Word lengths are optimized for efficient communication

Steven T. Piantadosi, Harry Tily, and Edward Gibson

- Research Question: **Are word lengths optimized for efficient communication?**
- Starting point: a theory by George Kingsley Zipf
  - *not* the commonly known Zipf's law: **the frequency of a word is inversely proportional to its rank**. The second most-common word appears half as many times as the first, the third most-common word one third as many times as the first, etc.
  - Zipf's "other thoery": **the length of a word is inversely proportional to its frequency**
  - Derived from Zipf's principle of least effort. Decrease effort by making most frequently uttered words shorter.
  - Authors test this and the correlation is not strong. In English: about 0.1
- Proposal: Update Zipf's law by stating that word lengths are a property of its *information content*. Defined as:

$$-\sum_{c} P(C=c|W=w) log P(W=w|C=c)$$

  The left term represents the probability of a context given a randomly chosen word; the logarithm on the right is the information content of the word given that context. It is the total informativeness of a word across all contexts, weighted by the probability of that context. Note that this is equivalent to:

$$-\frac{1}{N}\sum_{i=1}^{N} log P(W=w|C=c_i)$$

  which is the sum of the negative log-probability for each instance of the word, given the context. In this paper, a "context" is the preceding n words.
  - The cute, fuzzy ...
  - *dog*: high probability, low negative log probability, low surprisal, low informativeness
  - *scorpion*: low probability, high negative log probability, high surprisal, high informativeness
- Using Google ngrams, authors calculated information content of each word. Correlation with word length is much higher: 0.3 for English
  - Superiority of information context over frequency demonstrated in all 11 languages when $n = 2$, 10/11 languages when $n = 3$ (exception Polish), 7/11 languages when $n = 4$. (lower performance with higher error due to estimation error?)
- Frequency and Information are not unrelated. Using a technique called *partial correlation*, authors show that frequency is better understood as a consequence of information content.
  - More frequenty spoken words tend to have lower information content.
  - By modeling word length from information content, one gets a clearer picture
- Why this is the case: *the principle of uniform density*
  - This principle, which is attested to by much previous research (see paper citations), holds that when communicating, humans tend to keep the rate of information per time constant.
  - Let the length of the word be a proxy for how long it takes to produce the word audibly. High-information words are longer because it allows the speaker to "spread the information out" over time.

  - Since short words are low information-content, alternating short and long words keeps the information density constant

- **Conclusion:** the title of the paper - yes, word lengths *are* optimized for efficient communication, such that they keep the information density per unit time uniform.

- **Discussion**

  1. Bias in language sample: preference for European languages. 5 Romance, 4 Germanic, 2 Slavic. Agglutantive languages, like Turkish? Syllable languages, like Mandarin?
  2. Can languages be artificially optimized (for good, not evil)? Esperanto; common scientific language a la Lazebnik
  3. Can *artificial* languages be optimized? C, Java, Python
  4. How has the principle of uniform information density transferred into the written world?
  5. Let us assume this optimization is the result of a process analogous to evolution (recalling meme paper from last week). How can different forms of speaking be subject to selection pressures? What is the fitness function?