AMSI VACATION RESEARCH SCHOLARSHIPS 2022–23

SET YOUR SIGHTS ON RESEARCH THIS SUMMER

Analysis of Humpback Whale Songs using Information Theory Techniques

Macey Lawson
Supervised by Professor Matthew Roughan
The University of Adelaide

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

1 Introduction **2**

1.1 Statement of Authorship . . . 3

2 Information Theory **3**

2.1 Shannon Entropy . . . 3

2.2 Zipf-Mandelbrot Distribution . . . 4

3 The Process: Humpback Whale Songs **4**

3.1 Data Collection . . . 4

3.2 Preparing the Data . . . 5

3.3 Generating Results . . . 6

4 The Process: English Texts **7**

4.1 Data Collection . . . 7

4.2 Preparing the Data . . . 7

4.3 Generating Results . . . 7

5 Results **8**

5.1 Shannon Entropy . . . 8

5.2 Zipf-Mandelbrot Distribution . . . 10

6 Discussion and Conclusion **12**

7 Acknowledgements **12**

Abstract

This project explored humpback whale songs through the lens of information theory, and compared the vocalisations to the English language. Using unit sequences of humpback songs recorded from 2001 to 2005 off the coasts of Gabon and Madagascar, Shannon entropy was estimated. The combined dataset conveyed approximately 3 to 4 bits per unit. These entropy estimates were lower than that of words in English text, but were of the same order of magnitude. The majority of the arbitrarily chosen English texts, including a play, album, novel, and song, were estimated to convey approximately 8 to 9 bits per word.

Furthermore, the units were observed to be closely described by a Zipf-Mandelbrot distribution, suggesting humpback units are perhaps the equivalent of words in the English language. The Zipf-Mandelbrot approximation was much steeper than all four of the English texts. This result potentially highlights the inclination within English to employ elaborate and infrequent vocabulary, unlike the straightforward and uncomplicated quality of humpback songs.

1 Introduction

Humpback whales *(Megaptera novaeangliae)* are a species of baleen whale that can grow up to 17 metres long, with hearts that weigh just short of 200 kilograms (Clapham, 2000; Young People's Trust for the Environment, 2014). They have existed and thrived for 11 million years, and are known for their long migration routes and remarkably complex songs (McDonald, 2019).

Only male humpbacks sing, and they do so on breeding and feeding grounds (Parsons et al., 2008). Often all the male humpbacks within a population will sing the same song, which is modified gradually over the season (Mercado and Handel, 2012; Garland et al., 2017). Researchers are still unsure the definitive purpose of these songs, but they are commonly assumed to be used as a mating call, or display of dominance (Parsons et al., 2008).

Humpback songs can be thought to follow a hierarchical structure, as first proposed by Roger Payne and Scott McVay in 1971 (Payne and McVay, 1971). Units are the smallest building block of their songs, which are sounds separated by silence. For instance, units might sound like a 'squeak' or 'woop'. A combination of units creates a phrase, and a combination of phrases creates a theme. Themes go on to create a song, and a combination of songs, or more often so the repetition of one song, creates the song session (Payne and McVay, 1971). Song sessions can last anywhere from ten minutes to several hours (Miksis-Olds et al., 2008; McDonald, 2019).

Interestingly, humpbacks have been observed to exchange their songs with other populations, as a result of crossover migratory routes (Parsons et al., 2008; Garland et al., 2011). These trending themes being spread world-wide parallels that of viral songs humans enjoy. This begs the question, to what extent are humpback whale songs comparable to human language?

This project sought to explore the possible similarities between humpback whale song unit frequencies and word frequencies in the English language. The study also aimed to estimate a quantitative measure of the information able to be transmitted within humpback whale song units.

1.1 Statement of Authorship

- Professor Matthew Roughan supervised the project, provided guidance with interpreting the results, and wrote the Zipf-Mandelbrot fit function in MATLAB.
- Macey Lawson produced this report, generated all included results, wrote the Python code used to analyse the data, and wrote all other MATLAB code used to examine the English texts and humpback data.

