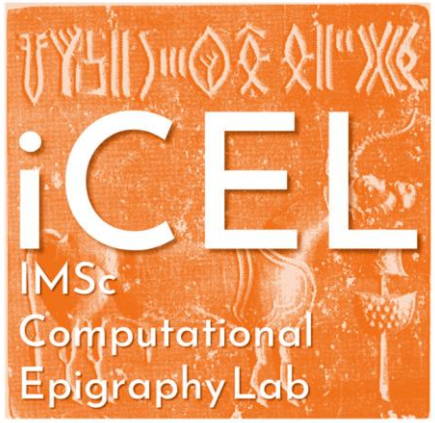
Bits & Scripts



Markov, Zipf, Shannon

Statistical Approaches to Analyzing Inscriptions

Part II

Sitabhra Sinha

Information

C E Shannon
A Mathematical Theory of Communication
Bell System Technical Journal (1948)

Shannon's Coding Theorem

Introduces the concept of information entropy to provide a fundamental limit of data compression for efficient transmission

Claude Shannon (1916-2001)

A Communications Primer (1953)

A film by Ray and Charles Eames

https://www.youtube.com/watch?v=kaDQlyXilSw

Information

C E Shannon
A Mathematical Theory of Communication
Bell System Technical Journal (1948)

$\Rightarrow$ information is a fundamental physical variable and provides a quantitative measure for it:

the bit (binary digit)

a state that can take any one of two possible values

Image:www.flowjuggle.com

Claude Shannon

Not Bitcoin, but

The Coin "Bit"

0

1

What is the information content of specifying the outcome of 4 successive coin tosses ?

Possible outcomes

Entropy: Information as "Surprise"

- a positive real number associated with an arbitrary finite probability space.
- measures the information carried by a random variable, or equivalently by a probability distribution

Let X be a random variable that can take values $\{a_1, \ldots, a_N\}$ having probabilities of occurrence $P(a_i) = P(X = a_i)$

$\Rightarrow$ The entropy of X (or equivalently of P) is

$$H(X) = H(P) = -\sum_{i=1}^{N} P(a_i) \log_2 P(a_i)$$

measures the average number of bits of information contained in a sample of the random variable X

Example: If a process always goes to the same final state, H = 0

Example: If a process has equal probability to converge to one of N possible states, H = ?

Entropy of Language

If a language is expressed in binary digits (0 or 1) in the "most efficient manner" its entropy H is the average number of bits required per letter of the original language

But,

natural languages have lots of redundancies

Example: "CAREFUL" can still be guessed if we lose the alternate letters "C_R_F_L" especially if we know the context

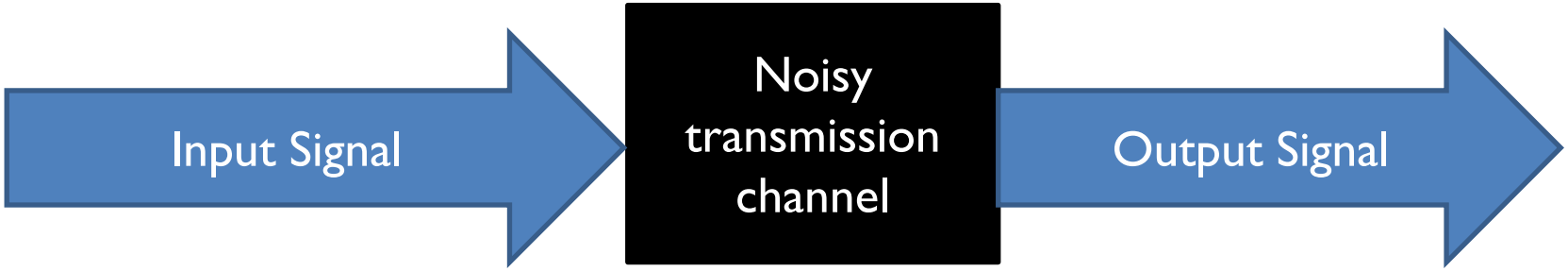
In fact, languages like Arabic or Hebrew using *abjad* writing systems (Daniels, 1990) use this principle to represent only consonants, leaving vowels to be inferred

So why do alphabets use vowels ? Stupidity ?

“Chinese whispers”

Why redundancy is useful

In reality any signal being transmitted is subject to “noise” (fluctuations, disturbances) with a finite probability of corruption



Example:

0100100 → 0_00100 → 0_01100 → 0_011001

Deletion mutation

Substitution mutation

Insertion mutation

To protect the information

000111000000111000000

Each bit repeated thrice

Redundancies allow robustness against information loss via noise

Problem : To transmit information through many generations robustly in pre-literate societies

E.g., in early Iron Age India, forms of chanting to ensure that Vedic texts were orally transmitted from generation to generation with high fidelity (F Staal)

Myths encode reality

"Our oldest written records date to 5,200 years ago, but we have been speaking and mythmaking for perhaps 100,000."

"Myths originally transmitted real information about real events and observations, preserving the information sometimes for millennia within nonliterate societies […] a quite reasonable way to convey important messages orally over many generations — although reasoning back to the original events is possible only under rather specific conditions."

