Take-home Points from Michael Vsauce's Video "Zipf Mystery" [1]

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

[1] available at: https://www.youtube.com/watch?v=fCn8zs912OE

Zipf's Law

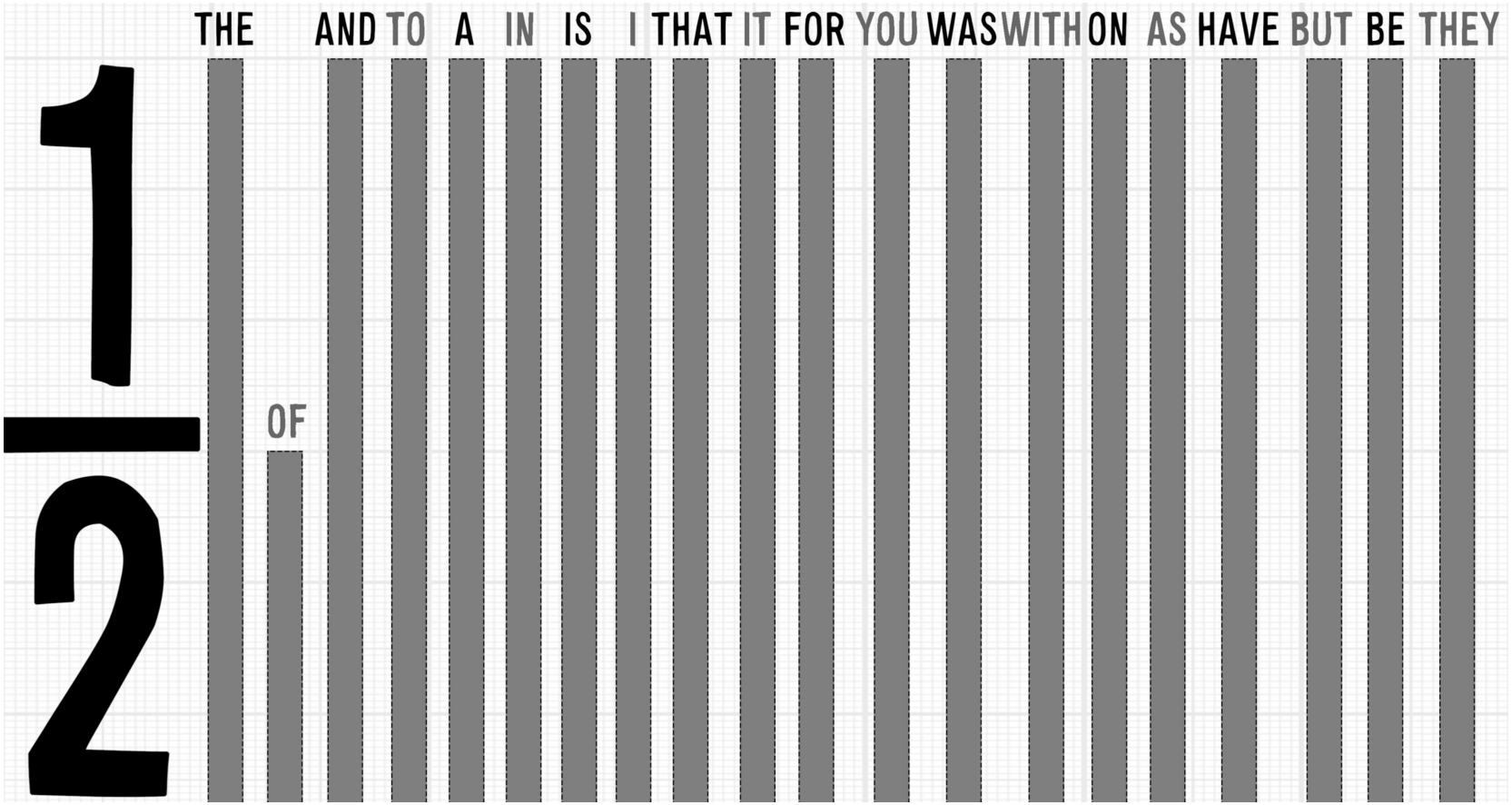
named after the American linguist George Kingsley Zipf, who was a

