What Makes Hemingway Hemingway?

A statistical analysis of the data behind Hemingway's style

by Justin Rice, published on 12/13/2016

Introduction

In 1954, Ernest Hemingway won the Nobel Prize in Literature. According to nobelprize.org, "The prize was awarded for his mastery of the art of narrative... and for the influence that he has exerted on contemporary style."

If you're reading this, chances are you're pretty familiar with Hemingway. You probably have a sense of his style. You may have read authors who themselves read Hemingway, and seen in them the strength of his influence. When you look at the quote above, you may think: "Passive voice. Not very Hemingwayesque."

Whatever you know of Hemingway's writing, though, is limited by the fact that you're only human: you can only read so fast; you can only keep track of so many words at a time. Your experience with Hemingway is qualitative, as is your experience with anything you read in a traditional, linear way.

What if, however, you supplement your reading with some computational heft? Instead of treating words as a linear progression, what if you think of them as atoms you can re-arrange and re-examine under different lenses looking for interesting patterns? **Can you start to quantify Hemingway's style and influence?**

Our goal here is to do just that. We'll take Hemingway's prose and treat it as data. We'll tally his words, calculate his choices, and try to come up with a statistical understanding of what makes Hemingway Hemingway.

Hemingway's Style

I. Sentence Length

> "Hemingway evolved his style in the herd school of journalistic reporting. In the editorial office of the Kansas City newspaper where he served his apprenticeship, there was a kind of pressman's catechism, the first dictum of which was: 'Use short sentences.'" — Anders Österling, Nobel Prize award speech, 1954

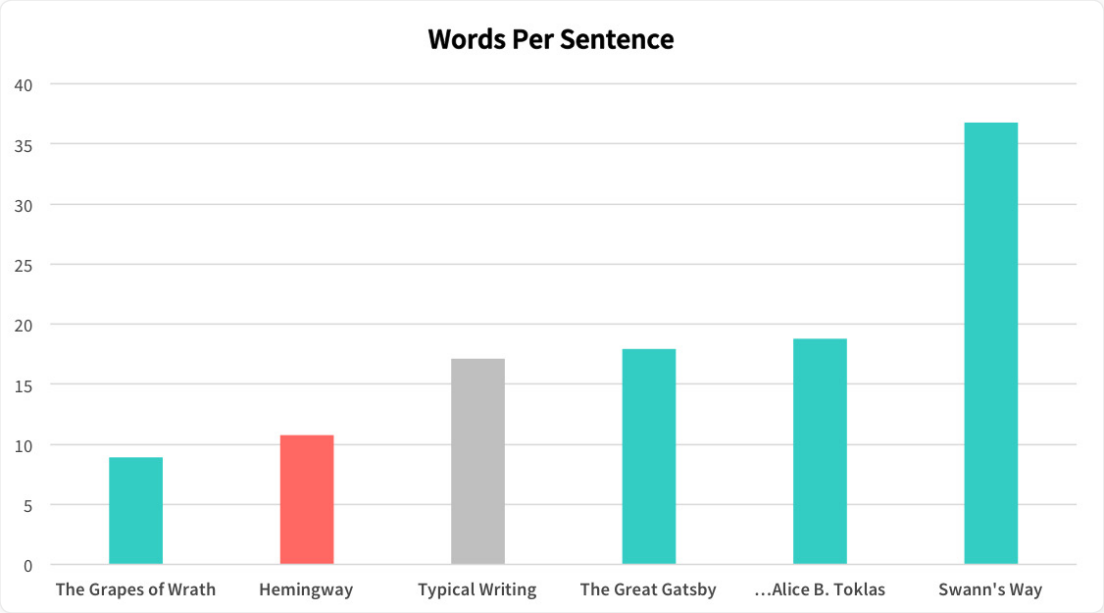
Is it true that Hemingway's sentences are especially short? Let's see what happens when we compare Hemingway's writing to typical writing, and to some of his contemporaries' most widely read novels (John Steinbeck's *The Grapes of Wrath*, F. Scott Fitzgerald's *The Great Gatsby*, Marcel Proust's *Swann's Way*, and Gertrude Stein's *The Autobiography of Alice B. Toklas*):

About Analitics

Analitics is a series of original articles from LitCharts that use math and data to analyze and illuminate literature.