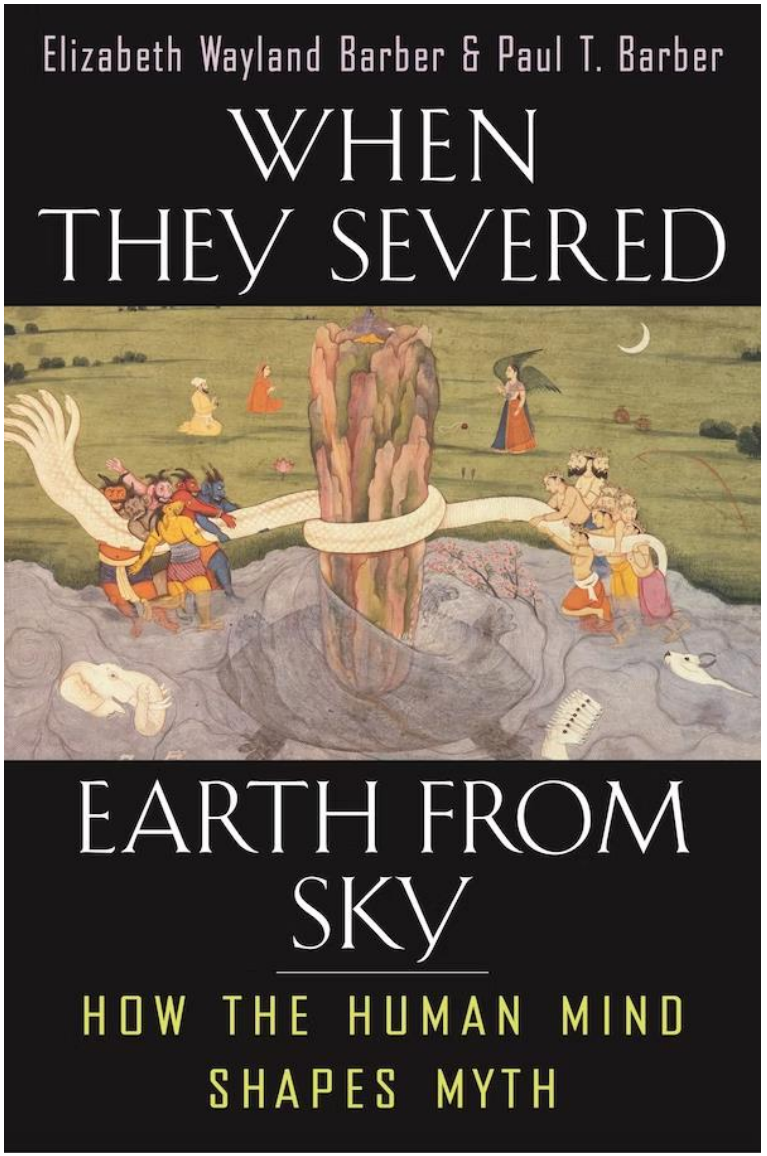


Image:Princeton University Press

Lord of the Rings : Lighting of the Beacons

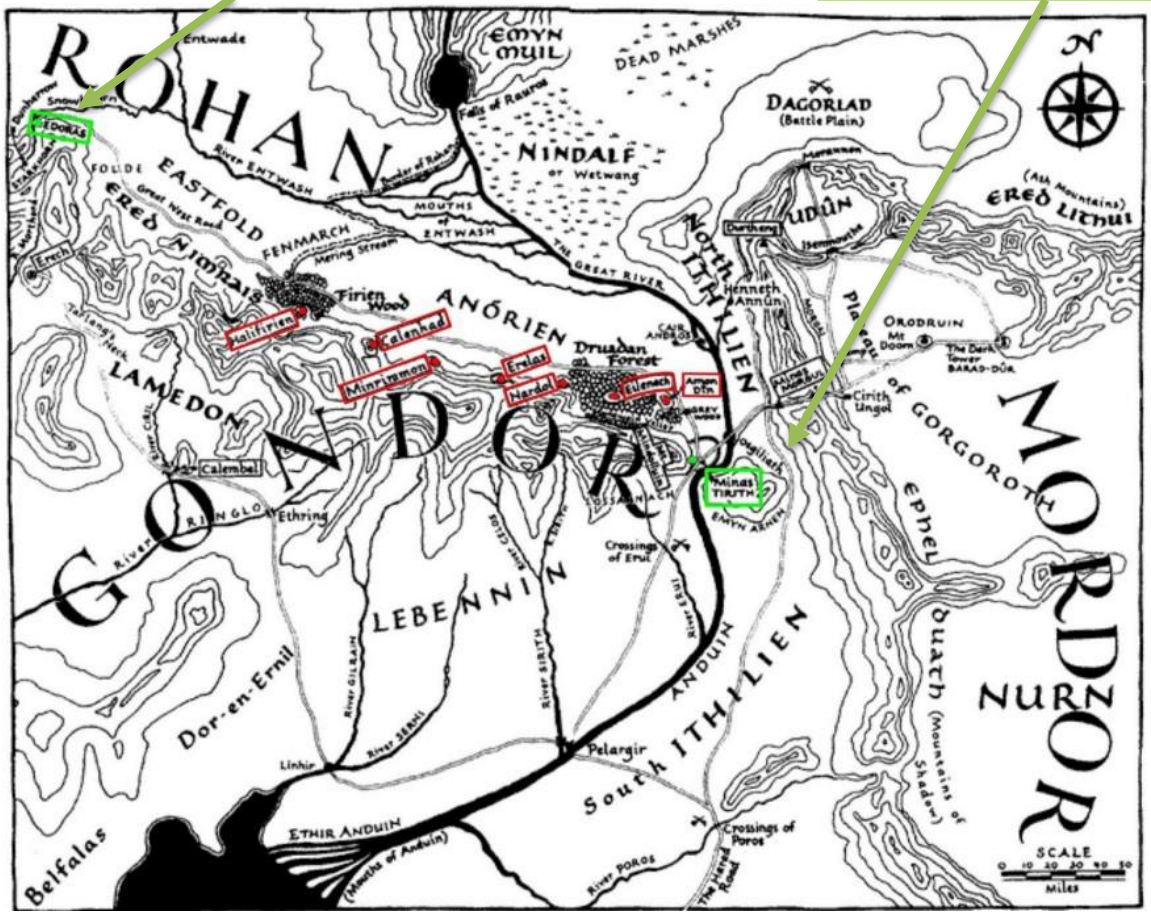
Information content of the message sent from Gondor to Rohan ?