2 Information Theory

Information theory is a branch of applied mathematical study, concerned with the transmission of information, often in the form of discrete symbol sequences generated by a source (Suzuki et al., 2006). In this case, the source is a humpback whale, and the discrete symbols are the recorded sequences of units (sounds separated by silence) produced by the humpback. Information theory offers techniques, like entropy estimation, which enables the exploration of the characteristics of a source, without requiring any prior knowledge of the meaning of its output (Suzuki et al., 2006).

2.1 Shannon Entropy

Information entropy is fundamentally a measure of uncertainty within a dataset. The more 'surprising' the given sequence is, the higher the entropy estimation will be.

In this project, the Shannon entropy of humpback whale units, $H(X)$, was estimated using Claude Shannon's formula (Shannon, 1948):

$$H(X) = -\sum_{k=1}^{N} p(k) \log p(k),$$

where X is the unit sequence, N is the number of unique units present in sequence X, and $p(k)$ is the estimated probability of unit k being used. Here, $p(k)$ is approximated based on each unit k's observed frequency in the given sequence.

The base of the logarithm decides the units of the resulting entropy quantity. For example, the natural log calculates an entropy estimate measured with nats. For this study, base 2 was implemented to produce entropy approximations measured with bits.

Shannon entropy estimates the average amount of information per unit that the humpback songs transmit over a given long unit sequence, provided units are independently and identically distributed. Note that this measurement ignores rhythm and semantics of the humpback songs. It purely takes into account the frequencies of different kinds of units.

Normally if the length of the unit sequence is increased, it will result in a more accurate estimation of the underlying entropy. However, this may not always be the case if there are substantially more unique units

within the longer sequence. Essentially, more unique units increase the entropy estimate, due to producing greater uncertainty within the data.

2.2 Zipf-Mandelbrot Distribution

Zipf's law models the inverse relationship present in natural languages between the frequency and ranking of words (Zipf, 1949). A Zipf-Mandelbrot distribution is a generalisation of Zipf's law, with the inclusion of an extra parameter 'q' to better capture the observed relationship (Mandelbrot, 1965).

Zipf-Mandelbrot distribution is governed by the following equation:

$$P(Z_{N,s,q} = k) = \frac{1}{H_{N,s,q}} \times \frac{1}{(k+q)^s},$$

where N is the number of unique units, k is the ranking of a unit (most to least frequent), s and q are data-tuned parameters, and $H_{N,s,q} = \sum_{i=1}^{N} \frac{1}{(i+q)^s}$.

3 The Process: Humpback Whale Songs

3.1 Data Collection

The dataset used during this project consisted of unit sequences of humpback songs. These sequences were recorded from 2001 to 2005, off both the East coast of Africa, in Gabon, and the West coast, in Madagascar.

These unit sequences were collated from two tables in the paper titled 'Culturally transmitted song exchange between humpback whales in the southeast Atlantic and southwest Indian ocean basins', published November 28, 2018, by Rekdahl *et al.* (Rekdahl et al., 2018). As detailed in the article, the humpback songs were recorded in three locations: Antongil Bay, Madagascar (16°00'S, 49°55'E), Iguela, Gabon (1°51'S, 9°20'E) and Mayumba, Gabon (3°26'S, 10°39'E). At all three sites, songs were recorded at a 44.1 kHz sampling rate and 16-bit resolution, using a suspended hand-held hydrophone. These hydrophones were attached to preamplifiers and connected to a Sony TCD-D100 digital audiotape recorder.

The recordings were converted to digital wav files using Avisoft-SASLab Pro, then converted to spectrograms in Raven Pro 1.5. Finally, the songs were manually broken down into units by a human classifier, Gabriella A. Carvajal, and inspected by a second classifier, Melinda L. Rekdahl. Both classifiers determined the units through reference to visual and acoustic qualities of each sound.

