ANIMAL BEHAVIOUR, 2005, **69**, F9–F17
doi:10.1016/j.anbehav.2004.08.004

Available online at www.sciencedirect.com

SCIENCE DIRECT®

The use of Zipf's law in animal communication analysis

RYUJI SUZUKI*, JOHN R. BUCK† & PETER L. TYACK‡

*Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology
†Department of Electrical and Computer Engineering, University of Massachusetts Dartmouth
‡Biology Department, Woods Hole Oceanographic Institution

(Received 7 July 2003; initial acceptance 11 July 2003; final acceptance 31 August 2004; published online 11 November 2004; MS. number: AF-23)

Information theory has been discussed as a technique to analyse communicative processes or sequential behaviour of nonhuman animals, as in MacKay (1972), Slater (1973) and Bradbury & Vehrencamp (1998, chapters 13–15) among others. Recently, McCowan et al. (1999) proposed the use of information theory for their study of bottlenose dolphin, *Tursiops truncatus*, whistles. They discussed several aspects of their analysis techniques. Although we agree about the effectiveness of information theory to analyse unknown sources, we would like to further the discussion of one analysis method used in McCowan et al. (1999). Specifically, we wish to illustrate that Zipf's law is of little use in the analysis of communication signals. The presence or absence in dolphins and other animals of some features of human language remain intriguing and open questions (Tyack 1999). However, we assert that a Zipf-based technique is methodologically inappropriate to address these questions.

McCowan et al. (1999, page 410) noted that 'Few investigators of animal behaviour have examined the use of first-order entropic analysis known as Zipf's law or statistic'. In fact, Zipf's law has been discarded as a linguistic tool, strongly criticized by Miller (1957), Miller & Chomsky (1963) and more thoroughly by Rapoport (1982). McCowan et al. (1999, page 411) also cite the application of Zipf's law to DNA sequences by Mantegna et al. (1994) 'with varying interpretations and reliability (Flam 1994; Damashek 1995; Bonhoeffer et al. 1996; Israeloff et al. 1996; Voss 1996)'. These references' interpretations vary from strong criticisms of the use of Zipf's law in the Mantegna et al. study (Bonhoeffer et al. 1996; Israeloff et al. 1996; Voss 1996) to a short news item about the upcoming publication of Mantegna et al.'s article (Flam 1994), to no mention whatsoever of Mantegna or Zipf (Damashek 1995). Besides the original Mantegna et al. paper, none of the references cited by McCowan et al. support Zipf's law as methodologically appropriate to the DNA study (see also Martindale & Konopka 1996). Both the linguistics and DNA communities have roundly rejected Zipf's law as a diagnostic tool.

Tests based on Zipf's law are highly susceptible to false positives, both in theory and practice. Consequently, when Zipf's law is used as a test for linguistic, communicative or otherwise meaningful processes, as in McCowan et al. (1999), the results are uninterpretable, even if estimates of the Zipf statistic are appropriately obtained. We present two simple probabilistic examples illustrating this issue, one in which a meaningless process satisfies the test proposed by McCowan et al. (1999), and another in which a meaningful process (this manuscript) fails the test. Moreover, we will illustrate that the estimation procedure used by McCowan et al. (1999) and Zipf (1949) is underconstrained and produces results that are not internally consistent. A properly constrained and internally consistent estimation process for the Zipf parameter would degrade the R^2 values of Table 1 in McCowan et al. (1999). Finally, McCowan et al. (1999) use some important technical terms without clear explicit definition, or define them differently from the information theory literature upon which they draw. To avoid any confusion, we will explicitly define 'random', 'entropy' and 'entropies', and 'first-order' before embarking on a discussion of the problems in using Zipf's law as a diagnostic test.