A little poem by Michael Vsauce

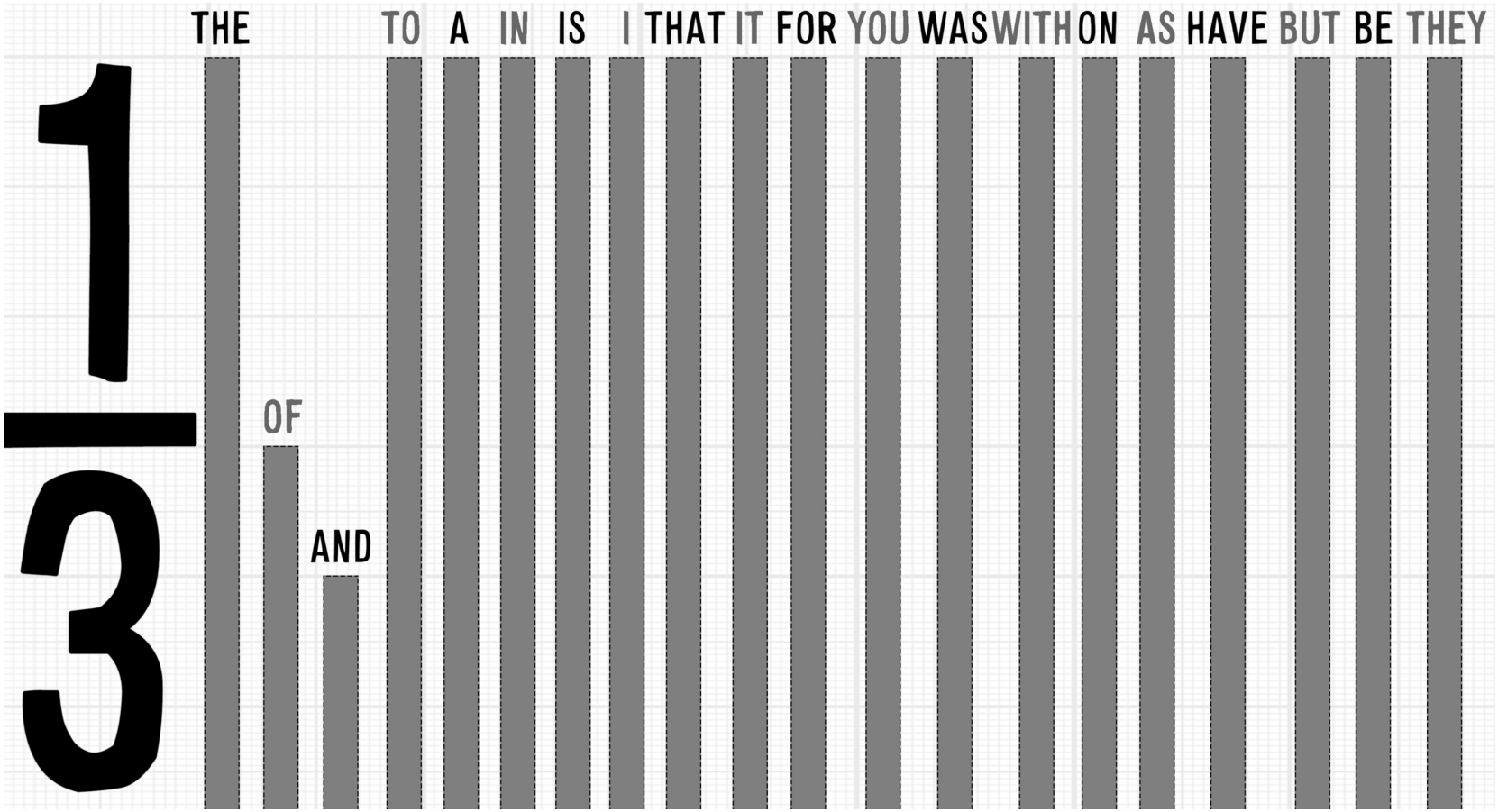
THE OF AND TO A IN IS I THAT IT FOR YOU WAS WITH ON AS HAVE BUT BE THEY

Question: What do these words have in common?

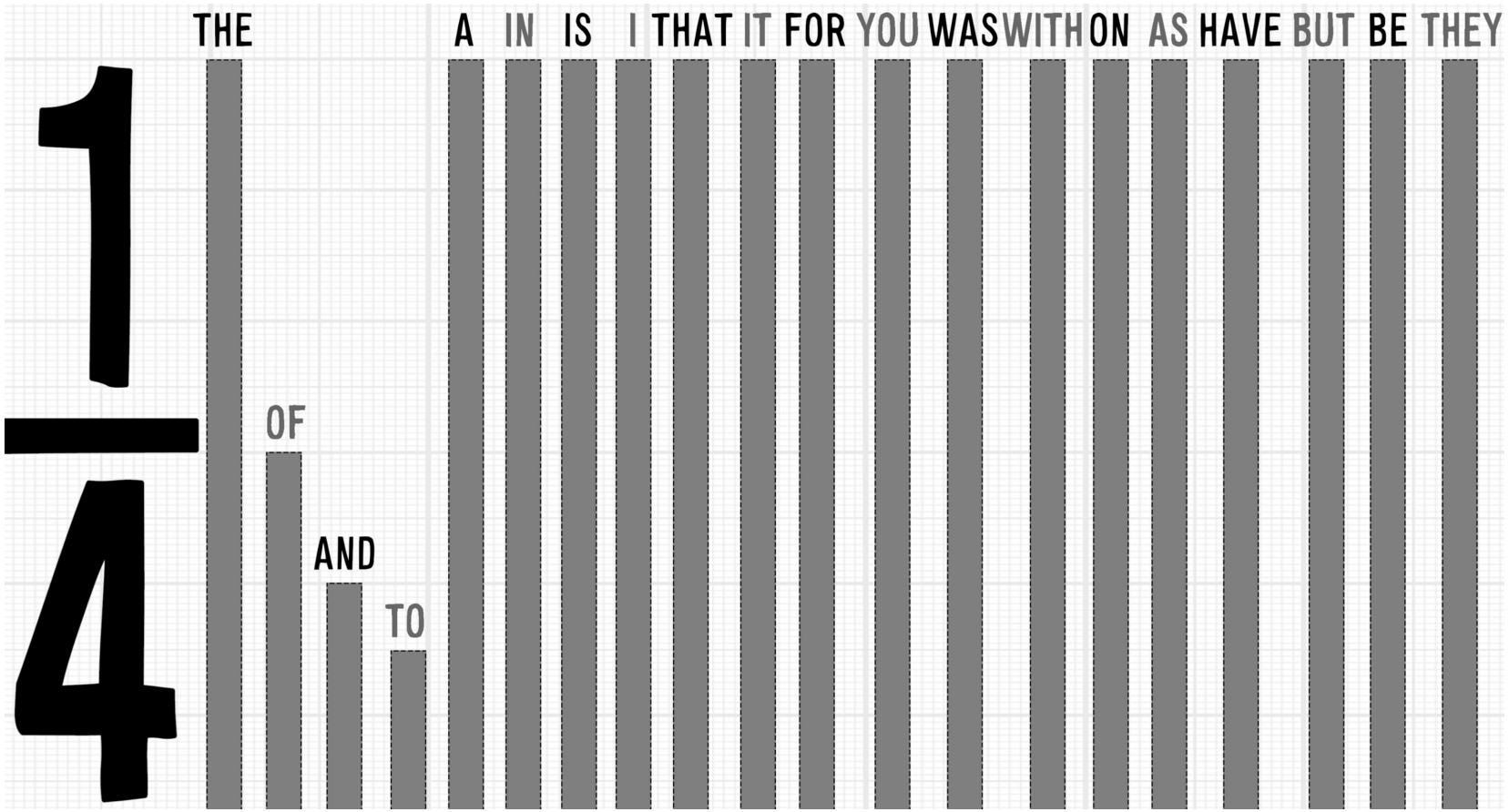
Frequencies for the top two most common words



