Methods in Ecology and Evolution

DR ARIK KERSHENBAUM (Orcid ID : 0000-0003-0464-0243)

Article type : Research Article

Editor : Veronica Zamora-Gutierrez

TITLE: Shannon entropy as a robust estimator of Zipf's Law in animal vocal communication repertoires

AUTHORS

Arik Kershenbaum[a,b], Vlad Demartsev[c], David E Gammon[d], Eli Geffen[e], Morgan L Gustison[f], Amiyaal Ilany[g], Adriano R Lameira[h]

[a]Department of Zoology, University of Cambridge, Cambridge, England

[b]Girton College, University of Cambridge, Cambridge, England

[c]Department of Biology, University of Konstanz, Konstanz, Germany

[d]Department of Biology, Elon University, Elon, North Carolina, USA

[e]School of Zoology, Tel Aviv University, Tel Aviv, Israel

[f]Department of Integrative Biology, University of Texas at Austin, Austin, Texas, USA

[g]Faculty of Life Sciences, Bar Ilan University, Ramat Gan, Israel

[h]Department of Psychology, University of Warwick, Warwick, England

CORRESPONDING AUTHOR ADDRESS

arik.kershenbaum@gmail.com

Girton College, Huntingdon Rd, Cambridge, CB3 0JG, UK

RUNNING HEAD

Shannon entropy as an estimator of Zipf's Law

ABSTRACT

1. Information complexity in animals is an indicator of advanced communication and an intricate socio-ecology. Zipf's Law of least effort has been used to assess the potential information content of animal repertoires, including whether or not a particular animal communication could be "language-like". As all human languages follow Zipf's law, with a power law coefficient (PLC) close to -1, animal signals with similar probability distributions are postulated to possess similar information characteristics to language. However, estimation of the PLC from limited empirical datasets (e.g. most animal communication studies) is problematic because of biases from small sample sizes.

2. The traditional approach to estimating Zipf's law PLC is to find the slope of a log-log rank-frequency plot. Our alternative option uses the underlying equivalence between Shannon entropy (i.e. whether successive elements of a sequence are unpredictable, or repetitive), and PLC. Here, we test whether an entropy approach yields more robust estimates of Zipf's law PLC than the traditional approach.

3. We examined the efficacy of the entropy approach in two ways. First, we estimated the PLC from synthetic data sets generated with *a priori* known power law probability distributions. This revealed that the estimated PLC using the traditional method is particularly inaccurate for highly stereotyped sequences, even at modest repertoire sizes. Estimation via Shannon entropy is accurate with modest sample sizes even for repertoires with thousands of distinct elements. Second, we applied these approaches to empirical data taken from 11 animal species. Shannon entropy produced a more robust estimate of PLC with lower variance than the traditional method, even when the true PLC is unknown. Our approach for the first time reveals Zipf's law operating in the vocal systems of multiple lineages: songbirds, hyraxes, and cetaceans.

4. As different methods of estimating the PLC can lead to misleading results in real data, estimating the balance of a communication system between simplicity and complexity is best performed using the entropy approach. This provides a more robust way to investigate the evolutionary constraints and processes that have acted on animal communication systems, and the parallels between these processes and the evolution of language.

KEYWORDS: Animal communication, Information theory, Language, Shannon Entropy, Zipf's Law

1. INTRODUCTION

All animal species have evolved communication systems to suit their ecological requirements, but these systems vary greatly in their complexity, and the volume of information that can be transmitted, received, and interpreted. Humans appear to be an extreme case, with the capability of conveying essentially unlimited information in our language, while relying on a finite set of signal elements. Non-human animals, on the other hand, are generally thought not to possess any true language (Cheney and Seyfarth, 1998; Fitch, 2005). The origin of human language has long been a controversial topic (Dunbar, 2003; Hauser et al., 2002; Jackendoff, 1999), but current thinking is that this apparently unique human ability arose from complex communication in our non-human ancestors (Bergman et al., 2019; Boë et al., 2019; Crockford et al., 2017; Lameira, 2017; Lameira and Call, 2018; Schlenker et al., 2016), rather than arising *de novo* as an adaptation to the unique social and ecological challenges that our ancestors faced (de Boer et al., 2020; Martins and Boeckx, 2019). To address these unknowns, comparative studies between human language and non-human animal communication systems continue to be important for identifying the conserved features of these systems and explaining any differences that emerged (Fitch, 2005).

It has long been observed that word frequency (i.e., prevalence) in human languages appears to conform to Zipf's Law of least effort (Zipf, 1949) henceforth, simply Zipf's Law, across most, if not all, languages in the world (Yu et al., 2018). Zipf's Law states that word frequency, ranked in descending order of frequency, follows a power law distribution

$$P_i \propto i^k$$

Where P_i is the probability (number of instances divided by total number of words) of the i^{th} most common word, and k is a constant. Specifically, in Zipf's Law, the word probabilities follow a power law where k=-1. Many suggestions have been proposed to explain this peculiar observation, from a statistical artefact (Suzuki et al., 2005) to a reflection of intentional information content (Lestrade, 2017; McCowan et al., 1999). However, there are reasons to suspect that the existence of Zipf's Law in human communication reflects a fundamental trade-off between signal complexity (potential information content), and signal cost (processing power required to generate and interpret) (Ferrer-i-Cancho and Solé, 2003). It has been postulated (Ferrer-i-Cancho, 2005) that this balance of trade-offs is fundamental to complex communication, and language in particular. If that is the case, then observing a value of k=-1 in the power law distribution of an animal communication system would imply that similar constraints on the co-evolution of complexity and signal cost have been operating in that species, as in humans. This hypothesis has led some researchers to suggest further that k=-1 may be used as a potential indicator of language-like communication in a particular species (McCowan et al., 2005). While such a claim may be subject to

debate, the converse claim is much stronger: that a communication system where *k* deviates strongly from -1 is unlikely to be a true language, because it represents a system where signals are either highly predictable, or "stereotyped" (*k*<<-1, a steeper negative slope) or highly random (*k*~*0*, a shallow negative slope). As *k*~-1 is likely a universal requirement for an efficient language system, a case has also been made for using the power law coefficient (PLC) *k* to screen nonhuman signal systems, including signals from outer space for signs of intelligent life (Doyle et al., 2011).

Although Zipf's Law and the traditional approach of PLC has merit in 'big data' contexts, e.g., an entire language corpus (while taking care not to "saturate" the Zipf plot with too much data), there remain practical challenges of measuring this entity in the real-world datasets of animal communication. The standard method for calculating the PLC '*k'* is to measure the negative slope of a log-log plot of word rank against word prevalence (Figure 1) (Corominas-Murtra et al., 2018). This technique has been used successfully in studies of Zipf's Law in human language (Yu et al., 2018), as well as other domains such as city sizes (Decker et al., 2007), and family name popularity (Zanette and Manrubia, 2001). However, while corpora of human language may run into hundreds of thousands or even millions of words, animal data sets are generally far smaller, and collecting full repertoires requires long term studies with automated recorders (e.g., (Ilany et al., 2013)). Small sample sizes means that calculations of empirical slopes vary strongly when elements (words in the case of human language) occur at low prevalence, e.g. when many elements occur just once or twice (Figure 1, region c). In addition, deviations from a power law distribution are known to occur for words that are particularly common or rare (Figure 1, regions a, c as compared to region b), known as the broken power law distribution (Newman, 2005). Furthermore, a communication system with a large vocabulary repertoire presents additional challenges to the ability of existing analyses to estimate *k* with a small sample size because the potential counts per call type are very low. Therefore, claims of particular values of *k* for animal communication systems with small sample size datasets may be unreliable and/or difficult to replicate.