First, McCowan et al. (1999) used 'random' as synonymous with 'independently uniformly distributed' when describing information sources or communication systems. In the engineering and mathematics communities, which gave birth to the information theory McCowan et al. promote for the study of animal communication, 'random' is used to denote any process or event with a nondeterministic or stochastic component, and not the highly restrictive meaning McCowan et al. assign to it. For example, probability and statistics texts commonly

Correspondence: R. Suzuki, Speech and Hearing Bioscience and Technology, Harvard-MIT Division of Health Science and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A. (email: rsuzuki@mit.edu). J. R. Buck is at the Department of Electrical and Computer Engineering and the School for Marine Science and Technology, University of Massachusetts Dartmouth, 285 Old Westport Road, North Dartmouth, MA 02747-2300, U.S.A. P. L. Tyack is at the Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, U.S.A.

0003–3472/04/$30.00/0
© 2004 The Association for the Study of Animal Behaviour. Published by Elsevier Ltd. All rights reserved.

discuss 'random variables' that are not independent or uniformly distributed (Feller 1968, page 212; Papoulis 1984, page 63). Loève (1963, page 497) notes 'In the literature, the terms ''random function'', ''random process'' and ''stochastic process'' are treated as synonymous'. The potential for confusion arises when McCowan et al. use 'nonrandom' as synonymous with 'possibly stochastic but not independently uniformly distributed' rather than the common definition of 'deterministic'. McCowan et al.'s (1999, page 411) statement that 'Such a function nevertheless remains a valid indication of both the nonrandomness of a system as well as the potential capacity for communication transfer of such a system' appears to be a stronger conclusion than it is. In common terminology, they are stating that Zipf's statistic can test whether the sample points deviate from being equiprobable.

Second, McCowan et al. (1999) do not clearly distinguish between the entropy of a process, which is a fundamental property of the source, and the estimates of this entropy derived from various-order Markov model assumptions. (A stochastic process, information source, or simply a process or source, produces a sequence of random variables, to which animals' vocalization signals are abstracted.) Shannon (1948) clearly defined the entropy H of an information source, but he did not define the 'entropic orders' nor the multiple 'Shannon entropies' that McCowan et al. (1999, e.g. page 410) frequently reference. Although Shannon's discussion of approximating English text through a sequence of successively higher-order Markov models might superficially be considered entropic ordering, neither Shannon (1948) nor McCowan et al. (1999) define 'entropic orders'. If McCowan et al. used 'entropic orders' to refer to the succession of estimates of the true entropy H obtained from these models, then their discussion confounds the distinction between the source's true entropy H and the model-based estimates of this quantity, because entropy is a fundamental property of an information source, and as such is unordered.

Third, McCowan et al. (1999, page 410) argued that Zipf's law is 'a first-order entropic analysis' without defining the term 'first-order'. This term has different meanings in different contexts. Since 'a first-order entropic analysis' is not a well-known analysis among information theorists, the term must be clearly defined for the readers to understand and judge the argument. Because several arguments about their use of 'first-order' are strongly linked to their use of Zipf's law, we will discuss this issue in detail in the course of discussing Zipf's law.

Zipf (1949) observed that, for many human languages, a plot of the frequency of words against the rank of occurrence on doubly logarithmic axes (log frequency versus log rank) is well approximated by a straight line with a slope around -1. We will call this slope the 'Zipf's statistic', denoted by α. Zipf postulated that this relationship held for all human languages, and this relation became known as Zipf's law. Although McCowan et al. (1999, page 411) assert that $\alpha \approx -1$ 'optimizes the amount of potential communication that can be carried through a channel from speaker to receiver', neither they nor Zipf (1949) provided a mathematical proof to support this assertion. As noted in Rapoport (1982, page 3), Zipf's law's purported theoretical basis 'the principle of Least Effort is stated in vague, connotation-ridden language, precluding rigorous deduction'. Zipf's law thus is an empirical fit of many observations to a vaguely defined theory of a simple form. It is not derived from or proved on a set of assumptions about the intrinsic properties of languages. As such, the term 'law' should be loosely interpreted.