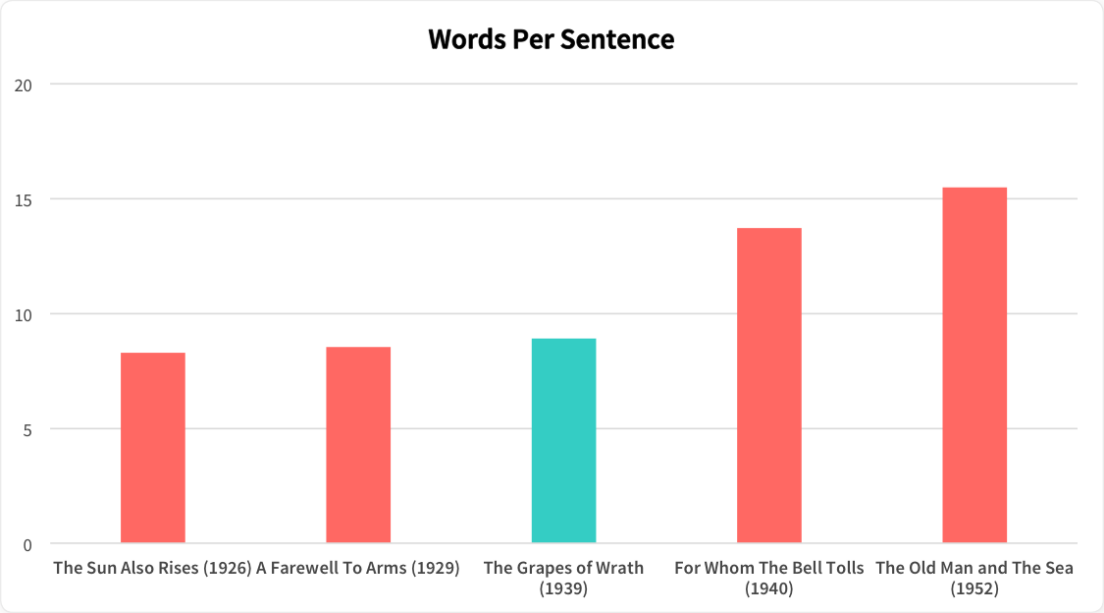
About LitCharts

LitCharts are the world's best literature guides. Over 1.5 million students and teachers read LitCharts every month.



Hemingway's sentences clock in about 7 words shorter than average, so yes: his sentences are short. Proust's sentences, meanwhile, are really, really long.

Surprisingly, the average sentence in *The Grapes of Wrath* is shorter than the average sentence in Hemingway's writing. This made us curious, so we decided to dig a little bit deeper and see what happened if we focused on each of Hemingway's books individually. Take a look:

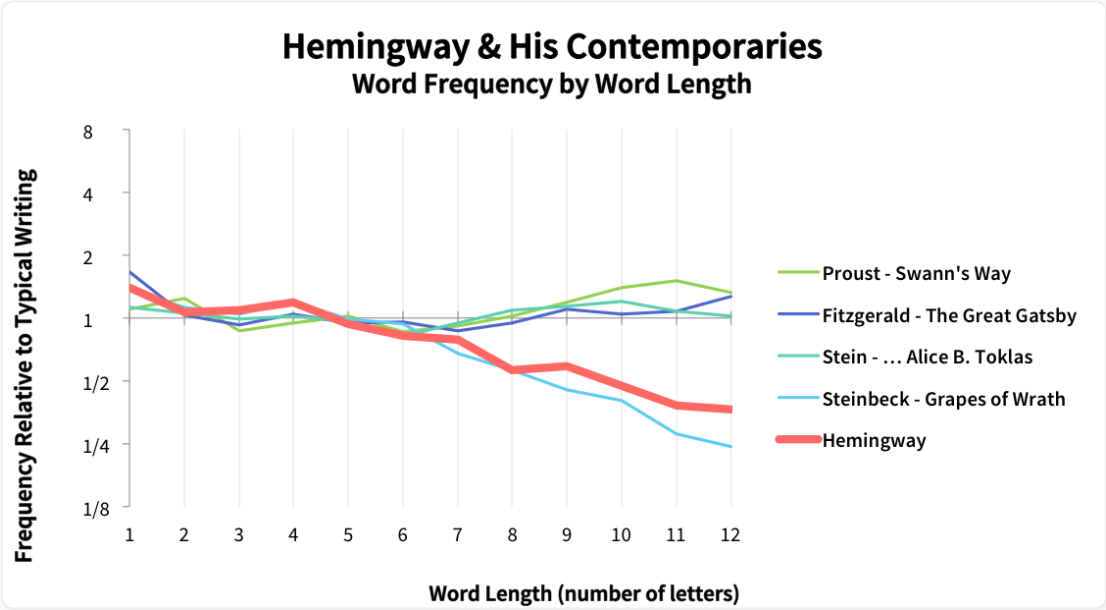


In his early novels, Hemingway out-shorts Steinbeck. As Hemingway gets older, however, his sentences get longer. **So while short sentences are characteristic of Hemingway, they define his work less and less as his career progresses.**

II. Word Length

> "Poor Faulkner. Does he really think big emotions come from big words? He thinks I don't know the ten-dollar words. I know them all right. But there are older and simpler and better words, and those are the ones I use." — Hemingway quoted in *Papa Hemingway: A Personal Memoir* by A. E. Hotchner, 1966

Let's investigate Hemingway's claim to "older and simpler and better words." Did Hemingway favor non-big (aka short) words? Here's what it looks like when we plot word-length frequency of Hemingway and his contemporaries compared with typical or "average" writing:

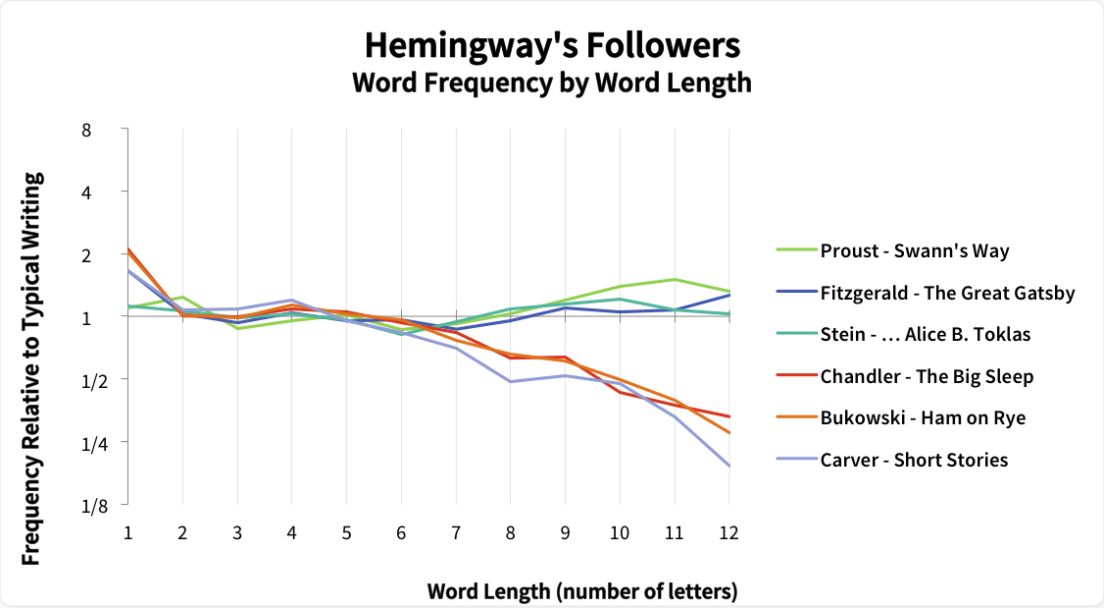


As you can see, several texts have an exceptional number of 1-letter words. Any guess as to why? (Hint: they've got a lot of "I.")

After the 1-word bump, things cluster pretty tightly for 2- through 6-letter words. At 7-letter words, however, frequencies in Hemingway and Steinbeck plummet, while frequencies in Proust, Fitzgerald, and Stein ascend. 7-letter-and-up words: those must be the "ten-dollar words" Hemingway mentioned. He eschews them. Steinbeck, you will notice, eschews them even more.

There's Steinbeck again, more Hemingway than Hemingway. It's worth noting that while he and Hemingway were contemporaries, Hemingway started publishing ten years earlier. Steinbeck read Hemingway, and in the manuscript of *East of Eden* acknowledged that Hemingway "was imitated almost slavishly by every young writer, including me." Are these Steinbeck numbers evidence of Hemingway's influence?

If so, we'd hope to see that influence elsewhere, so let's look at other authors who are self-proclaimed Hemingway admirers. Do we see a tendency to avoid ten-dollar words? If we sub in texts from three Hemingwayesque writers (*The Big Sleep* by Raymond Chandler, *Ham On Rye* by Charles Bukowski, and a collection of Raymond Carver's short stories) here's what we get:



III. Lexical Richness

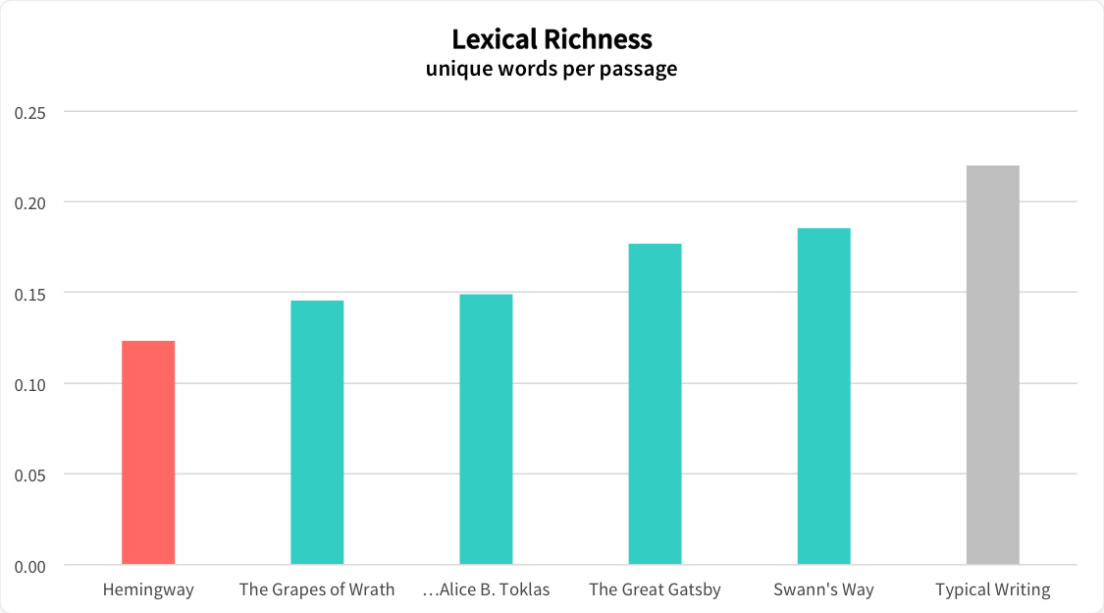
In addition to thinking about the length of Hemingway's words, we can also think about how many different words he uses. Does he use the same words over and over, or does he utilize synonyms to avoid repetition?