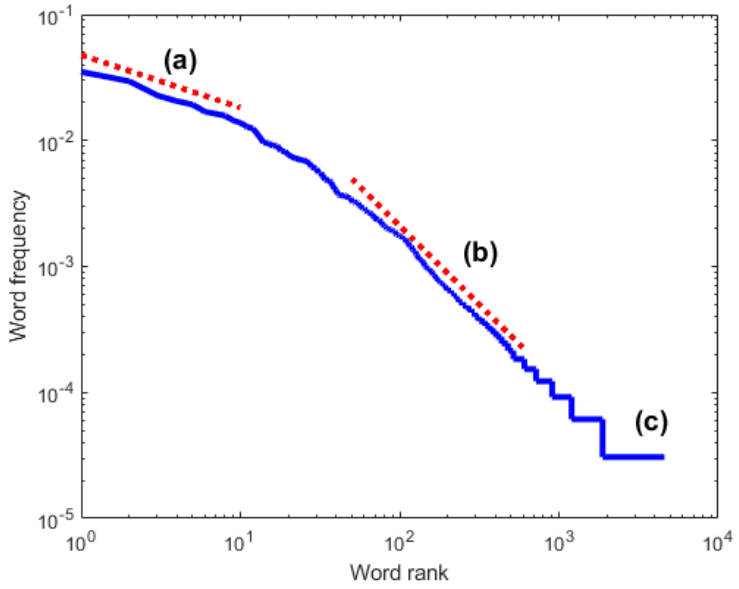


Figure 1. A log-log plot of the word prevalence vs. word prevalence rank of more than 4600 words in Shakespeare's Hamlet (solid blue). The markedly different slopes demonstrate the broken power law distribution. The dotted red lines indicate straight line fits to (a) the 10 commonest words, and (b) words ranked 50-600. Region (c) illustrates deviation from the power law for low prevalence (single occurrences of words, which for the purpose of best-fit could be treated either as sequential points of increasing rank, as here, or as a single point). Of the three regions, only region (b) approaches a slope of -1, as predicted by Zipf's Law.

Given the challenges associated with using a traditional log-log frequency approach to estimate PLC in small animal communication datasets, it is important to design an alternative approach that circumvents these challenges. A promising alternative is to use an approach that relies on the information theory concept of Shannon entropy. Shannon entropy is a measure of "uncertainty" in a group of discrete elements (Shannon, 1948), and is fundamentally related to Zipf's law PLC (see Section 2.1). Communication systems with a highly repetitive, or stereotyped, repertoire (k<<-1) have low information entropy (signals are highly predictable), whereas systems with more random repertoires (k~0) have high entropy (signals are unpredictable). Shannon entropy is often used to describe the information potential in animal communication systems (Da silva et al., 2000; Doyle et al., 2008; Freeberg and Lucas, 2012; Suzuki et al., 2006). Yet, it remains to be seen whether an entropy approach could be used to estimate PLC in situations in which the traditional approach is unreliable.

In this study, we examine the efficacy of the entropy approach in estimating Zipf's law PLC. First, we lay out a mathematical proof showing that PLC can be robustly estimated via the normalised Shannon entropy *H* of a communication repertoire, due to this fundamental equivalence between *H* and the probability distribution of the communication elements (Corominas-Murtra and Solé, 2010). Estimating *k* via *H* reduces the sensitivity to deviations from the power law distribution, and therefore represents a promising approach to testing animal communication conformity to Zipf's Law. Then, we demonstrate the practical utility of an entropy approach in artificial and real-world datasets. We show the utility of the entropy approach in artificial data by generating a set of communication sequences with *a priori* known probability distributions, using sample sizes typical of animal communication studies. Then, we show the utility of the entropy approach in real-world datasets by comparing entropy- and traditional-derived PLC estimates in vocal repertoire datasets for 11 vertebrate species.

2. METHODS

2.1 Equivalence of Zipf's Law and Shannon entropy

Throughout this manuscript, we use "log" to refer to natural logarithm (base *e*). Assuming that the element frequencies follow a power law, then for *W* distinct element types, arranged in order of decreasing probability, the probability of the i^{th} type is given by:

$$P_i = Ai^k \tag{1}$$

where *k* is the PLC of the power distribution, $k=[-\infty,0]$, i.e. $-\infty \leq k \leq 0$. The normalisation constant *A* ensures that P_i sums to one, and is given by

$$A = \frac{1}{\sum_{i=1}^{W} i^k} \tag{2}$$

The normalised entropy *H* (taking values between [0,1]) of a sequence of elements selected from this distribution is then given directly by:

$$H = -\sum_{i=1}^{W} P_i \log_W P_i \tag{3}$$

where using the logarithm to the base of the repertoire size $\log_W P = \frac{\log_e P}{\log_e W}$ ensures that the normalised entropy *H* will take values between 0 and 1, independent of *W*.

Substituting Equation (1), the expression for H becomes:

$$H = -\sum_{i=1}^{W} Ai^{k} \log_{W} Ai^{k} \quad (4)$$

Crucially for animal communication studies where repertoire sizes are typically small compared to human language, it is clear that the relationship between the PLC k and normalised entropy H depends on the repertoire size W (Figure 2). In the limiting case as $W \rightarrow \infty$ and the slope of the curve becomes steeper, Zipf's Law, i.e. k=-1, applies to all values of H. For realistic repertoire sizes, however, even a signal with H=0.5, i.e. half way between stereotypy (H=0) and randomness (H=1), and would show a PLC of a substantially lower magnitude than -1. In addition, estimating the entropy H gives equal weight to rare elements with equal rank, whereas estimation of PLC via Zipf's Law may or may not, depending on the method used to perform linear fitting.

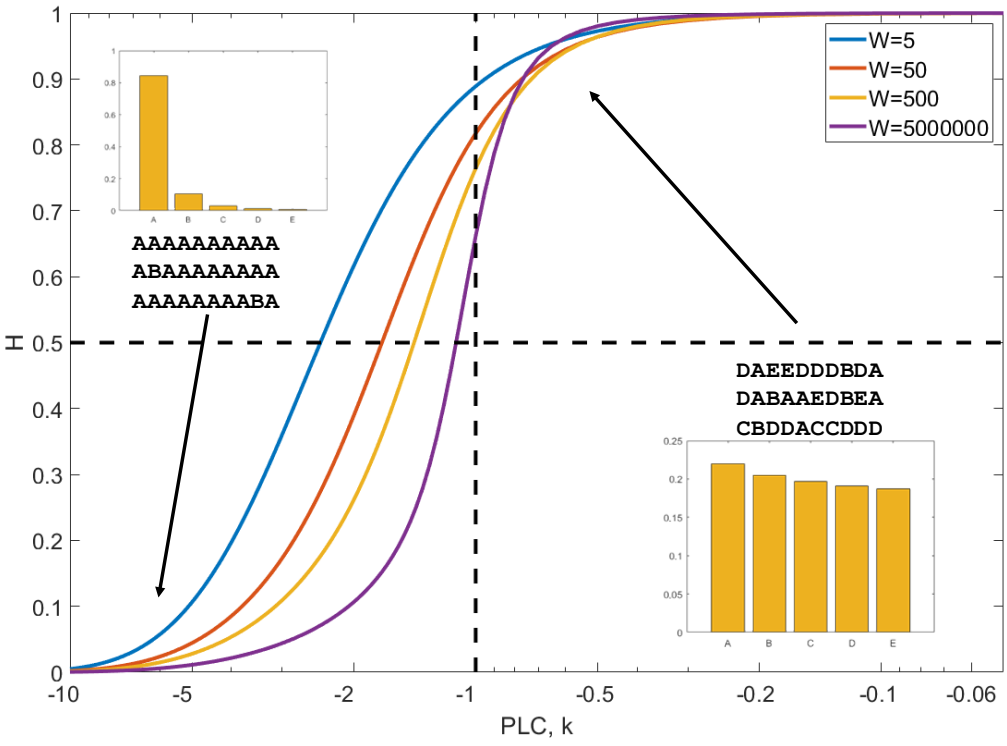


Figure 2. Shannon entropy H as a function of k, for different repertoire sizes W. The case where k=-1 represents Zipf's Law, and corresponds to H=0.5 as $W \rightarrow \infty$. The two inset boxes show examples of probability histograms for W=5 for a low entropy system (left), where successive elements are highly predictable (almost always "A"), and for a high entropy system (right), where each element has a similar

probability of occurrence, and so the sequence is highly unpredictable (note that for clarity, only a short sequence is shown, so that not all possible elements necessarily appear).