Mandelbrot (1952) considered the optimal transmission of information in a word-by-word manner. His premise was that the cost of sending a single word is proportional to the number of letters in the word (including the spaces separating the words), and he considered the information conveyed per unit cost. He found that in order to achieve optimal efficiency under this operating condition, the source must obey Zipf's distribution. 'There is a strong temptation to reverse the implication and to argue that because we obey Zipf's law we must therefore be communicating word-by-word with maximal efficiency' (Miller 1954, page 415). Although natural languages are observed to obey Zipf's distribution, 'they are far less efficient than they could be under the constraints Mandelbrot imposes' (Miller 1954, page 415). This counterexample indicates that the converse of Mandelbrot's result does not hold. That is, under Mandelbrot's condition, Zipf's distribution is necessary for optimal information transmission, but not sufficient. Note that McCowan et al. (1999, page 411) claim the slope of $\alpha \approx -1$ 'optimizes the amount of potential communication' and not the amount of actual communication. In the absence of evidence that dolphins transmit whistles under Mandelbrot's constraints, neither potential nor actual communication is optimized by $\alpha \approx -1$, and thus there is no basis for this claim of optimality. Basic information theory results show that $\alpha = 0$ maximizes the amount of potential communication (Cover & Thomas 1991, page 27) in the absence of Mandelbrot's word-by-word encoding and minimum cost constraint conditions.

Note that a Zipf's statistic of $\alpha \approx 0$ is neither a necessary nor a sufficient condition to conclude that a source contains no communicative or linguistic content. For example, the purpose of data compression is to transform input data into the most compact form from which the original data can be later faithfully retrieved, optimizing the amount of communication per symbol to be transmitted. As a result of this transformation, the compressed data appear as close to independent and equiprobable as possible (Visweswariah et al. 1998), and thus will have $\alpha \approx 0$. When communication is not restricted by Mandelbrot's conditions, $\alpha = 0$ maximizes the potential communicative ability of the source. Data compression algorithms are consistent with this principle.

Example (Data Compression)

Analysing this manuscript, including punctuation, as a text file on a character by character basis with a Zipf's statistic yields $\hat{\alpha} = -1.95$ ($\hat{\alpha}$ denotes a statistical estimate of true α from sample data). After this text is compressed

by a standard gzip data compression utility, $\hat{\alpha} = -0.12$ is obtained in the same procedure. Both files contain the same information and communicative content, but provide radically different values of $\hat{\alpha}$. The strongest statement that can be made about a source with $\alpha \approx 0$ is that researchers without external knowledge of the data format or encoding are unlikely to detect or interpret any linguistic content that may be present.

Zipf's distribution is recognized to be a simple consequence of taking rank as the independent variable under some general conditions that are widely observed in, or underlying, many processes (Miller 1957; Miller & Chomsky 1963; Li 1992). By so doing, many sources with different distributions show an inverse power or Zipf distribution, 'without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations' (Miller 1957, page 314). This article repeats two major points from past discussions of Zipf's law: (1) the Zipf's distribution is observed in many noncommunicative and nonlinguistic processes; therefore, it cannot discriminate communicative processes from meaningless and purposeless processes such as noise; (2) the persistent slope of roughly $\alpha \approx -1$ for Zipf's plots does not imply that processes are similar, but rather that the slope is insensitive to large changes in the underlying mechanism or probabilistic description of the processes.

Many noncommunicative systems are known to exhibit the Zipf's distribution. Although McCowan et al. (1999, page 411) claim that the Zipf's statistic 'measures the potential capacity for information transfer' in a system, even a simple stochastic mechanism, such as a die-rolling trial, satisfies Zipf's law. The following example demonstrates that a stochastic process devoid of semantic or communicative content may still satisfy Zipf's law.