Let's start by comparing raw vocabulary size:

Title	Total Words	Unique Words
The Old Man and the Sea	25747	2402
The Great Gatsby	44436	5337
The Sun Also Rises	66846	4548
A Farewell to Arms	88371	5142
…Alice B Toklas	91669	6395
For Whom the Bell Tolls	162815	7894
The Grapes of Wrath	175477	8330
Swann's Way	193468	12154
Typical Writing	981716	40234

As you can see, there's a strong correlation between total words and unique words. That makes sense: a 5-word sentence is going to have fewer unique words than a 1,000-page book.

What we're interested in isn't actually raw vocabulary size: it's the portion of unique words in a given passage, which is a measure called **lexical richness**. Higher lexical richness means less repetition. (This sentence, for instance, has a lexical richness of 1.00 because no word is repeated.) Lower lexical richness means more repetition. How does Hemingway's lexical richness compare?



It's low. His word choice is repetitive. He not only uses shorter words and shorter sentences, he also chooses to use the same words over and over.

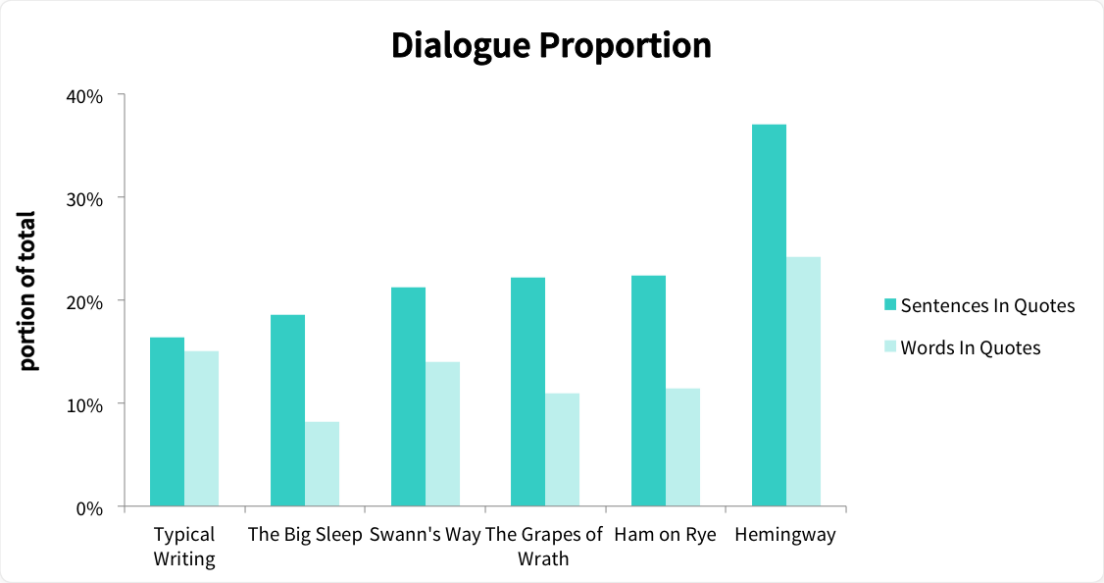
You may notice that our average, or "Typical Writing"— a corpus of 500 texts by 500 different authors—has a higher lexical richness than any individual author. Why is that? While every author uses a limited set of unique words, each author's set is slightly different. A given set is like a fingerprint: not only does it include its own characters, dialect, and special vocabulary, but it reflects a pattern of choices characteristic of the author. When we look at an individual author, we're looking at one fingerprint. When we look at our average, we're looking at 500 fingerprints overlapping.

This argument is part of what led the editors of *The New Oxford Shakespeare* to list Christopher Marlowe as co-author of three Shakespeare plays, and we'll return to it when we look at characteristic words.