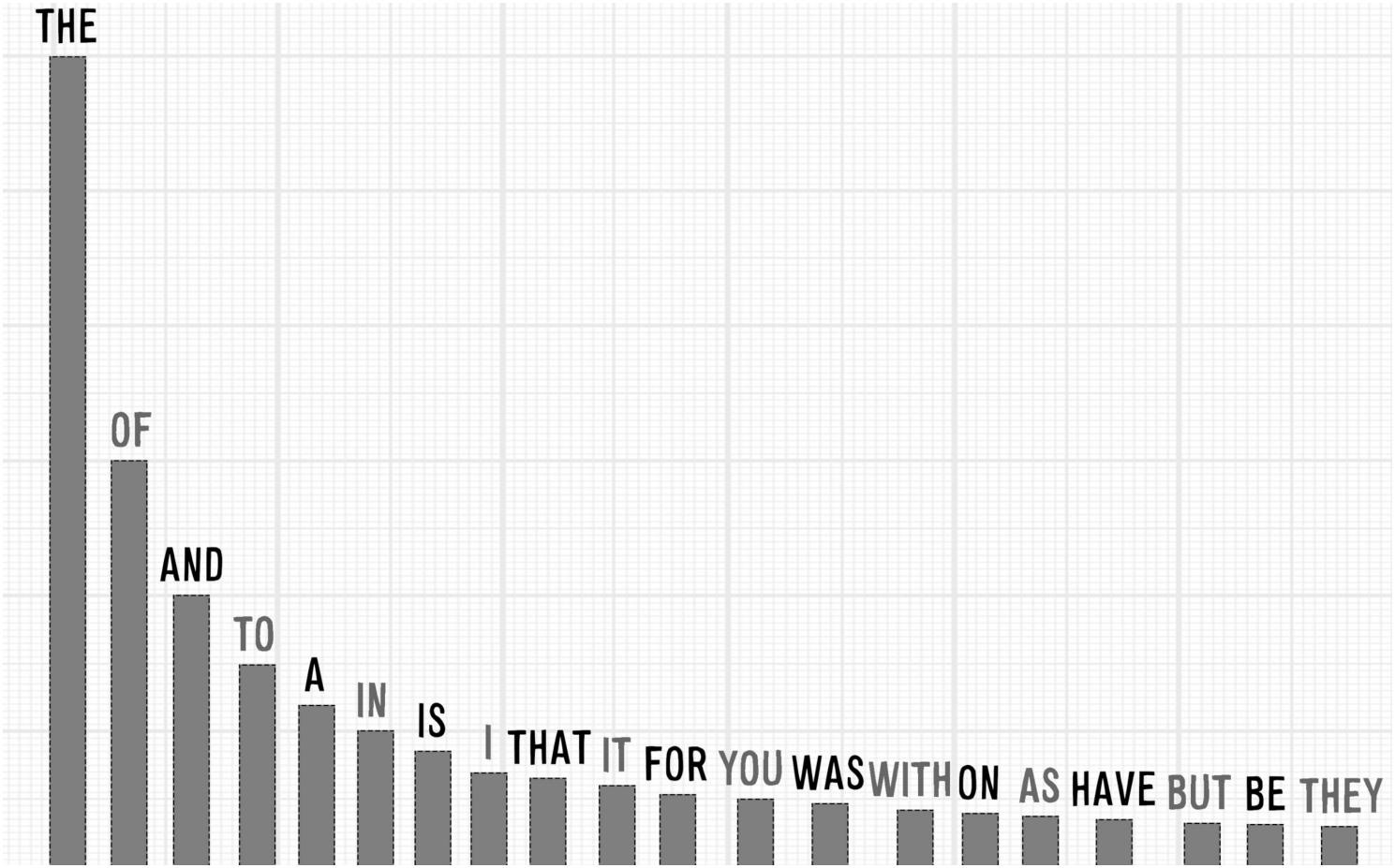
Frequencies for the top three most common words



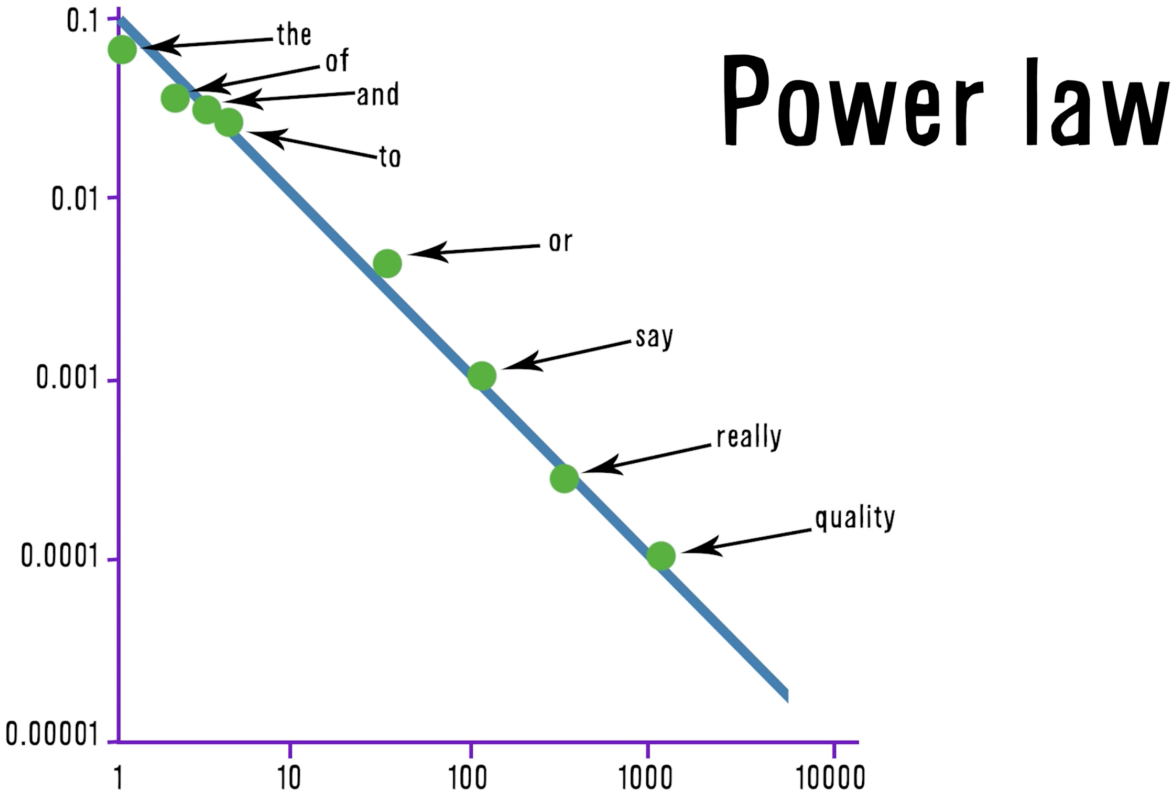
Frequencies for the top four most common words