Example (Die-Rolling)

We roll a fair cubic die repeatedly. We treat the number resulting from each roll as a letter, which we write down after each roll. We arbitrarily choose 6 to represent a space forming the break between successive words. A resulting string might begin '5 _ 42 22 5133 _ 2 4…' where '_' denotes the null word (made visible), which occurs when we see two adjacent spaces (i.e. 6 is rolled twice in a row). The analysis below is based on Li (1992), but allows the null word with length zero to keep the analysis simpler. Li (1992) excludes the null word but shows a very similar result. Miller (1957) and Miller & Chomsky (1963) present a similar example based on equiprobably and identically distributed random digits.

Let n_ℓ denote the number of distinct possible words of length ℓ, p_ℓ the probability of each of these words, and $R(\ell)$ the set of ranks of words of length ℓ, one of whose elements is represented by $r(\ell)$. In this example, all words of the same length are equiprobable.

At the beginning of the experiment, the most probable word is the null word, which occurs when we roll a 6 first. Thus, the null word has probability $p_0 = 1/6$, $n_0 = 1$, and $R(0) = \{1\}$. If the first die roll is not a 6, then the next most probable words are $n_1 = 5$ words each containing one letter, all obtained by rolling a number from 1, 2,…, 5 on the first roll, followed by a 6 on the second roll. The probability of any one of these words is $(1/6)(1/6) = 1/36$, and the ranks for these one-letter words, $R(1)$, are {2, 3, 4, 5, 6}. Generalizing to words of length ℓ, we have

$$n_\ell = 5^\ell \tag{1}$$

$$p_\ell = \left(\frac{1}{6}\right)^\ell \frac{1}{6} = 6^{-(\ell+1)} \tag{2}$$

and

$$R(\ell) = \left\{ n : \sum_{k=0}^{\ell-1} 5^k < n \leq \sum_{k=0}^{\ell} 5^k \right\}. \tag{3}$$

Note that the analysis for all successive words following the space or a 6 that ends the previous word is identical to that above for the beginning of the experiment.

It is clear from equation (2) that $0 < p_\ell < 1$ for all ℓ. It is also clear that each event consists of a single sample point. Let us verify that the probabilities sum to one. There are $n_\ell = 5^\ell$ different words of length ℓ each of whose individual probability is $p_\ell = 6^{-(\ell+1)}$, for each $\ell = 0, 1, 2, \ldots$ Therefore,

$$\begin{aligned} \sum_{\ell=0}^{\infty} n_\ell p_\ell &= \sum_{\ell=0}^{\infty} 5^\ell 6^{-(\ell+1)} \\ &= \frac{1}{6} \sum_{\ell=0}^{\infty} \left(\frac{5}{6}\right)^\ell \\ &= \frac{1}{6} \frac{1}{1-\frac{5}{6}} \\ &= 1, \end{aligned}$$

and thus equations (1) and (2) satisfy the axiom of probability (Loève 1963, pp. 8–9, 15–17; Feller 1968, page 22).

Now, we find the slope of the straight line matching this rank–probability plot on doubly logarithmic axes. Computing the geometric series, equation (3) simplifies to

$$\frac{5^\ell - 1}{4} < r(\ell) \leq \frac{5^{\ell+1} - 1}{4}.$$

Rearranging and taking the base 5 log of all terms yields

$$\ell < \log_5[4r(\ell) + 1] \leq \ell + 1.$$

Raising 1/6 to the power of each term and substituting equation (2),

$$p_\ell \leq 6^{-\log_5[4r(\ell)+1]} < p_{\ell-1}.$$

Further manipulation using properties of exponents and logarithms shows that the middle term is

$$4^{-\frac{\log 6}{\log 5}} \left(r(\ell) + \frac{1}{4} \right)^{-\frac{\log 6}{\log 5}}.$$

This expression represents the same value for any arbitrary but fixed base of the logarithms ($\log_2$, $\log_e$, $\log_{10}$,…),

because ratios of logs are independent of the base of the logs.