IV. Amount of Dialogue

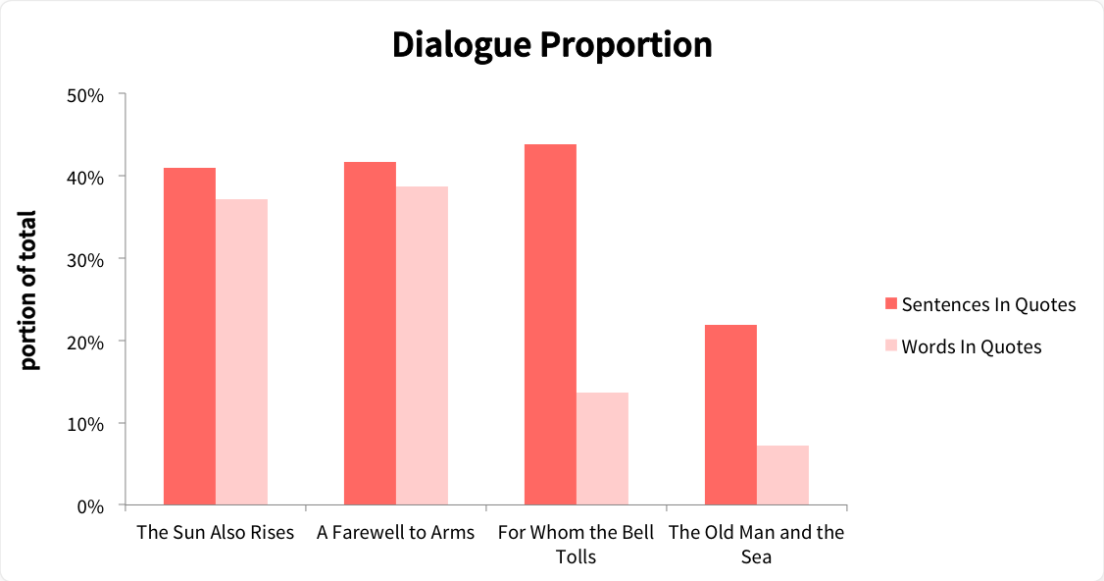
> "Hemingway's significance as one of this epoch's great moulders of style is apparent...chiefly in the vivid dialogue and the verbal thrust and parry, in which he has set a standard as easy to imitate as it is difficult to attain." — Anders Österling, Nobel Prize award speech, 1954

When we look at the amount of dialogue in Hemingway, here's what we find:



Not only does he use twice as much dialogue as an average writer, but he uses far more than any of the Hemingwayesque writers we've considered. Including Steinbeck. So while short sentences and short words define Hemingway's style, what really sets him apart from his admirers is his decision to let his characters speak.

Hemingway's tendency to avoid long words is consistent across his books, but like sentence length, his use of dialogue also changes. Take a look:



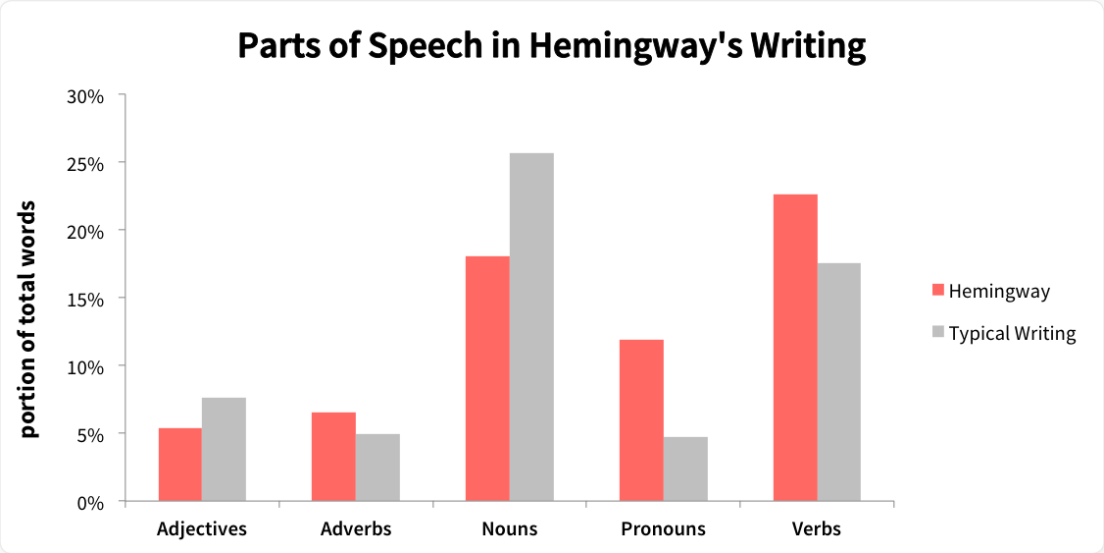
The characters in *The Sun Also Rises* and *A Farewell to Arms* talk about the same amount and say about the same amount. By the time we get to *For Whom the Bell Tolls*, the characters talk more but say less. Finally, in *The Old Man and the Sea*, the characters barely talk at all.

Hemingway's Word Choices

1. Parts of Speech

Now that we have a sense of what defines Hemingway's style—**short sentences, short words, lots of dialogue, lots of repetition**—let's see if we can hone in on the words themselves. Is there a vocabulary characteristic of Hemingway? What patterns can we find in his choice of words?

To start with, we'll break his words into smaller categories. If we tag each word by part of speech, here's what we find:



The biggest difference is that Hemingway uses fewer nouns and more pronouns than average. What does this choice suggest? Think about the number of characters in Hemingway's books, and the amount of ink given to each one. When it comes to subjects, does he favor breadth or depth? How does his use of pronouns factor into that choice?