It is therefore straightforward to calculate the equivalent Shannon entropy H for any particular value of k. Conversely, however, no closed-form solution exists for the inverse problem of finding k given a value of the entropy H. Estimating k must therefore be done iteratively using a least-squares approach, assuming that the repertoire size W is known.

2.2 Estimating the PLC

Observed probabilities of sequence elements are unlikely to follow a power law distribution precisely. This is because of sampling errors (e.g., rare elements) and the broken power law effect (e.g., element generation operates differently at extreme high and low frequencies), as shown in Figure 1. As a result, estimation of PLC in empirical data is sometimes performed using only those elements that occur with middling probability (Corominas-Murtra et al., 2018). In this study, we use the log-log slope to estimate PLC in two ways: with the full set of element probabilities and with the coefficient calculated with the central 50% of the log-transformed ranks (Figure 1, region b).

2.3 Artificial data sets with known entropy

As the information complexity of empirical animal data sets is not known *a priori*, we generated random artificial sequences with three known distinct PLCs, k=[-0.5, -1, -2]. At the lowest value, k=-0.5, the sequences would show high entropy and would appear relatively random, at k=-1 the sequences should correspond to those expected by Zipf's Law, and at k=-2 the sequences would be more stereotyped than the others, with the more common elements being repeated much more often.

We also generated sequences with each of these values of k for multiple vocabulary (repertoire) sizes, W=[8, 32, 256, 1024, 8192, 65536]. The smallest repertoire sizes (8, 32) are typical of the number of calls

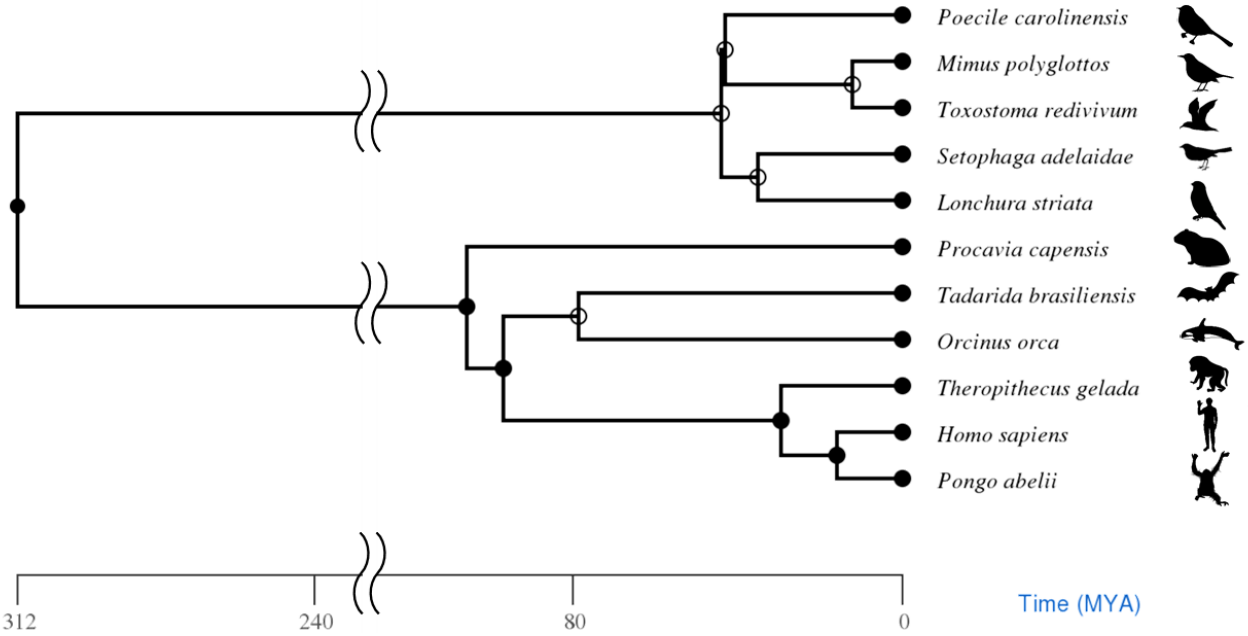
in many animal communication systems (Kershenbaum et al., 2016). W=256 and 1024 are unusual for animals, but not unheard of (e.g. Figure 3 California thrasher, 182 call types). Higher repertoire sizes are not found in animal communication, but begin to approximate the number of words in human language. The highest repertoire size, W=65536 approached the maximum size possible because of computational constraints.

For each of the sequences, we bootstrap sampled elements randomly with different sample sizes, S=[10, 40, 80, 320, 5120], representing the different sampling efforts possible in realistic field data collection scenarios. Each sampling size was repeated for each data set 1,000 times.

As an indication of the accuracy and precision of estimation of the information entropy, we calculated the mean and standard deviation of the estimated PLC using three methods. (a) Fitting a straight line to the log(probability) vs. log(rank) graph for all the sequence elements ("Full Zipf"). (b) Fitting a straight line to the middle 50% of the graph in (a), i.e. between ranks $e^{¼\log(W)}$ and $e^{¾\log(W)}$ ("Central Zipf"). (c) Estimating the PLC from the Shannon entropy of the probability distribution of sequence elements, from the relationship described in Equation 4, using the Matlab function *fmincon* to find the value of k that best predicts the measured entropy H for a particular vocabulary W.

2.4 Real animal data sets

Finally, we estimated the PLCs for empirical animal communication datasets taken from other studies (Figure 3), using each of the methods above. As the role of vocalisations varies greatly between the different species, e.g. male display song (songbirds, hyrax) vs. contact and other social calls (orca, bats) vs. complex messaging (humans), the units of analysis also varied greatly, and this breadth of signal purpose adds to the general applicability of the method. Although the true PLCs for these data are not known *a priori*, we bootstrap subsampled the data for each of the study sets 1000 times using each of the sample sizes used with artificial data sets in Section 2.3, i.e. S=[10, 40, 80, 320, 5120] (up to the maximum size of the empirical data set). Using these analyses, we calculated the variance, and hence the precision, in the estimate of the PLC using each of the three methods above: Full Zipf, Central Zipf, and Shannon entropy.



Species	Repertoire size	Repertoire unit	Number of samples	Source
Carolina chickadee *Poecile carolinensis*	10	Notes	44764	Freeberg, 2008
Northern mockingbird *Mimus polyglottos*	100	Song (mimicked)	3466	Gammon, 2014
California thrasher *Toxostoma redivivum*	182	Phrase	2800	Sasahara et al., 2012
Adelaide warbler *Setophaga adelaidae*	86	Song types	1808	Schraft et al., 2017
Bengal finch *Lonchura striata*	7	Syllable	29645	Jin and Kozhevnikov, 2011
Rock hyrax (male) *Procavia capensis*	8	Syllable	117746	Demartsev et al., 2019
Free-tailed bat *Tadarida brasiliensis*	3	Phrase	2385	Bohn et al., 2009
Orca *Orcinus orca*	31	Call	773	Crance et al., 2014
Gelada *Theropithecus gelada*	6	Call	4746	Gustison, Semple, Ferrer-I-Cancho, & Bergman, 2016
Human (Hamlet, English) *Homo sapiens*	4602	Word	32598	
Orangutan *Pongo abelii*	29	Syllables	2501	Lameira et al., 2013

Figure 3. Empirical animal vocalisation datasets used, and the phylogenetic relationship between the species. The 11 species include five birds and six mammals. Timeline from http://timetree.org/ (Kumar et al., 2017)

As an additional test of the practical relevance of the PLC estimates, we used the estimated PLCs calculated from the full empirical data for each species as *k*-values for artificial sets, as in 2.3. Using these power law distributions with known PLC, we generated 100 random data sets according to the probability distribution using Equation 1, but with sizes (number of sampled elements) corresponding to the actual size of the empirical datasets. For example, for the orangutan data set, we estimated the PLC, k_{orang}, and used that value to generate 100 artificial data sets of 570 elements each, taken from a repertoire of 7, using a power law distribution $P_i = Ai^{k_{orang}}$. For each of these random data sets, we compared the element frequencies to those in the empirical data set, and calculated the root mean squared error between the artificial and empirical distributions. This provides an estimate of the extent to which an approximation of the PLC is a reliable predictor of the element distribution in the animal's repertoire.