Our straight line fitting the rank–frequency data is in the form

$$p(r) = cr^{\alpha}, \tag{4}$$

where α and c are given as

$$\alpha = -\frac{\log 6}{\log 5} \tag{5}$$

and

$$c = \frac{1}{\sum_{r=1}^{\infty} r^{-\frac{\log 6}{\log 5}}}, \tag{6}$$

respectively. Note that c is defined in equation (6) in such a way that it makes $\sum_{r=1}^{\infty} p(r) = 1$.

In this expression, we see that α, the power of the rank, is $-\log 6/\log 5$, which is approximately -1.11. This α will be the slope when the data are plotted on doubly logarithmic axes. Consequently, this simple stochastic process devoid of any semantic or linguistic content has produced data closely resembling the Zipf's statistic for human language. Observing the output of this process without knowledge of the generating mechanism, we would be unable to distinguish the output from these die rolls from human language using Zipf's statistic. Since even very simple stochastic processes exhibit Zipf's distribution with slope $\alpha \approx -1$, compatibility with Zipf's law is not an appropriate route to conclude anything about the linguistic nature or potential capacity for communication transfer. Furthermore, if we used a (26 + 1)-face die to generate random words, $\alpha = -1.01$. The difference in α between a 6- and 27-face dice is small, and in both cases the values are surprisingly close to Zipf's predicted value, -1. More generally, if we had an m-face die, $\alpha = -\log m/\log(m-1)$, which approaches -1 as m increases, and is never too far from -1 for any $m > 2$. It is especially striking considering that a very simple stochastic process model produced these results, and that a large change in the number of faces did not substantially change the Zipf's distribution parameter α.

Figure 1 shows the probability versus rank in log–log coordinates for the six-sided die case. The staircase line represents the theoretical distribution of equations (1–3), the straight line is the fit obtained in equations (4–6), and the dots are a simulation result with 800 rolls. The simulation result fits a straight line better than the theoretical (staircase) result, because the random nature of the process and sampling make the empirical distribution noisy. Thus, the empirical distribution fills the gaps between the steps of the theoretical curve.

In this die-rolling example, a single output or externally observable event (word) is produced from possibly multiple internal or underlying events (individual digits). Note, however, in behavioural studies, what observers can record is output events and not underlying events, which cannot be identified without knowledge of the internal mechanism. In this die-rolling example, we have exact

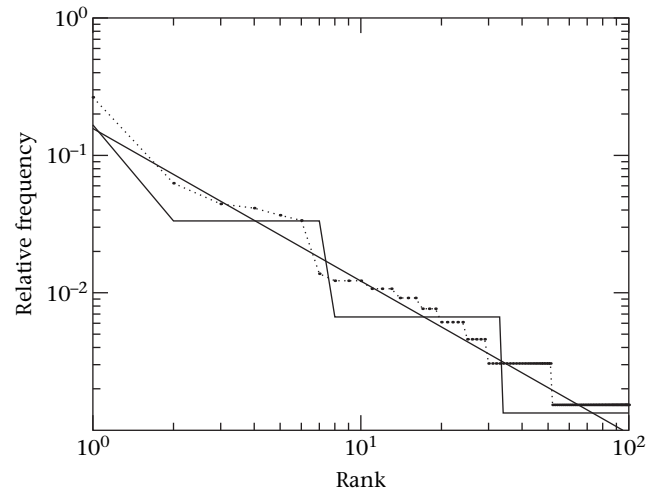

Figure 1. Rank–frequency plot of words produced by the die-rolling example. The staircase represents the theoretical distribution, the straight line the fitted line of equations (4–6), and the dotted line a simulation result with 800 rolls.

knowledge on the internal mechanism and we know that a sequence of independent identically distributed random variables can fully describe this process.