Lord of the Rings : Return of the King
dir. Peter Jackson, music: Howard Shore

https://www.youtube.com/watch?v=i6LGJ7evrAg

Destination: Edoras

Origin: Minas Tirith

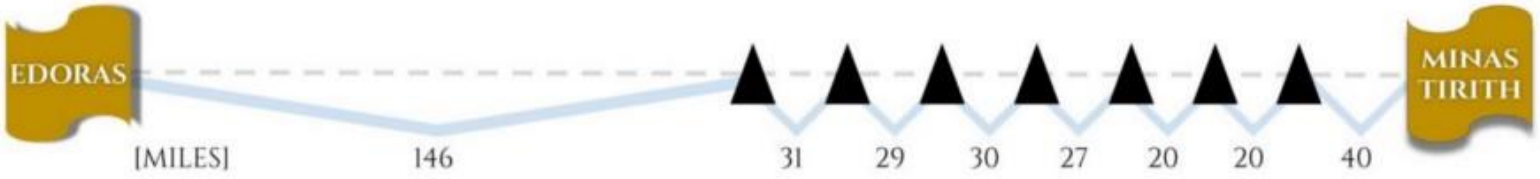


"The meaning of the message had to be effectively condensed into a single bit. A binary choice, *something* or *nothing*. …To transmit this one bit required immense planning, labor, watchfulness, and firewood"

James Gleick, *Information*

Map of Rohan, Gondor, and Mordor (J R R Tolkien)

Beacon locations marked in red



https://abrown18-68137.medium.com/how-long-did-it-take-for-gondor-to-call-for-aid-1e5997fede74

Interpreting entropy as a measure of information

Image: ethw.org

Ralph Hartley

Hartley (1928): for a message of n symbols, each chosen from an alphabet of size s, amount of information sent is

$$H = n \log s$$

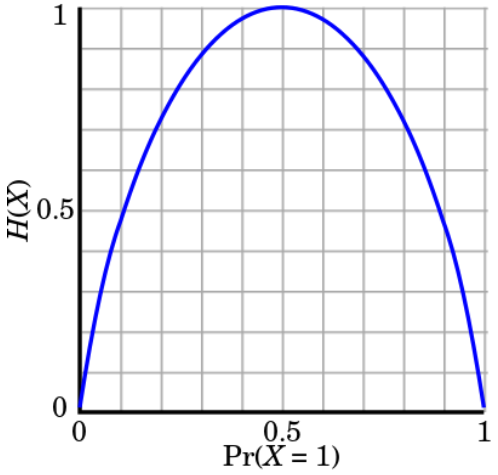
Why log ?