3. RESULTS

3.1 Artificial data sets with known PLC

All three methods, Full Zipf, Central Zipf, and Entropy generally converged on the correct value of *k* for large sample sizes and small vocabulary sizes (Figure 4). However, in the low entropy data set (*k*=-2.0), representing relatively stereotyped sequences, the Full Zipf method failed to converge on the correct value even at modest repertoire sizes (*W*>256). The Central Zipf and Entropy methods converged similarly well, although the Entropy method showed a lower error than Central Zipf at low sample sizes (SI Figure 1). For the low entropy data set, the Central Zipf method showed a substantially higher error at low sample sizes than either Entropy of Full Zipf. The Central Zipf method was also substantially more variable than either of the other methods, except in the high entropy data set (SI Figure 2).

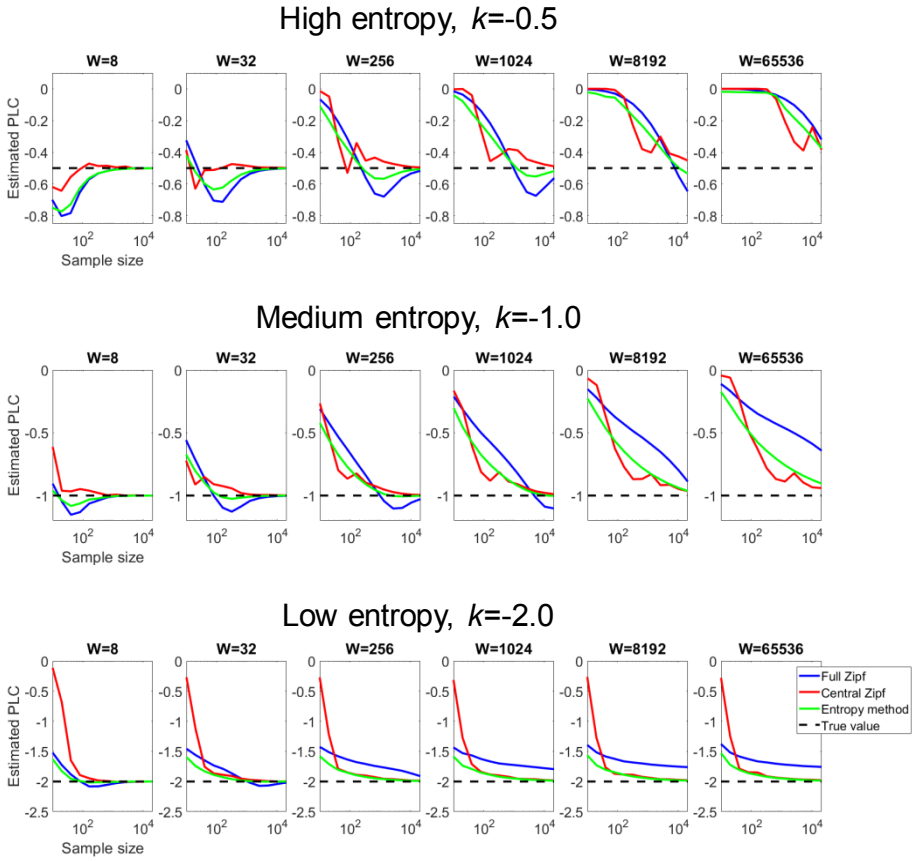


Figure 4. Estimated PLC for different simulated data sets. Each panel shows the true PLC (dashed line), and the estimated PLC for the three methods, Full, Central, and Entropy, for increasing sample size (within panel), and increasing vocabulary size (between panels). The upper panels show data generated for low PLC (k=-0.5), high entropy, producing random-like sequences, the middle panels for medium PLC (k=-1.0), medium entropy, producing sequences with statistical properties similar to human language, and the bottom panels for high PLC (k=-2.0), low entropy, stereotyped sequences.

3.2 Real animal data sets

On real data sets, although the true value of the PLC was unknown, the Central Zipf method showed substantially higher variation than the other methods (Figure 5). High variation (low precision) indicates that the Central Zipf method is likely a poor choice for estimating vocal complexity with realistic samples. Although the Full Zipf method generally showed variation similar to the Entropy method, most of the sampled species also showed increased variability at higher sample sizes.

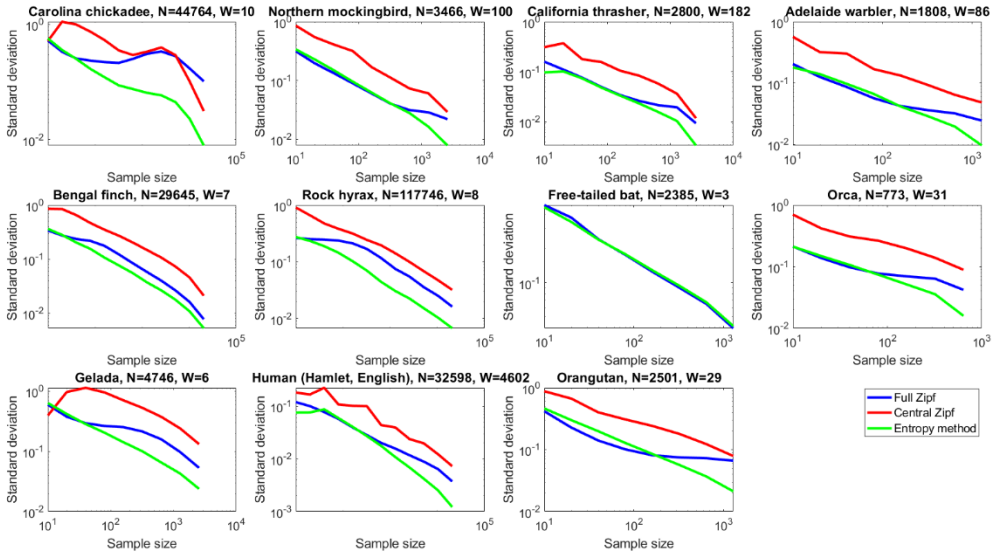


Figure 5. Variation in the estimated PLC when subsampling empirical animal data sets, for the three methods, Full, Central, and Entropy. The x-axis indicates the number of calls sampled from the data set to estimate PLC.

Examples of the probability distributions obtained by synthesising data sets based on the estimated PLC are shown in Figure 6. Both the Entropy and the Central Zipf methods generated similar distributions, except when the repertoire size was small (e.g. orangutan). Clearly, the Central Zipf method could not be used for the free-tailed bat, where W=3. The Entropy method generally produced distributions more similar to the empirical data than the other methods, except in the cases of the Carolina chickadee, the rock hyrax, the orca, and the California thrasher. In each of these, the Central Zipf method performed best, although the results of both the Central Zipf and Entropy methods were substantially better than the Full Zipf method (Figure 7). These four species were prominently characterised by the most infrequent elements deviating strongly from the power law distribution.