Next, note that **Hemingway uses fewer adjectives and more verbs than average**. Those numbers make sense given what we observed earlier: adjectives complicate sentences and make them longer; verbs make things happen, and every sentence needs one. Fewer adjectives means less description. More verbs mean more action. The parts of speech we've looked at so far are the perfect ingredients for short sentences and simple words.

But what about the adverbs? Based on the chart above, it looks like Hemingway uses more than average. Wouldn't that suggest more complicated sentences? What's going on there?

To answer that question, let's look at the 20 most frequent adverbs in Hemingway. The words on this list alone constitute 70% of the total adverbs in Hemingway's writing:

Word	Frequency
up	0.0813...
out	0.0695
then	0.0670
now	0.0632
down	0.0548

when	0.0481
back	0.0431
where	0.0348
how	0.0271
here	0.0252
again	0.0239
never	0.0234
just	0.0228
only	0.0220
well	0.0211
off	0.0189
away	0.0188
yes	0.0186
still	0.0157
always	0.0153

Adverbs modify other words by specifying time, place, frequency, or manner. The list above includes time adverbs ("then", "now"), place adverbs ("up", "out"), and frequency adverbs ("again", "never"), but it doesn't include any manner adverbs. Manner adverbs tend to end in "ly," and when we think of adverbs, manner adverbs are usually, reasonably, or perhaps presumptively what come to mind.

When we count up words that end in "ly," we find that Hemingway actually uses manner adverbs much, much less than the average writer (42% as often).

II. Characteristic Words

> "I tried to make a real old man, a real boy, a real sea and a real fish and real sharks. But if I made them good and true enough they would mean many things. The hardest thing is to make something really true and sometimes truer than true." — Hemingway quoted in *Time* magazine, 1954

When you look at the adverbs listed above, you may notice that they're not very distinctive. They'd probably appear frequently in almost any piece of writing. To find words characteristic of Hemingway, we can't just look at words Hemingway uses most: we need to look at words Hemingway uses more than the average writer. **Statisically speaking, here are the most Hemingwayesque verbs, adjectives, and nouns:**

Hemingway's defining...		
Verbs	**Adjectives**	**Nouns**
furled	rotten	absinthe
motioned	disgraceful	vermouth
flatter	groggy	sacks
lashing	khaki	fiesta
baited	shiny	helmet
slung	oblong	shirts
galloping	sallow	armoire

rowed	uphill	gorge
loosen	womanly	ambulances
circling	sleepy	bombardment
sweating	shady	sniper
joked	unfaithful	plateau
stroked	jealous	concierge
dipping	acrid	bulls
misunderstood	ruddy	tiredness
commenced	bloody	bait
punched	upturned	candlelight
dipped	taut	mules
tipped	gloomy	arcade
oiled	repugnant	capes

It's an evocative list, and the words on it certainly feel Hemingwayesque. What else does it tell us? While there's too much to go into here, we'll make a few observations and think about the questions they raise.

- Hemingway's verbs ("punched", "stroked," "galloped" and so on) are visceral and active. What does that suggest about his characters's tendency to reflect vs. their tendency to act? About masculinity in his writing?
- Most of Hemingway's adjectives are pessimistic ("rotten," "disgraceful," "unfaithful," "jealous," "gloomy", "repugnant"). What does that say about his characters' worldview? About the notion of "the lost generation"?
- Hemingway's nouns focus on drinking, war, bullfighting, and travel. How do those subjects define his characters's day-to-day lives? How do they relate to each other?

We looked at Hemingway's most frequent adverbs earlier, but we've saved his characteristic adverbs for last. That's because, in addition to Hemingwayesque words, **we can also look at words uncharacteristic of Hemingway**—words the average writer uses a lot, but Hemingway uses very little, if at all—and the difference between the two in the adverb category is striking:

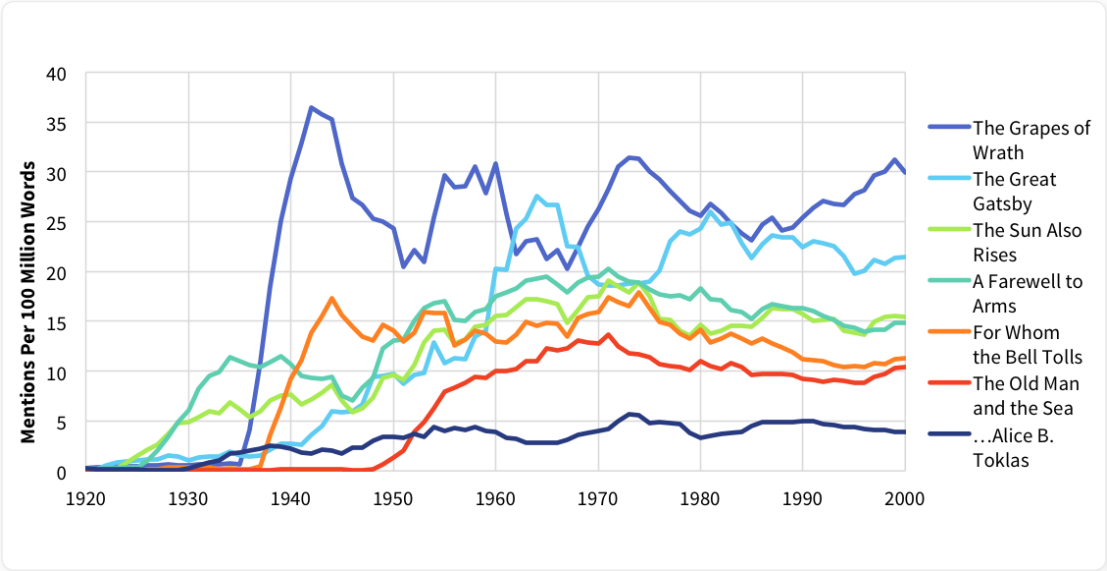
Hemingway Adverbs	UnHemingway Adverbs
steeply	generally
apiece	immediately
sideways	daily
delicately	apparently
frightfully	moreover
lovingly	approximately
mockingly	primarily
cleanly	largely
lazily	abroad
sarcastically	precisely
imperceptibly	prior

