THE STATISTICAL FOUNDATION OF WHALE SONG

Andrew Longo

THE MYTH AND BIG-PICTURE REASEARCH QUESTION

- The myth: Whether animal communication "behavior really involves *Language* in a meaningful way" (Kaplan 2016, 52)
- The study: Inbal Arnon et al analyzed eight years of whale recordings
- Big Picture research question: Does humpback whale song contain statistically coherent, repeatable parts?

MEASURABLE RESEARCH QUESTION

- Do the subsequences of humpback whale song follow a logarithmic Zipfian power law? (do some 'words' occur much more often than others)
- Further, do they follow Zipf's Law of brevity, with a negative correlation between length of utterance and frequency of utterance?

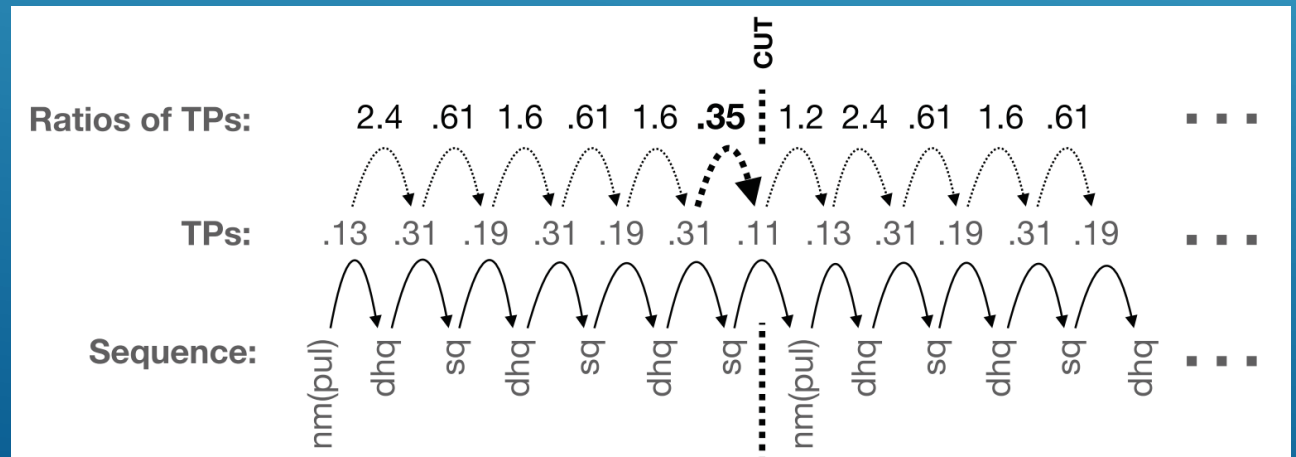
(Sharp. 2024)

THE PROBLEM AND SOLUTION

- A Zipfian Power Law and Zipf's Law of Brevity are based on human language words
- Whale communication would have to be divided into analogous structures
- Human words are separated mentally based on transitional probability
- For example, [b]->[a] are far more likely to occur within the same word than [b]->[[ð]
- The researchers applied this same principle to whale communication

METHODOLOGY: TRANSITIONAL PROBABILITY

- The hypothesis of repeated meaningful units was proven by observing newly learned song segments being "spliced" into songs at places of "highest structural similarity" (Arnon et. al 2025, 1)
- As with human words, transitional probabilities between subsequences were lower than within them
- Sequences were separated by the researchers at points where the TP ratio was <0.5 (the ratio in human experimental data). This ratio was found to be robust!
- Thus, whales learn song the same way that humans acquire language: through cultural statistical learning



(Arman et al 2025, 1)

WHAT IS A ZIPFIAN DISTRIBUTION?