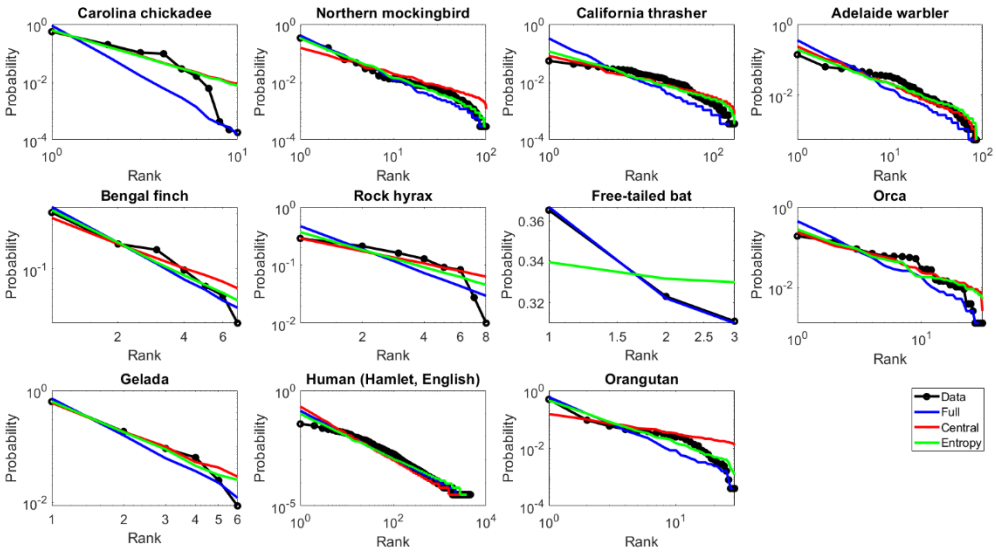


Figure 6. Probability of each element in the empirical animal data sets vs. rank (black line and points) on a log-log graph. Also shown are the probability-rank plots for sample data sets generated from a power law distribution using the estimate of the PLC by each of the three methods, Full, Central, and Entropy.

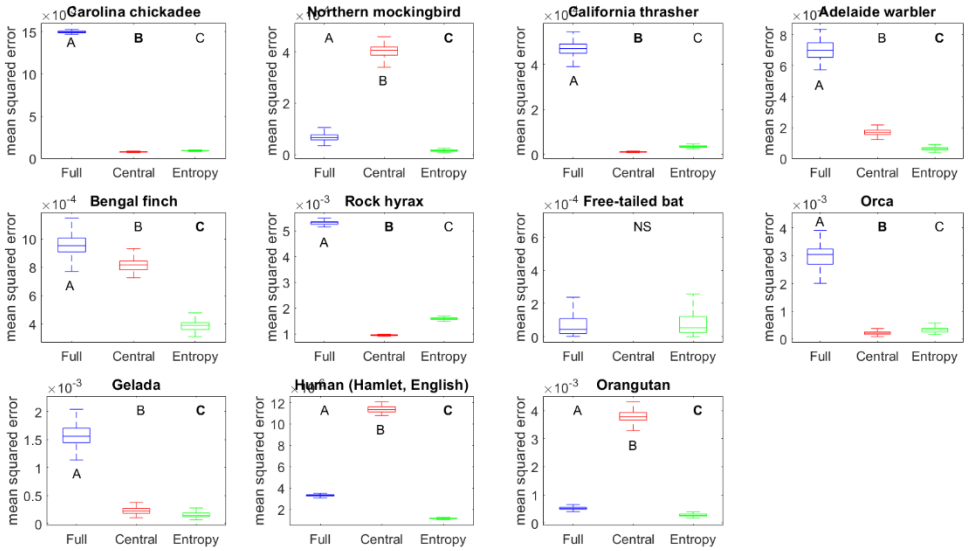


Figure 7. Mean squared error between the element probabilities in the empirical data sets, and the element probabilities in artificial datasets generated from a power law distribution using the estimate of the PLC by each of the three methods, Full, Central, and Entropy. Letters indicate post-hoc Tukey test classification, with the letter in bold showing the lowest error.

Comparing the estimates of PLCs for the full animal datasets (Figure 8), we can see that the Full Zipf method generally estimated a lower value of *k* relative to the other two methods. The Central Zipf and Entropy methods mostly agreed on which species fell within the region around *k*=-1, corresponding to a balance between complexity and simplicity, although the hyrax and the thrasher were considered by the Central Zipf method to be more random than by the Entropy method. It is worth noting that all three methods estimated that human language corresponds to *k*=-1, although the Central Zipf method ranked it as somewhat more stereotyped than the other methods.

Although many of the species we examined showed PLC estimates that were similar using each of the three methods, some species gave very differing results according to the different approaches. For example, the orangutan (one of the smallest sample sizes in our collection) was estimated to be highly stereotyped by the Central Zipf method (*k*<-2.5), whereas both the Full Zipf and Entropy methods suggested it to be much closer to the linguistic range (*k*=-1.6). Similarly, the orca, another dataset with a small sample size, was found to have a PLC very close to -1 by both the Entropy and Central Zipf methods, although the Full Zipf method predicted much more stereotyped behaviour (*k*=-1.6). If anything, the Entropy method resulted in a "flatter" curve in Figure 8, with the mean PLC across all species closer to *k*=-1, implying that more species conformed more closely to Zipf's Law than previously thought.

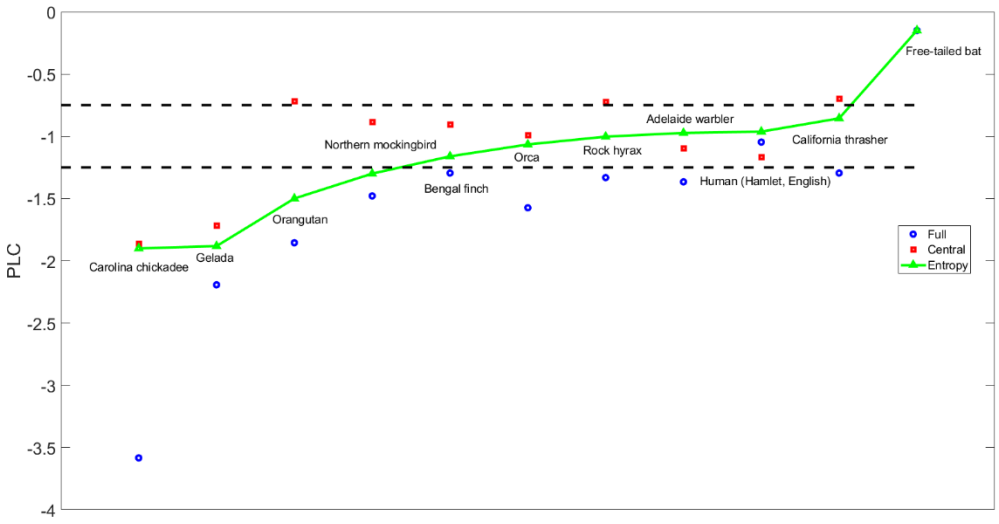


Figure 8. Animal data sets ordered by PLC as estimated by the Entropy method (green line + triangles), as well as coefficient estimates by the Full Zipf (blue circles) and Central Zipf (red squares) methods. Dashed black lines indicate the region around PLC k=-1 (± an arbitrary amount: 0.25) similar to human language.

4. DISCUSSION

Field zoologists are, by the nature of their study systems, often constrained to drawing conclusions from much smaller sample sizes than is the case in other fields of science. In the case of animal vocal communication, comparison with human language, and language's almost unlimited size of corpora available for testing, has led to considerable interest in "Zipf's Law" as an indicator of balance between complexity and simplicity. However, with smaller sample sizes than the human corpora (often millions of words long) that available for linguistic analysis, animal communication studies require a more robust metric than estimating PLC from the slope of a log-log graph. Theoretical analysis shows that Zipf's Law can be transformed into a formalisation measured in terms of information entropy, and we have shown that such an approach appears to be more robust for small sample sizes.

4.1 Overall suitability of the entropy method

We examined the suitability of traditional methods of estimating the PLC of animal vocal repertoires, calculating the slope of the rank-probability log-log graph using all the data points (Full Zipf) or only the central portion of data (Central Zipf), and compared them to a novel method, that of normalised Shannon entropy. Normalised entropy performed generally better than either of the traditional methods: converging on the true value of artificial data sets more quickly and more reliably and showing less variation within subsamples. The entropy method also showed substantially less variation in estimating the PLC from empirical animal data sets. Although the accuracy cannot be measured (i.e., the true PLC is not known for these real data), the entropy method was more consistent than the Central Zipf method. The Central Zipf method, on the other hand, was highly variable on animal data sets. The Entropy method also has the advantage that it can be calculated from data sets of any size, whereas the Central Zipf method in particular, relies on assumptions about what sized central region of the probability distribution should be chosen for measuring the rank-frequency slope. We have provided simple Matlab scripts to perform the Entropy PLC calculations from field data.

