Historical trends in English sentence length and syntactic complexity

Mark Liberman
University of Pennsylvania
https://www.ling.upenn.edu/~myl

ABSTRACT

It's easy to perceive clear historical trends in the length of sentences and the depth of clausal embedding in published English text. And those perceptions can easily be verified quantitatively.

Or can they? Perhaps the title should be "Historical trends in English punctuation practices", or "Historical trends in English conjunctions and discourse markers."

The answer depends on several prior questions: What is a sentence? What is the boundary between syntactic structure and discourse structure? How is message structure encoded in speech (spontaneous or rehearsed) versus in text? This presentation will survey the issues, look at some data, and suggest some answers – or at least some fruitful directions for future work.

Illustrating the Obvious

Older texts in English
tend to have longer sentences
with greater depth of syntactic and conceptual embedding.

Since different genres and styles also make a difference,
we'll start with a simple case:

The inaugural addresses of American presidents.

The first paragraph of George Washington's Inaugural Address, 1789:

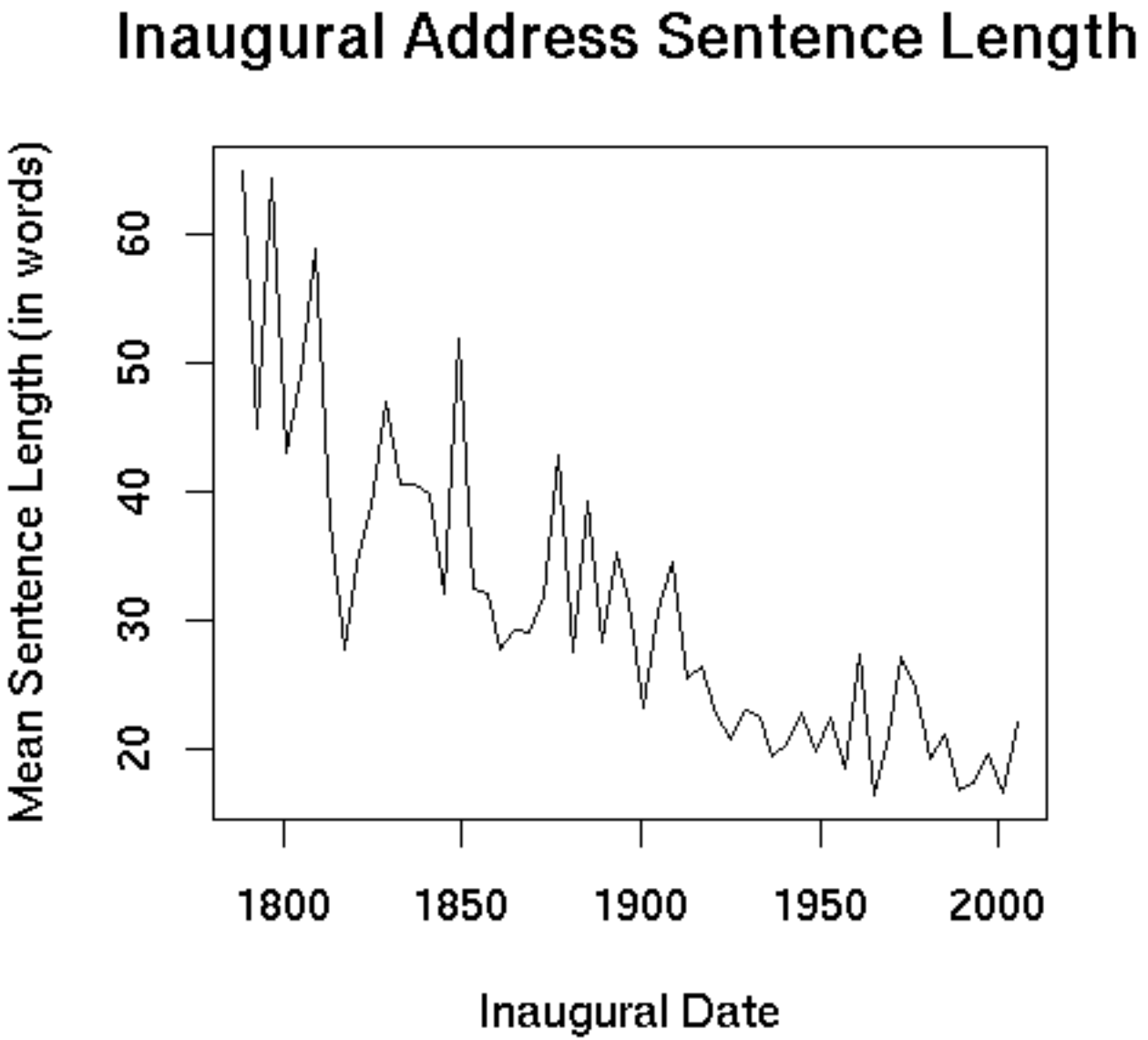
Among the vicissitudes incident to life no event could have filled me with greater anxieties than that of which the notification was transmitted by your order, and received on the 14th day of the present month. On the one hand, I was summoned by my country, whose voice I can never hear but with veneration and love, from a retreat which I had chosen with the fondest predilection, and, in my flattering hopes, with an immutable decision, as the asylum of my declining years—a retreat which was rendered every day more necessary as well as more dear to me by the addition of habit to inclination, and of frequent interruptions in my health to the gradual waste committed on it by time. On the other hand, the magnitude and difficulty of the trust to which the voice of my country called me, being sufficient to awaken in the wisest and most experienced of her citizens a distrustful scrutiny into his qualifications, could not but overwhelm with despondence one who (inheriting inferior endowments from nature and unpracticed in the duties of civil administration) ought to be peculiarly conscious of his own deficiencies. In this conflict of emotions all I dare aver is that it has been my faithful study to collect my duty from a just appreciation of every circumstance by which it might be affected. All I dare hope is that if, in executing this task, I have been too much swayed by a grateful remembrance of former instances, or by an affectionate sensibility to this transcendent proof of the confidence of my fellow-citizens, and have thence too little consulted my incapacity as well as disinclination for the weighty and untried cares before me, my error will be palliated by the motives which mislead me, and its consequences be judged by my country with some share of the partiality in which they originated.

The first two paragraphs of George W. Bush's Inaugural Address, 2005:

On this day, prescribed by law and marked by ceremony, we celebrate the durable wisdom of our Constitution, and recall the deep commitments that unite our country. I am grateful for the honor of this hour, mindful of the consequential times in which we live, and determined to fulfill the oath that I have sworn and you have witnessed.

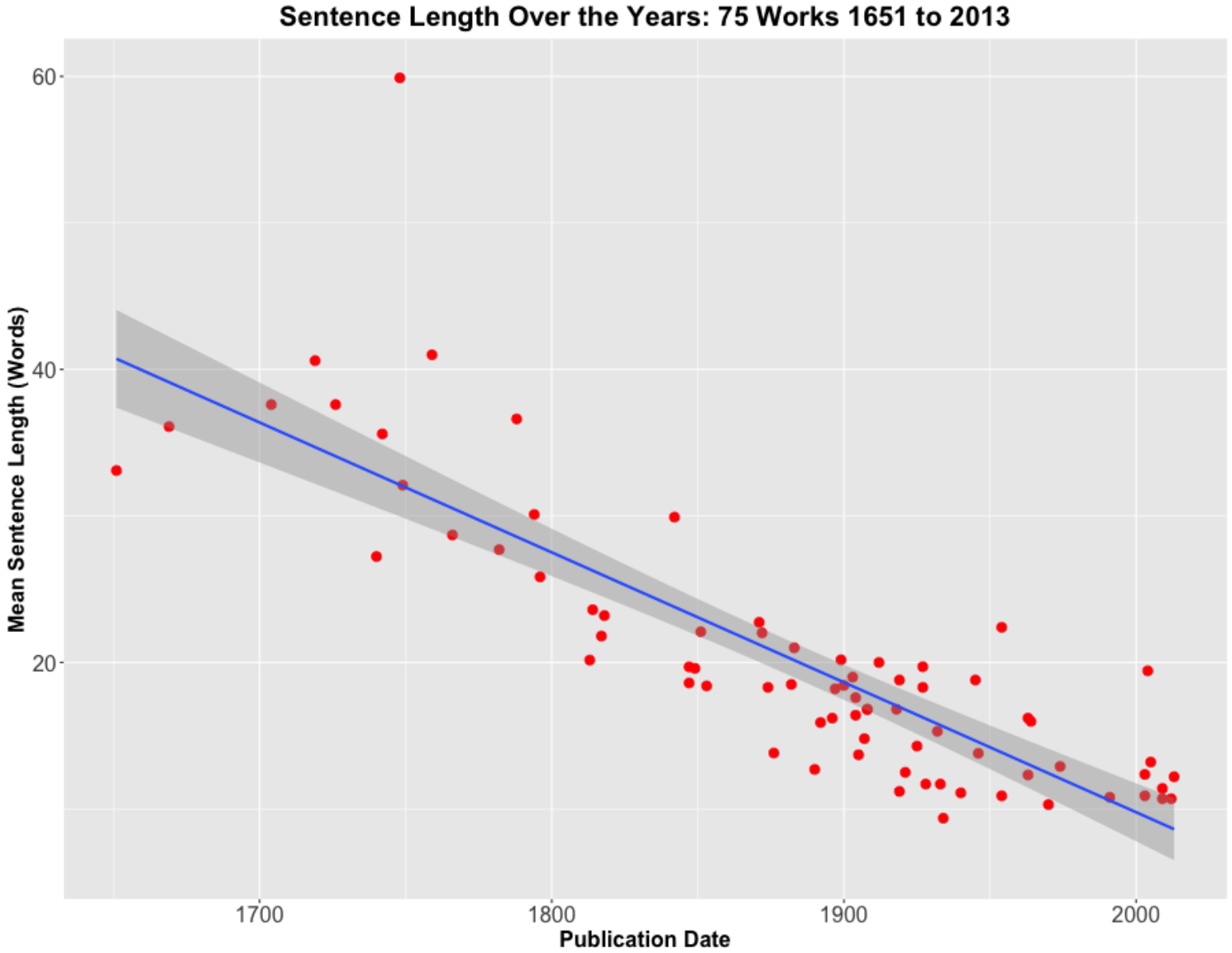
At this second gathering, our duties are defined not by the words I use, but by the history we have seen together. For a half a century, America defended our own freedom by standing watch on distant borders. After the shipwreck of communism came years of relative quiet, years of repose, years of sabbatical—and then there came a day of fire.

