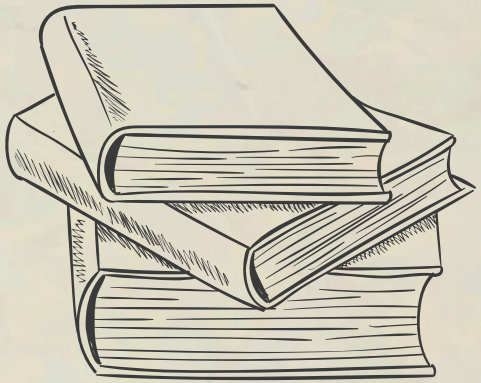
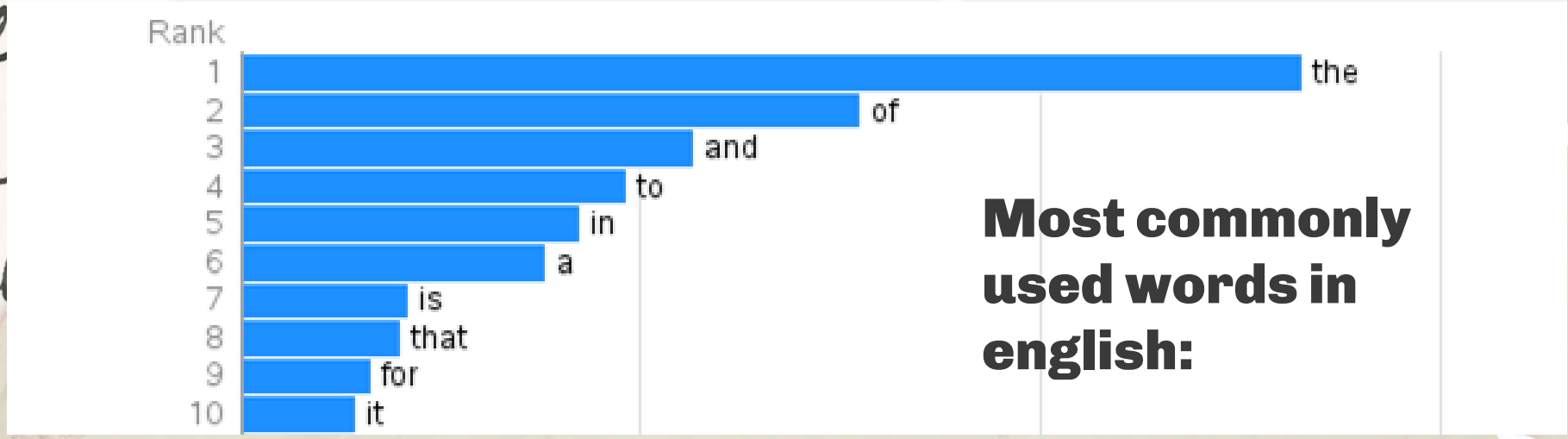
Zipf's Law

By Imogen



Why are words like ‘the’ ‘and’ and ‘of’ used so much more than other words?



Most commonly used words in english:

Most rarely used:

Agelast → a person who never laughs

Epizeuxis → a rhetorical term for the repetition of a word for emphasis.

What is Zipf's Law?

Formulated by American linguist George Kingsley Zipf, the law relates to the frequency of different words in the English language.

It states that:

The frequency of a word is inversely proportional to its rank.

Mathematical Formula

If you count the frequency of each word in a body, and then list the word types in decreasing order of their frequency, the relationship between frequency (f) and its position in the list (r) can be shown as:

$$f(r) = \frac{C}{r^s}$$

Where C is the frequency of the most ranked word and S is close to 1.

$$f(r) \propto \frac{1}{r}$$

General case where s = 1

So in theory:

Rank (r)	Example Word	Frequency f(r)
1	the	1
2	of	1/2
3	and	1/3
4	to	1/4
5	in	1/5

Example

Two friends were met by a bear. One climbed a tree, abandoning the other. The other played dead, and the bear left him unharmed.