McCowan et al. (1999, page 410) claimed that analysis using Zipf's law is only valid for first-order transitions. However, they did not define the term first-order in this context, nor did they provide mathematical basis supporting this claim. In our view, there are four possibilities for McCowan et al.'s definition of 'first-order'.

(1) The condition or assumption that the observed sequence consists of symbols that are independently identically distributed. This is the definition Shannon (1948, 1951) used in his examples.

(2) The condition that the analysis is applied on the sequence of symbols that represent the smallest events that occurred during observation.

(3) The condition that the analysis is applied on the sequence of symbols that represent the fundamental perceptual units of the study animal.

(4) The condition that the analysis is applied on the sequence of symbols that represent the internal events that lead the animal to produce the behaviour or vocalization of the observer's interest.

If definition (1) is adopted, the quoted claim by McCowan et al. above simply means that the Zipf analysis is only valid when there is no probabilistic dependence between events observed at different times. This is equivalent to claiming that the Zipf-based method is only valid for the analysis of the first model in an infinite sequence of models approximating the source in Shannon's experiment, and is not applicable to the source directly. Because the first-order model is the exact model of our die-rolling process, it is a valid application of Zipf's law based on this definition of first-order.

If definition (2) is adopted, an objection may be raised against our die-rolling example, because the output events were produced from multiple internal events. Another objection may claim that the die-rolling example

produced an output sequence with words of variable length consisting of elements from a fixed-size alphabet. This contrasts with McCowan et al.'s analysis, which applied Zipf's law to a fixed-length (1 whistle) output with a variable, growing alphabet size (the number of distinct whistles observed). However, the analysis is applied to the output events (words), not the internal events (individual numbers). Since we kept our example simple, it is easy to look at an output event and identify what internal events produced it. If the output event was produced through some other mapping procedure, an observer probably could not identify what internal events were responsible for the output event being observed. Therefore, this objection has no basis.

Ferrer i Cancho & Solé (2002) raised one potential objection to explaining Zipf's law by using randomly generated texts such as our die-rolling example above. They argue that the probability of a word is immediately determined by the word length, and the word length distribution is unrealistic compared with empirically obtained word length distributions for natural languages. They also argue that the rank–frequency distribution of five-letter-long English words fits Zipf's distribution, whereas that of five-digit-long words in randomly generated texts shows a flat rank–frequency distribution. Their questions are concerned about the statistics of a particular aspect (length) of word forms in natural languages, and not the frequency of words. In other words, their argument requires prior knowledge about the text to select a group of words for analysis. Such a preselection criterion is rarely justified in the analysis of unknown texts or sequences of animal vocalizations or behaviours.

Definitions (3) and (4) present a deep and pervasive issue for many studies of animal communication because of the lack of reliable knowledge about the perceptual boundaries or the internal mechanisms of the animals that control behaviour or vocalization.

McCowan et al. (1999) chose an individual dolphin whistle as the unit of their analysis to estimate the Zipf's statistic, where they defined the interwhistle interval as a minimum of 300 ms of silence. They analysed each whistle within the context of a whistle sequence. However, they provided no evidence that the whistles thus segmented and categorized coincide with the categorization of bottlenose dolphins' perceptual boundaries or with internal events that lead bottlenose dolphins to produce each whistle.

If definition (3) is adopted, several objections may be raised against our die-rolling example. The same objection that the analysis is not based on perceptual boundaries also applies to McCowan et al.'s study. It would be inconsistent to use definition (3) of 'first-order' to criticize the die-rolling example while supporting McCowan et al.'s study, since they also provide no perceptual data supporting their arbitrary choice of 300 ms of silence as the boundary between whistles (i.e. their unit of analysis). Under definition (3), 'first-order' analyses of the die-rolling example and McCowan et al.'s whistle sequences are equally valid or invalid.