Each unique type of unit was given a short abbreviation by the classifiers to match their acoustic description. For instance, the abbreviation 'acr' represented a distinct type of 'ascending cry', and 'm' represented a distinct kind of 'moan'. These labels were based on previous interpretations of humpback noises in marine studies (Payne and McVay, 1971). A table matching the acoustic descriptions to their abbreviations was included as supplementary material to the article online (Rekdahl et al., 2018).

3.2 Preparing the Data

The data was accessed online in the form of two tables. The first table specified different themes, and listed the sequence of units that made up each theme. The second table pinpointed the sequence of themes that then created each song.

The tables were read as CSV files into Python. The following steps were repeated for both the Gabon data and the Madagascar data separately.

1. From the first table, the column of theme numbers was broken up into entries for each of the five years accordingly (2001 through 2005). Likewise, the column containing the unit sequences making up each theme was split into five separate lists for each year.

2. From the second table, the columns of song numbers and corresponding theme sequences were converted to lists. Missing values were then removed, due to the Gabon and Madagascar data featuring a different number of recorded songs. These lists were then split into five separate lists containing the theme sequences of 2001 through to 2005 accordingly.

3. A function was called, which inserted the correct unit sequences into their corresponding theme, to then output the merged unit sequence of all the songs recorded for that year. The function took inputs of theme numbers, unit sequences for each theme, and theme sequences for each song as separate lists. It then transformed these inputs into arrays, and created a dictionary mapping each unit sequence to its correct theme. The output list was generated by reading through the theme sequences and concatenating each appropriate unit sequence using the dictionary. This function was called for 2001 through to 2005 using the separated data.

4. Output lists of unit sequences were concatenated to create a long unit sequence of all the humpback songs recorded in that region.

Finally the Gabon and Madagascar unit sequences were joined together to create a combined set of all humpback unit sequences. As seen in Table 1, the combined dataset consisted of 903 units, made up of 44 different types of units.

NOTE: The reason this process was followed with the data split into years 2001 to 2005, rather than the five years combined, was due to repetition of theme labels over the years. For example, a certain theme labelled as '7s' in 2003, differed slightly from the theme labelled '7s' in 2004. The original paper, from which the data was acquired, focused on song similarity between the Gabon and Madagascar humpback populations over the years (Rekdahl et al., 2018). Thus, it was beneficial for that study to group together highly comparable themes. The researchers labelled matching themes with an 'm', shifting themes with an 's', and evolutionary themes with an 'e'.

Table 1: Table summarising humpback whale song data as unit sequences.

DATA	Total Units	Unique Units
Gabon	451	34
Madagascar	452	30
Combined	903	44

3.3 Generating Results

Results were generated using both Python (Jupyter Notebook), and MATLAB (version R2021b).

Shannon Entropy

After preparing the humpback unit sequence data in Python, Shannon entropy was estimated using a function. The function took a unit sequence as input, and estimated the probability of each type of unit occurring by dividing its frequency by the total number of units in the sequence. The function then used those probabilities to estimate entropy with Shannon's formula. That is, it calculated the negative of the sum of each units' estimated probability of occurring, multiplied by the binary logarithm of that same probability.

The same code was produced in MATLAB to verify these entropy estimates for the humpback data, and for later use with the English texts.

Zipf-Mandelbrot Distribution

In MATLAB, the built-in 'unique' function was used to create an array of all the unique types of units present in the combined sequence. Those unique units were then ordered by how frequently they occurred, from most to least frequent. Then, a function was called, which fit the frequency and rank arrays to a Zipf-Mandelbrot distribution using least squares. The function output parameters s, q and N (the number of unique units), which were in turn used as input to a second function that calculated the Zipf-Mandelbrot curve. The resultant distribution was plotted on a log-log plot, with unit frequencies versus the ranking of unique units.