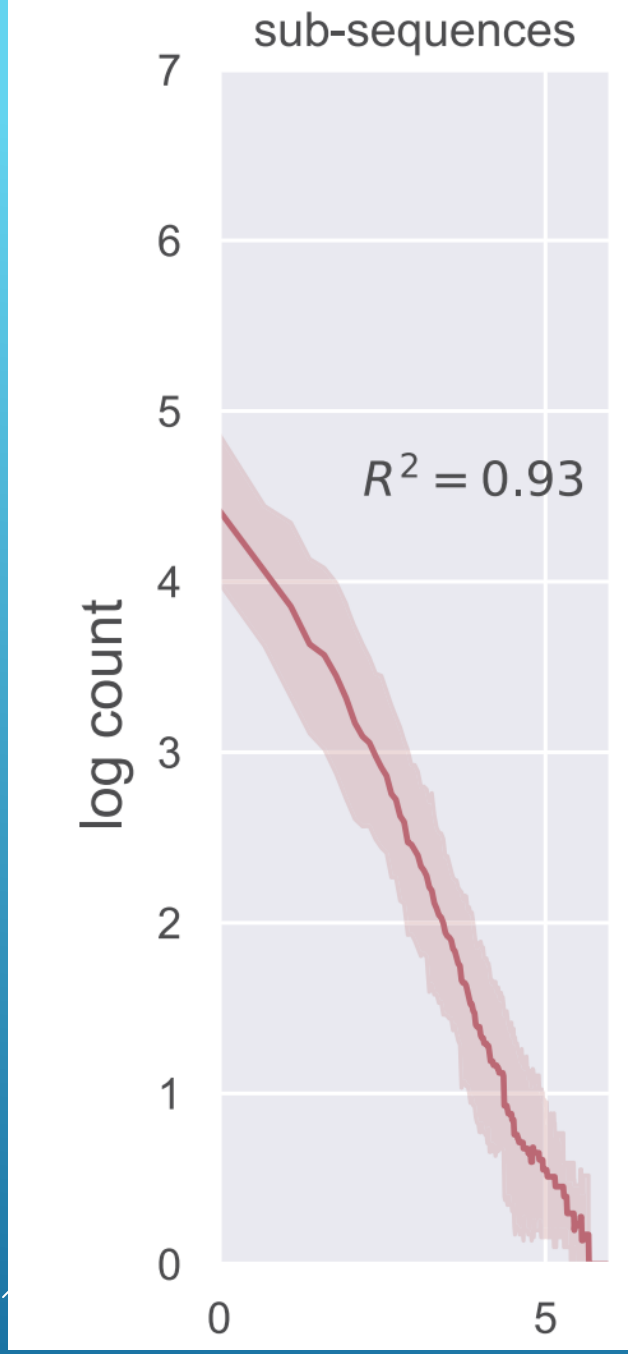
- A Zipfian Distribution is a universal feature of human language in which the most frequent segments are used substantially more than less frequent ones (Arman et al 2025, 1)
- Following this power law, the most frequent word in English, "the" constitutes ~7% of spoken words, twice as often as the second most, "of", at ~3.5%, and thrice as often as "and" at ~2.4%

$$\text{word frequency} \propto \frac{1}{\text{word rank}}$$

(Wikipedia)

PARSING THE ZIPFIAN DISTRIBUTION

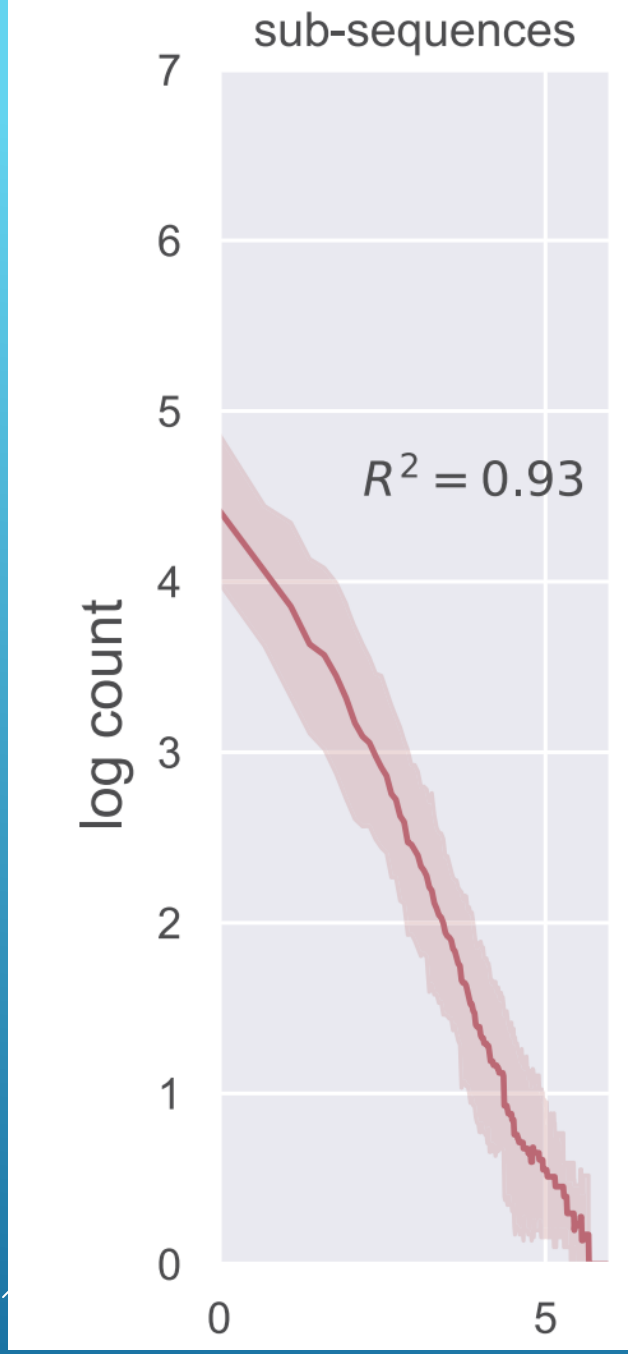
- The discovered subsequences were plotted onto the X-axis in the order of their frequency
- The Y-axis represents the log count of how many times they appeared in the data files
- The red line maps the mean log count value of each sub-sequence over eight years of recordings
- The pink zone around it represents the 95% bootstrapped confidence interval
- A bootstrapped confidence interval means that 95% of randomly re-sampled mean curves would fall within this zone



