the first 90%, and then spikes enormously in the last 10%, the resolution of the marriage plot. Whereas, in 1980, marriage *begins* narratives more than it ends them – it is relatively frequent at the very beginning and declines slowly and consistently across the text, maybe still ticking up slightly at the end, though nothing like the huge spike in the 19th century.

Marriage *migrates* across the narrative interval – the post-war novel increasingly begins with marriage, the novel becomes preoccupied with marriage itself, and less the courtship that leads up to it. But, how do we reason about the size or significance of this change? In the same way that we needed a way to capture the degree to which a word deviates from the expected (flat) distribution inside any given slice of texts – the score that allowed us to pick out "a" as the most statistically irregular word across the text – now we need a way to inductively identify the words that show the most significant movements over time. Just eyeballing the plots, it looks like "marriage" moves much more significantly than "a," but how do we formalize this? What moves the

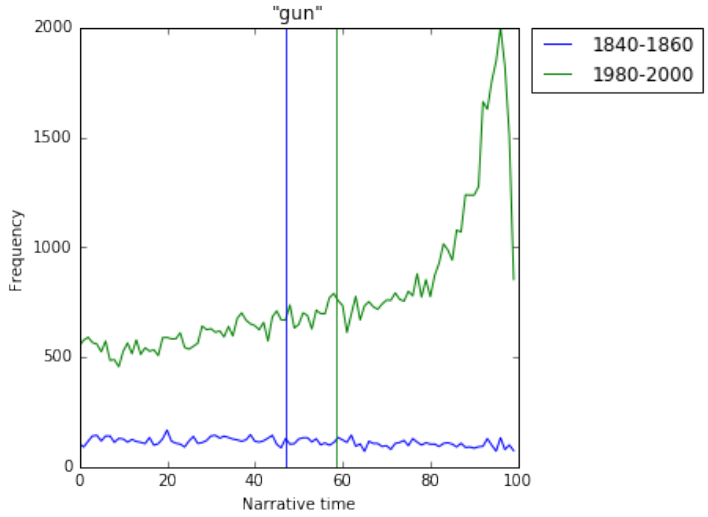
most, what are the most significant narratological shifts in the American novel since 1840?

This turns out to be a somewhat tricky statistical question, mainly because of the large differences that we often see in the overall volume of a word across the different time periods – for example, the corpus contains about twice as many instances of "marriage" in 1980 than in 1840. This means, essentially, that distributions pulled from different time periods will often have very different levels of statistical power, which makes it hard to compare them directly. We need some way to measure the degree to which distributions change over time, but in a way that factors in the different levels of statistical confidence that we have in the data from each of the periods.

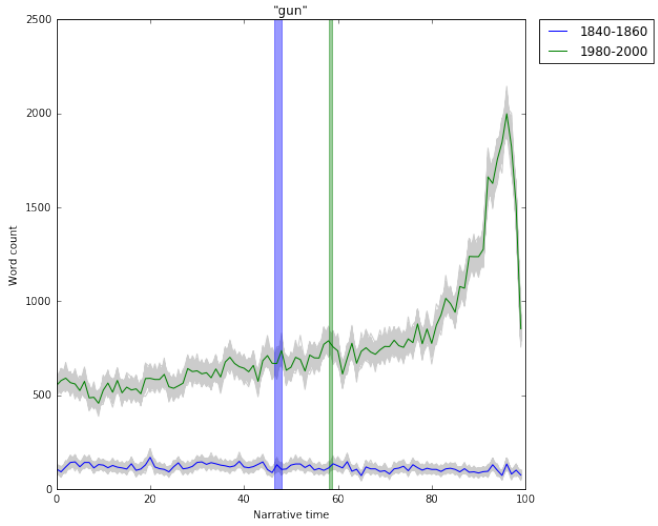
For example, take the word "gun." When we plot out the raw distributions for the earliest and latest novels in the corpus, we can see a really significant change in the trend – at the end of the 20th century, "gun" spikes massively around the 95% marker, whereas in the middle of the 19th century it's more or less level across the text, maybe with a slight upward trend – but also that "gun" has a much higher frequency overall in 1980 than in 1840:



How to compare these? One way to model the overall profile of each of the distributions is to compute its "center of mass," the line that divides the area below each curve into two equally sized regions:



So, in this case, the big spike at the end of the 1980 distribution pulls the center of mass to the right, to just shy of the 60% marker, whereas in the 1840 the center of mass is closer to the center, a bit to the left, since the curve is somewhat higher at the beginning. This gives us a way to capture the extent to which this distributions are different, the degree to which "gun" shifts to the right over the course of the 20th century – we can just take the distance between the two lines on the X-axis, in this case about 10 percentiles. And, to account for the fact that we have much less data for 1840 than 1980, we can again draw 1,000 random samples for each curve, take a new center of mass for each sample, and then use what's called an "effect size" metric to capture the statistical strength of the movement:



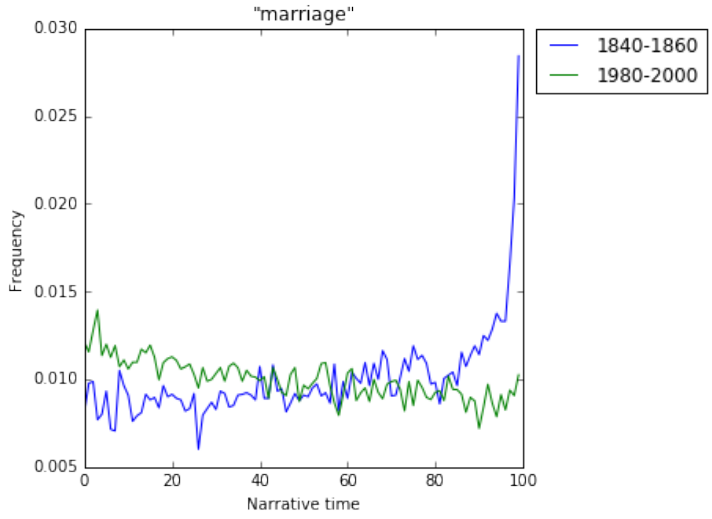
So, for "gun," the effect size between the two distributions is $\tilde{5}4$, whereas for "a" it is $\tilde{3}0$ – "gun" moves farther than "a."

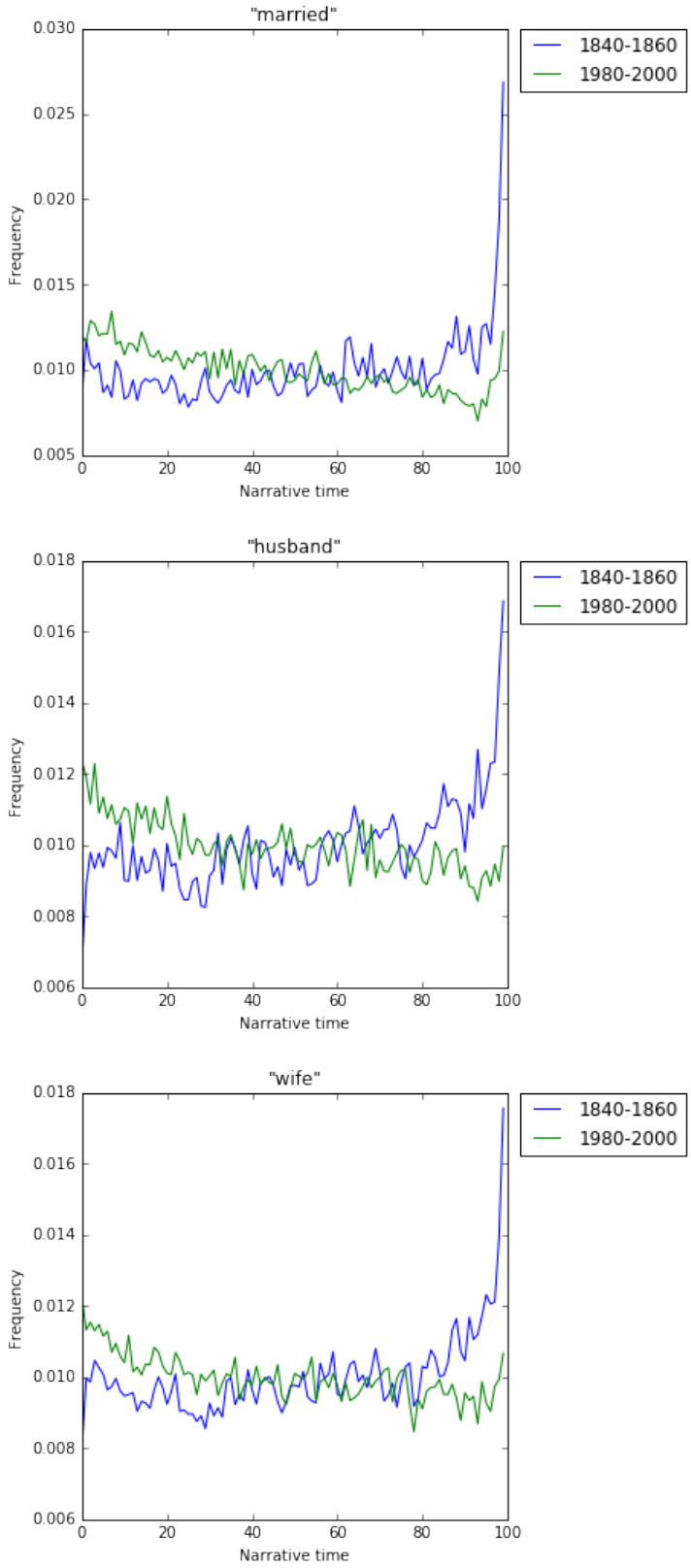
With this in hand, we can now probe through the dictionary again and inductively pick out the words that move most significantly, that are the most *narratologically unstable* over time, and then try to use these words as signposts that point to large-scale changes in narrative structure over the last 150 years. Here are the 500 words with the largest movements:

i, had, you, t, gun, her, of, which, don, said, ll, think, will, father, marriage, we, months, friends,

married, since, it, crime, what, husband, years, day, wife, been, fire, often, name, want, who, get, most, do, down, by, year, woman, us, court, going, during, in, out, about, up, she, victim, a, once, can, for, very, head, public, law, ve, me, their, whom, know, lived, go, looking, won, little, from, prison, his, are, days, that, right, met, an, something, prisoner, city, pale, kill, death, why, try, paper, intelligence, old, week, ago, late, snow, interview, also, private, mother, sure, figure, sister, please, afraid, god, there, letter, if, got, thing, last, whose, every, york, daily, american, wrote, fine, them, side, weeks, see, summer, friend, paul, social, wouldn, priest, just, way, love, yes, news, when, cabin, light, state, didn, than, parties, interest, water, thus, take, letters, wind, new, papers, m, joy, shot, after, rock, feet, past, such, hill, trees, living, alice, former, read, chief, learned, stop, over, month, property, meet, gate, arrival, uncle, guilty, mouth, written, duty, hotel, business, distance, life, slave, length, subject, care, quietly, storm, plan, happy, office, with, girl, here, dollars, looked, health, next, walter, saw, running, kitchen, officer, nose, believe, gentleman, excitement, usual, door, before, made, native, air, known, margaret, named, behind, stairs, result, clear, guess, boston, along, party, united, governor, write, british, s, murder, hold, christian, servant, manner, must, couldn, parents, women, into, strength, sometimes, have, so, history, note, george, family, own, dying, heaven, blessed, killed, firm, coming, then, l, usually, earth, deal, cheek, daughter, rocks, season, brown, road, nice, taken, cried, three, my, brother, early, slowly, long, previous, toward, good, ex, carriage, myself, robert, did, shoulders, winter, open, first, began, offered, face, make, event, required, spring, report, informed, england, jack, america, tired, mr, secret, huge, arrived, case, small, female, town, became, union, english, born, major, art, understand, help, presence, darkness, herself, marry, bar, brief, passed, class, memory, ready, whenever, top, let, dr, loved, around, few, dead, supposed, suppose, put, above, immediately, shouted, jane, wanted, front, chance, cousin, six, received, given, expression, man, the, example, fortune, stood, discovered, anxiety, sweet, edward, pain, not, servants, affairs, sorry, evidence, more, says, fate, money, one, coat, society, hall, short, stay, patient, proved, peter, on, words, home, sky, pardon, grown, john, opening, standing, officers, hard, hour, either, arm, paid, amount, states, henry, boat, blue, stand, off, ran, claim, thinking, sir, like, ground, mark, features, even, call, changed, laughing, happened, meeting, occasion, tell, really, run, entered, forest, harry, angry, keep, fled, washington, bill, end, talk, crowd, post, proper, pretty, died, thought, across, everything, mystery, writing, breast, deck, bore, am, produced, visited, table, therefore, request, noise, quite, impression, rough, walk, cold, moving, shame, give, religious, remained, although, aunt, unhappy, less, holy, covered, holding, several, change, is, never, language, self, forehead, mountain, stopped, well, de, idea, talking, twelve, kate, too, brain, steps, mercy, live, shadows, household, work, boys, suffered, young, enough, knew, asleep, refused, band, respect, below, having, sought, knife, wish, st, away, grass, captain, fact

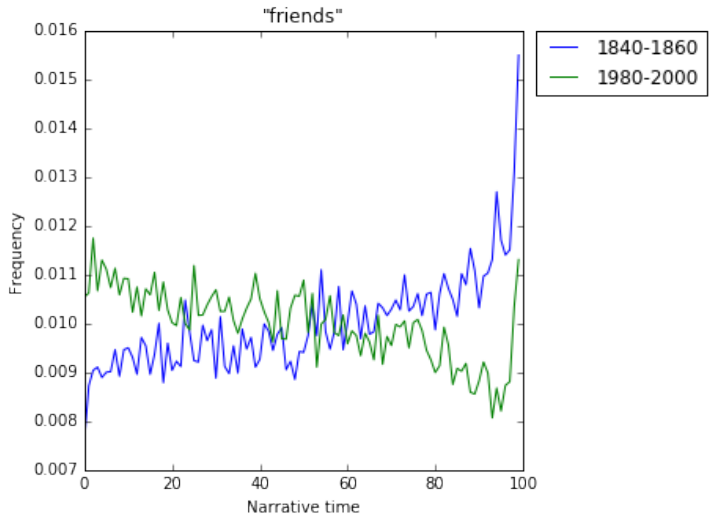
Again, it's far beyond the scope of this essay to look at this comprehensively, but some threads immediately pop out. Marriage comes in as the 11th most "migratory" word, and other related words are nearby:



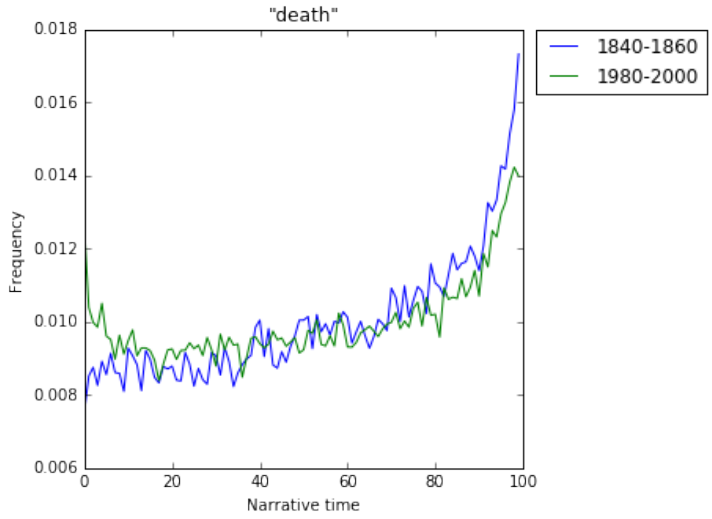


All of these follow the same pattern – in 1840 they spike at the end, the conclusion of the marriage plot; but, by 1980, they have shifted to the left, and concentrate at the beginning. Or, take "friends," another word about the coming together of people, in a sense, and presumably marking the resolution of some kind of narrative tension that

complicates the path to friendship / marriage:



And likewise with "death":

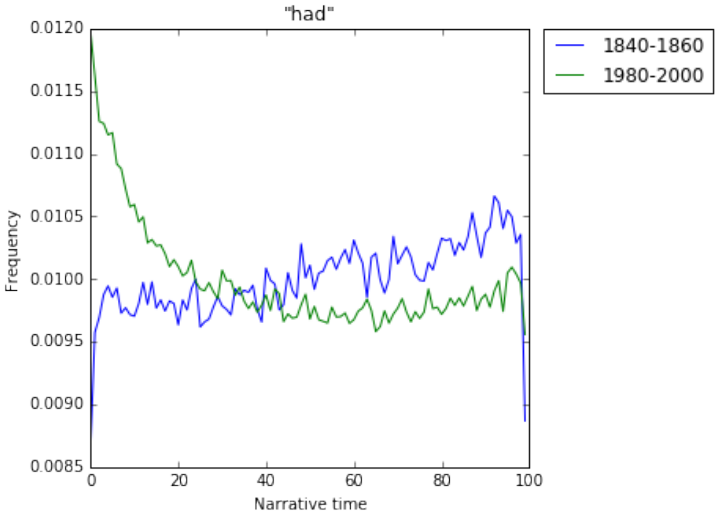


Again, a spike at the end, as we would expect, but, by 1980, we can see that it also develops a pronounced concentration at the very beginning – in addition to ending with death, novels begin with it, death becomes a starting point, an initial event that sets the plot into motion, what Brooks might call the "irritation" into narrative. It provides the energy that opens up the interval of the narrative, not just the ("death"-like, literally in this case) closure that seals off the narrative at the end.

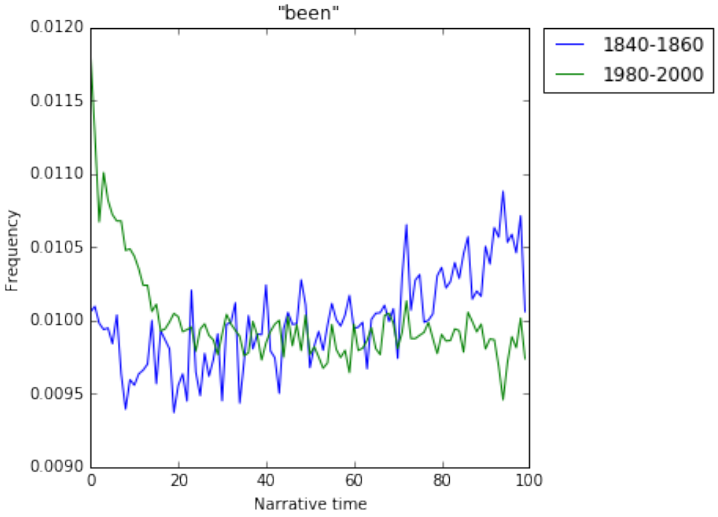
Marriage, friendship, death – all shift from the end towards the beginning, become semantic signals at the start of the text, not just the end. It seems then, in a general sense, as if the post-war American novel is migrating into spaces that, in the past, marked the *boundaries* of *terminuses* of narrative. Narrative ends are repurposed as new types of narrative beginnings. The postwar novel spills over the edges of narratives from the 19th century, it moves into what was previously the post-narrative space – marriage itself, not just courtship; friendship itself, not just the making of friends; the aftermath of death, not just the life that came before it.