4 The Process: English Texts

4.1 Data Collection

Four different English texts were selected arbitrarily to compare to the humpback whale unit sequence data. The texts included a play, an album, a novel, and a song, which were all obtained online.

- *Henry VI, Part I* is a historical play written by William Shakespeare in 1591. The script was located online using Massachusetts Institute of Technology's Shakespeare archives (Shakespeare, 1591).
- *Midnights (3am Edition)* is singer-songwriter Taylor Swift's 2022 extended studio album, released by Republic Records. The lyrics to all songs were obtained through a website titled 'Lyrics on Demand' (Swift, 2022).
- *The Hunger Games* is a dystopian young-adult fiction novel by Suzanne Collins, published in 2008 by Scholastic. The text was accessed through a site called 'All Books Hub' (Collins, 2008).
- "Never Gonna Give You Up" is the 1987 dance-pop hit by Rick Astley. The lyrics were accessed through the website 'Genius' (Astley, 1987).

4.2 Preparing the Data

The English texts were converted to text files and imported into MATLAB (version R2021b). Using the Text Analytics Toolbox, each text was processed as followed to prepare for analysis:

1. The text file was read into MATLAB and extracted.
2. The extracted text was tokenised by separating the data into meaningful pieces. Sentence and part-of-speech details were updated to the tokenized document.
3. Punctuation was removed.
4. The document was lemmatized. That is, variants of words were reduced to their root word.

4.3 Generating Results

Results were generated using MATLAB.

Shannon Entropy

Once the English texts were pre-processed, each of the four tokenised documents were converted into strings using the built-in 'string' function. The entropy of the words in these strings was then estimated using Shannon's entropy formula, via a function as discussed earlier.

Zipf-Mandelbrot Distribution

After pre-processing the texts, the frequency of each unique word was found using the 'bagOfWords' function, and arranged into descending order in an array. The rank and frequency arrays were employed as input into a function, which fit the data to a Zipf-Mandelbrot distribution. The function output the parameters s, q and N (the number of unique words), which were then used in another function to calculate a Zipf-Mandelbrot curve. Finally, the resultant Zipf-Mandelbrot distributions, fit to each of the texts accordingly, were plotted on a log-log plot with word frequencies versus the ranking of unique words.

5 Results

5.1 Shannon Entropy

As seen in Table 2, the Shannon entropy estimates of units in the yearly sequences ranged from 2.47 to 3.75 bits per unit. For the Gabon dataset, the median entropy estimate is 3.34 bits per unit. For the Madagascar dataset, the median entropy estimate is similarly 3.09 bits per unit. For the combined Gabon and Madagascar dataset, the median entropy across individual years is slightly higher, at an estimated 3.62 bits per unit. This increase is presumably due to the joint dataset seeing a greater number of unique units, as compared to the Gabon and Madagascar sequences separately (see Table 1).

Table 2: Table containing Shannon entropy estimates of the humpback units across each year, measured in bits per unit.

DATA	2001	2002	2003	2004	2005
Gabon	3.33	2.98	3.55	3.34	3.46
Madagascar	2.47	2.66	3.09	3.61	3.18
Combined	3.39	3.08	3.62	3.75	3.64

Shannon entropy was also estimated for the data of all five years combined, rather than separately. Again, likely due to that increased number of unique units, an increase in entropy is seen (Table 3). The combined dataset had a Shannon entropy estimate of approximately 4.63 bits per unit.

Table 3: Table containing Shannon entropy estimates of the humpback units over the combined five years (2001-2005), measured in bits per unit.

DATA	ENTROPY ESTIMATE 2001-2005
Gabon	4.42
Madagascar	4.33
Combined	4.63

As shown in Table 4, the English texts were seen to have Shannon entropy estimates of around 8 or 9 bits per word. While noticeably higher than that of the humpback data, these results are still within the same realm and have the same order of magnitude. "Never Gonna Give You Up" has a lower estimated entropy than the other English texts, at 4.89 bits per word. This may be due to the smaller dataset, or a consequence of the repetitive nature of the song.