(Arman et al 2025, 2)

ANALYZING THE ZIPFIAN DISTRIBUTION

- Notice the narrowness and evenness of the confidence interval and that it mirrors the mean curve
- R^2=0.93, meaning that 93% of the variation would be explained by a straight line Zipfian Distribution
- Notice the increasing jaggedness of the line and confidence interval as utterances become more uncommon
- This shows that the less-common utterances inherently have more variability due to the unpredictable nature of their inclusion in annual datasets



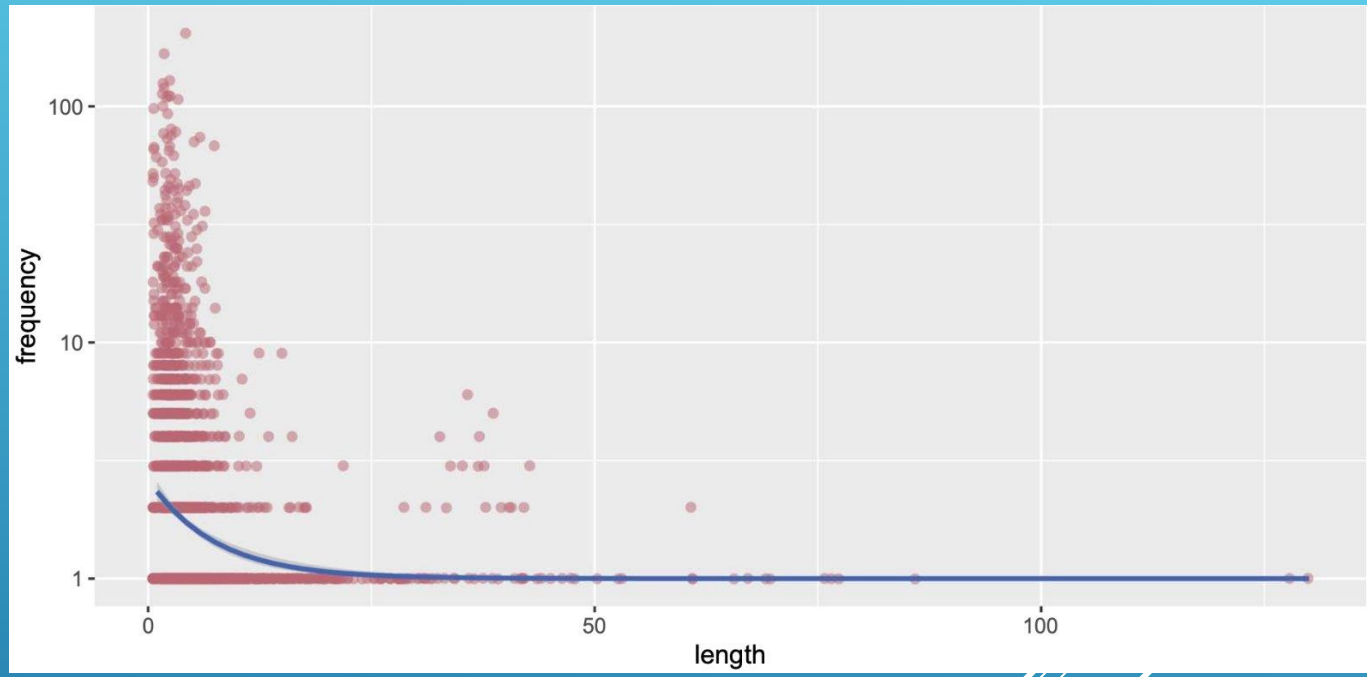
(Arman et al 2025, 2)

WHAT IS ZIPF'S SECOND LAW?