4

Indeed, when we look closer, we can make out another dynamic along the same lines that seems to play out at an even lower and more fundamental level – at a register that, once again, appears to sit below any kind of surface-level genre convention. Here is the movement between 1840 and 2000 of the word "had," which has the second most significant movement of any word in the dictionary. In the middle of the 19th century, "had" was more or less flat across the text, rising slowly to a peak near the end. By the end of the 20th century, though, it has shifted enormously towards the beginning of the text – it is highest at the very beginning, falling off quickly over the first half of the narrative:

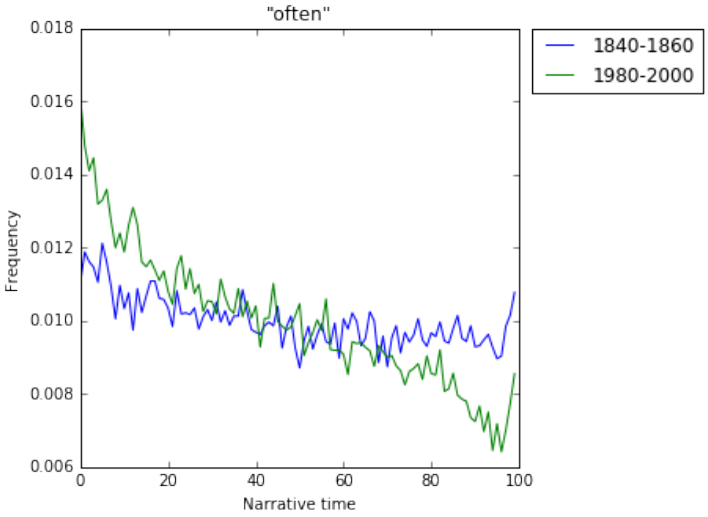


Now, at a grammatical level, "had" is two things – it is the past tense of "to have," but, more frequently, it forms the past perfect tense, which we can think of as the past tense within the past tense, used to describe things that, at some point in the past, *had* already been finished, completed, perfected – "had lived," "had worked," "had thought," etc. Or, as part of the past perfect progressive, with "been," to mark actions that had been taking place at some point in the past – "had been living in ...," "had been working at ...," "had been thinking about ...," and so on and so forth. Indeed, with "been," we see the same movement:

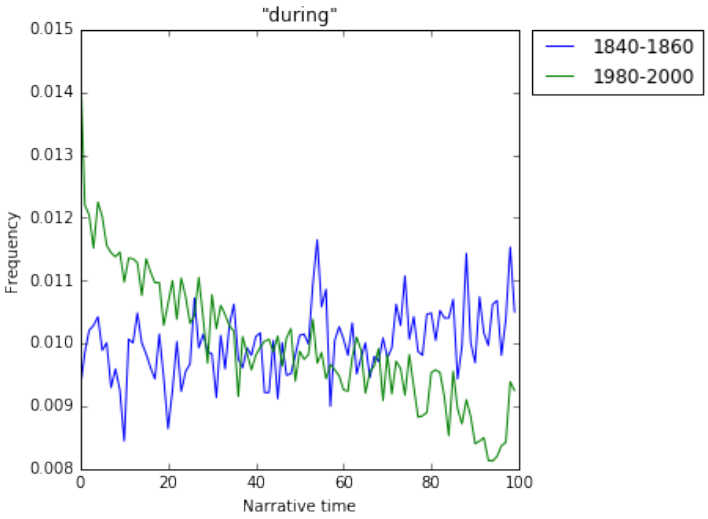


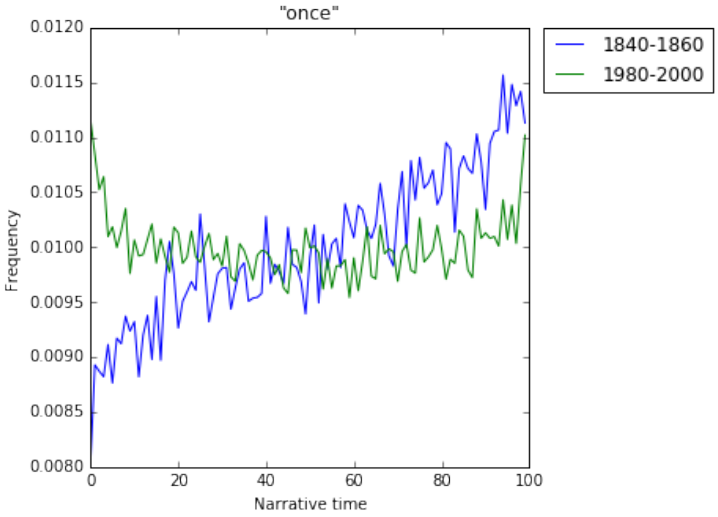
Novels, in other words, increasingly begin in the past perfect.[11] At the beginning of the text, the 20th century narrative is reaching backwards into the pre-narrative space – instead of charging directly forward out of the gates into the narrative interval of the story, the syuzhet begins by first swiveling around and looking backwards into what came before the fabula. The beginning, in a sense, becomes concerned with what came before the beginning, the pre-beginning, the beginning of the beginning.

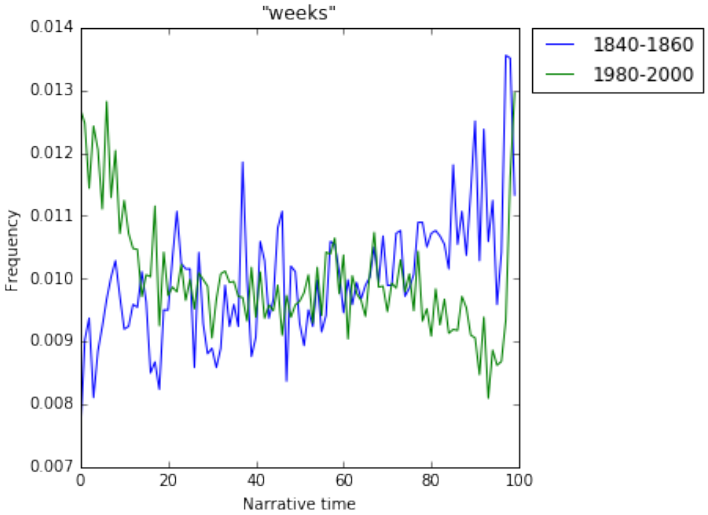
Likewise, we can see a cohort of words related to temporality and duration that show the same pattern – that move leftwards over the course of the 20th century, that migrate to the beginning:

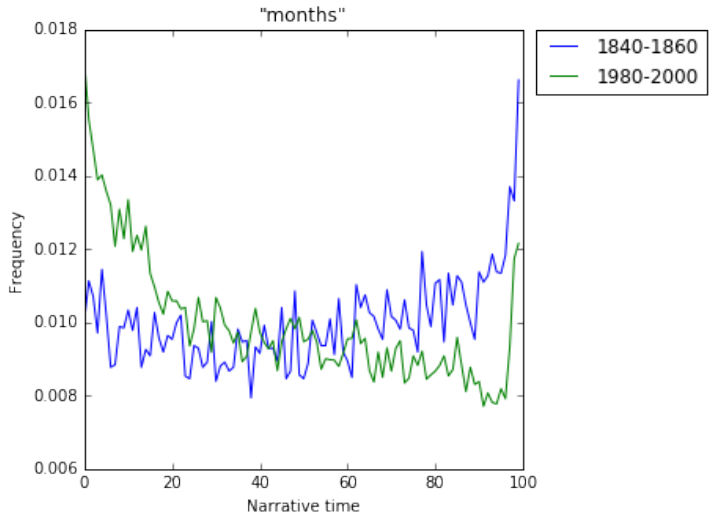


[11] Part-of-speech taggers have a hard time distinguishing between the two functions of "had" – as a regular past tense verb ("I had five dollars") and as an auxiliary verb in a past perfect ("I had lived in New York"). To confirm that this trend is being driven by the past perfect and not the regular past tense, I hand-coded a random sample of 1,000 occurrences of "had" drawn from the first 1% of narrative time in novels between 1980 and 2000. Of that sample, 816 (81.6%) were past perfects and 184 (18.4%) were past tenses, which suggests that the past perfect is the main force behind this. Going forward, it should be possible to write a classifier to do this automatically.