The median Shannon entropy estimation was also calculated from a similar previous study (Suzuki et al., 2006), as a reference point for the results. The median entropy, at 4.38 bits per unit, appears to be higher than the results for sequences from each year separately (Table 2), but is extremely similar to that of the combined years' estimation, at 4.63 bits per unit (Table 3). Longer sequences provide more data, allowing for more accurate approximations of the underlying entropy. Thus, this higher estimate may be closer to the true entropy of humpback units, or perhaps the consequence of increased unique units.

Table 4: Comparison table of Shannon entropy approximations of humpback units and English words.

HUMPBACK DATA	MEDIAN ENTROPY (bits per unit)	ENGLISH DATA	ENTROPY (bits per word)
Gabon	3.34	***Henry VI*** Play by William Shakespeare	8.62
Madagascar	3.09	***Midnights*** Album by Taylor Swift	8.00
Combined	3.51	***The Hunger Games*** Novel by Suzanne Collins	9.00
Data from Previous Study (Suzuki *et al.*, 2006)	4.38	**"Never Gonna Give You Up"** Song by Rick Astley	4.89

5.2 Zipf-Mandelbrot Distribution

Figure 1: Plot of humpback unit frequency versus ranking distribution, with fitted Zipf-Mandelbrot approximation for combined humpback unit sequence data. Plot features some examples of where certain humpback song units land in the ranking.

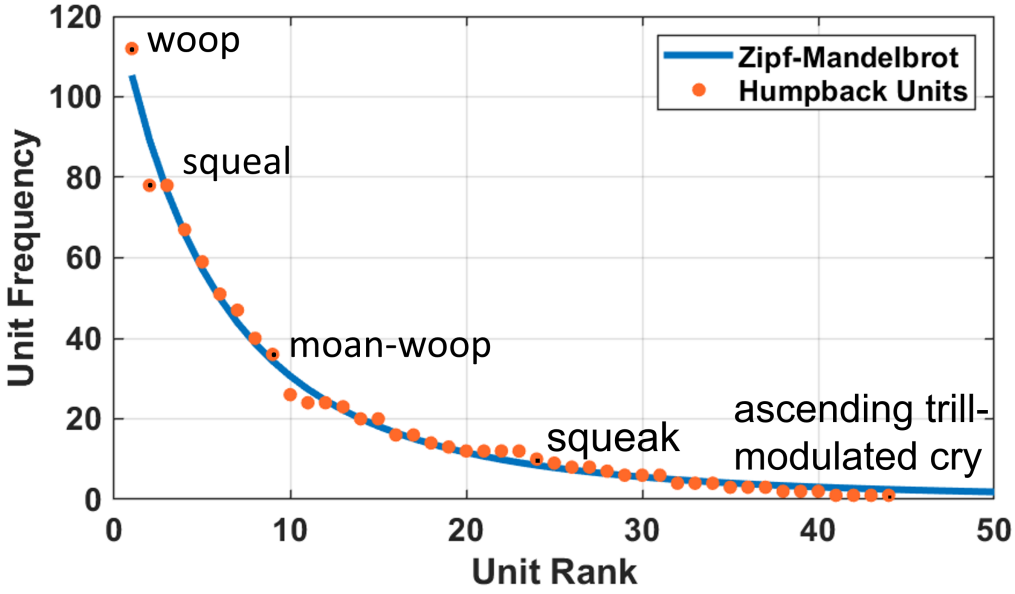
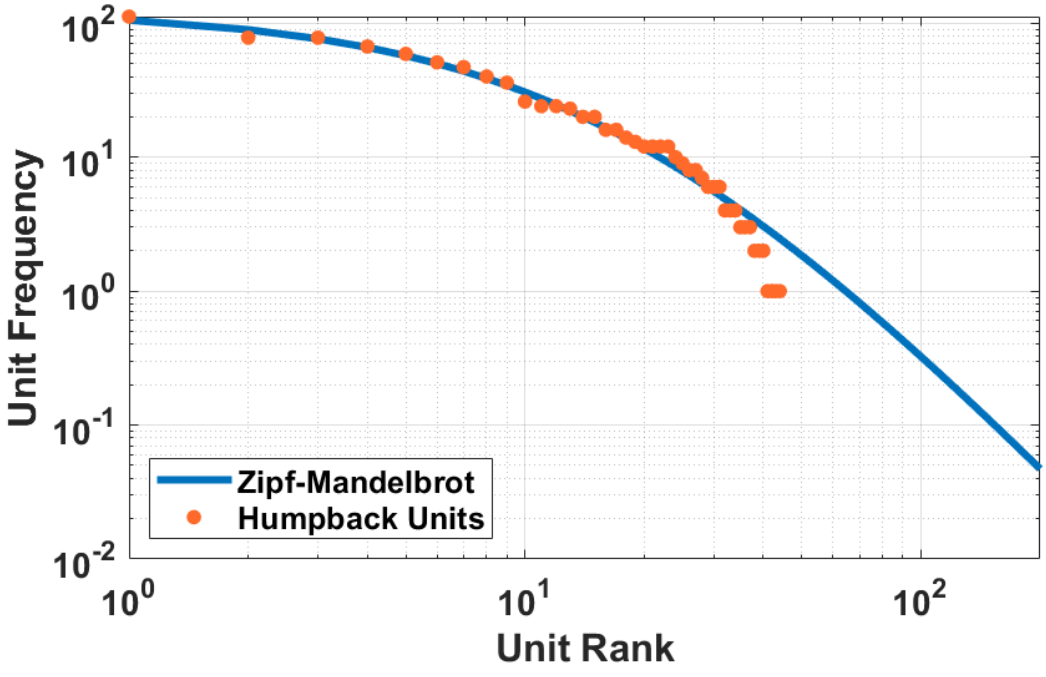