If definition (4) is adopted, our die-rolling example fails to comply with McCowan et al.'s first-order condition because a single output word is a concatenation of possibly more than one internal event. However, as noted above, in behavioural studies, what observers can record is output events and not underlying events, which cannot be identified without knowledge of the internal mechanism. Therefore, this definition of 'first-order' is impractical for behavioural studies. This is also true for McCowan et al. (1999), since they did not attempt to provide evidence that each whistle is probabilistically generated by a single internal event. Therefore, their first-order condition fails for both our die-rolling example and their study. If definition (4) is adopted, their claim negates their own analysis technique.

Concerning the segmentation and classification of animal vocalizations, we wish to make it explicit that almost every study of animal communication necessitates some arbitrary decision of how to segment, group and classify the observed data for analysis. This decision made by scientists may or may not agree with the study animal's internal processes that are responsible for the generation of externally observable events. It also may or may not agree with the perceptual boundaries of signal categories of the study animal. Consequently, one should be leery of analysis methods that make assumptions or limitations on the underlying mechanisms generating the observed data, or how the observed data would be perceived by the receiver.

The objection raised with definition (2) can be further clarified by constructing an invertible mapping from the words generated by the die-rolling game to another set of words (e.g. Israeloff et al. 1996). Briefly, consider an alphabet consisting of countably infinite elements $\{a_1,\ldots, a_5, a_{11}, a_{12}\ldots\}$. Mapping each die-rolling word d to the single element a_d produces an output sequence that has fixed-length words (one element) constructed from an alphabet of variable size, as in McCowan et al.'s (1999) analysis. Although the sequence of words generated by this mapping will satisfy Zipf's law, it lacks semantic content.

Imagine numbering each word in a dictionary, then consider mapping each numerical word in the die-rolling example to the corresponding word in an English dictionary (i.e. replace each occurrence of '1' by 'a', each '3' by 'aardvark', each '12' by 'abaft', and so on). Additionally, each null word generated by two successive rolls of 6 is replaced with 'um'. Thus 'um' would be chosen with probability 1/6, 'a' and 'aardvark' would be chosen with probability 1/36, 'abaft' with probability 1/216, and so on. Each output word is now unmistakably viewed as a single random event, and has a simple probability mass function describing its generation. This is one possible bijective mapping between the die-rolling example and an independent identically distributed process producing English words, whose rank–frequency relation satisfies Zipf's law.

The external observer viewing only the words and not the die rolls would find that the word frequencies met Zipf's law, although the sequence of words has no semantic content. Moreover, the external observer cannot distinguish the sequence of words produced by this mapping of the die rolls from the sequence of words

produced by an independent identically distributed source that assigns probability $6^{-(\ell+1)}$ to each dictionary word whose corresponding die-rolling word has length ℓ. In other words, different internal mechanisms can generate output sequences that are indistinguishable by an external observer using Zipf's statistic. Thus, our die-rolling example is a first-order process in the sense of definition (1). This mapping refutes any objection arguing that the die-rolling example obeyed Zipf's law as a consequence of its higher-order structure in generating a sequence devoid of semantic content.

Summarizing the discussion of the die-rolling example, there are a large class of processes that show the Zipf's distribution, including both linguistic processes such as English, and nonlinguistic processes such as our die-rolling example. Zipf himself mentioned that the size of cities followed his law. Mandelbrot (1952) subsequently attempted to apply Zipf's law using a variety of techniques (see also Miller 1954), and found that Zipf's law was 'universal' in the sense that even a simple stochastic process like our die-rolling example obeys the law, reaching the conclusion that 'Zipf's law is linguistically very shallow' (Mandelbrot 1982, page 346). Miller & Chomsky (1963, page 463) surveyed the studies on the relation of language to the Zipf's distribution, concluding that 'its occurrence does not constitute evidence that the signal analysed must have come from some linguistic or purposeful source'. Compatibility with Zipf's law may be a necessary condition for a language, but it is by no means a sufficient condition. As Miller & Chomsky (1963, page 463) noted, Zipf's law 'has something of the status of a null hypothesis, and like many null hypotheses, it is often more interesting to reject than to accept'. This is partly because, if Zipf's distribution is indeed a necessary condition for a language, then rejecting the fit to Zipf's distribution immediately implies that the observed data did not come from a language, whereas accepting the fit to Zipf's distribution does not imply that the data came from a language. In fact, our data compression example above demonstrates that Zipf's law is not even a necessary condition for a data sequence to have semantic content, because compressing this commentary produces $\alpha \approx 0$ but preserves all of the information. Note also that Troll & Graben (1998) pointed out that hierarchical grammar or long-range correlation is unnecessary in order for Zipf's law to hold, despite the claims made in some early works.