4.2 Why are some data sets mis-characterised by the entropy method?

For empirical data sets, the Entropy method also showed more consistency when measuring the PLC of a particular species, i.e. less variation across bootstrapped samples than other methods. The PLC estimated by the Entropy method generally allowed us to simulate artificial power law data sets that more closely corresponded to the empirical data in terms of element frequencies. However, the Entropy method was less successful than the Central Zipf method in simulating artificial data sets based on four of the species examined. These four species, the Carolina chickadee, the rock hyrax, the orca, and the California thrasher, all empirically showed element distributions that deviated strongly from a power law model. In particular, the most infrequent elements were substantially less frequent than predicted by a straight line model (Figure 9).

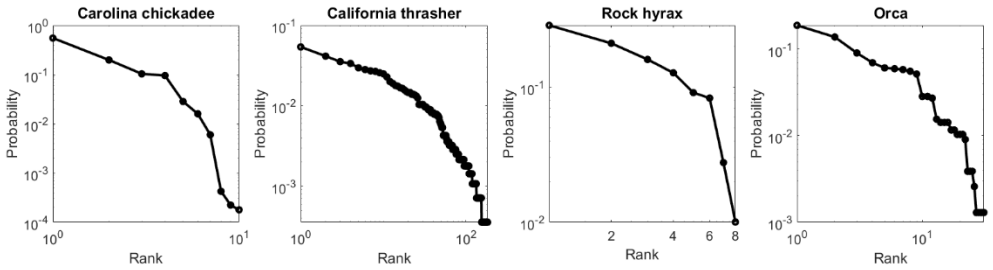


Figure 9. log-log plots of rank against element probability for the four species for which the Entropy method generated inaccurate simulated datasets. All of these species show particularly low probabilities for the least frequent elements.

We can only speculate at this stage whether the deviation from the power law in these species represents a true result of evolutionary trade-offs in the cognitive and production mechanisms governing communication, or whether it is an artefact of data collection, or a combination of both. Closer examination of the empirical distribution of element probabilities, compared to the probabilities generated by sampling a power law, illustrates well the problems of estimating complexity in animal repertoires. Figure 10 shows three of the species analysed: the gelada, with a small repertoire (W=6) and a small sample size (N=4,746); the rock hyrax, with a small repertoire (W=8) but a very large sample size (N=117,746); and the northern

mockingbird, with a very large repertoire (W=100) but a small sample size (N=3,466). Although for the gelada the sample size is small, the empirical data (red) correspond well to a power law (straight line), and so estimates of the PLC are likely to be accurate. However, it is unclear whether this correspondence is truly a reflection of an underlying power law nature of the communication mechanism in the animal, or merely an artefact of the small sample size. In contrast, the rock hyrax, with a similar repertoire size but much larger sample, shows the two least frequent elements far less common than predicted by a power law model. As a result, both the Full Zipf and the Entropy methods (which are calculated using all the data), deviate from the Central Zipf method, which excludes the infrequent elements. Again, however, it is unclear whether these extremely uncommon sounds are genuinely a part of the animal's evolutionarily adaptive communication strategy, or an artefact of data sampling (e.g. inter-individual variation, including due to abnormal behaviour, errors in coding, etc). The northern mockingbird, on the other hand, has an extremely large repertoire and relatively small sample, but nonetheless corresponds well to the power law model. It is probably safest to say that the optimal method used to quantify the communicative complexity of animal vocalisations via PLC requires a preliminary examination of how and whether the collected data correspond to the power law in the first place. As with any analysis involving linear assumptions, visual examination of the supposedly linear relationship (the log-log plot in this case) should be performed before applying quantitative methods.

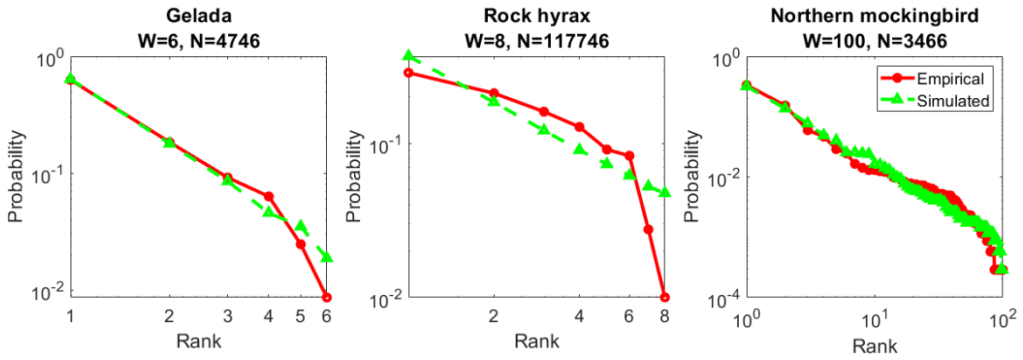


Figure 10. Empirical measurements (red) and a single instance of an artificial data set (green) of the same repertoire and sample size, generated from a power law distribution of similar PLC. Three species are shown: gelada (left) with small repertoire size and small sample size, rock hyrax (centre) with small repertoire size and large sample size, and mockingbird (right) with large repertoire size and small sample size.

4.3 Why estimate the power law coefficient of animal vocalisations?

Like all traits, animal communication is subject to evolutionary trade-offs, including effects of environmental transmission (Henry et al., 2011), the requirement for interspecific (species recognition) and intraspecific (individual recognition) specificity of the signaller (Kershenbaum et al., 2016), information content, and cognitive and motor constraints on the production of sound (Kipper et al., 2006). The question of the trade-off between information content vs. cognitive complexity has attracted particular interest because of its link to the evolution of human language (Ferrer-i-Cancho, 2005). Any communicative channel must increase in complexity if it is to contain more information, but complex communication places higher cognitive and motor demands on the animal, both for meaningful signal production, and accurate signal interpretation (Fedurek et al., 2017; Seyfarth and Cheney, 2010). There is considerable evidence that many animal species balance these trade-offs according to empirical laws of information encoding, such as the Menzerath-Altmann law (Gustison et al., 2016; Gustison and Bergman, 2017; Heesen et al., 2019), Zipf's law of brevity/abbreviation (Demartsev et al., 2019; Semple et al., 2010), and Zipf's law of least effort (this study;(Allen et al., 2019; Ferrer-i-Cancho and McCowan, 2009)). In itself, conforming to such laws does not indicate that an animal possesses true linguistic abilities, rather, that the evolutionary history of the species has had to contend with conflicting pressures of information content and cognitive-motor constraints. However, it does seem plausible that our own pre-human ancestors themselves met such balanced constraints in the pre-linguistic phase of human evolution, upon which other factors, such as social cognitive demands (Freeberg et al., 2012), could act to drive the evolution of language. As such, the balance shown by, amongst other metrics, the PLC, may indicate the extent to which an animal's communicative system has been subjected to pressures to increase information content, and therefore inescapably balancing information with cognitive and biomechanical simplicity.