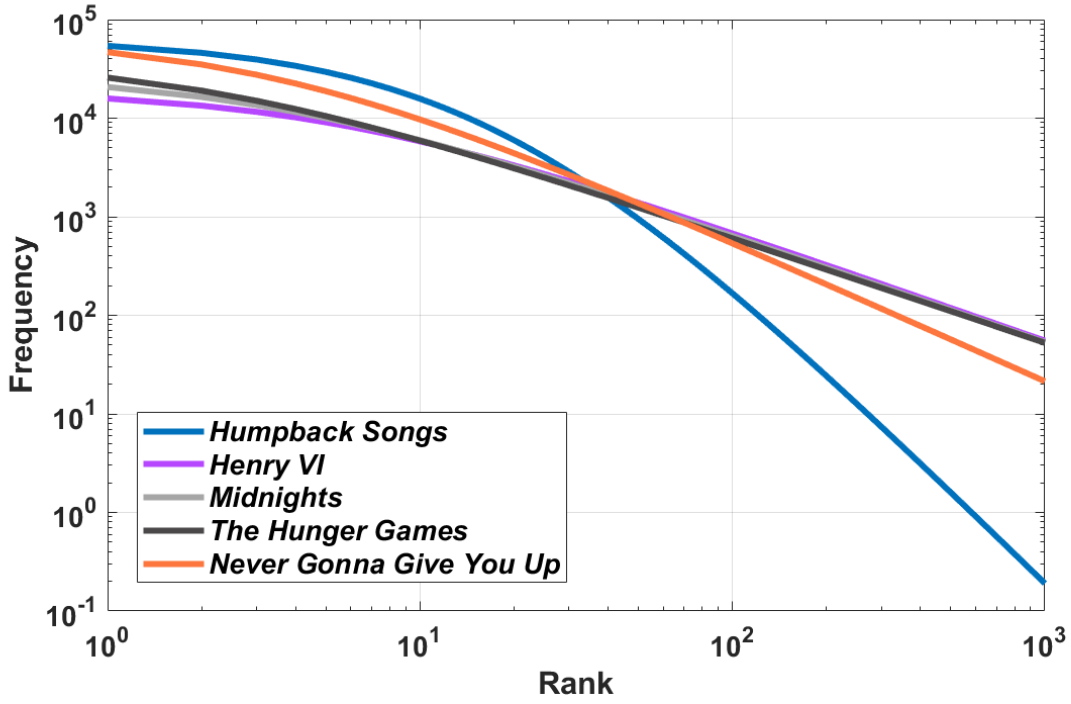
Figure 2: Log-log plot of humpback unit frequency versus ranking distribution, with fitted Zipf-Mandelbrot approximation for combined humpback unit sequence data.