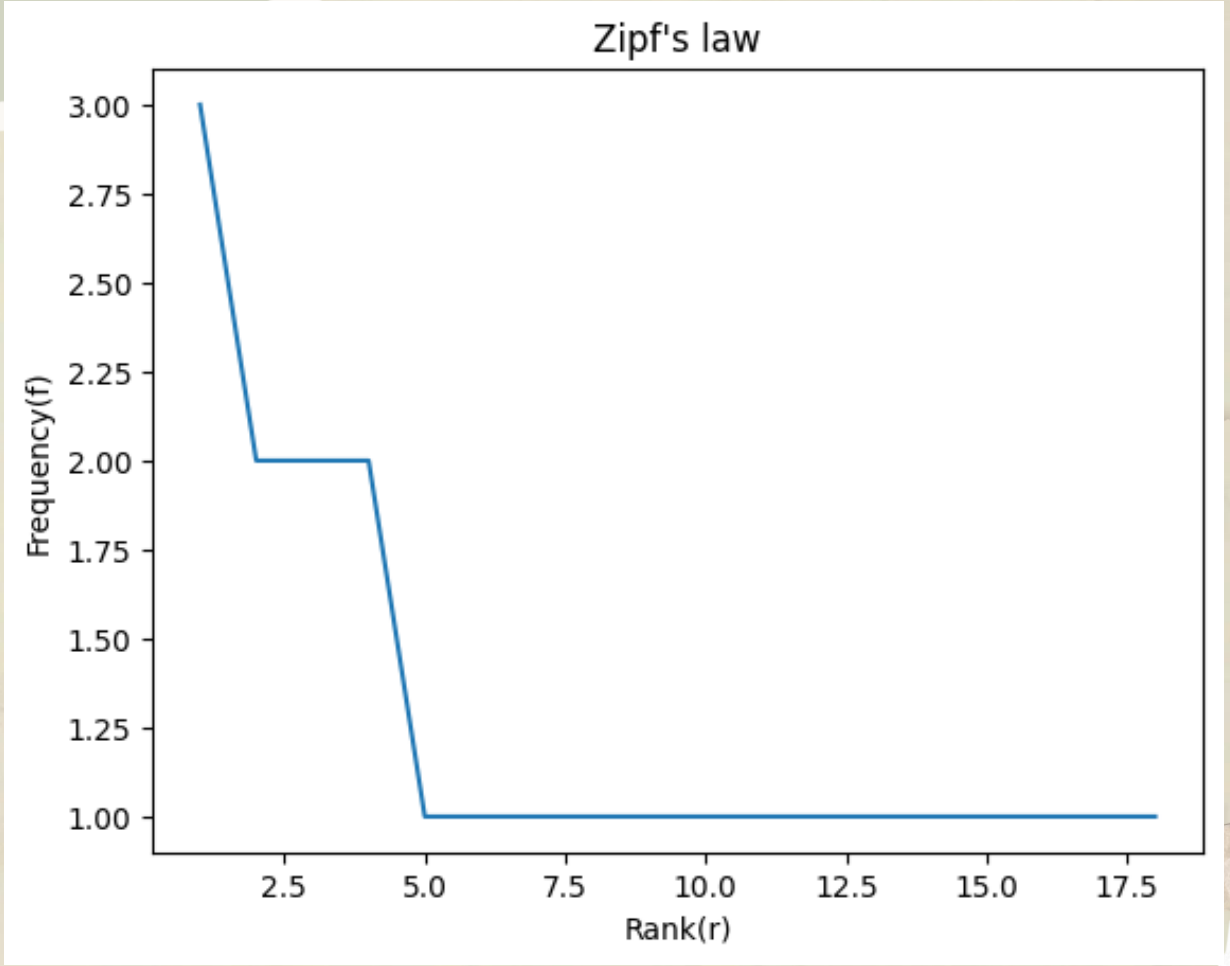
1. **‘The’** f = 3
2. **‘A’** f = 2
3. **‘Bear’** f = 2
4. **‘Other’** f = 2
5. **‘Two’** f = 1
6. **‘Friends’** f = 1
7. **‘Were’** f = 1
8. **‘Met’** f = 1
9. **‘By’** f = 1
10. **‘One’** f = 1
11. **‘Climbed’** f = 1
12. **‘Tree’** f = 1
13. **‘Abandoning’** f = 1
14. **‘Played’** f = 1
15. **‘Dead’** f = 1
16. **‘And’** f = 1
17. **‘Left’** f = 1
18. **‘Him’** f = 1
19. **‘Unharmed’** f = 1

Highest rank: ‘The’ (r = 1)

Lowest rank: ‘unharmed’ (r = 19)

The most common word has almost 2x the frequency of the second most common word.