huskily	elsewhere
dreadfully	virtually
admiringly	presumably
contemptuously	specifically
authoritatively	briefly
insolently	inevitably
skillfully	regardless
arrogantly	recently
smoothly	partly

Most of Hemingway's adverbs make action more specific ("steeply," "delicately," "mockingly"), while most of the adverbs he avoids hedge certainty ("generally," "apparently," "approximately"). How does the choice to avoid "hedging" adverbs relate to Hemingway's stated goal: to make things "truer than true"?

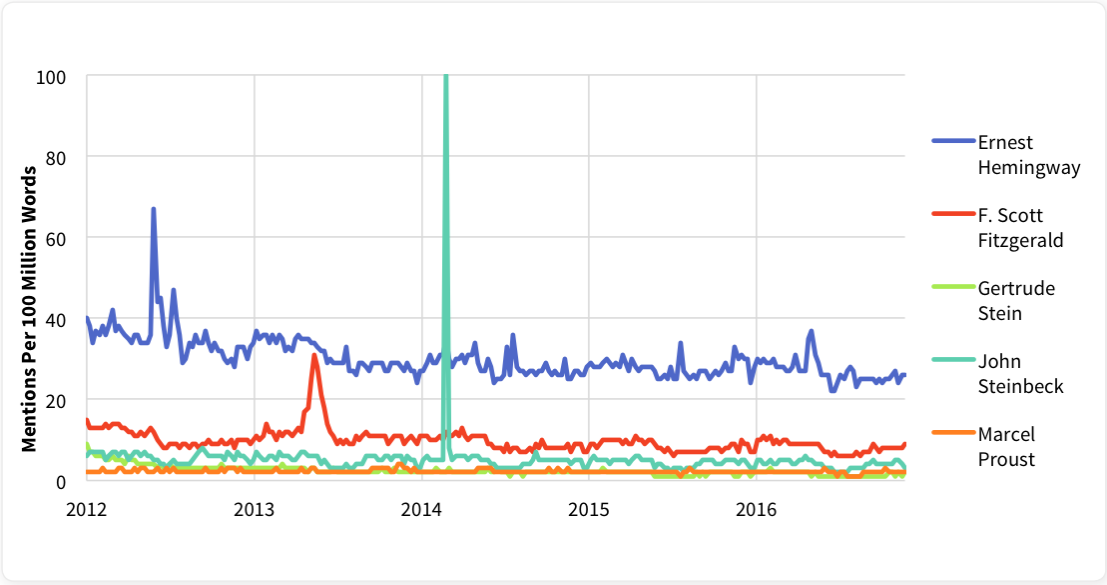
Hemingway's Influence

We've analyzed part of what makes Hemingway's style remarkable. For our final section, let's look at his legacy, and at how that style has endured over the years. One way to chart influence is by comparing references to his work via Google Books:



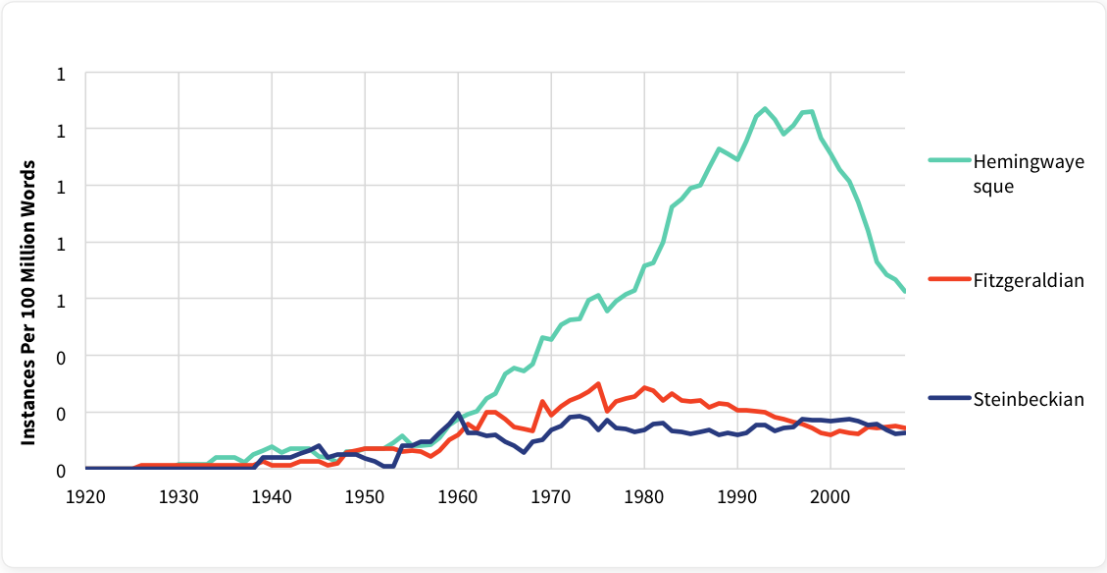
Both *The Grapes of Wrath* and *The Great Gatsby* chart higher than any and all of Hemingway's novels. There's a case that those are more enduring than anything Hemingway wrote.