Mean sentence length of Inaugural Addresses from G. Washington (1789) to G.W. Bush (2005):

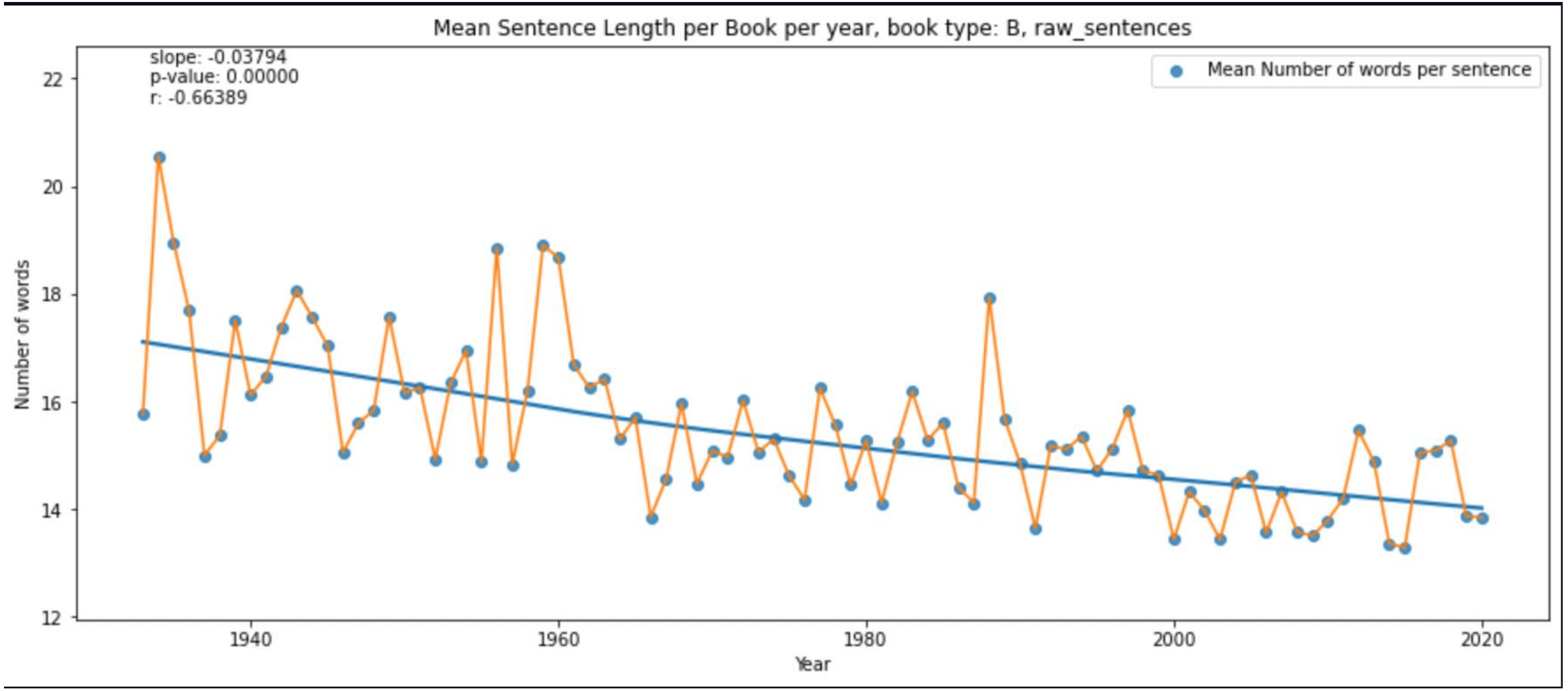


There's a longer-term trend.

Here's a sample of novels
(and novel-like works)
over 3 ½ centuries:

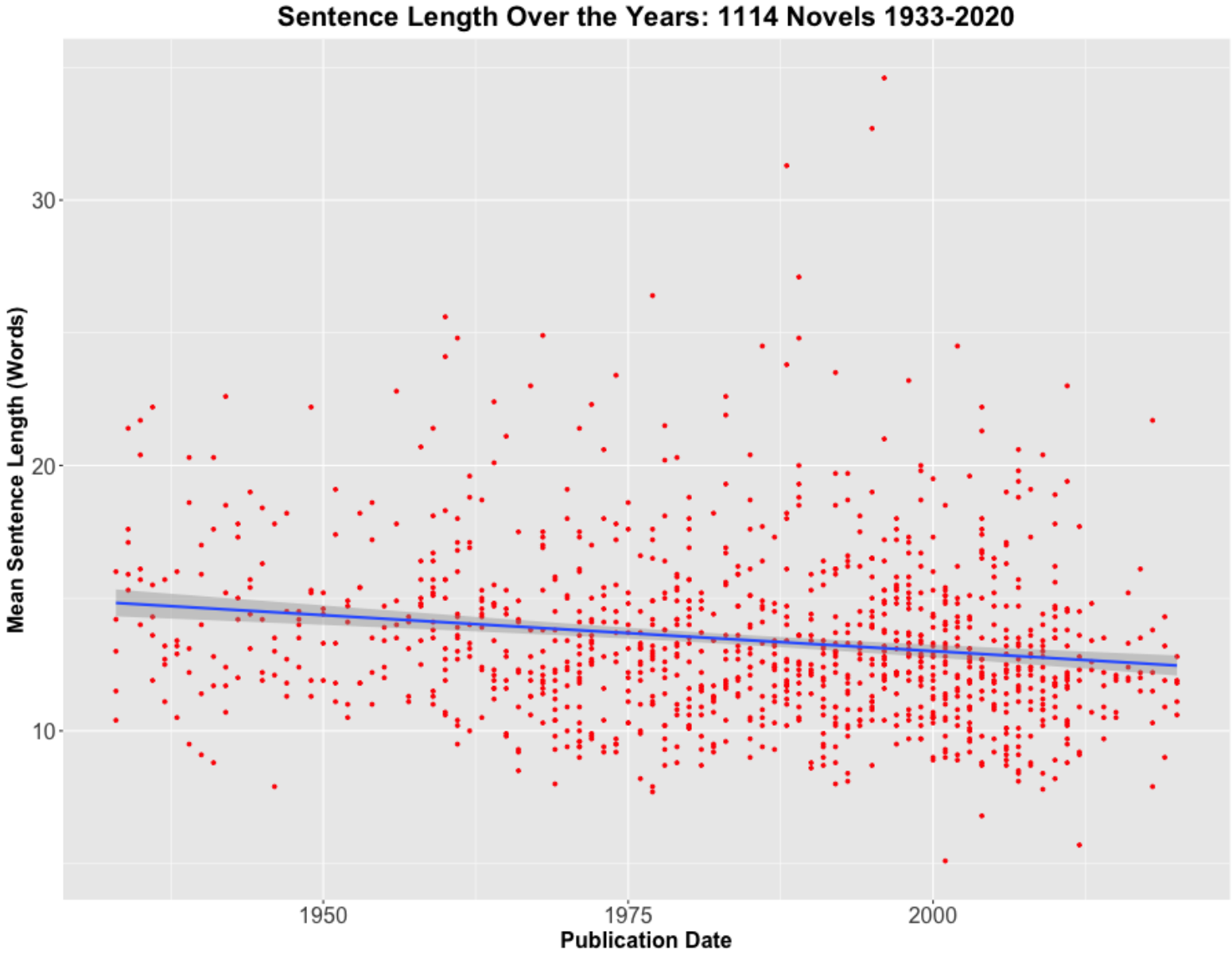


A gradual tend continues -- Top five NYT best-sellers per year, 1933-2020:



And there's a lot of variation!

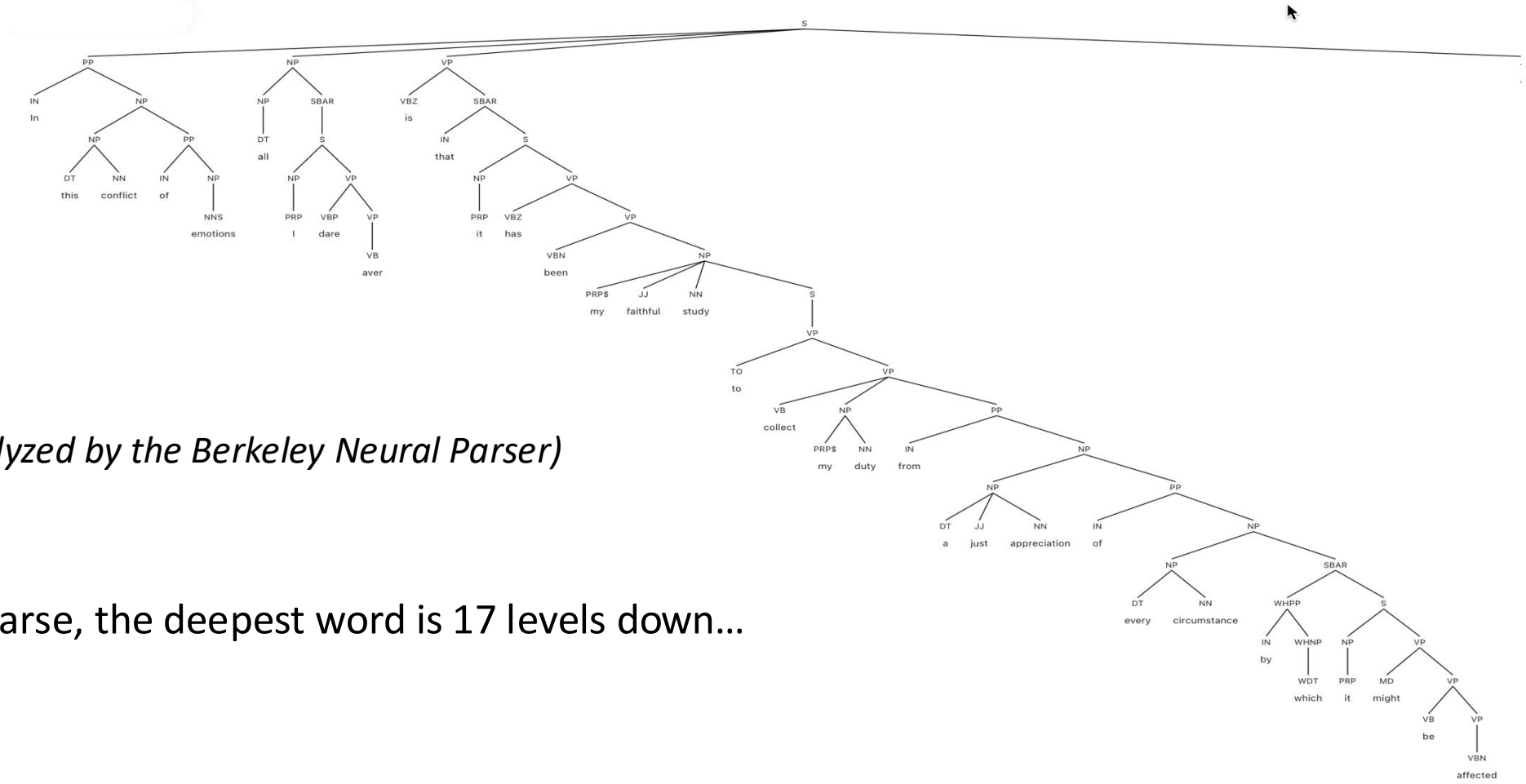
But the same gradual trend is visible in a larger collection over the same time period:



And sentences in older English-language texts
are not only (mostly) longer,
they're also (mostly) deeper…

A 34-word sentence from G. Washington 1789:

In this conflict of emotions all I dare aver is that it has been my faithful study to collect my duty from a just appreciation of every circumstance by which it might be affected.

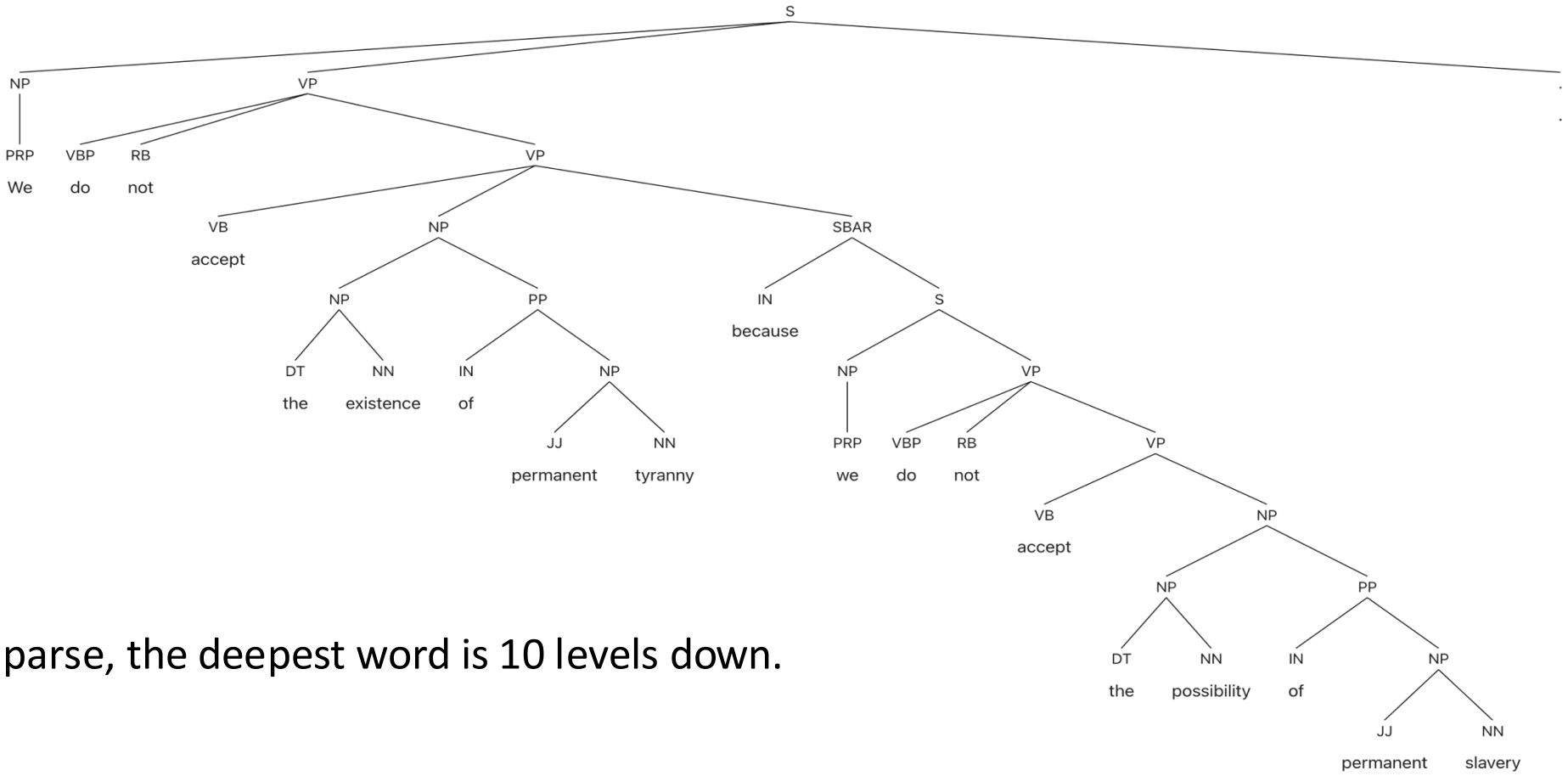


(As analyzed by the Berkeley Neural Parser)

In this parse, the deepest word is 17 levels down…

A 19-word sentence from G.W. Bush 2005:

We do not accept the existence of permanent tyranny because we do not accept the possibility of permanent slavery.



In this parse, the deepest word is 10 levels down.

On the theory that embedding of finite clauses is a key feature,
I analyzed the levels of finite (= tensed) clauses in each Inaugural:

0 In this conflict of emotions all I dare aver is
1 [that it has been my faithful study to collect my duty from a just appreciation of every circumstance
2 [by which it might be affected.]]

0 We do not accept the existence of permanent tyranny
1 [because we do not accept the possibility of permanent slavery.]

…and counted the number (and proportion) of words in each text
at each level of embedding.

Illustrative Results for Four Inaugural Addresses: Mean Sentence Length and Depth of Embedding by Word Count

	0	1	2	3	4	Mean Sentence Length
Washington 1789	629 **(44%)**	554 (39%)	206 (14%)	36 (3%)	5 (<1%)	**60**
Lincoln 1865	440 **(63%)**	222 (32%)	38 (5%)	0	0	**26**
Bush 2005	1842 **(88%)**	244 (12%)	4 (<1%)	0	0	**22**
Trump 2017	1264 **(87%)**	178 (12%)	15 (1%)	0	0	**15**

But “deeper” (more embedding, clausal or otherwise)
is not the only way for sentences to get longer…

Hypotaxis:

ὑπο- (hypo) "below" + τάξις (taxis) "placing"
Syntactic subordination of one clause or construction to another.

Parataxis:

Παρά- (para) "beside" + τάξις (taxis) "placing“
The placing of propositions or clauses one after another,
without indicating by connecting words the relation
(of coordination or subordination) between them,
as in “Tell me, how are you?” or “I came; I saw; I conquered;

Note that for present purposes, I'll (mis-)use the term ***parataxis***
to include stringing things together,
both with and without conjunctions:

Thus

I came; I saw; I conquered.
I came; I saw; and I conquered.

are both examples of ***parataxis*** in my sense,
since no explicit syntactic subordination is involved.

But this distinction leaves open a crucial question –
or rather, it opens the door to a large set of difficult questions.

Is a paratactic sequence just one thing after another,
whether within a “sentence” or across “sentences”?

Or is there an implicit hierarchical relationship,
with specific semantic (or discourse-structural) content?

This is not a new question.

Among the OED's glosses for the noun ***period*:**

16. a. *Rhetoric*. A grammatically complete sentence, *esp.* one made up of a number of clauses formed into a balanced or rhythmical whole; (more generally) a series of sentences seen as a linguistic unit.

So maybe the relevant structures should be rhetorical sequences,
whether or not they contain internal punctuational "periods"
(in the typographical sense of the word)?

(And even if the units are divided by punctuational "periods",
textual variants abound, as we'll see later…)

No matter how we answer these difficult questions,
we shouldn't be surprised that they get tangled up in social stereotypes…

From Otto Jespersen's 1922 book *Language: Its Nature Development and Origin,* Chapter XIII, *The Woman*:

> In learned terminology we may say that men are fond of hypotaxis and women of parataxis. Or we may use the simile that a male period is often like a set of Chinese boxes, one within another, while a feminine period is like a set of pearls joined together on a string of *ands* and similar words.