If different parts of a message are independently produced,

Information should be additive, i.e., $H(ab) = H(a) + H(b)$

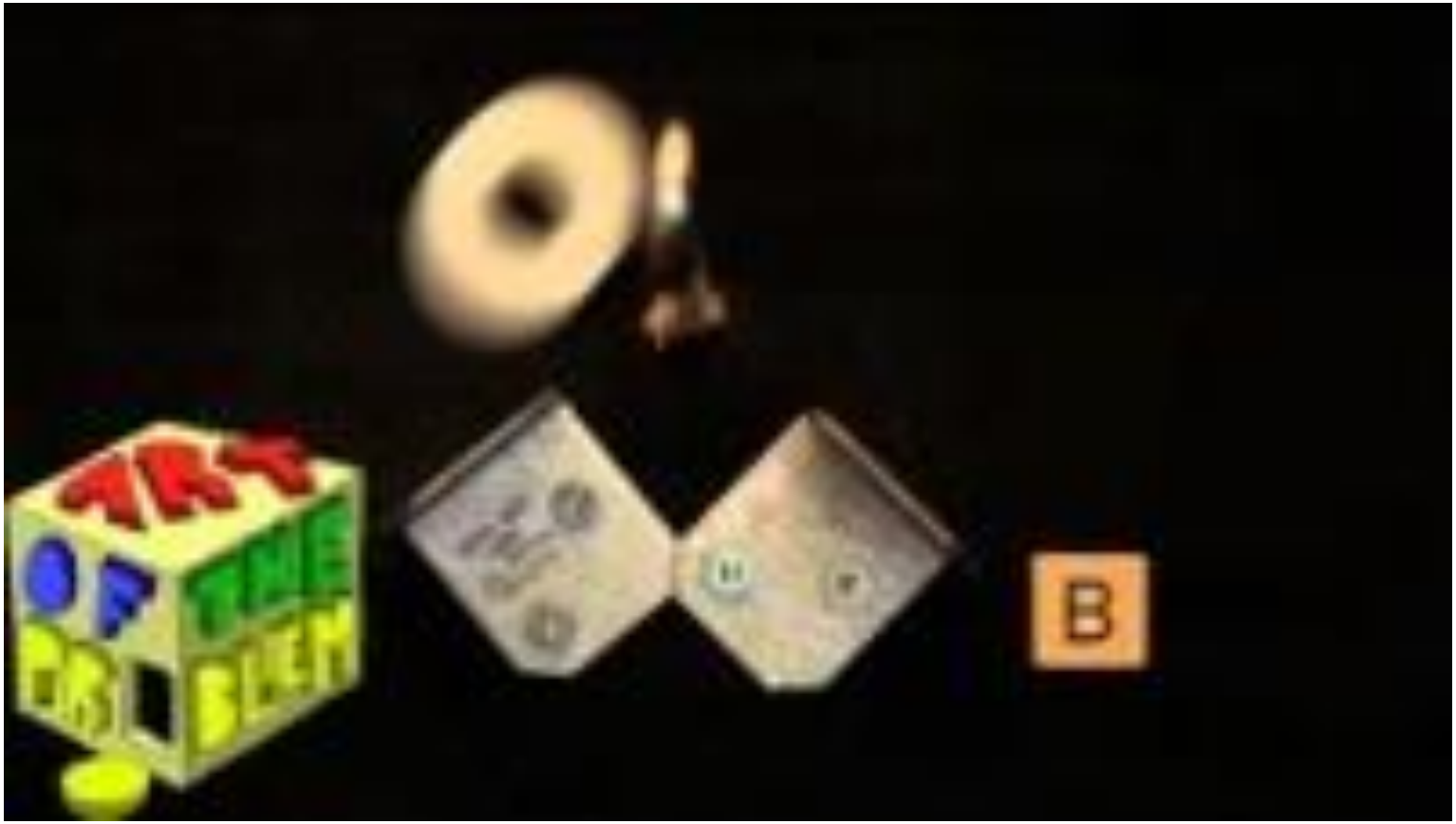
Shannon (1948): $H = -\sum_i p_i \log p_i$

Entropy is the minimum number of YES/NO questions one needs to ask to determine completely the content of the signal



Shannon's Information Entropy

A video by *Art of the Problem*



https://www.youtube.com/watch?v=R4OlXb9aTvQ

Entropy of English

Prediction and Entropy of Printed English
By C. E. SHANNON
THE BELL SYSTEM TECHNICAL JOURNAL, JANUARY 1951

If only the 26 letters of English are considered, naïve estimate of entropy is

$log_2 26 = 4.7$ bits per letter

But all letters do not occur with same probability !

Important for cryptanalysis
used by Sherlock Holmes to crack the cipher messages in "The Adventure of the Dancing Men"