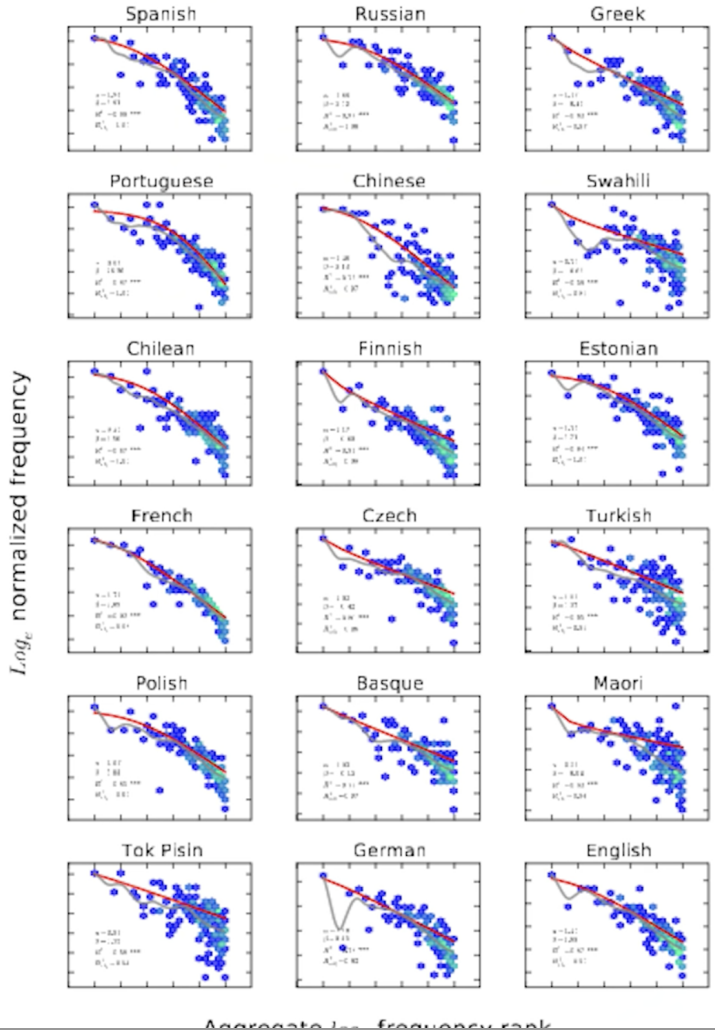
Frequencies for the top twenty most common words



Power Law Distribution



Comparing Word Distributions Across Languages



Some Memorable Quotes from Michael Vsauce's Video

the is the most used word in the English language. About one of every 16 words we encounter on a daily basis is **the**.

The twelve top most common English words are *the*, *of*, *and*, *to*, *a*, *in*, *is*, *I*, *that*, *it*, *for*, *you*

More on Word Frequencies

- *One word accounts for 6 percent of what we say.*
- *The top 25 most used words make up about a third of everything we say.*
- *Roughly speaking, and this is mind blowing, nearly half of any book, conversation or article will be nothing but the same 50 to 100 words.*
- *And nearly the other half will be words that appear in that selection only once.*

Top hundred most common words

1	the	21	this	41	so	61	people	81	back
2	be	22	but	42	up	62	into	82	after
3	to	23	his	43	out	63	year	83	use
4	of	24	by	44	if	64	your	84	two
5	and	25	from	45	about	65	good	85	how
6	a	26	they	46	who	66	some	86	our
7	in	27	we	47	get	67	could	87	work
8	that	28	say	48	which	68	them	88	first
9	have	29	her	49	go	69	see	89	well
10	I	30	she	50	me	70	other	90	way
11	it	31	or	51	when	71	than	91	even
12	for	32	an	52	make	72	then	92	new
13	not	33	will	53	can	73	now	93	want
14	on	34	my	54	like	74	look	94	because
15	with	35	one	55	time	75	only	95	any
16	he	36	all	56	no	76	come	96	these
17	as	37	would	57	just	77	its	97	give
18	you	38	there	58	him	78	over	98	day
19	do	39	their	59	know	79	think	99	most
20	at	40	what	60	take	80	also	100	us

Zipf's Law

A pattern emerges: The second most used word will appear about half as often as the most used. The third one third as often. The fourth one fourth as often. The fifth one fifth as often. The sixth one sixth as often, and so all the way down.

The amount of times a word is used is just proportional to one over its rank:

$$\frac{1}{rank} \propto \frac{1}{1}, \ldots \frac{1}{2} \ldots \frac{1}{3} \ldots \frac{1}{4} \ldots \tag{1}$$

This rank-frequency-based distribution goes by the name Zipf's Law. It holds widely across languages and across different corpora of a given language.

Beyond language ...

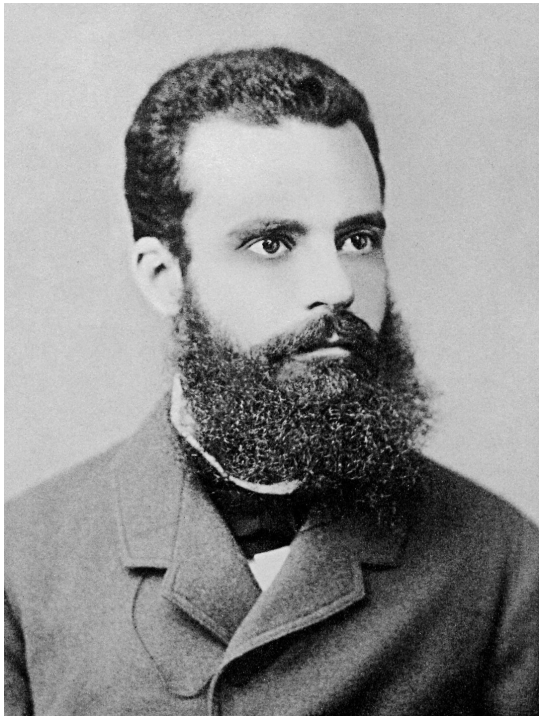
Moreover, Zipf's Law doesn't just mysteriously describe word use. It's also found in city populations, solar flare intensities, protein sequences and immune receptors, the amount of traffic websites get, earthquake magnitudes, the number of times academic papers are cited, last names, the firing patterns of neural networks, ingredients used in cookbooks, the number of phone calls people received, the diameter of Moon craters, etc.