And yet—barring a brief spike in interest in Steinbeck surrounding the release of the 75th-anniversary edition of *The Grapes of Wrath*—Hemingway himself is more popular than Fitzgerald, Proust, Steinbeck, and Stein, as we can see through Google Trends, which plots search popularity:



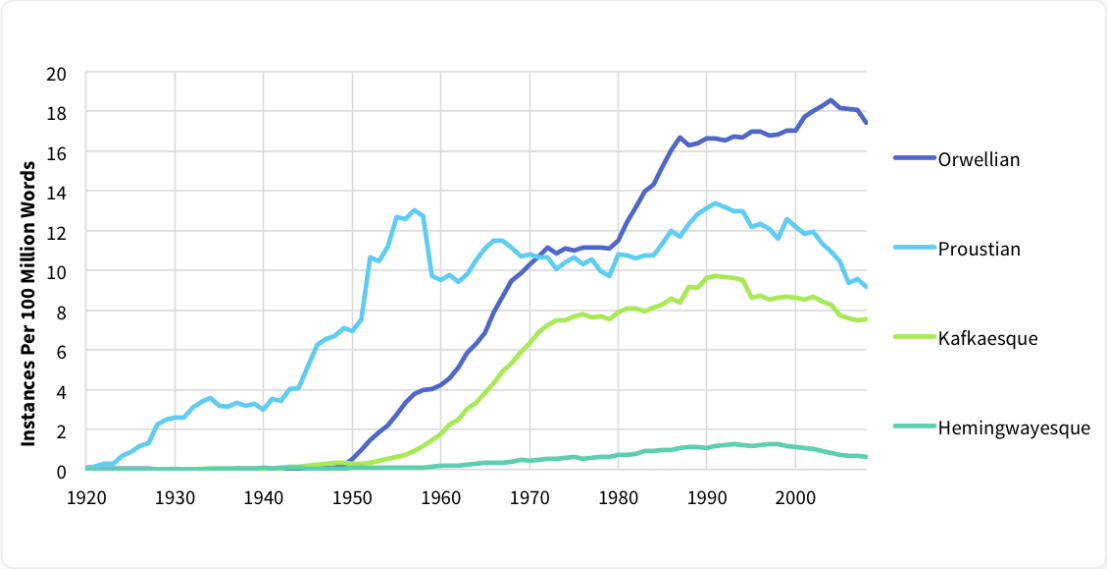
Does the higher popularity of "Ernest Hemingway" and the lower popularity of "The Old Man and the Sea" mean people are more interested in Hemingway the man and less interested in Hemingway the writer? Or are we looking at a tendency to do exactly what we've done, which is to write about Hemingway's work in aggregate, but about Fitzgerald's and Steinbeck's individually? Can you think of ways we might try to answer those questions?

Take, for instance, the eponym "Hemingwayesque." Its very existence implies that Hemingway's works have enough in common to create an aggregate impression. We haven't just relied on that impression: we've tested it, and found that there are indeed characteristics that define Hemingway's writing. It makes sense to write about a Hemingway corpus. Does it make sense to write about a Fitzgerald or Steinbeck corpus? Are "Fitzgeraldian" and "Steinbeckian" even words people use? Not really:



How does the popularity of "Hemingwayesque" factor into our assessment of Hemingway's relative influence? What about the fact that its popularity has been in steep decline since the mid-1990s? Is Hemingway's influence on the wane?

If we add a few more eponyms for context, here is what we find:



"Orwellian" and "Kafkaesque" seem to resonate more than "Hemingwayesque." Perhaps surprisingly, so does "Proustian."

How, then, do we assess Hemingway's influence? Is there some calculation we can come up with based on the relative popularity of the man, his books, and his eponym? Maybe we just to go back to the beginning, seek external validation, and note that Hemingway won the Nobel Prize in Literature. Then again: so did Steinbeck.

Justin Rice graduated from Harvard with a degree in comparative literature. He is a software developer specializing in Python programming and is a writer for LitCharts.