From Ursula K. Le Guin's 1953 essay "Introducing Myself",
as published in her collection *The Wave in the Mind*:

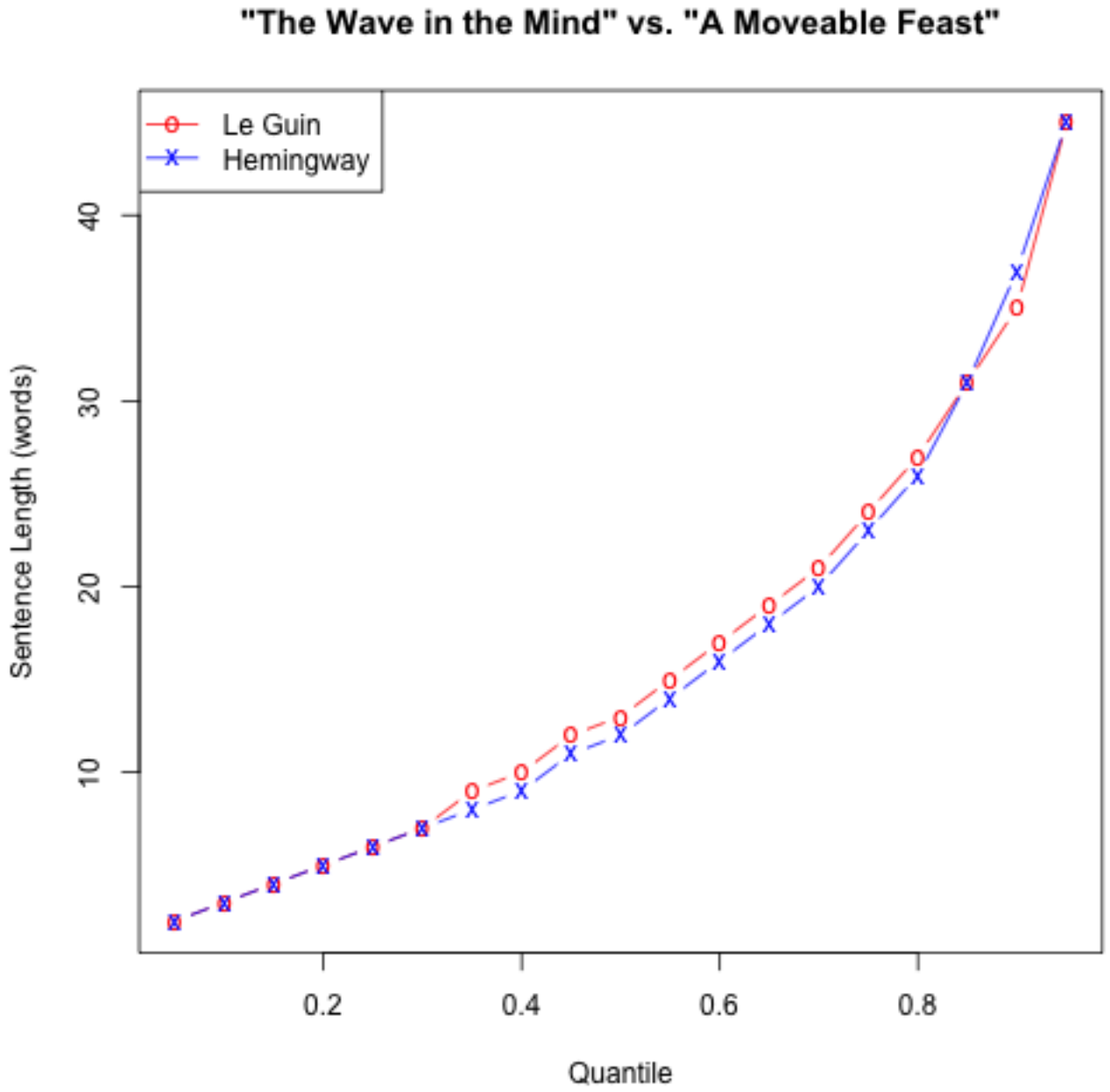
What it comes down to, I guess, is that I am just not manly. Like Ernest Hemingway was manly. The beard and the guns and the wives and the little short sentences. I do try. I have this sort of beardoid thing that keeps trying to grow, nine or ten hairs on my chin, sometimes even more; but what do I do with the hairs? I tweak them out. Would a man do that? Men don't tweak. Men shave. Anyhow white men shave, being hairy, and I have even less choice about being white or not than I do about being a man or not. I am white whether I like being white or not. The doctors can do nothing for me. But I do my best not to be white, I guess, under the circumstances, since I don't shave. I tweak. But it doesn't mean anything because I don't really have a real beard that amounts to anything. And I don't have a gun and I don't have even one wife and my sentences tend to go on and on and on, with all this syntax in them. Ernest Hemingway would have died rather than have syntax. Or semicolons. I use a whole lot of half-assed semicolons; there was one of them just now; that was a semicolon after "semicolons," and another one after "now."

But the distribution of sentence lengths in Le Guin's essay collection

The Wave in the Mind

is almost identical to the distribution in Hemingway's memoir

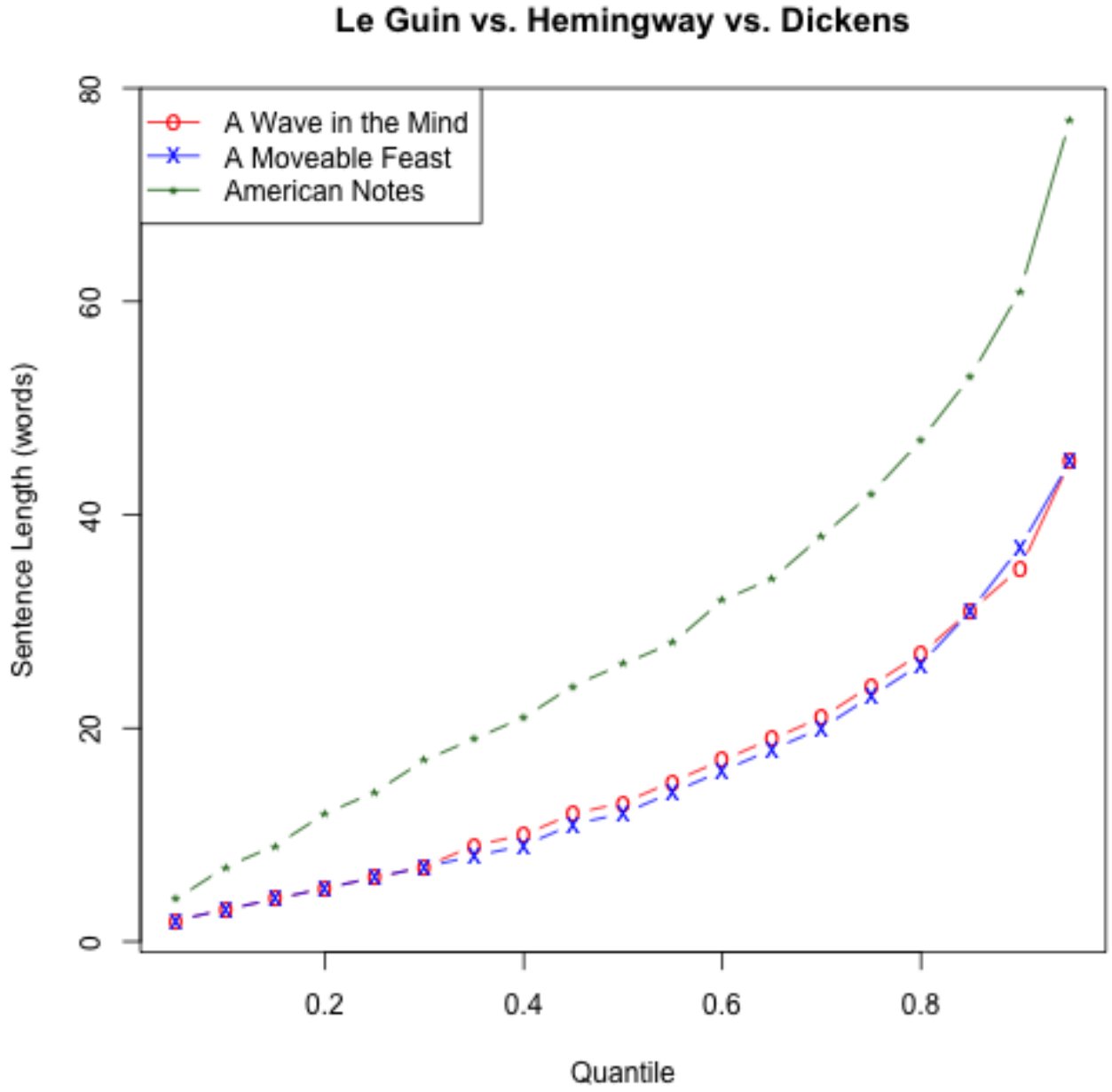
A Moveable Feast.



This is really a surprising coincidence, not just a general fact about text.

Authors **can** use longer sentences!

Compare the distribution
of sentence lengths
in Charles Dickens' *American Notes:*

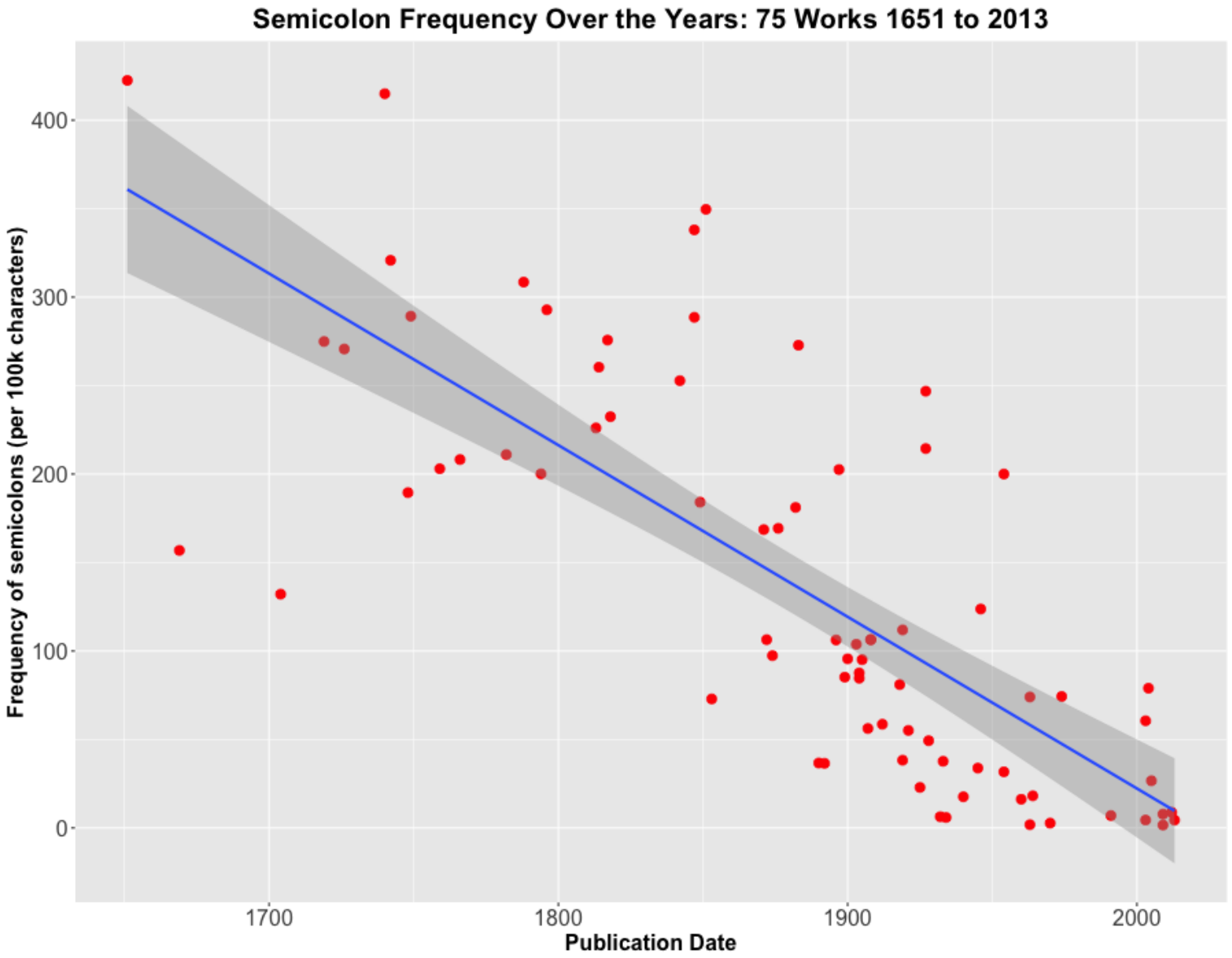


Le Guin is right about the semicolons,
at least in comparison to Hemingway.