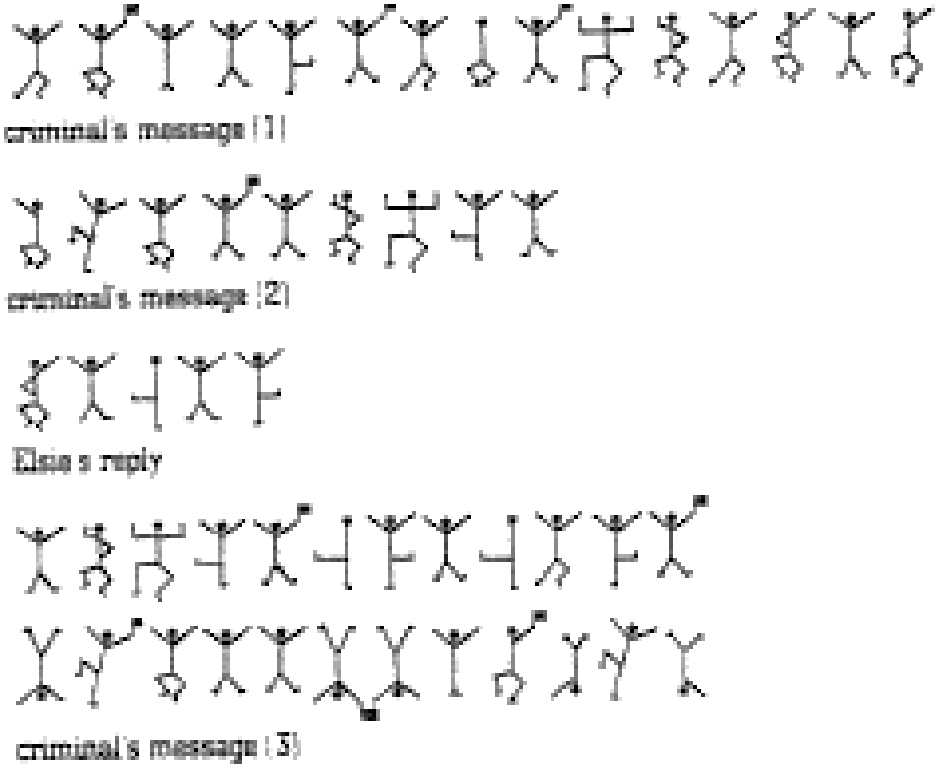
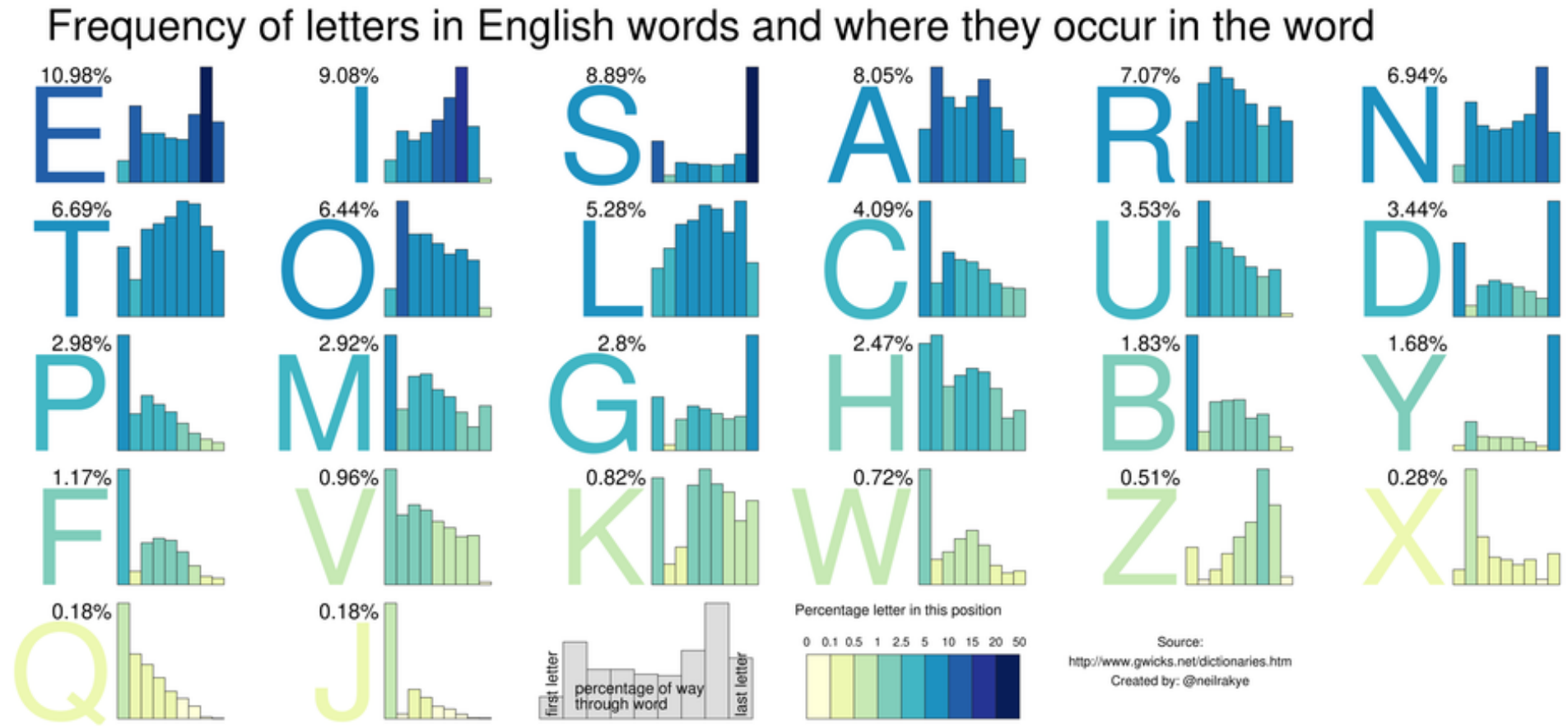


Image: www.shodor.org



Source: www.reddit.com/r/dataisbeautiful/comments/lot486/frequency_of_letters_in_english_words_and_where/

Entropy of English

Prediction and Entropy of Printed English

By C. E. SHANNON

THE BELL SYSTEM TECHNICAL JOURNAL, JANUARY 1951

If only the 26 letters of English are considered, naïve estimate of entropy is

$log_2 26 = 4.7$ bits per letter

Including letter frequencies in our calculation we get a improved estimate

$$F_1 = -\sum_{i=1}^{26} p(i) \log_2 p(i) = 4.14 \text{ bits per letter}$$

But even this ignores pairwise correlations between letters, e.g., the fact that vowels occur with high probability next to consonants !

To take into account such correlations between n letter clusters or n-grams

Entropy of English

One method of calculating the entropy H is by a series of approximations $F_0, F_1, F_2, \cdots$, which successively take more and more of the statistics of the language into account and approach H as a limit. F_N may be called the N-gram entropy; it measures the amount of information or entropy due to statistics extending over N adjacent letters of text. F_N is given by[1]

$$F_N = -\sum_{i,j} p(b_i, j) \log_2 p_{b_i}(j)$$
$$= -\sum_{i,j} p(b_i, j) \log_2 p(b_i, j) + \sum_i p(b_i) \log p(b_i) \quad (1)$$

in which: b_i is a block of N-1 letters [(N-1)-gram]

j is an arbitrary letter following b_i

$p(b_i, j)$ is the probability of the N-gram b_i, j

$p_{b_i}(j)$ is the conditional probability of letter j after the block b_i, and is given by $p(b_i, j)/p(b_i)$.

