

Lay Summary

In most languages, words that are used more frequently tend to be shorter than words that are used less frequently. In English, for example, “the”, “it”, and “and” are among the most frequent words, while words like “enhancement”, “hypotenuse”, and “carbuncle” are among the least frequent. The linguist George Zipf hypothesised that this inverse relationship between word length and frequency is the result of an optimisation process. Speakers are lazy, and inclined to produce the shortest utterances possible, even if that means risking ambiguity or potentially being misheard. Hearers, on the other hand, prefer as little ambiguity as possible, to ensure they can understand what is being said, even if it means the speaker has to use longer, less easily confusable words. An optimal solution for balancing these two competing pressures would be to keep the most frequently used words short, while allowing less frequent words to be longer. This minimises the average length of any given utterance, while still mostly keeping distinct words for distinct meanings.

In this thesis, I present the results of experiments that test Zipf’s hypothesis. Participants learn miniature made-up languages and then I observe how they reshape their input as they communicate with other participants. I find that, when under pressures to be both efficient and accurate in their communications, participants tend to use shorter words for more frequent or predictable meanings. When these pressures are removed, optimised languages do not emerge. This suggests that these competing communicative pressures are directly responsible for inciting optimisation behaviour. I then investigate a large data set of text from books published over the last 100 years to see whether these effects accumulate over time, producing the inverse frequency-length relationships observed in so many languages. I find that, indeed, for words whose frequency is increasing over time, a shorter variant of that word gains in popularity faster than it does for other words. Overall, these results support the theory that factors related to the communicative use of language are instrumental in shaping certain universal features of language structure.