- Zipf's second law, the Law of Brevity, states that there is a *negative correlation* between the **length** and **frequency** of a word (Arman et al 2025, 3)
- For instance, "be" and "the" appear far more frequently in English than "distribution" or "communicate"

PARSING THE LAW OF BREVITY

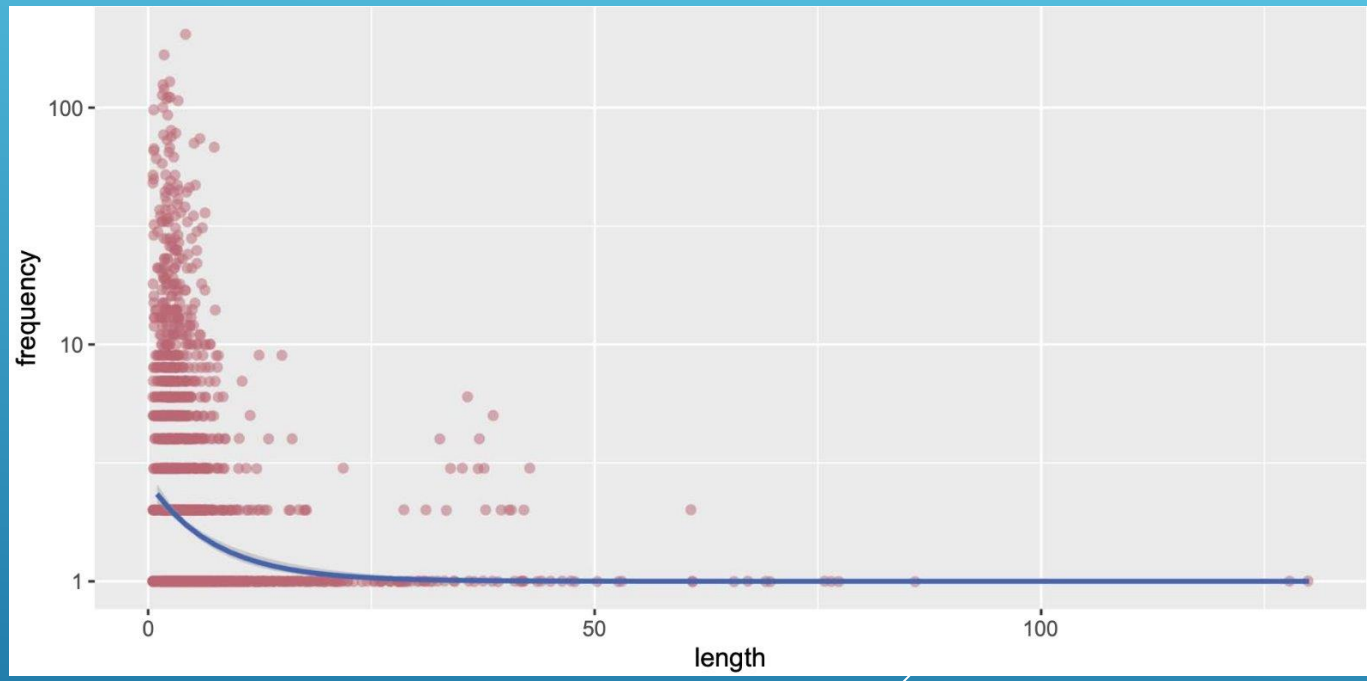
- Each point represents one subsequence
- The X-Axis represents the length of a sub-sequence in repeatable groups of sounds (groans, chirps, etc.)
- the Y-Axis represents how many times each subsequence appears in the datasets
- The blue line represents the Poisson regression line, which is the predicted average frequency of an utterance based on X length



(Arman et al 2025, 3)

ANALYZING THE LAW OF BREVITY

- Notice that every point above 10 occurrences appears to the far left of the graph
- The Regression line tapers off quickly; after foundational short length utterances, all utterances are comparatively rare
- The confidence interval around the blue line is barely visible, showing strong predictability
- Frequency can be predicted by length with a P value of <0.00001.
- If there was no correlation between length and frequency, there would have been a <0.001% chance of getting a fit of this strength



(Arman et al 2025, 3)

IS IT LANGUAGE?

- Whale songs were found to follow both a Zipfian Distribution and Zipf's Law of Brevity.
- Whale communication is cultural, learned through transitional probability.
- These findings **Disprove** the once-held myth that these components are unique to human language
- Human language has semantic content that is *likely* absent in whale song
- Whale song may be more akin to music, which also follows a Zipfian distribution
- "Once thought of as the hallmark of human uniqueness, it may transpire that foundational aspects of human language are shared across species." (Arnon et al, 2025, 4)

WORKS CITED

- Arnan, Inbal, Simon Kirby, & Jenny Allen. 2025. Whale Song Shows Language-like Statistical Structure. *Science* 387: 649-653.
- Kaplan, Abby. 2016. *Women Talk More Than Men … And Other Myths about Language Explained.* Cambridge University Press.
- Sharp, Charles. 2024. *Mother with calf off Moorea, French Polynesia*. Wikipedia.