Conditional entropy of the next letter j when the preceding N – 1 are known

Thus, trigram entropy

$$F_3 = -\sum_{i,j,k} p(i, j, k) \log_2 p_{ij}(k)$$
$$= -\sum_{i,j,k} p(i, j, k) \log_2 p(i, j, k) + \sum_{i,j} p(i, j) \log_2 p(i, j)$$ = 3.3 bits per letter

For longer N-grams, approximate by looking at word frequency distributions

Entropy of English and Zipf's law

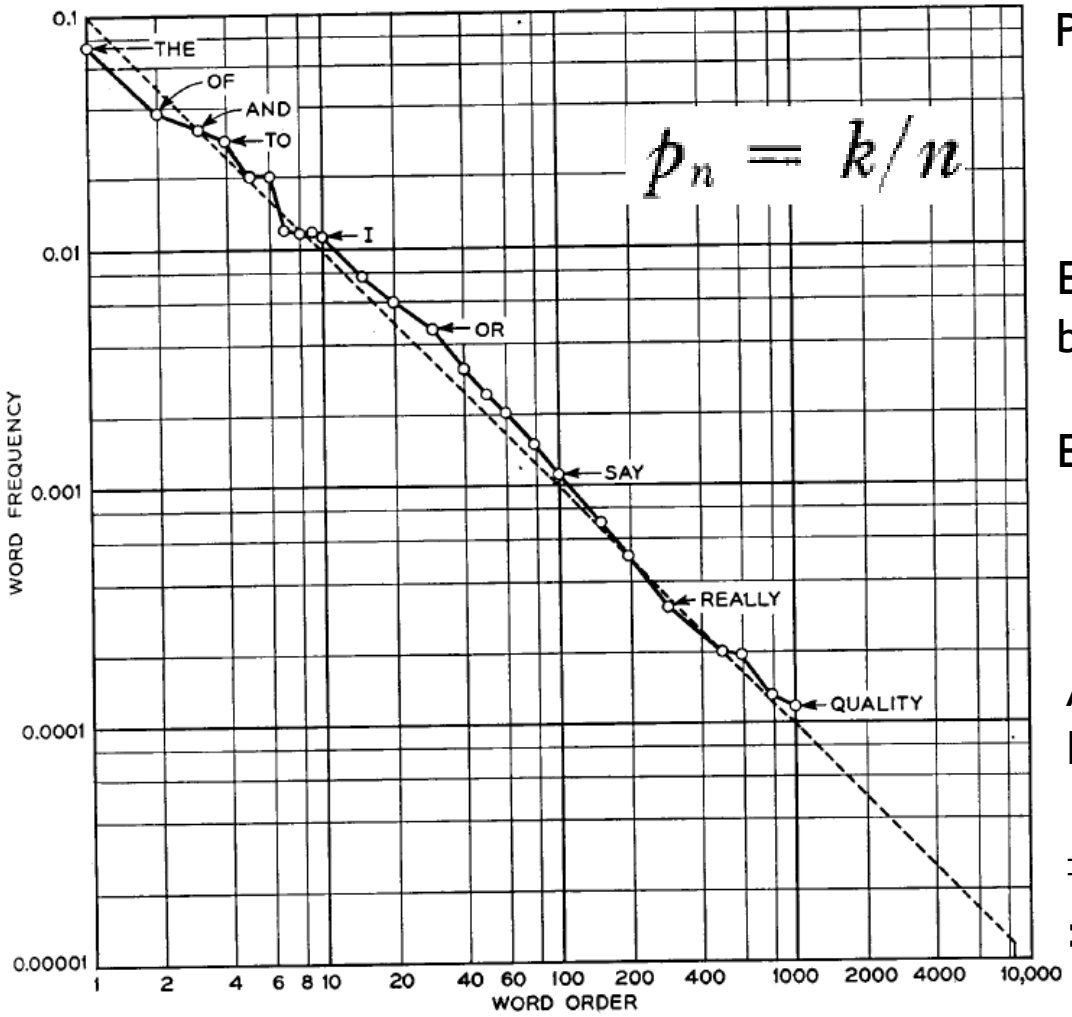


Fig. 1—Relative frequency against rank for English words.