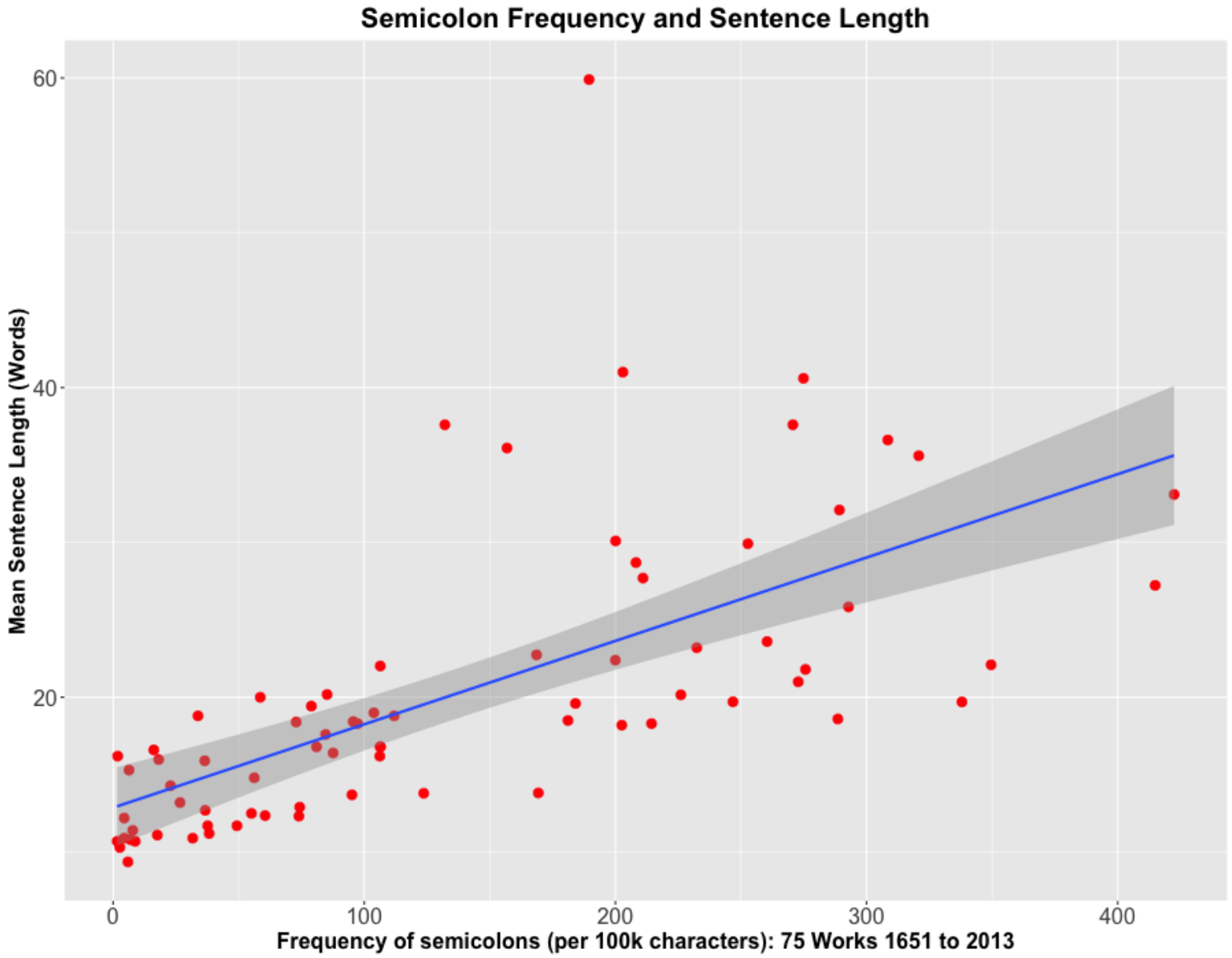
The relative frequency of semicolons
in her essay collection *The Wave in the Mind*
is more than 4 times greater than the semicolon frequency
in Hemingway's memoir *A Moveable Feast*:

Source	Semicolons	Total Characters	Semicolons per 100k Characters
The Wave in the Mind	411	520,607	78.95
A Moveable Feast	58	319654	18.14

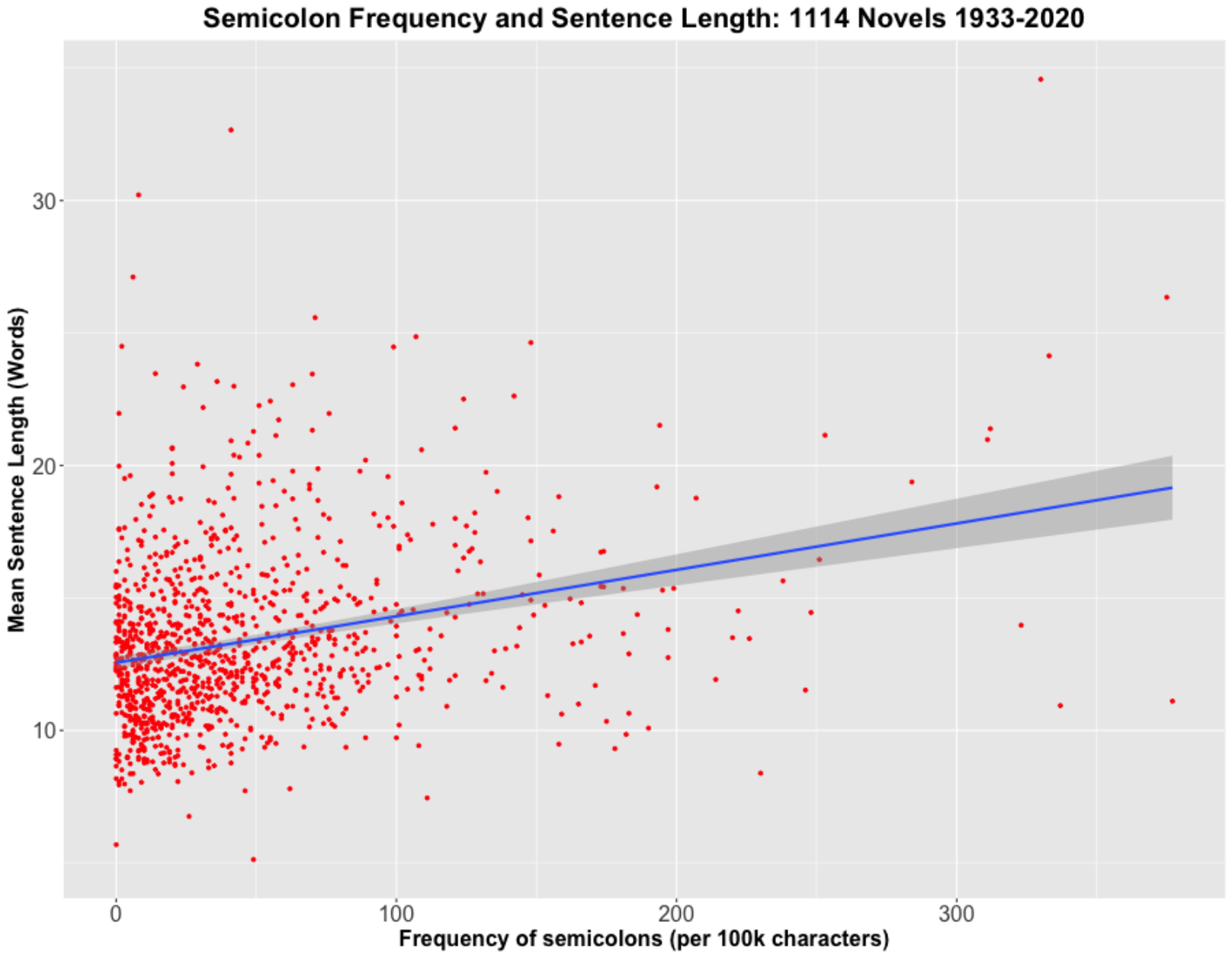
And there's definitely
a secular trend
in semicolon frequency:



And semicolon frequency correlates with sentence length over the centuries:



And also in a larger
and more recent sample:



But does semicolon frequency correlate with gender?

As this sample suggests, not so much.