I cannot remember the books I've read anymore than the meals I have eaten, even so, they have made me.

Pareto Principle

Zipf's Law is a continuous form of the Pareto distribution.

Pareto Principle: 20 % of the causes are responsible for 80 % of the outcome.



Some Examples

In 1880, Pareto showed that appr. 80% of the land in Italy was owned by just twenty percent of the population.

The richest 20% of humans have 82.7% of the world's income. In the US, 20% of patients use eighty percent of the health care resources.

Like in language, where the most frequently used 18 percent of words account for over 80% of word occurrences.

The Principle of Least Effort

Balancing act between ease of language production and ease of comprehension:

A few words are used often and many many many words are used rarely.

This balancing act helps dissipate information load density on listeners, spacing out important vocabulary so that the information rate is more constant.

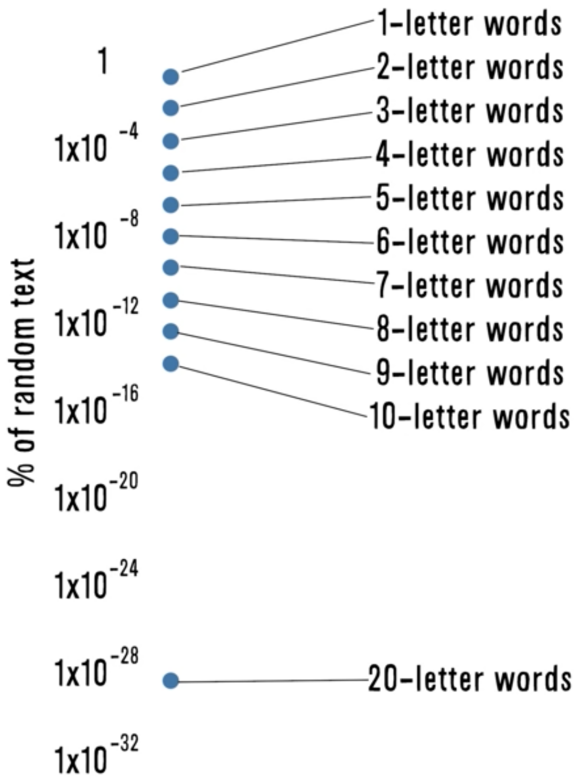
An Even Simpler Explanation

due to Benoit Mandelbrot

If you just randomly type on a keyboard you will produce words distributed according to Zipf's Law.

There are exponentially more different long words than short words. For instance, the English alphabet can be used to make 26 one letter words, but 26 squared two letter words. Also, in random typing, whenever the space bar is pressed, a word terminates. Since there is always a certain chance that the spacebar will be pressed, longer stretches of words are exponentially less likely than shorter words. The combination of these exponentials is pretty "Zipf-y".

Probability of the letter v



$$\frac{(26/27)^{(1-1)} \quad \text{x } (1/27)}{26^{(1)}} = .142\%$$

"V"

More generally ...

If we divide by the number of unique words of each length, we get the frequency of occurrence expected for any particular word given its length.

$$\frac{\frac{26}{27}^{(wordlength-1)} \times \frac{1}{27}}{26^{(wordlength)}}$$

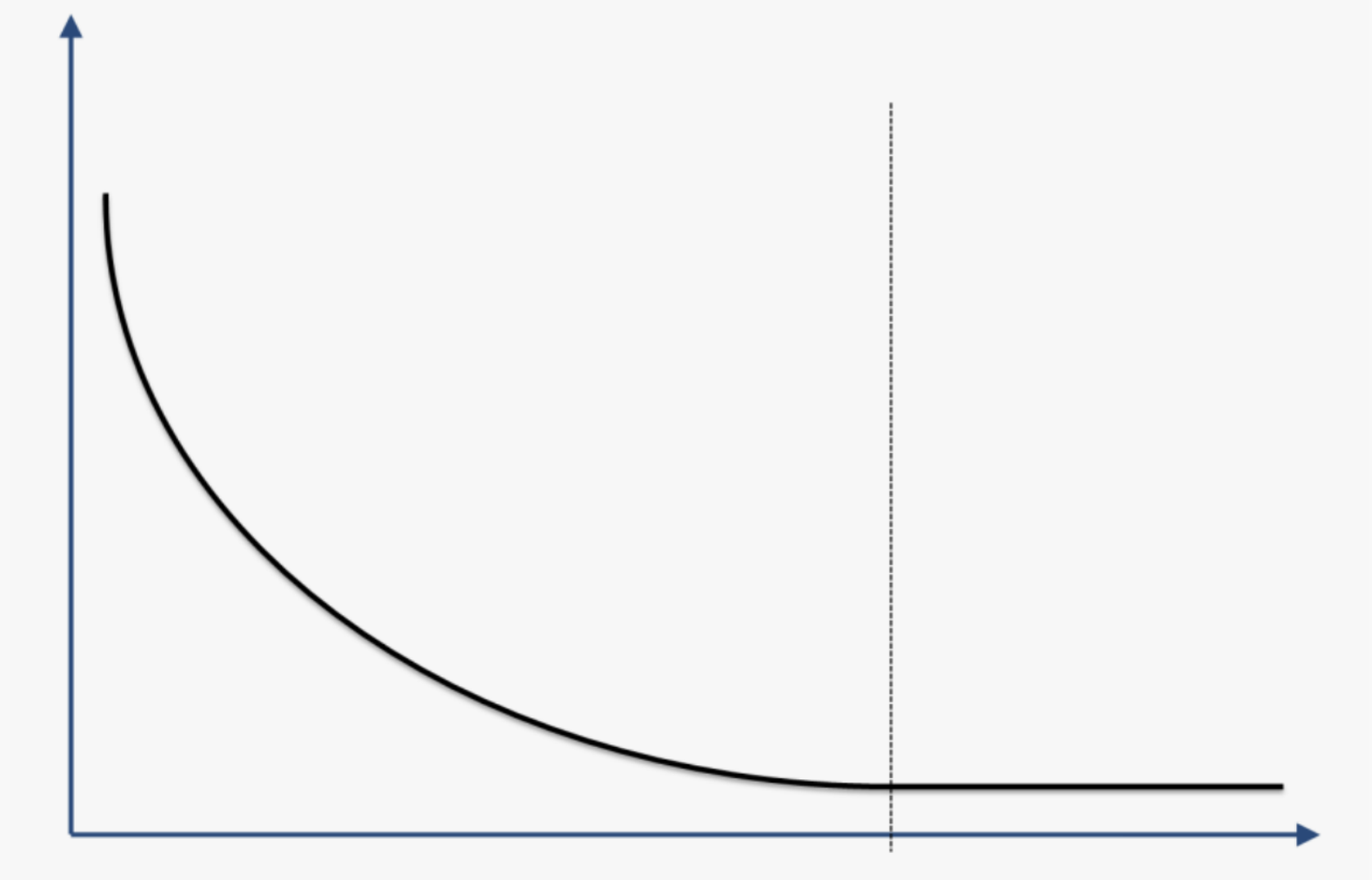
For example, the letter V will make up about 0.142 percent of random typing. The word "Vsauce" 0.0000000993 percent.