Probability of the n^{th} most frequent word

$$p_n = \frac{.1}{n}$$

Estimating from data that the largest n beyond which $p_n = 0$ is n ~ 8727,

Entropy $-\sum_{1}^{8727} p_n \log_2 p_n = 11.82$ bits per word

Approximating average word length in English to be 4.5 letters,

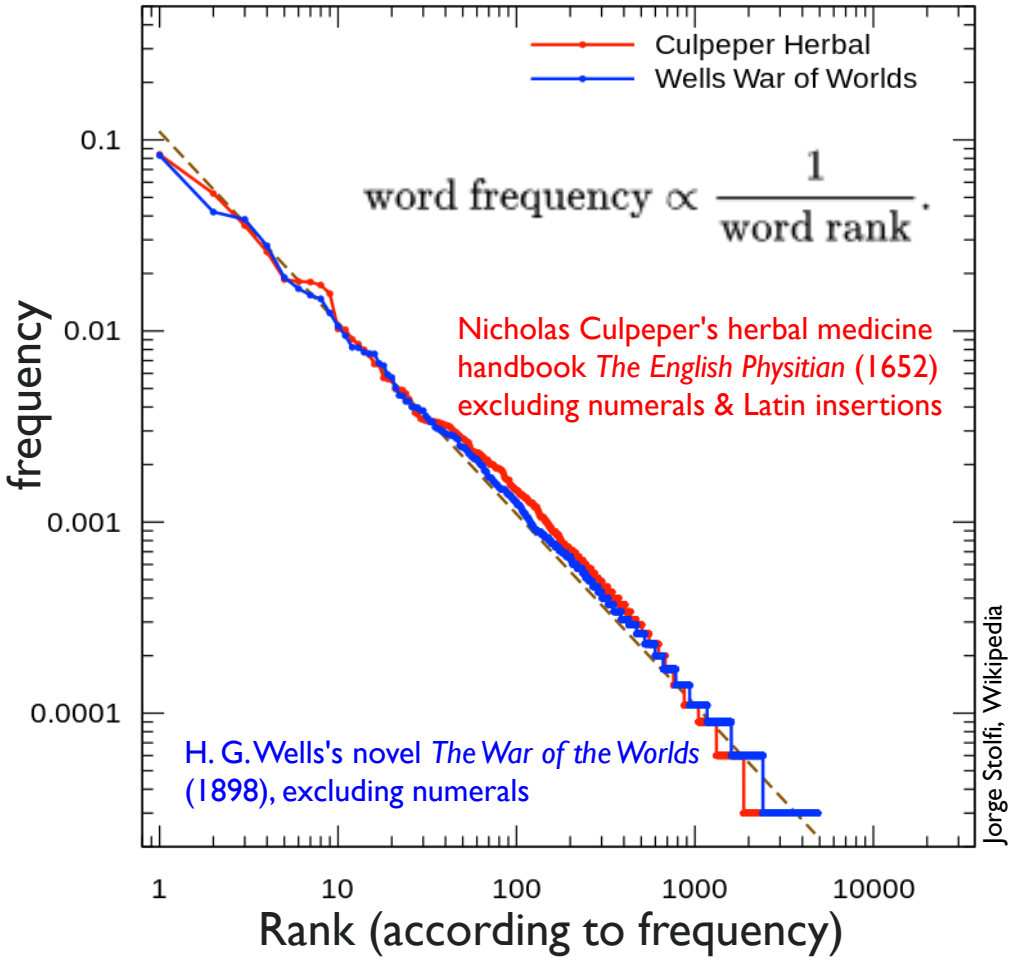
$\Rightarrow$ Entropy = 11.82/4.5 bits per letter

= 2.62 bits per letter

Zipf & The Principle of Least Effort

1935: American linguist George K Zipf stated that a universal property of human languages is that "the magnitude of words stands in an inverse (not necessarily proportionate) relationship to the number of occurrences"

George Kingsley Zipf (1902-1950)



Zipf's law of abbreviation:

the more frequent a word is, the shorter it tends to be

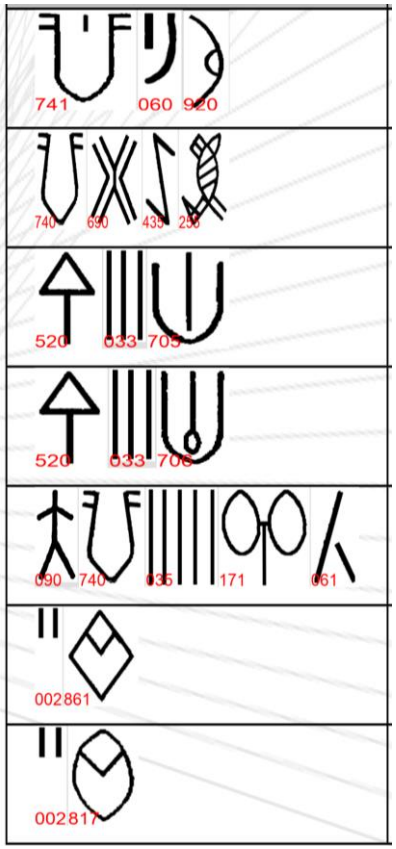
Zipf (1949) → this universal design feature is a result of optimizing form-meaning mappings under competing pressures to communicate accurately as well as efficiently