"during"
1840-1860
1980-2000
Frequency
Narrative time
0.015
0.014
0.013
0.012
0.011
0.010
0.009
0.008
0
20
40
60
80
100

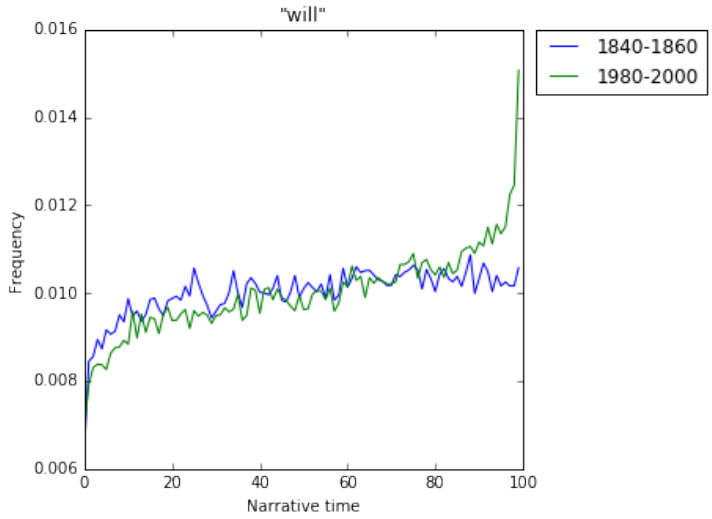

"once"
1840-1860
1980-2000
Frequency
Narrative time
0.0120
0.0115
0.0110
0.0105
0.0100
0.0095
0.0090
0.0085
0.0080
0
20
40
60
80
100


"weeks"
1840-1860
1980-2000
Frequency
Narrative time
0.014
0.013
0.012
0.011
0.010
0.009
0.008
0.007
0
20
40
60
80
100



The beginning is looking backwards into the "weeks" and "months" that preceded it. In the opening shot the temporal "camera" of the narrative is facing backwards, away from the contents of the interval that actually "belongs" to the present narrative, that is enclosed by the text. The beginning is in the past tense relative to itself, situating itself in the context of what came before.

And, at the end of the text? The opposite happens. At the end of the 20th century, "will" – the future tense – spikes massively at the very *end* of the text:[12]



If the beginning is increasingly looking back into the pre-narrative space, then ends are looking forward into the post-narrative space, towards what will come after the terminus of the narrative. The ending becomes less of an ending and more of a threshold of transition, a second beginning, the narrator and characters gazing past the present narrative and into the future. In the post-war American novel, then – *the beginning is in the past perfect tense, and the ending is in the future tense.*

[12]Unlike with "had," the part-of-speech tagger is able to distinguish "will" as a modal verb from the various other meanings of "will" as a noun, making it easy to isolate the future tense.

5

So – over the course of the 20th century, and especially since 1945, the edges of the American novel become permeable, porous, leaky. The beginning starts to look back into the time before the beginning, and the end looks forward to what will come next. It is as if the narrative becomes a *sub*-narrative, embedded inside of an implied larger narrative that surrounds it on either side. What to make of this? Lots could be said here, and it's beyond the scope of this essay to provide a definitive answer. But, briefly – maybe the most striking thing about this is the extent to which is *doesn't* correspond to some basic assumptions about the structure and purpose of narrative that were developed over roughly the same period of time during which we see this empirical change in the structure of the novel. Namely, the idea developed by theorists like Peter Brooks, Frank Kermode, Walter Benjamin, Jean Paul Sartre, and Henry James that narratives exist, in no small part, to provide a sense of *enclosure* – and specifically that narratives depend on an ending that is similar to a "death," a moment of total closure that converts the text into something unified and meaningful. Beginnings and ends are useful, in this view, precisely to the extent that they aren't permeable, aren't porous, aren't leaky.

The most concise statement of this is probably Peter Brooks' 1977 essay *Freud's Masterplot*,[13] in which Brooks takes up Freud's account of the human lifecycle in *Beyond the Pleasure Principle*[14] and repurposes it as a metaphor for narrative, which becomes a fractal of a human life. The beginning is a structural birth:

> For plot starts (must give the illusion of starting) from that moment at which story, or "life," is stimulated from quiescence into a state of narratability, into a tension, a kind of irritation, which demands narration. Any reflection on novelistic beginnings shows the beginning as an awakening, an arousal, the birth of an appetency, ambition, desire or intention. (291)

After which:

> The ensuing narrative – the Aristotelean "middle" – is maintained in a state of tension, as a prolonged deviance from the quiescence of the "normal" – which is to say, the unnarratable – until it reaches the terminal quiescence of the end. (291)

The narrative interval, then, is surrounded on both ends by the "unnarratable," the literary equivalent of the nothingness before birth and after death. And, meanwhile, the death-like ending provides the full stop, a moment of total closure that makes it possible comprehend the text, in the same way that a life only becomes "transmissible" at the moment of death:

> All narration is obituary in that life acquires definable meaning only at, and through, death. In an independent but convergent argument, Walter Benjamin has claimed that life assumes transmissible form only at the

[13] Brooks, Peter. "Freud's Masterplot." Yale French Studies, no. 55/56, 1977, pp. 280–300. www.jstor.org/stable/2930440.