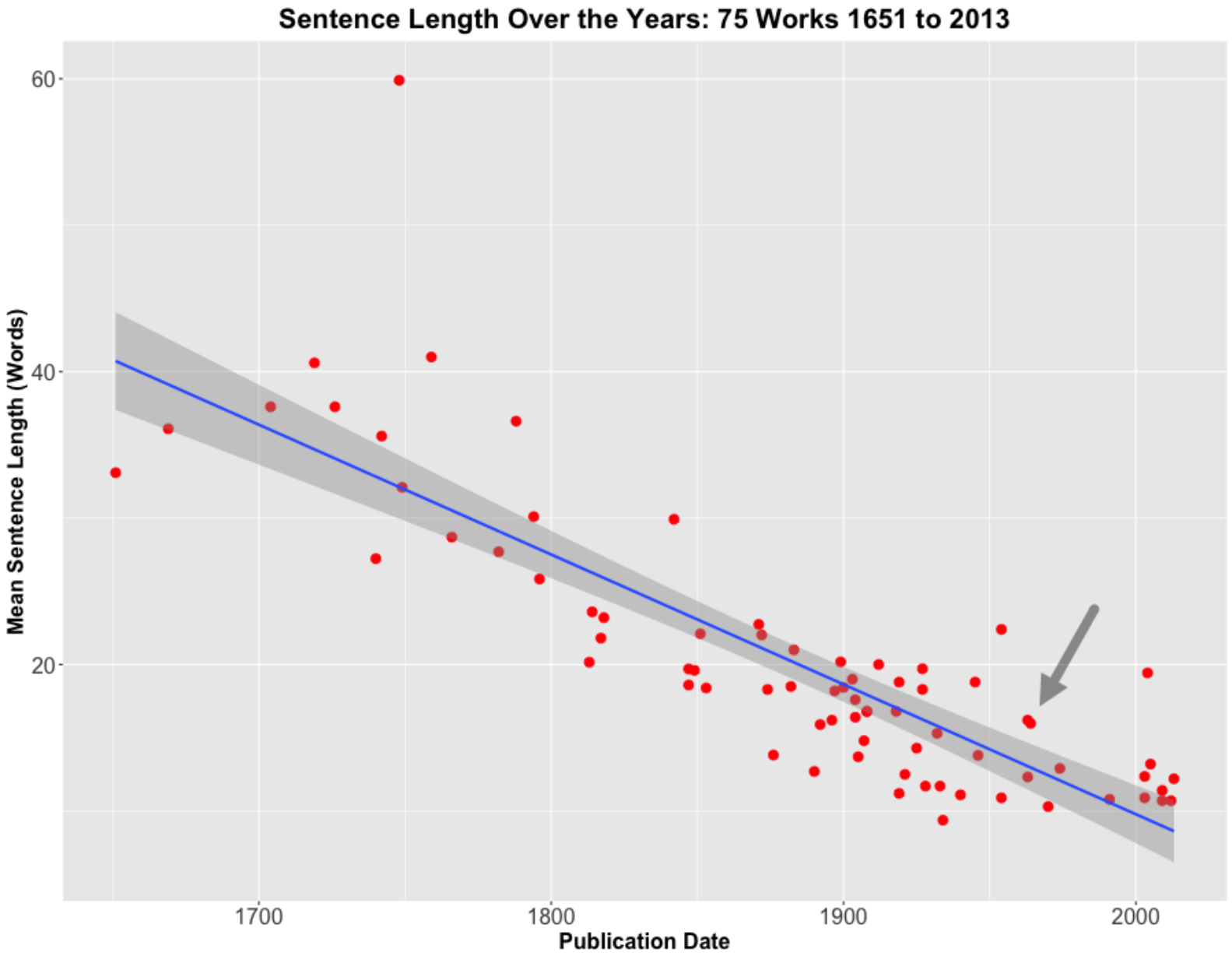
(*female authors in red*, *male authors in blue*…)

Source	Date	Semicolons	Total Chars	Semicolons/100k Chars
Pamela	1740	4,676	1,126,913	414.94
Decline and Fall of the Roman Empire	1788	39907	12936452	308.48
Camilla	1796	5,786	1,975,887	292.83
Pride and Prejudice	1813	1538	680359	226.06
American Notes	1842	1464	579209	252.76
Little Men	1871	925	548,683	168.59
Middlemarch	1872	1,874	1,761,476	106.39
Tom Sawyer	1876	642	379,164	169.32
The River War	1899	629	738,261	85.20
The Wonderful Wizard of Oz	1900	194	202966	95.58
The Great Gatsby	1925	60	162,323	22.87
To the Lighthouse	1927	941	381,272	246.81
Murder Must Advertise	1933	241	639,852	37.66
Murder on the Orient Express	1934	20	338,879	5.90
V.	1963	761	1,028,507	73.99
A Moveable Feast	1964	58	319,654	18.14
Oryx and Crake	2003	362	597,829	60.55
The Wave in the Mind	2004	411	520,607	78.95

Semicolons aside, Hemingway's short-sentence stereotype is a bit of a mystery.

His writings are generally a bit behind the sentence-length trend –

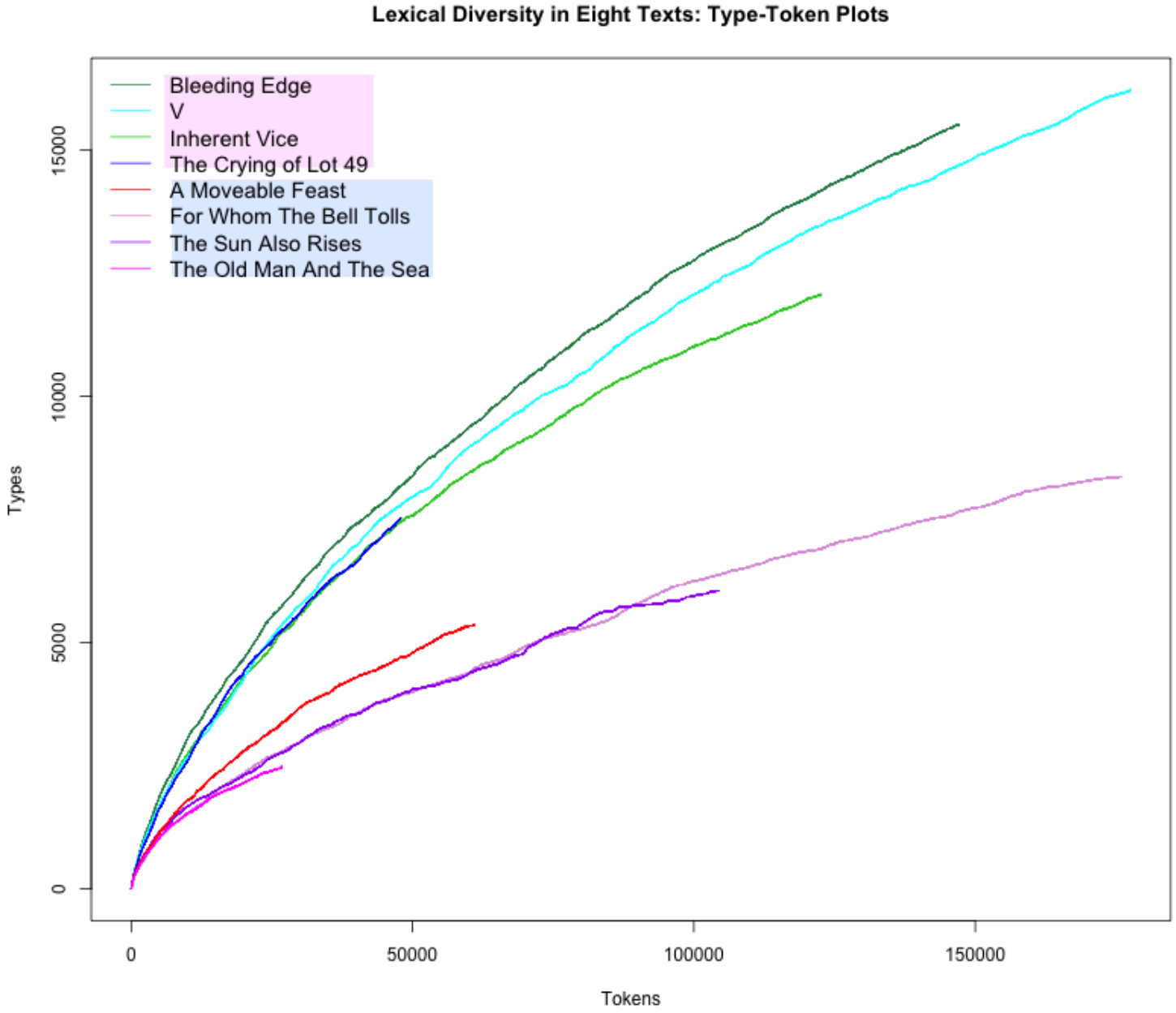
Here's the point for *A Moveable Feast*:



Maybe ”lexical diversity”
(really the rate of lexical display)
plays a role?

Here are type-token plots
for four works by Thomas Pynchon
and four by Hemingway:

Of all the authors I’ve checked,
Pynchon wins the type-token race,
and Hemingway comes in last…