Principle of Least Effort

Ashraf: Constructing a kernel lexicon of recurring segments

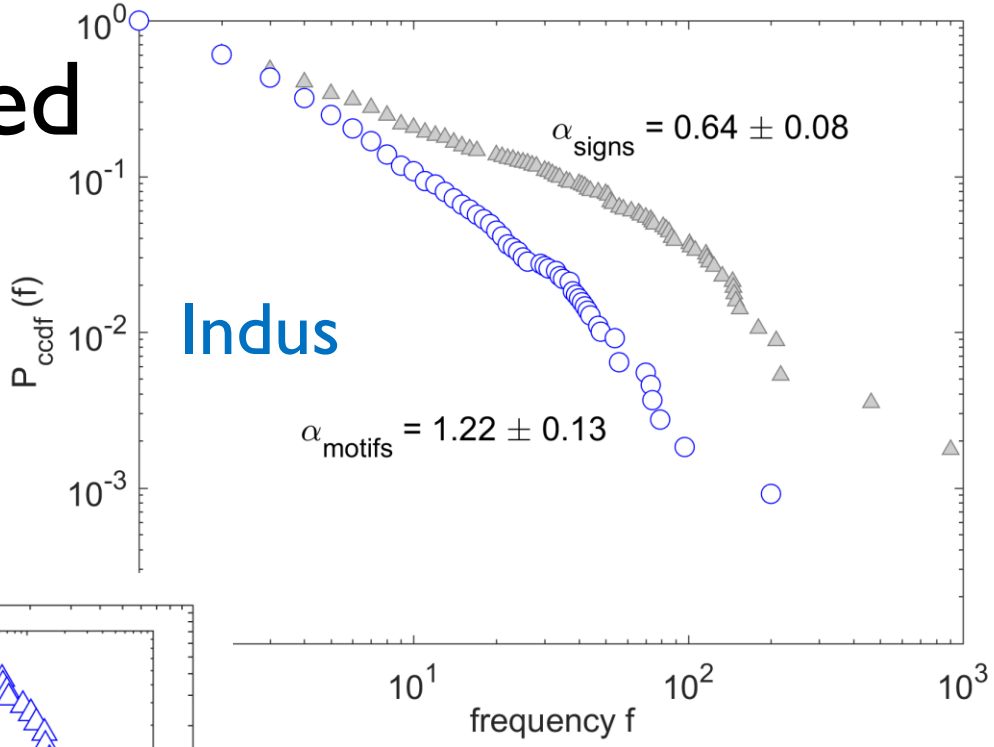
English

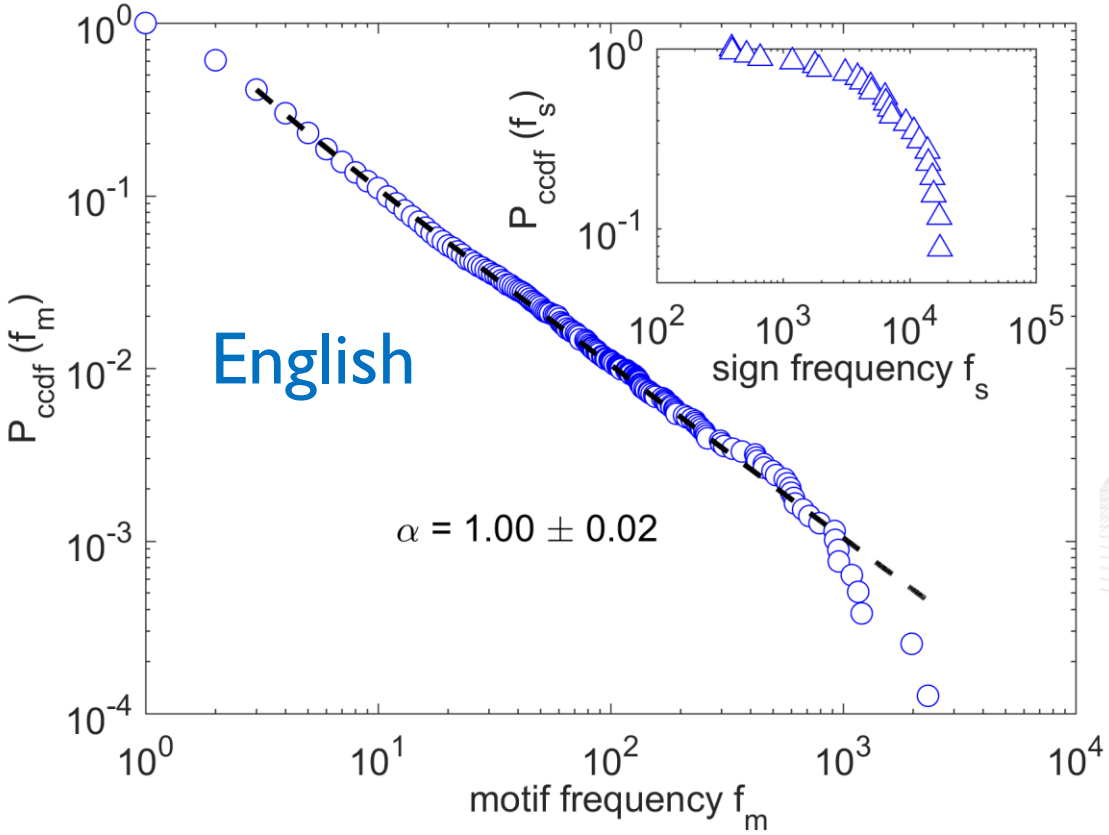
fragment of the reconstructed Indus kernel lexicon

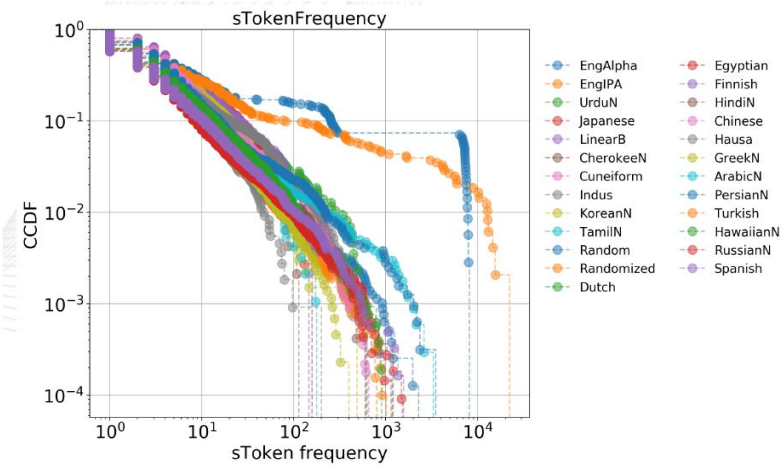


Zipf's law generalized

The distribution of frequencies with which the motifs occur in words within a linguistic corpora appear to follow Zipf's law very closely across various languages & writing systems







An Universal Pattern

how "words" are composed using motifs

Words in various languages using different writing systems – despite large differences in their length measured in terms of the number of individual signs they involve – are composed of almost the same number (~ 3 on average) of motifs [significant sign clusters] which are used in various combinations to make up different words