[14] Freud, Sigmund. Beyond the Pleasure Principle; Trans. by C. J. M. Hubback. London, Vienna: International Psycho-Analytical, 1922

> moment of death. For Benjamin, this death is the very "authority" of narrative: we seek in fictions the knowledge of death, which in our own lives is denied to us. (284)

The curtain drops, the narrative dies out, and it is only in this moment of total closure that what Kermode calls the "successive" structure of the narrative congeals into a sense of unity and meaning – the ending seals off the text, closes it out, converts the metonymic chaos of the chronological experience into metaphorical unity, provides a structural enclosure for meaning that is denied to the reader in real life. Which, unlike the novel, is always experienced en route, partially, in the middle of the successive piling-up of circumstance and experience. In *The Sense of an Ending*,[15] Kermode calls this the experience of life "in the middest":

> Men, like poets, rush 'into the middest', in media res, when they are born; they also die in medis rebus, and to make sense of their span they need fictive concords with origins and ends, such as give meaning to lives and to poems. They fear is, and as far as we can see have always done so; the End is a figure for their own deaths. [...] In the middest, we look for a fullness of time, for beginning, middle, and end in concord. (7)

Narratives are expressions of this desire for a "fullness of time" – beginnings and ends are *useful* to us, Brooks and Kermode think, to the extent that they enforce this sense of demarcation, the degree to which they can measure off a discrete piece of narrative cloth and cut it cleanly at both sides. More than just a characteristic of narrative, this is presented as the fundamental competency of narrative, nothing less than the reason that we write and read. As James puts it famously, in the preface to *Roderick Hudson*:

> Really, universally, relations stop nowhere, and the exquisite problem of the artist is eternally but to draw, by a geometry of his own, the circle within which they shall happily *appear* to do so.

And yet, just as this understanding of beginnings and ends is developing during the beginning and middle of the 20th century, it seems as if the novel is *losing* precisely this sense of enclosure – the circle is breaking, the beginning and end cease to be rigid boundaries and turn into something much more fluid and partial. They become thin lines in the sand, bent corners on the pages of a larger narrative that implicitly surrounds the text, pointed to by "had" at the beginning and "will" at the end. It is as if the novel becomes more like a vignette, an episode, a novella, a short story, a fragment, a mimetic representation of "life in the middest" and not an escape from it.

This itself, though, is again more of a critical beginning than an end. So the edges of the novel become blurry, the membrane becomes permeable, the capsule begins to leak – and the novel, perhaps, ceases to do the psychological work that Brooks, Kermode, Sartre, and Benjamin imagined, if it ever did. But – why? We can imagine a range of answers to this, ranging from pragmatic historicisms to much more general and speculative comments about the cultural milieu of post-war novel, though, to support any of these, we would need to do a great deal more empirical and critical work. Perhaps – the "program era," as documented by McGurl,[16] led to a rise in the

[15] Kermode, Frank. The Sense of an Ending: Studies in the Theory of Fiction. Oxford: Oxford UP, 1966. Print.

[16] Mark McGurl. The Program Era: Postwar Fiction and the Rise of Creative Writing. Cambridge: Harvard UP, 2009.

intellectual currency of the short story, easier to workshop in the context of the MFA, and, as writers trained on this format, the novel starts to echo the vignette quality of shorter fiction? Or, in a more general sense – perhaps we're simply seeing the novel become involved in a kind of feedback loop with the various flavors of post-structuralism that flourished in the second half of the 20th century, which, to paint a broad stroke, became distrustful of the idea that a narrative might be able to provide a "fiction of concord" or a "unified meaning," or that such a thing would even be desirable or productive.[17]

At a more meta-critical level, though, it's worth asking – what exactly *should* an answer to this question look like? To take up a question that the Literary Lab is posing in a conference in the spring – what exactly constitutes a "result" in these types of projects? At a very pragmatic level, how do we know when we've reached a stopping point? Especially with this type of "three dimensional" data that plays out across both historical time and narrative time – where should the critical process finally "land," at the end of the day? It could go in two directions – we could use the structural changes in the novel as a point of entry to say something about the history of the 20th century. Or, we could think of the historical axis as simply a mechanism for better understanding narrative itself – change over historical time provides a kind of stereoscopic vision onto narrative, a continuously running set of experimental conditions and permutations that provide a triangulation onto the range of forms that the narrative is capable of exhibiting. Do narratological changes tell us something about history, or does historical change tell us something about narrative? Or is this a false choice – is it possible to say something about both at once?

[17] Or, to throw caution entirely to the wind – if we indulge Kermode's argument that the beginnings and ends are essentially modern stand-ins for the structure and meaning that was provided historically by prophesies of the apocalypse – the religious expectation of a literal end of history – then perhaps we can wonder if the approach of the millennium in the second half of the 20th century provided a real and external sense of closure that, for a moment, made it less important for fictional narratives to provide a substitute? If we look just at novels published since 2000, after the millennium – would we see a reversal of this trend, a swing back in the direction of narrative enclosure – the past perfect draining out of the beginning, "will" sloshing back away from the ending, a re-sealing of the text?