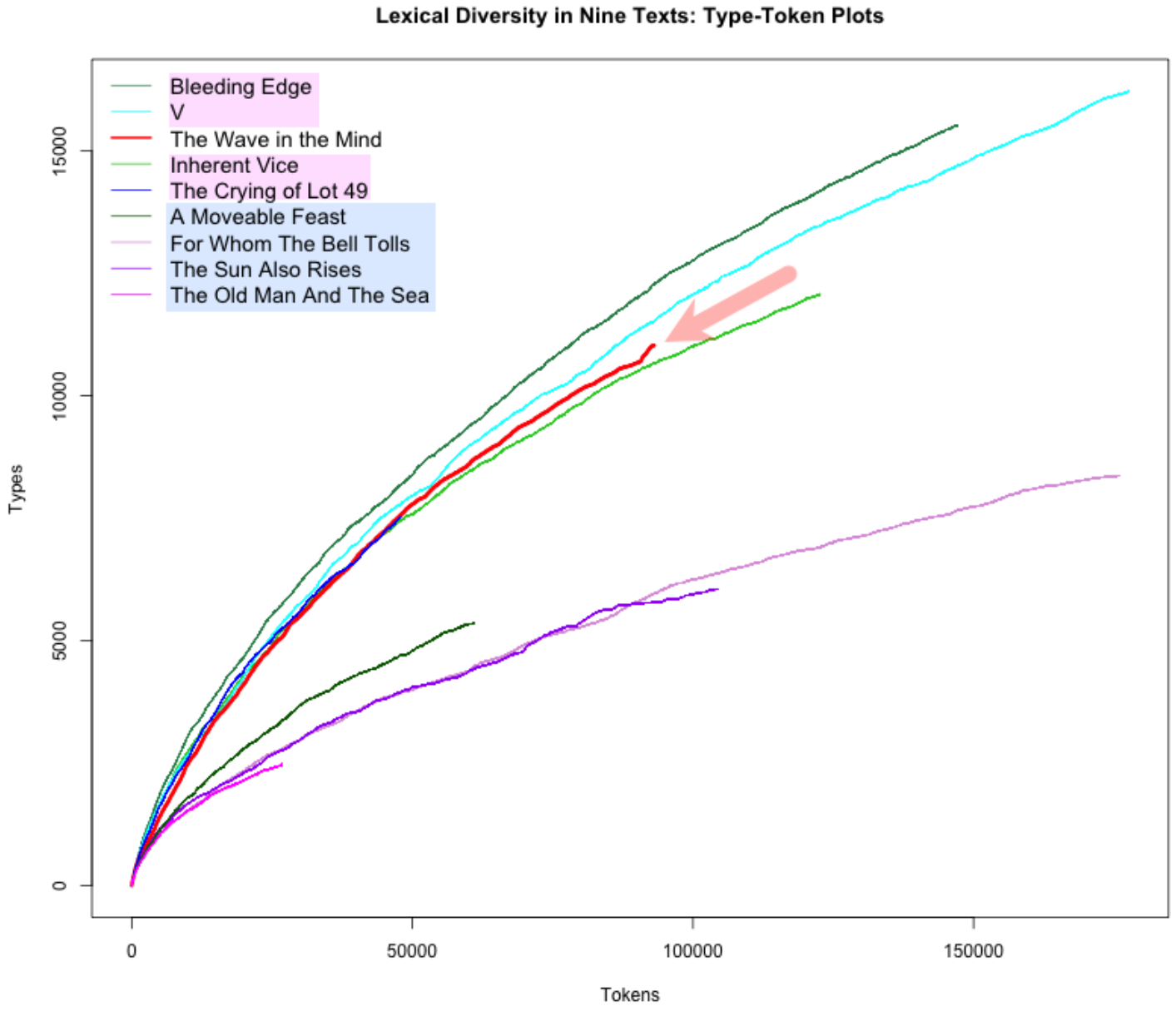


What about Le Guin?

The Wave in the Mind
is right up there
with Pynchon.

So maybe ***lexical bling*** influences
sentence-length judgment?

Though again,
gender doesn't seem crucial…



But this talk is not about lexical diversity,

so let’s get back to the length and structure of textual units…

Beyond mere sentence length, and the paratactic/hypotactic distinction,
there are obviously many other relevant dimensions of stylistic variation in structure:

- Amount (and type) of within-clause modification;
- Subordinate vs. parallel clauses/sentences;
- Number (and length and distribution) of parentheticals and appositives;
- Pronouns vs. definite descriptions;
- . . .

And these dimensions vary with many things other than historical time:

- Genre;
- Author (and editors);
- Intended audience;
- . . .

All these other dimensions aside,
“sentence length” is an interesting, relevant, and widely used variable –

For example, it’s central to the Flesch-Kincaid Grade Level Measure.

The Flesch-Kincaid “Grade Level” measure

Fifty years ago, the U.S. Navy was concerned that its technical manuals and instructional materials were too difficult for recruits to understand. So they contracted with J. Peter Kincaid to update Rudolf Flesch’s 1948 measure of “readability”, which was based on the plausible intuition that longer words and longer sentences are harder to read, other things equal.

This being the middle of the previous century, Flesch and Kincaid counted words, sentences, and syllables in a collection of graded educational materials, and then used multiple regression to estimate readability and grade level. This is Kincaid’s resulting “grade level” formula:

$$0.39\left(\frac{\text{total words}}{\text{total sentences}}\right)+11.8\left(\frac{\text{total syllables}}{\text{total words}}\right)-15.59$$

But “sentence length”
depends on how a discourse is divided into “sentences” –

which is a process with many layers
of orthographic, authorial, and editorial choice.

Here are the first two sentences of John Knox's infamous screed

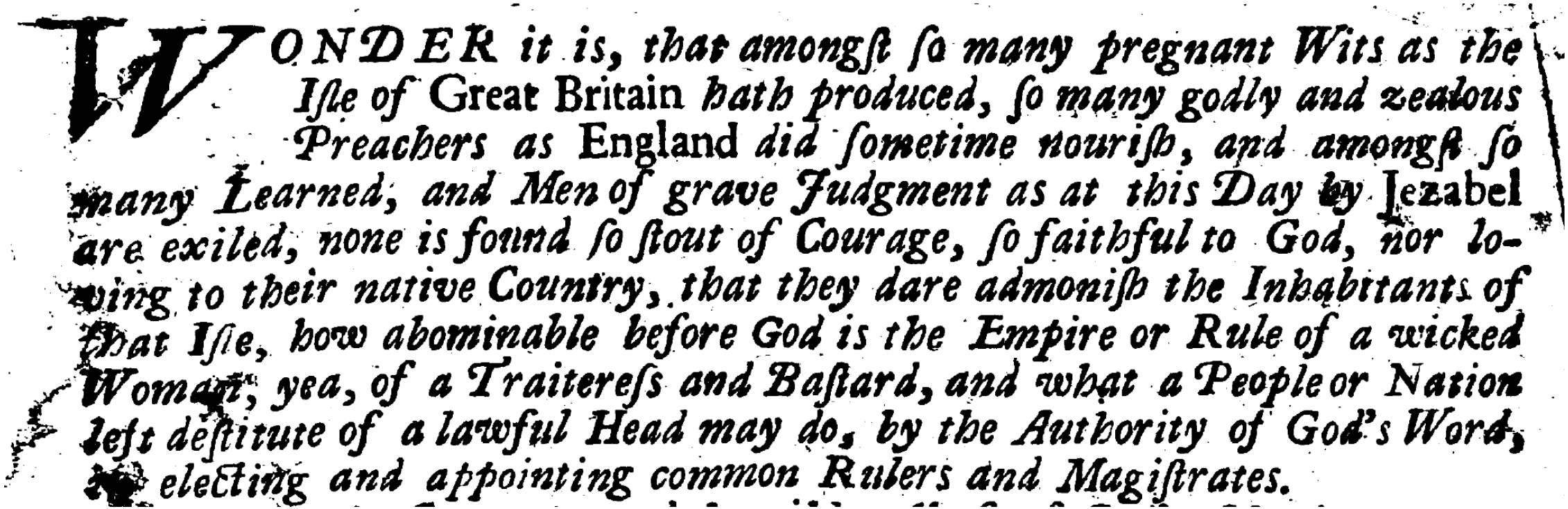
THE FIRST BLAST OF THE TRUMPET against the monstruous regiment of Women

…transcribed from the (original) 1558 edition:

Wonder it is, that amongest so many pregnant wittes as the Ile of greate Brittanny hath produced, so many godlie and zelous preachers as England did somtime norishe, and amongest so many learned and men of graue iudgement, as this day by Iesabel are exiled, none is found so stowte of courage, so faithfull to God, nor louing to their natiue countrie, that they dare admonishe the inhabitantes of that Ile how abominable before God, is the Empire or Rule of a wicked woman, yea of a traiteresse and bastard. And what may a people or nation left destitute of a lawfull head, do by the authoritie of Goddes worde in electing and appointing common rulers and magistrates.

In this version, the first sentence (in red) is 90 words long,
and the second one is 28 words long.

Here's an image from a 17th-century edition, a century later,
in which those two sentences are combined with a comma,
to make one sentence of 90+28 = 118 words:



WONDER it is, that amongſt ſo many pregnant Wits as the Iſle of Great Britain hath produced, ſo many godly and zealous Preachers as England did ſometime nouriſh, and amongſt ſo many Learned, and Men of grave Judgment as at this Day by Jezabel are exiled, none is found ſo ſtout of Courage, ſo faithful to God, nor loving to their native Country, that they dare admoniſh the Inhabitants of that Iſle, how abominable before God is the Empire or Rule of a wicked Woman, yea, of a Traitereſs and Baſtard, and what a People or Nation left deſtitute of a lawful Head may do, by the Authority of God's Word, in electing and appointing common Rulers and Magiſtrates.

WONDER it is, that amongſt ſo many pregnant Wits as the Iſle of Great Britain hath produced, ſo many godly and zealous Preachers as England did ſometime nouriſh, and amongſt ſo many Learned, and Men of grave Judgment as at this Day by Jezabel are exiled, none is found ſo ſtout of Courage, ſo faithful to God, nor loving to their native Country, that they dare admoniſh the Inhabitants of that Iſle, how abominable before God is the Empire or Rule of a wicked Woman, yea, of a Traitereſs and Baſtard, and what a People or Nation left deſtitute of a lawful Head may do, by the Authority of God's Word, by electing and appointing common Rulers and Magiſtrates.

Still, we generally see the division of written text
into “sentences” as given.

But in spontaneous speech,
division into “sentences” depends on transcribers’ choices,
which can and do vary a lot.

Matt Viser, "For presidential hopefuls, simpler language resonates“
(" Trump tops GOP field while talking to voters at fourth-grade level"),
Boston Globe 10/20/2015:

When Donald Trump announced his presidential campaign, he decried the lack of intelligence of elected officials in characteristically blunt terms.

“How stupid are our leaders?” he said. “How stupid are they?”

But with his own choice of words and his short, simple sentences, Trump’s speech could have been comprehended by a fourth-grader. Yes, a fourth-grader.

The Globe reviewed the language used by 19 presidential candidates, Democrats and Republicans, in speeches announcing their campaigns for the 2016 presidential election. The review, using a common algorithm called the Flesch-Kincaid readability test that crunches word choice and sentence structure and spits out grade-level rankings, produced some striking results.