Quantifying the "complexity" of animal communication remains an open area for debate. Metrics such as overall repertoire size have been used, particularly when comparing individuals of the same species, e.g. of varying nutritional state (Kipper et al., 2006), or members of closely related species (Freeberg et al., 2012). However, this cannot be a complete description, as two of the species in this study display repertoires of 100 element types or more, and yet neither the mockingbird nor the thrasher show signs of using this large information capacity to communicate the kind of complex concepts that could be supported by such a large repertoire. Previous studies have looked at the structure of transitions between elements in a communicative sequence as an indicator of either the non-randomness of sequences (Sayigh et al., 2013), or of the fundamental complexity itself (Kershenbaum, 2014). However, the probabilities of transition between different elements in a sequence is unlikely to be a good indicator of communicative complexity, because of the varied production mechanisms employed by different species (Kershenbaum et al., 2014). The PLC itself is not a measure of information complexity in communication, and certainly not an indicator

of intentional information content. Rather, PLC reflects maximum information potential in a way that is similar to entropy. However, a balanced PLC value does appear to point to a situation where animal communication has been evolutionarily optimised for efficient information content - although not necessarily optimised for maximum information content. Our observations that distantly related taxa (orca, hyrax, warbler, thrasher) show balanced PLC may indicate that optimised acoustic communication is a trait that has evolved independently in these taxa, rather than being conserved across vertebrates broadly, or being a trait found in primates specifically. In this work, we have looked only at the diversity of low level vocal elements (words, calls, syllables), but information theory techniques can also be used to analyse higher level syntactic structures, as well as hierarchical organisation in communication (Garland et al., 2017; McCowan et al., 1999).

4.4 Conclusion

Appealing as the concept of Zipf's Law and "a straight line graph with slope -1" are, they are nonetheless merely indicators of the extent to which a communicative repertoire is balanced between complexity and simplicity. The popular method of estimating the coefficient of a power law distribution, which we have dubbed the Central Zipf method, places an emphasis on the comparison to the strong linear relationship seen in human language. However, animal communication repertoire sizes are much smaller than corpora of human language, as are the available animal vocalization databases. Although the fundamental equivalence between the power law coefficient k, and the Shannon entropy H, mean that either could be used interchangeably, it may be more conservative to quantify the balance between complexity and simplicity in animal communication directly via entropy, rather than attempting to fit limited data to the model seen in human language. In any case, as an entropy-based approach is more precise than slope measurements across both animal and artificial data sets, we recommend the entropy estimate of the PLC as a more rigorous indicator of the extent to which an animal repertoire is balanced in information complexity. We provide the following guide (Table 1), based on our results shown in Figure 4, as a guide to choosing an appropriate method, according to the sample and vocabulary sizes.

Table 1. Recommended complexity method for data sets with varying sample and vocabulary sizes. Recommendations are given separately for stereotyped signals, consisting mostly of very common elements, and random signals, where element frequencies are more evenly distributed. Recommended methods are E (Entropy), F (Full Zipf), and C (Central Zipf).

		Stereotyped calls (*low entropy*)		Random calls (*high entropy*)	
		Sample size			
		Low	*High*	*Low*	*High*
Vocabulary size	*Low*	E/F	any	C	E/F
	High	E	E/C	none	E/C

5. ACKNOWLEDGEMENTS

We thank Todd Freeberg, David Logue, Dezhe Jin, and Kisi Bohn for permission to use their datasets.

6. AUTHORS' CONTRIBUTIONS

AK conceived the study, carried out the experiments and analysis, and prepared the manuscript. All authors (AK, VD, DG, EG, MG, AI, AL) gathered data, and contributed to refining the analysis and the manuscript.

7. DATA AVAILABILITY

Animal vocal datasets were obtained from the published literature and the various authors of those studies. Matlab scripts detailing the analysis carried out are available via Zenodo (Kershenbaum et al., 2020) https://doi.org/10.5281/zenodo.4288793

8. REFERENCES

Allen, J.A., Garland, E.C., Dunlop, R.A., Noad, M.J., 2019. Network analysis reveals underlying syntactic features in a vocally learnt mammalian display, humpback whale song. Proc. R. Soc. B Biol. Sci. 286, 20192014. https://doi.org/10.1098/rspb.2019.2014

Bergman, T.J., Beehner, J.C., Painter, M.C., Gustison, M.L., 2019. The speech-like properties of nonhuman primate vocalizations. Anim. Behav. 151, 229–237. https://doi.org/10.1016/j.anbehav.2019.02.015

Boë, L.-J., Sawallis, T.R., Fagot, J., Badin, P., Barbier, G., Captier, G., Ménard, L., Heim, J.-L., Schwartz, J.-L., 2019. Which way to the dawn of speech?: Reanalyzing half a century of debates and data in light of speech science. Sci. Adv. 5, eaaw3916. https://doi.org/10.1126/sciadv.aaw3916

Cheney, D.L., Seyfarth, R.M., 1998. Why Animals Don't Have Language. Tann. Lect. Hum. Values 38.

Corominas-Murtra, B., Seoane, L.F., Solé, R., 2018. Zipf's Law, unbounded complexity and open-ended evolution. J. R. Soc. Interface 15, 20180395. https://doi.org/10.1098/rsif.2018.0395

Corominas-Murtra, B., Solé, R.V., 2010. Universality of Zipf's law. Phys. Rev. E 82, 011102. https://doi.org/10.1103/PhysRevE.82.011102

Crockford, C., Wittig, R.M., Zuberbühler, K., 2017. Vocalizing in chimpanzees is influenced by social-cognitive processes. Sci. Adv. 3, e1701742. https://doi.org/10.1126/sciadv.1701742

Da silva, M.L., Piqueira, J.R.C., Vielliard, J.M.E., 2000. Using Shannon Entropy on Measuring the Individual Variability in the Rufous-bellied Thrush Turdus rufiventris Vocal Communication. J. Theor. Biol. 207, 57–64. https://doi.org/10.1006/jtbi.2000.2155

de Boer, B., Thompson, B., Ravignani, A., Boeckx, C., 2020. Evolutionary Dynamics Do Not Motivate a Single-Mutant Theory of Human Language. Sci. Rep. 10, 1–9. https://doi.org/10.1038/s41598-019-57235-8

Decker, E.H., Kerkhoff, A.J., Moses, M.E., 2007. Global Patterns of City Size Distributions and Their Fundamental Drivers. PLoS ONE 2, e934. https://doi.org/10.1371/journal.pone.0000934

Demartsev, V., Gordon, N., Barocas, A., Bar‐Ziv, E., Ilany, T., Goll, Y., Ilany, A., Geffen, E., 2019. The "Law of Brevity" in animal communication: Sex-specific signaling optimization is determined by call amplitude rather than duration. Evol. Lett. 3, 623–634. https://doi.org/10.1002/evl3.147

Doyle, L.R., McCowan, B., Hanser, S.F., Chyba, C., Bucci, T., Blue, J.E., 2008. Applicability of information theory to the quantification of responses to anthropogenic noise by southeast Alaskan humpback whales. Entropy 10, 33–46.

Doyle, L.R., McCowan, B., Johnston, S., Hanser, S.F., 2011. Information theory, animal communication, and the search for extraterrestrial intelligence. Acta Astronaut., SETI Special Edition 68, 406–417. https://doi.org/10.1016/j.actaastro.2009.11.018

Dunbar, R.I.M., 2003. The social brain: mind, language, and society in evolutionary perspective. Annu. Rev. Anthropol. 32, 163–181.

Fedurek, P., Zuberbühler, K., Semple, S., 2017. Trade-offs in the production of animal vocal sequences: insights from the structure of wild chimpanzee pant hoots. Front. Zool. 14, 50. https://doi.org/10.1186/s12983-017-0235-8

Ferrer-i-Cancho, R., 2005. Zipf's law from a communicative phase transition. Eur. Phys. J. B - Condens. Matter Complex Syst. 47, 449–457. https://doi.org/10.1140/epjb/e2005-00340-y

Ferrer-i-Cancho, R., McCowan, B., 2009. A Law of Word Meaning in Dolphin Whistle Types. Entropy 11, 688–701. https://doi.org/10.3390/e11040688

Ferrer-i-Cancho, R., Solé, R.V., 2003. Least effort and the origins of scaling in human language. Proc. Natl. Acad. Sci. 100, 788–791. https://doi.org/10.1073/pnas.0335980100

Fitch, W.T., 2005. The Evolution of Language: A Comparative Review. Biol. Philos. 20, 193–203. https://doi.org/10.1007/s10539-005-5597-1

Freeberg, T.M., Dunbar, R.I.M., Ord, T.J., 2012. Social complexity as a proximate and ultimate factor in communicative complexity. Philos. Trans. R. Soc. B Biol. Sci. 367, 1785–1801. https://doi.org/10.1098/rstb.2011.0213