As seen in Figure 1, a Zipf-Mandelbrot curve appears to accurately describe the relationship between the unit frequency and rank. To better examine the fit of the data a log-log plot was employed.

As shown in Figure 2, there is some slight trailing of data points near the end on the right, but these points indicate types of units only used only once or twice throughout the songs. That is, there is not enough data to properly model the distribution around these points. Overall, this Zipf-Mandelbrot distribution is a strong fit for the humpback data.

Figure 3: Comparison of Zipf-Mandelbrot distributions of humpback whale song units and words in arbitrary English texts.



As visible in Figure 3, the Zipf-Mandelbrot curves for the three larger bodies of English texts (*Henry VI*, *Midnights* and *The Hunger Games*) appear to be extremely similar. "Never Gonna Give You Up" has a steeper slope, highlighting a proportionally higher usage of top-ranking units. This might be an outcome of the repetitiveness of Astley's song, or a consequence of the more limited dataset.

The Zipf-Mandelbrot approximation for the humpback data evidently has a much steeper decreasing slope on the log-log plot than that of the English texts, even more so than "Never Gonna Give You Up". That longer, flatter slope for the English data suggests that the rare words are not quite as rare comparatively to whales. This might indicate simplicity within whale songs. Perhaps, if it is a form of communication, humpbacks are more direct and straightforward, whereas humans make more regular use of their rare words, making text and speech more flamboyant or flowery.

6 Discussion and Conclusion

This project sought to compare humpback whale song units with words in English texts. It was estimated that the combination of all recorded humpback songs in the dataset conveyed approximately 3 to 4 bits per unit. This was much lower than that of English texts at 8 to 9 bits per word, but is still of the same order of magnitude.

The lower entropy estimates may suggest that humpback whale songs are more predictable, or have a higher level of structure than English. Potentially a more organised language, paired with the observed smaller vocabulary size, makes it simpler for humpbacks to comprehend and communicate songs. Furthermore, the lower entropy may reflect a restricted scope of topics needed to be communicated, as compared to the vast range of topics explored within English text.

It was also seen that the humpback song units conformed well to a Zipf-Mandelbrot distribution, potentially suggesting units in whale language behave similarly to words in human language. A much steeper decreasing Zipf-Mandelbrot slope was observed for the humpback songs on the log-log plot. The longer, more uniform slope for the English data indicates that rare English words are used more frequently, proportionally to that of rare whale units. This result perhaps highlights a more direct, simplistic nature of whale songs, or a tendency within English text to make common use of complex, rare words.

Further research is required to estimate more meaningful measures of entropy, since Shannon entropy assumes units to be independently and identically distributed. With whale songs evidently adhering to certain structural patterns, such as repetition of themes and phrases, this is not a valid assumption. Entropy could instead be estimated through treating the unit sequences as a Markov model or a non-parametric process.