Power Law Distributions

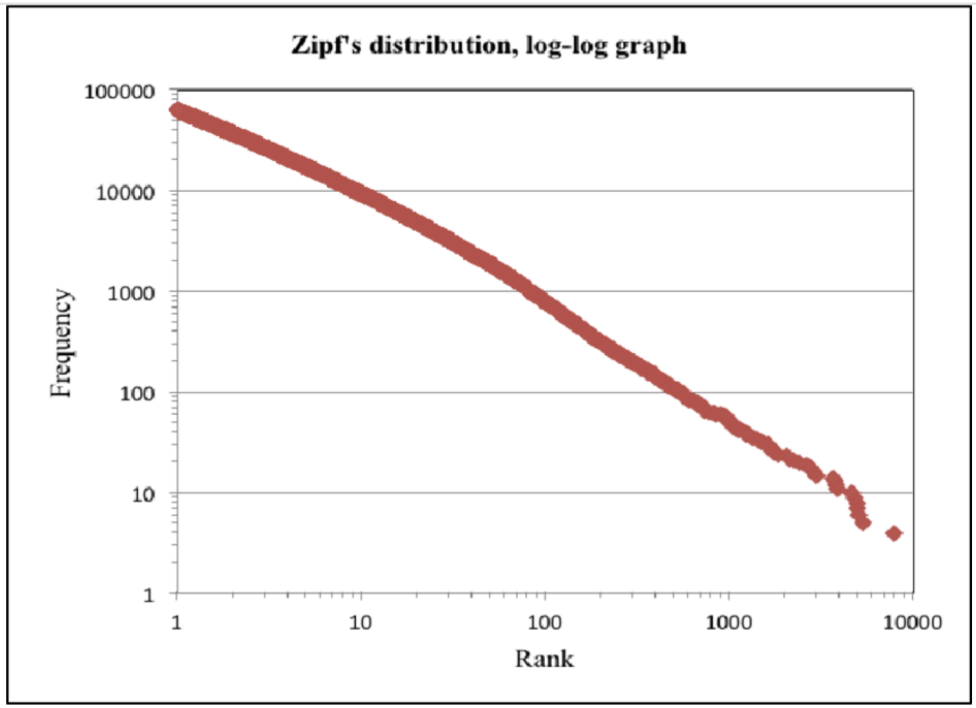
There are 26 possible one letter words, so each of the top 26 ranked words are expected to occur about this often. The next 676 ranks will be taken up by two letter words that show up about this often. If we extend each frequency according to how many members it has, we get a power law distribution:

$$\frac{1}{27} \cdots \frac{1}{27^2} \cdots \frac{1}{27^3} \cdots \frac{1}{27^4}$$