Freeberg, T.M., Lucas, J.R., 2012. Information theoretical approaches to chick-a-dee calls of Carolina chickadees (Poecile carolinensis). J. Comp. Psychol. 126, 68–81. https://doi.org/10.1037/a0024906

Garland, E.C., Rendell, L., Lilley, M.S., Poole, M.M., Allen, J., Noad, M.J., 2017. The devil is in the detail: Quantifying vocal variation in a complex, multi-levelled, and rapidly evolving display. J. Acoust. Soc. Am. 142, 460–472. https://doi.org/10.1121/1.4991320

Gustison, M.L., Bergman, T.J., 2017. Divergent acoustic properties of gelada and baboon vocalizations and their implications for the evolution of human speech. J. Lang. Evol. 2, 20–36. https://doi.org/10.1093/jole/lzx015

Gustison, M.L., Semple, S., Ferrer-i-Cancho, R., Bergman, T.J., 2016. Gelada vocal sequences follow Menzerath's linguistic law. Proc. Natl. Acad. Sci. 113, E2750–E2758. https://doi.org/10.1073/pnas.1522072113

Hauser, M.D., Chomsky, N., Fitch, W.T., 2002. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? Science 298, 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Heesen, R., Hobaiter, C., Ferrer-i-Cancho, R., Semple, S., 2019. Linguistic laws in chimpanzee gestural communication. Proc. R. Soc. B Biol. Sci. 286, 20182900. https://doi.org/10.1098/rspb.2018.2900

Henry, K.S., Gall, M.D., Bidelman, G.M., Lucas, J.R., 2011. Songbirds tradeoff auditory frequency resolution and temporal resolution. J. Comp. Physiol. A 197, 351–359. https://doi.org/10.1007/s00359-010-0619-0

Ilany, A., Barocas, A., Kam, M., Ilany, T., Geffen, E., 2013. The energy cost of singing in wild rock hyrax males: evidence for an index signal. Anim. Behav., Including Special Section: Behavioural Plasticity and Evolution 85, 995–1001. https://doi.org/10.1016/j.anbehav.2013.02.023

Jackendoff, R., 1999. Possible stages in the evolution of the language capacity. Trends Cogn. Sci. 3, 272–279.

Kershenbaum, A., 2014. Entropy rate as a measure of animal vocal complexity. Bioacoustics 23, 195–208. https://doi.org/10.1080/09524622.2013.850040

Kershenbaum, A., Blumstein, D.T., Roch, M.A., Akçay, Ç., Backus, G., Bee, M.A., Bohn, K., Cao, Y., Carter, G., Cäsar, C., Coen, M., DeRuiter, S.L., Doyle, L., Edelman, S., Ferrer‐i‐Cancho, R., Freeberg, T.M., Garland, E.C., Gustison, M., Harley, H.E., Huetz, C., Hughes, M., Bruno, J.H., Ilany, A., Jin, D.Z., Johnson, M., Ju, C., Karnowski, J., Lohr, B., Manser, M.B., McCowan, B., Mercado, E., Narins, P.M., Piel, A., Rice, M., Salmi, R., Sasahara, K., Sayigh, L., Shiu, Y., Taylor, C., Vallejo, E.E., Waller, S., Zamora‐Gutierrez, V., 2016. Acoustic sequences in non-human animals: a tutorial review and prospectus. Biol. Rev. 91, 13–52. https://doi.org/10.1111/brv.12160

Kershenbaum, A., Bowles, A.E., Freeberg, T.M., Jin, D.Z., Lameira, A.R., Bohn, K., 2014. Animal vocal sequences: not the Markov chains we thought they were. Proc. R. Soc. B Biol. Sci. 281, 20141370. https://doi.org/10.1098/rspb.2014.1370

Kershenbaum, Arik, Demartsev, Vlad, Gammon, David, Geffen, Eli, Gustison, Morgan, Ilany, Amiyaal, & Lameira, Adriano. (2020). Shannon entropy as a robust estimator of Zipf's Law in animal vocal communication repertoires. Methods in Ecology and Evolution. Zenodo. http://doi.org/10.5281/zenodo.4288794

Kipper, S., Mundry, R., Sommer, C., Hultsch, H., Todt, D., 2006. Song repertoire size is correlated with body measures and arrival date in common nightingales, Luscinia megarhynchos. Anim. Behav. 71, 211–217. https://doi.org/10.1016/j.anbehav.2005.04.011

Kumar, S., Stecher, G., Suleski, M., Hedges, S.B., 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol. Biol. Evol. 34, 1812–1819. https://doi.org/10.1093/molbev/msx116

Lameira, A.R., 2017. Bidding evidence for primate vocal learning and the cultural substrates for speech evolution. Neurosci. Biobehav. Rev. 83, 429–439. https://doi.org/10.1016/j.neubiorev.2017.09.021

Lameira, A.R., Call, J., 2018. Time-space–displaced responses in the orangutan vocal system. Sci. Adv. 4, eaau3401. https://doi.org/10.1126/sciadv.aau3401

Lestrade, S., 2017. Unzipping Zipf's law. PLOS ONE 12, e0181987. https://doi.org/10.1371/journal.pone.0181987

Martins, P.T., Boeckx, C., 2019. Language evolution and complexity considerations: The no half-Merge fallacy. PLOS Biol. 17, e3000389. https://doi.org/10.1371/journal.pbio.3000389

McCowan, B., Doyle, L.R., Jenkins, J.M., Hanser, S.F., 2005. The appropriate use of Zipf's law in animal communication studies. Anim. Behav. 69, F1–F7. https://doi.org/10.1016/j.anbehav.2004.09.002

McCowan, B., Hanser, S.F., Doyle, L.R., 1999. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. Anim. Behav. 57, 409–419. https://doi.org/10.1006/anbe.1998.1000

Newman, M.E.J., 2005. Power laws, Pareto distributions and Zipf's law. Contemp. Phys. 46, 323–351. https://doi.org/10.1080/00107510500052444

Sayigh, L., Quick, N., Hastie, G., Tyack, P., 2013. Repeated call types in short-finned pilot whales, Globicephala macrorhynchus. Mar. Mammal Sci. 29, 312–324. https://doi.org/10.1111/j.1748-7692.2012.00577.x

Schlenker, P., Chemla, E., Zuberbühler, K., 2016. What Do Monkey Calls Mean? Trends Cogn. Sci. 20, 894–904. https://doi.org/10.1016/j.tics.2016.10.004

Semple, S., Hsu, M.J., Agoramoorthy, G., 2010. Efficiency of coding in macaque vocal communication. Biol. Lett. 6, 469–471. https://doi.org/10.1098/rsbl.2009.1062

Seyfarth, R.M., Cheney, D.L., 2010. Production, usage, and comprehension in animal vocalizations. Brain Lang., Special Issue on Language and Birdsong 115, 92–100. https://doi.org/10.1016/j.bandl.2009.10.003

Shannon, C.E., 1948. A Mathematical Theory of Communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Suzuki, R., Buck, J.R., Tyack, P.L., 2006. Information entropy of humpback whale songs. J. Acoust. Soc. Am. 119, 1849–1866. https://doi.org/10.1121/1.2161827

Suzuki, R., Buck, J.R., Tyack, P.L., 2005. The use of Zipf's law in animal communication analysis. Anim. Behav. 69, F9–F17. https://doi.org/10.1016/j.anbehav.2004.08.004

Yu, S., Xu, C., Liu, H., 2018. Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. ArXiv180701855 Cs.

Zanette, D.H., Manrubia, S.C., 2001. Vertical transmission of culture and the distribution of family names. Phys. Stat. Mech. Its Appl., Proceedings of the IUPAP International Conference on New Trends in the Fractal Aspects of Complex Systems 295, 1–8. https://doi.org/10.1016/S0378-4371(01)00046-2

Zipf, G.K., 1949. Human behavior and the principle of least effort, Human behavior and the principle of least effort. Addison-Wesley Press, Oxford, England.