Further exploration into other kinds of linguistic laws within human language could increase our understandings in the similarities between humpback and human communication. Observed patterns that could be examined include Heap's law, Brevity law, or Menzerath's law.

Overall, this project observed slightly lower entropy estimates and a steeper Zipf-Mandelbrot approximation for humpback whale songs compared to English texts, but a much larger dataset would need to be employed to increase accuracy of the results. The study could also be improved through adopting automated classifiers, or additional human classifiers, to remove as much human-interpretive bias as possible during unit classification.

7 Acknowledgements

The author would like to thank Professor Matthew Roughan for his ongoing support and guidance throughout the course of this project.

References

Astley, R. (1987). Never Gonna Give You Up. RCA Records. https://genius.com/Rick-astley-never-gonna-give-you-up-lyrics. Accessed on Jan 31, 2023.

Clapham, P. J. (2000). The humpback whale. *Cetacean Societies, Field Studies of Dolphins and Whales. Chicago: The University of Chicago*, pages 173–196.

Collins, S. (2008). *The Hunger Games*. Scholastic. https://allbookshub.com/the-hunger-games-pdf-ebook. Accessed on Feb 2, 2023.

Garland, E., Goldizen, A., Rekdahl, M., Constantine, R., Garrigue, C., Hauser, N., Poole, M., Robbins, J., and Noad, M. (2011). Dynamic horizontal cultural transmission of humpback whale song at the ocean basin scale. *Current biology : CB*, 21:687–91.

Garland, E., Rendell, L., Lamoni, L., Poole, M., and Noad, M. (2017). Song hybridization events during revolutionary song change provide insights into cultural transmission in humpback whales. *Proceedings of the National Academy of Sciences*, 114:7822–7829.

Mandelbrot, B. (1965). Information theory and psycholinguistics. *BB Wolman and E.*

McDonald, K. (2019). Data of the humpback whale. Medium. https://kcimc.medium.com/data-of-the-humpback-whale-9ef09c5920cd. Accessed on Jan 4, 2023.

Mercado, E. and Handel, S. (2012). Understanding the structure of humpback whale songs. *The Journal of the Acoustical Society of America*, 132:2947–50.

Miksis-Olds, J. L., Buck, J. R., Noad, M. J., Cato, D. H., and Dale Stokes, M. (2008). Information theory analysis of australian humpback whale song. *The Journal of the Acoustical Society of America*, 124(4):2385–2393.

Parsons, E., Wright, A., and Gore, M. (2008). The nature of humpback whale *(Megaptera novaeangliae)* song. *Journal of Marine Animals and Their Ecology*, 1.

Payne, R. S. and McVay, S. (1971). Songs of humpback whales: Humpbacks emit sounds in long, predictable patterns ranging over frequencies audible to humans. *Science*, 173(3997):585–597.

Rekdahl, M. L., Garland, E. C., Carvajal, G. A., King, C. D., Collins, T., Razafindrakoto, Y., and Rosenbaum, H. (2018). Culturally transmitted song exchange between humpback whales *(Megaptera novaeangliae)* in the southeast atlantic and southwest indian ocean basins. *Royal Society Open Science*, 5(11):172305.

Shakespeare, W. (1591). *Henry VI, Part I*. Massachusetts Institute of Technology Shakespearean Archives. http://shakespeare.mit.edu/1henryvi/full.html. Accessed on May 25, 2021.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Suzuki, R., Buck, J., and Tyack, P. (2006). Information entropy of humpback whale songs. *The Journal of the Acoustical Society of America*, 119:1849–66.

Swift, T. (2022). *Midnights (3am edition)*. Republic Records. `https://www.lyricsondemand.com/t/taylorswiftlyrics/midnights3ameditionalbumlyrics.html`. Accessed on Jan 31, 2023.

Young People's Trust for the Environment (2014). Whale (Humpback). `http://ypte.org.uk/factsheets/whale-humpback/overview`. Accessed on Jan 3, 2023.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press.