Ie. $(f_{the} \approx 2f_a)$


Zipf's law
Frequency(f)
3.00
2.75
2.50
2.25
2.00
1.75
1.50
1.25
1.00
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Rank(r)

Wikipedia example:

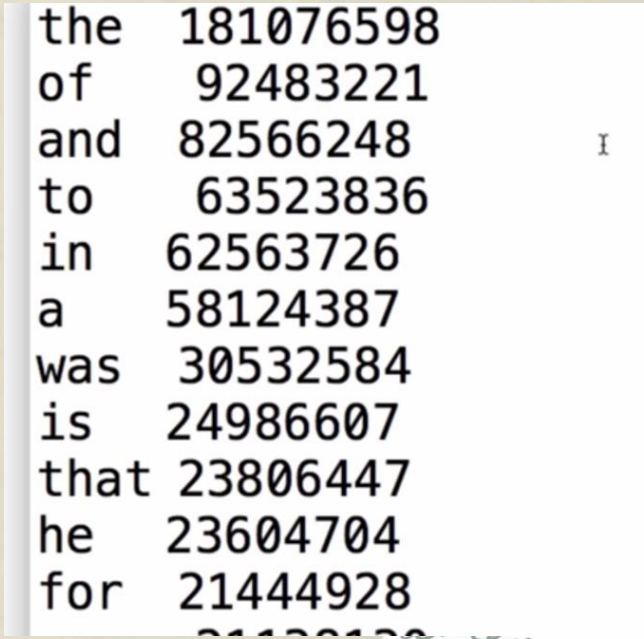
According to wordcount.org, ‘sauce’ is the 5555th most common english word.

This shows the frequency of words in Wikipedia and in the entire Gutenberg corpus of 1000s of public work books.

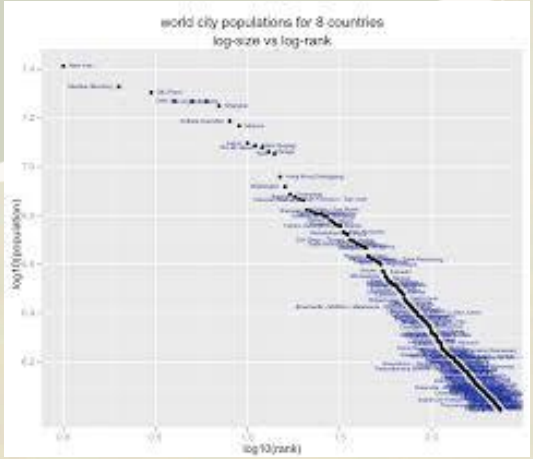
According to the formula:

181 million x 1/5555
= 181 million / 5555
= approx 30,000

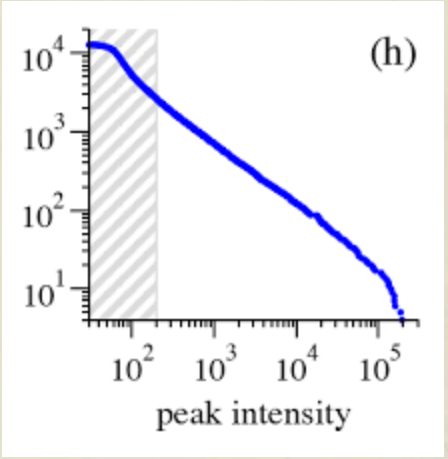
Therefore ‘sauce’ should appear approx 30,000 times.



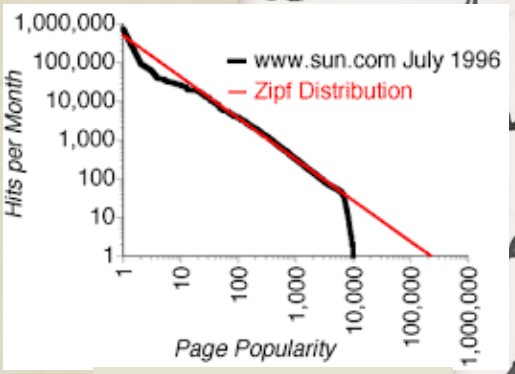
convoy 29622
parking 29611
gladly 29610
gerald 29608
bending 29604
clause 29595
decisive 29595
assumption 29594
sauce 29594
jose 29591
shapes 29580
whoever 29569