Here's a sample from the cited speech, starting with the version of the transcript used by the Globe, and then calculating "grade level" based on versions with the same word sequence but slightly different punctuation:

It's coming from more than Mexico. It's coming from all over South and Latin America. And it's coming probably — probably — from the Middle East. But we don't know. Because we have no protection and we have no competence, we don't know what's happening. And it's got to stop and it's got to stop fast. **[Grade level 4.4]**

It's coming from more than Mexico, it's coming from all over South and Latin America, and it's coming probably — probably — from the Middle East. But we don't know, because we have no protection and we have no competence, we don't know what's happening. And it's got to stop and it's got to stop fast. **[Grade level 8.5]**

It's coming from more than Mexico, it's coming from all over South and Latin America, and it's coming probably — probably — from the Middle East; but we don't know, because we have no protection and we have no competence, we don't know what's happening. And it's got to stop and it's got to stop fast. **[Grade level 12.5]**

There are two distinct questions about such (apparently) paratactic sequences:

1. What are the units?
2. What is the structure?

So maybe there are two sentences as in the transcript (broken at silent pauses):

> But we don't know.
> Because we have no protection and we have no competence, we don't know what's happening.

Maybe there are three, or four -- or maybe there's only one:

> But we don't know, because we have no protection and we have no competence – we don't know what's happening.

But no matter how we punctuate it, the meaning is the same –
and presumably the "syntax" should reflect that.

This comes up in cases where everybody agrees about the punctuation.

For example, what Barbara Partee calls “Baseball Conditionals”:

> "He could have been a little rusty early on, and then the inning he gave up four runs I think he kind of lost his composure a little bit," Orioles manager Sam Perlozzo said. **"He just did a little damage control in that situation, we're OK."**

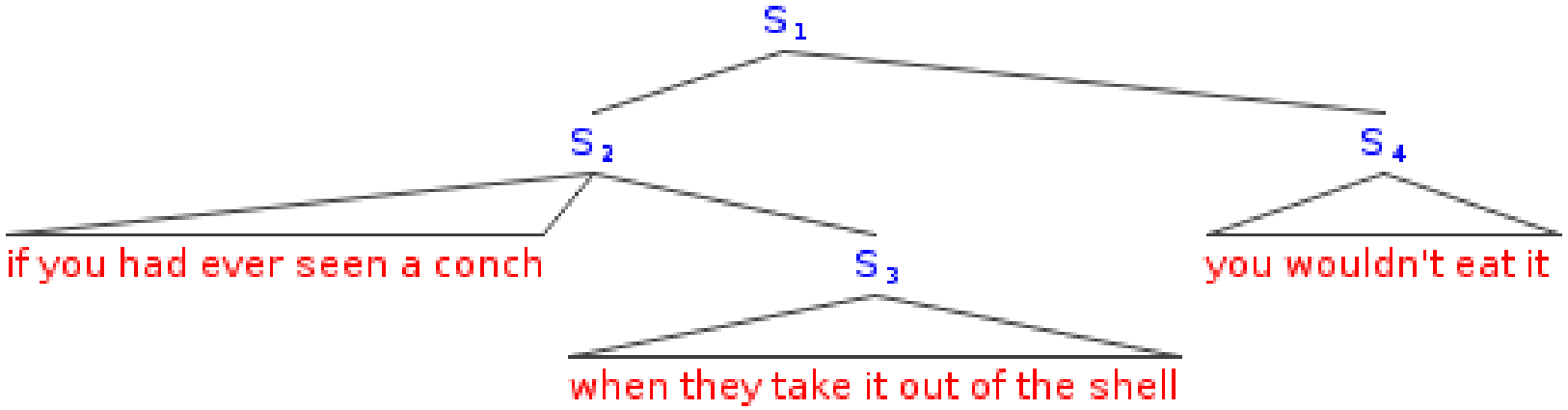
Or

> Manager Terry Collins on the Mets' victory over the Marlins last night: “**A year ago, we don’t win tonight.** It’s a different mentality in our clubhouse now."

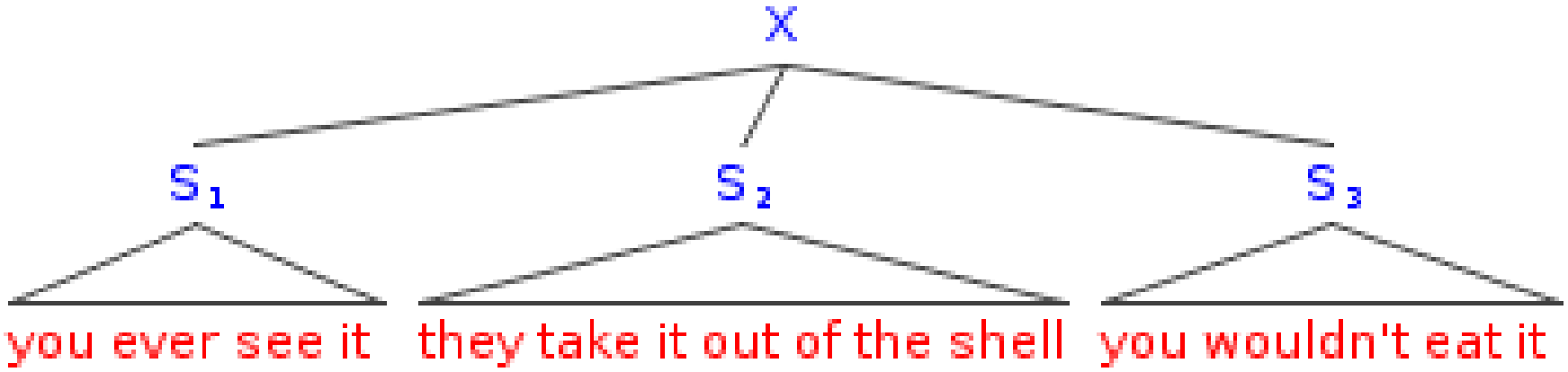
Similar paratactic constructions are common in informal English.
Here's a quote from Elmore Leonard's novel *LaBrava*:

"What're you having, conch?
You ever see it they take it out of the shell? You wouldn't eat it."

A hypotactic translation of the last two "sentences" in that quote might be

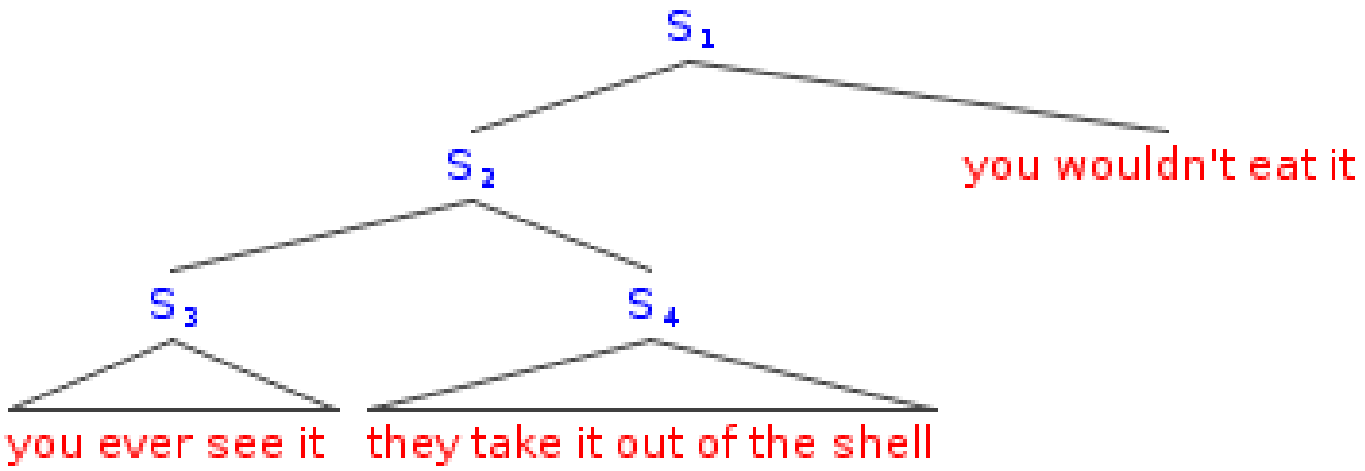


Should the three clauses just be strung together,
with the semantics to be inferred by discourse-level processes?



*(And the **X** node – which connects those clauses*
and separates them from what precedes and what follows –
is another representational choice…)

Or should the sequence be given the same structure as the hypotactic version



…with features on the nodes and links specifying their semantic relationships?

(i.e. S4 is a temporal modifier of S3… and S2 is a conditional clause… or something like that.)

In other words: is parataxis really covert hypotaxis?

This should remind us of The Great Recursion Debate.

Hauser, Chomsky, and Fitch 2002:

> We hypothesize that FLN [the Faculty of Language in the Narrow Sense]
> only includes recursion
> and is the only uniquely human component of the faculty of language.

Everett 2005:

> … the evidence suggests that Pirahã lacks embedding altogether.

Nevertheless, Everett presents and discusses examples that he translates as

"When I finish eating, I want to speak to you"; "If it rains, I will not go";
"I want the shirt that Chico sold"; "The woman wants to see you";
"He knows how to make arrows well"; "I said that Kó'oí intends to leave";
and so on.

These seem like transparent counterexamples to his “no embedding” claim.

But basically, his argument is that the Pirahã are like Elmore Leonard characters –

Their semantically complex examples, on his analysis, are paratactic sequences of units without any explicit syntactic relationships among them.

In my opinion, this situation poses a problem for all linguists.

If we deploy syntactic embedding and semantic relations to describe
“baseball conditionals”, Elmore Leonard’s dialogue, Donald Trump’s speeches
– and the paratactic aspects of much ordinary English –

can we rationally avoid providing a ”syntactic” analysis
for larger-scale discourse structures?

And can we stop at local rhetorical relations like
exemplification, concession, and generalization,
whether within or across ”sentences”?

Or do we need to give a syntactic account of the relations
among paragraph- and chapter-sized discourse chunks?

Conclusions:

- English-language text exhibits long-term trends in sentence length and complexity
- To some extent, these trends reflect changes in punctuational style
- Many relevant features besides “sentence” length remain to be explored,

  including the balance between *hypotaxis* and *parataxis*
- The syntax of paratactic sequences is an issue, within and across “sentences”
- The boundary between “syntax” and “discourse” is interestingly unclear

Some questions:

- Why does this trend exist?
  - Declining influence of Latin and Greek models?
  - General trend towards simpler/plainer style?
  - Random cultural evolution?
- What are the effects of devices and media?
- What about other languages?
  (In particular, French seems different…)

Note of acknowledgment:

The 1933-2020 text collection
comes from work in progress
with Angela Duckworth, Lyle Ungar, Benjamin Manning, and Jordan Ellenberg.