Power Law Graph



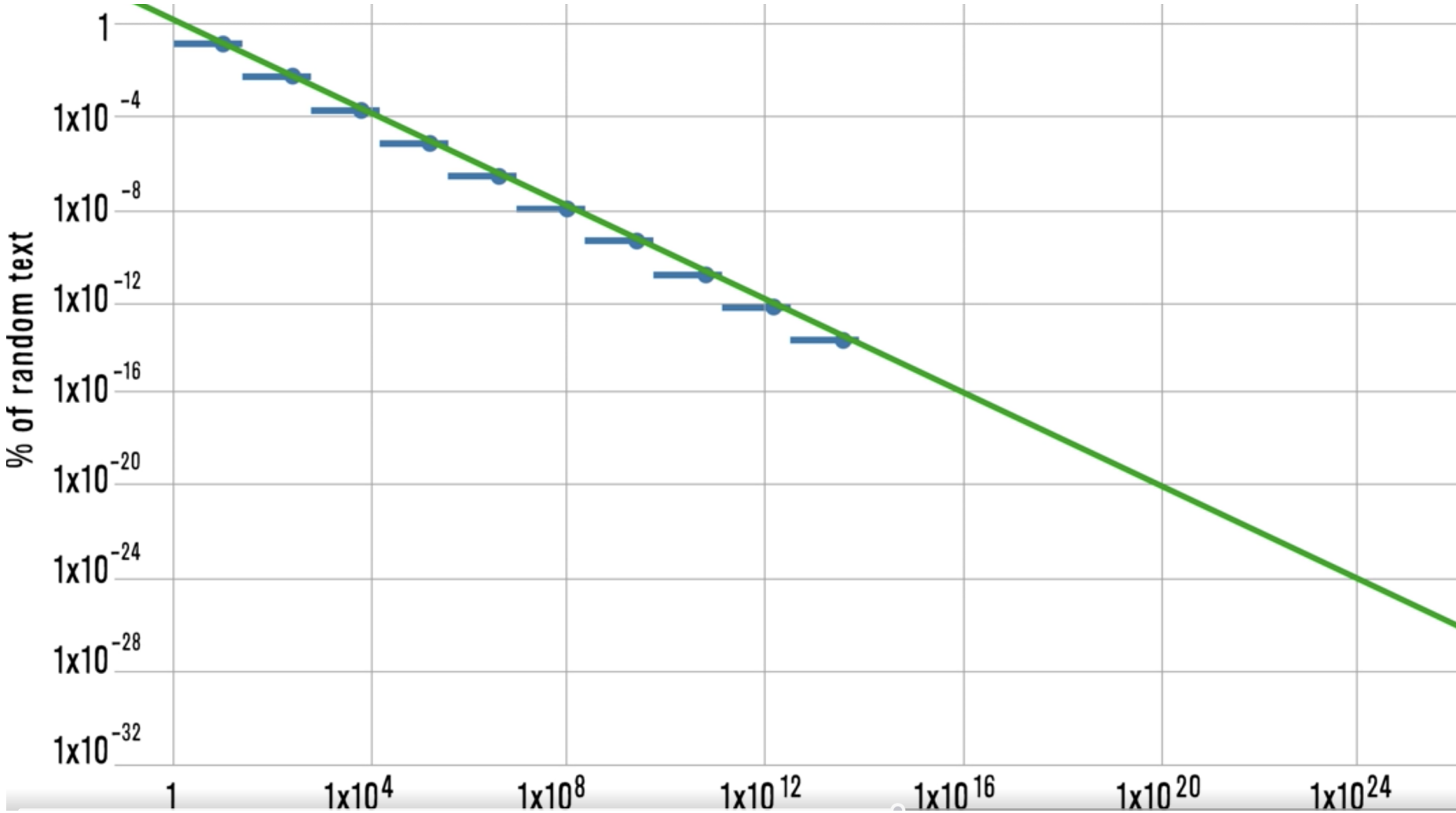
Visualizing Zipf's Law in log-log space



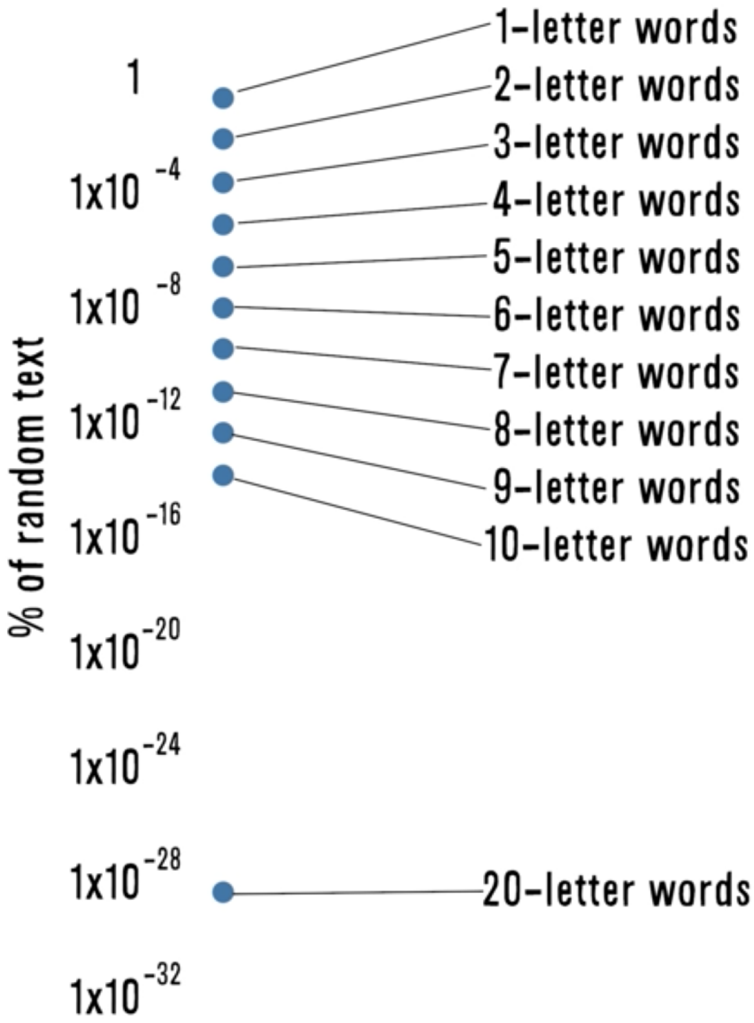
Zipf distribution log-log graph for types across the web-based corpus (Sharoff, 2006)

Word frequency and ranking in a log log graph follow a nice straight line, as is typical of a power-law.

Visualizing Zipf's Law in log-log space



Percentage of Random Text



$$\frac{(26/27)^{(\text{wordlength}-1)} \times (1/27)}{26^{(\text{wordlength})}}$$

Zipf's Law: General Formulation

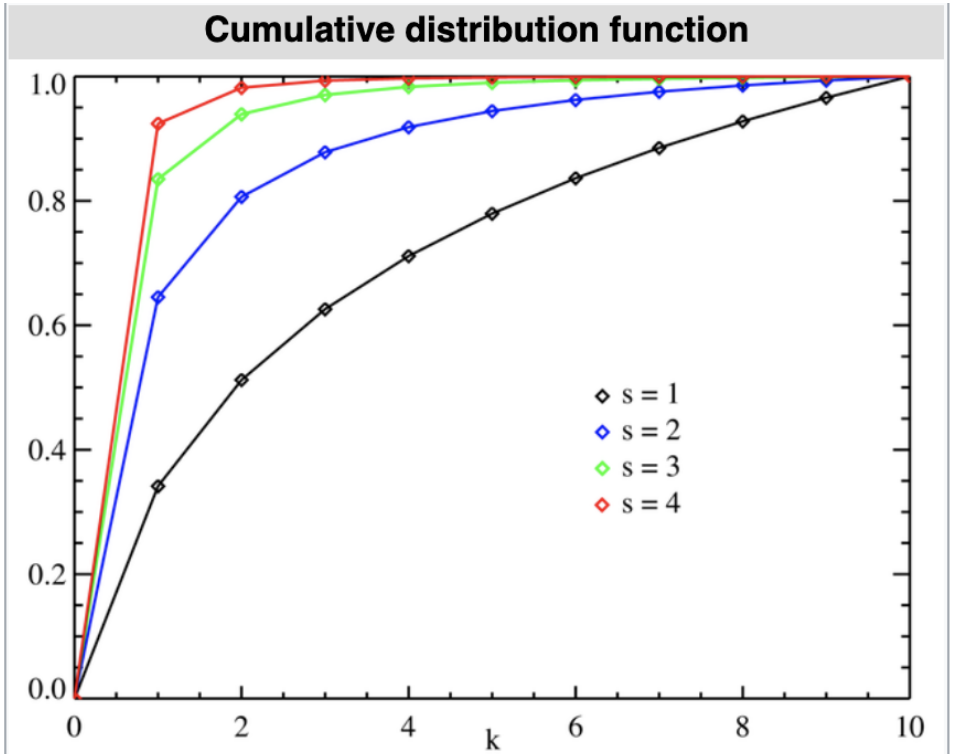
$f(k; s, N)$, where:

N	is the number of elements
k	their rank
s	the value of the exponent characterizing the distribution (at least 1).