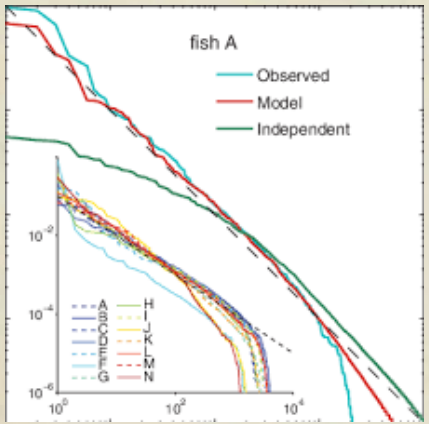
City populations



Solar flare intensities

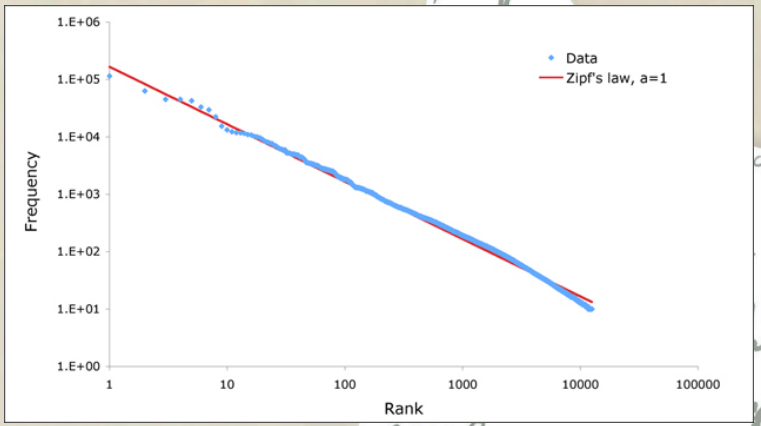


Website traffic



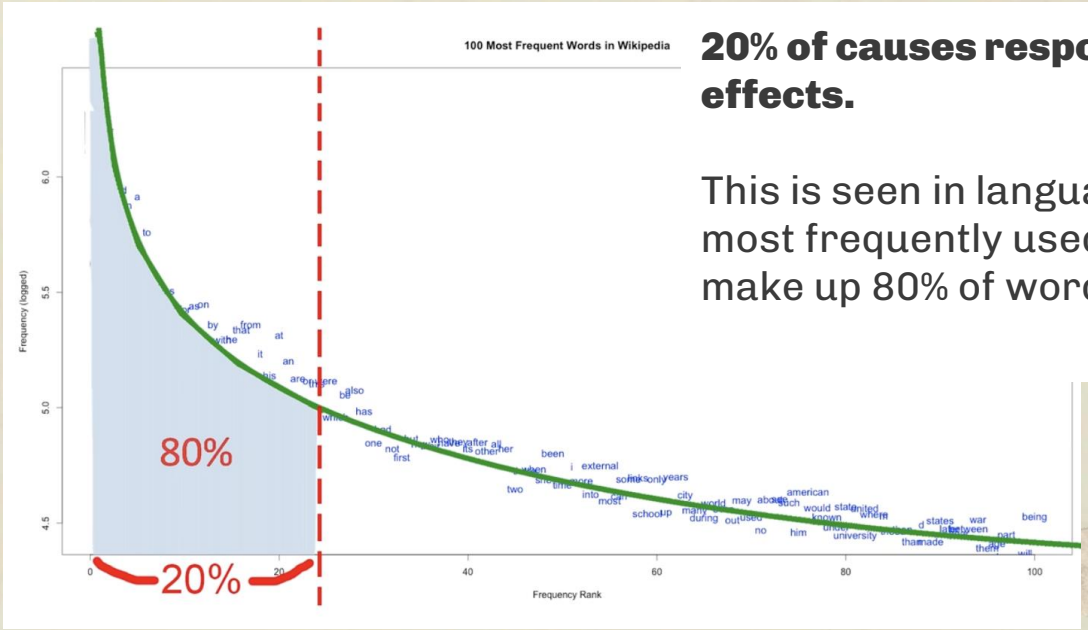
Protein sequences in immune receptors

Earthquake magnitudes



The Pareto Principle

Zipf's law, is a discrete form of the continuous Paleto distribution. It shows a cumulative total of the rankings seen in Zipf's Law.



20% of causes responsible for 80% effects.

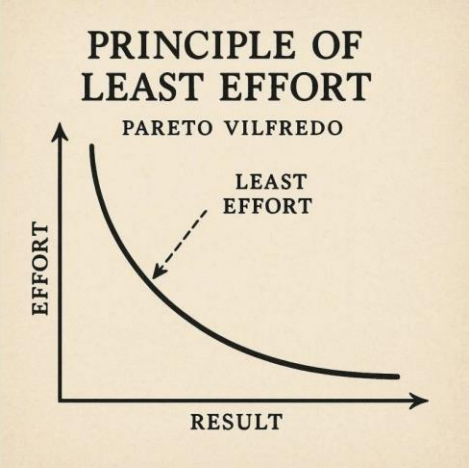
This is seen in language too! Where the most frequently used 18% of words make up 80% of word occurrences.

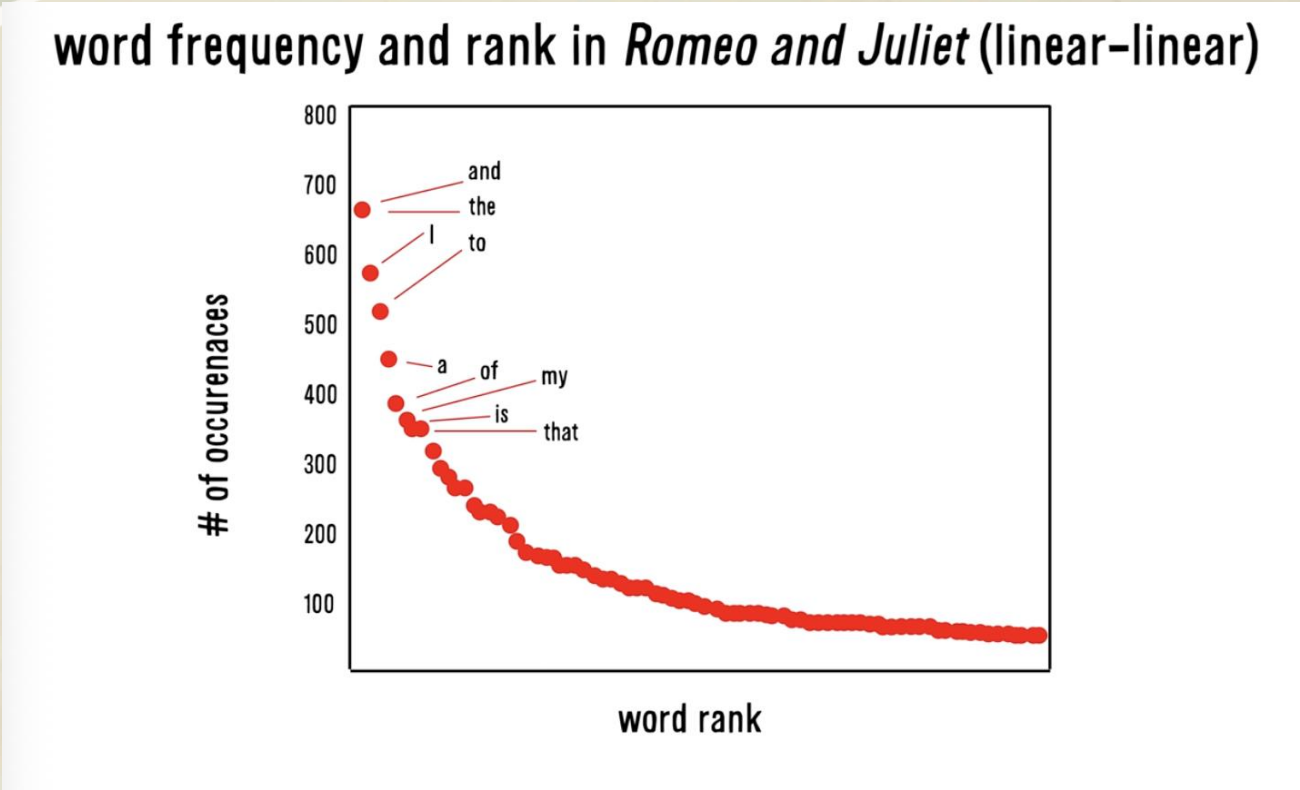
So why does it happen?

The Principle of Least Effort

Humans, systems and animals naturally gravitate towards the path of least resistance.

So speakers will tend to use as few words as possible, but listeners would prefer larger vocabularies with more specification.




word frequency and rank in *Romeo and Juliet* (linear–linear)
# of occurenaces
800
700
600
500
400
300
200
100
and
the
I
to
a
of
my
is
that
word rank

Tested with a small dataset:

A recent speech given by King Charles on his US state visit.

Results:

Rank (r)	Word	Frequency
1	the	209
2	in	206
3	of	164
4	and	129
5	two	93

The expected frequencies by Zipf's Law:

The second most common word should be about ½ as common as the most common word.

209 / 2 = 154.5

The third most common word should be about ⅓ as common as the most common word.

209 / 3 = 69.7

The reality:

Most common = the = 209

Second most common = in = 206, 206 > 154.5

Third most common = of = 164, 164 > 69.7

So why doesn’t Zipf’s law always work?

Deviation from the law

Variations from Zipf's law are common, and can be attributed to various factors:

- Too small a dataset / corpus (likely the case here)
- Grammar / content of writing
- The law observes a general trend rather than an exact rule
- Often East-Asian languages with many homophones deviate from the law.

Thank you for listening!