More specifically:

$$f(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}}$$

Visualizing Zipf's Law as a Cumulative Probability Function



Zipf CDF for $N = 10$. The horizontal axis is the index k . (The function is only defined at integer values of k. The connecting lines do not indicate continuity.)

Source: wikipedia article on Zipf's Law

Wait a Minute ...

Actual language is very different from random typing:

- *Communication is deterministic to a certain extent. Utterances and topics arrive based on what was said before.*
- *The vocabulary that we have to work with certainly isn't the result of purely random naming.*
- *For example, the names of planets have a Zipfian distribution. This a fact about the real world, not a fact about language.*

Alternative Explanations for Zipfian Distributions

- Preferential Attachment Processes: occurs when something (e.g. money, views, attention, friends) is given out according to how much is already possessed.

- Critical Points: Writing and conversation often stick to a topic until a critical point is reached and the subject is changed and the vocabulary shifts. Processes like these are known to result in power laws.

Summary

So, in the end, it seems tenable that all these mechanisms might collude to make Zipf's law the most natural way for language to be. Perhaps some of our vocabulary and grammar was developed randomly, according to Mandelbrot's theory. And the natural way conversation and discussion follow preferential attachment and criticality, coupled with the principle of least effort when speaking and listening are all responsible for the relationship between word rank and frequency.

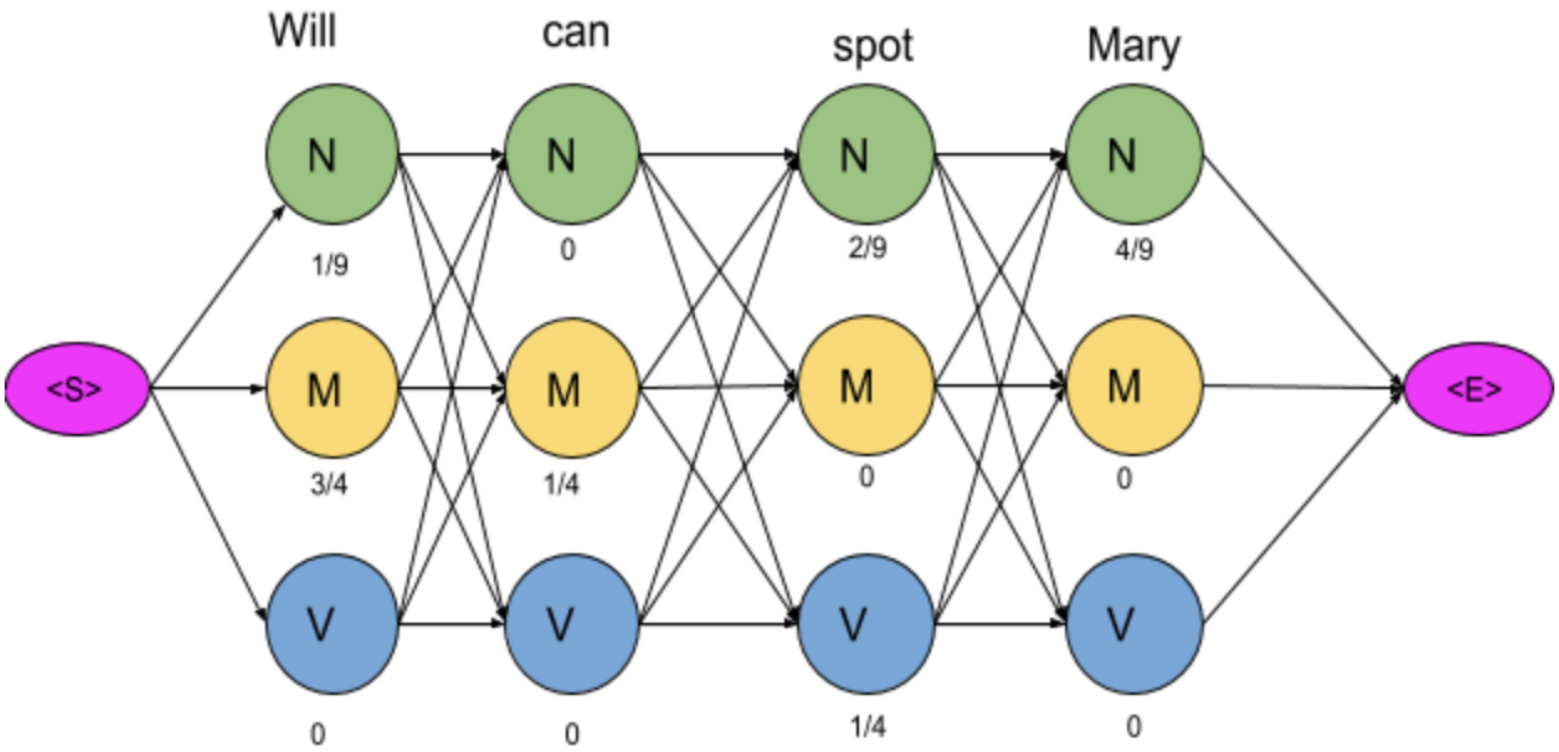
Andrey Markov

Source: https://de.wikipedia.org/wiki/Andrei_Andrejewitsch_Markow_(Mathematiker,_1856)

Russian Literature, Markov Chains, Markov Processes, and Hidden Markov Models

In his 1906 paper entitled: Распространение закона больших чисел на величины, зависящие друг от друга (rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga; English: Extension of the law of large numbers to dependent quantities), Andrey Markov examined 20,000 characters of Alexander Pushkin's novel Eugene Onegin by hand (!!) and encoded them in terms of Markov Chains in order to predict whether at any position in Pushkin's novel, the next letter would be a consonant or a vowel.

An Example of a Hidden Markov Model



Source:
https://www.mygreatlearning.com/blog/pos-tagging/