Figure 1 of McCowan et al. (1999) comparing randomly generated data and the hypothetical distribution for human languages raises another question. As noted above, McCowan et al. use random to mean independent and equiprobable. Consequently, the figure illustrates that the Zipf's statistics for both human languages and dolphin whistles differ significantly from the equiprobable distribution. Such nonuniform distributions are commonly found in natural systems. For example, Feller (1968, pp. 23–24) observed that 'The usefulness of sample spaces in which all sample points have the same probability is restricted almost entirely to the study of games of chance and to combinatorial analysis'.

There are largely two general contexts where Zipf's law has been related to language or other communicative processes. In one context, a study begins with a known language text or a mathematical model of communication, and it investigates the circumstances under which the data fit a Zipf's distribution. The studies of Zipf (1949), Mandelbrot (1952) and Ferrer i Cancho & Solé (2003) all belong to this context. Those studies indicate that Zipf's distribution appears when the processes are communicating under certain conditions. Serious problems arise when such results are misinterpreted to suggest their converse arguments. There is a broad range of stochastic mechanisms, regardless of communicative or linguistic content, which can give rise to Zipf's distribution when only the output events are observed. It is the vast nonspecificity of Zipf's statistic that makes it difficult to carry the findings from studies in this context to studies where unknown data are subjected to Zipf's model-based analysis. In this latter context, one would expect a high rate of false positive decisions on the statistical test, making this approach practically useless. This is the major reason why we disagree with McCowan et al.'s decision to use Zipf's model-based analysis. Similarly, it is best to remain cautious in interpreting results such as Mandelbrot (1952) and Ferrer i Cancho & Solé (2003), which may apply only under the specific conditions assumed in their respective studies.

Shannon's (1951) use of the Zipf's distribution provides an illustrative contrast in the proper application of this empirical relation. Shannon exploited Zipf's law as an engineering technique to approximate the entropy of English, but not as a test for linguistic features, communication efficiency, or as comparison of two communication systems. Specifically, knowing that English texts follow Zipf's distribution with reasonable accuracy, Shannon used the relation to approximate the probability mass function of words. These probability mass functions were then used to estimate the per word entropy of a first-order Markov model approximating human language. McCowan et al. (1999) do not cite this work by Shannon, nor do they make the distinction between the appropriate use of Zipf's law as an empirical approximation method and the inappropriate use of the law as a linguistic test or as a comparison of two communication schemes.

McCowan et al. (1999) noted that when the information source follows a Zipf's distribution, the parameter α is related to the information entropy. Because the entropy is the upper bound on the amount of information transmitted by the source (Cover & Thomas 1991, chapter 5), this relation is important for determining whether Zipf's law is useful for the analysis of animal communication. Specifically, the essential question is: 'How does the value of α relate to the Shannon entropy H for an information source following Zipf's distribution?' Mandelbrot (1952, chapter 6) analysed this question in detail with first-order (hereafter in the sense of definition (1) above) sources. Based on the Zipf's distribution, the probability of a word depends on the word's rank. The maximum rank is influenced by the length of observation and the vocabulary or repertoire size of the source. An observer typically has no prior knowledge about the true repertoire size. In the general theoretical case, the maximum rank can grow to be countably infinite. However, in practical

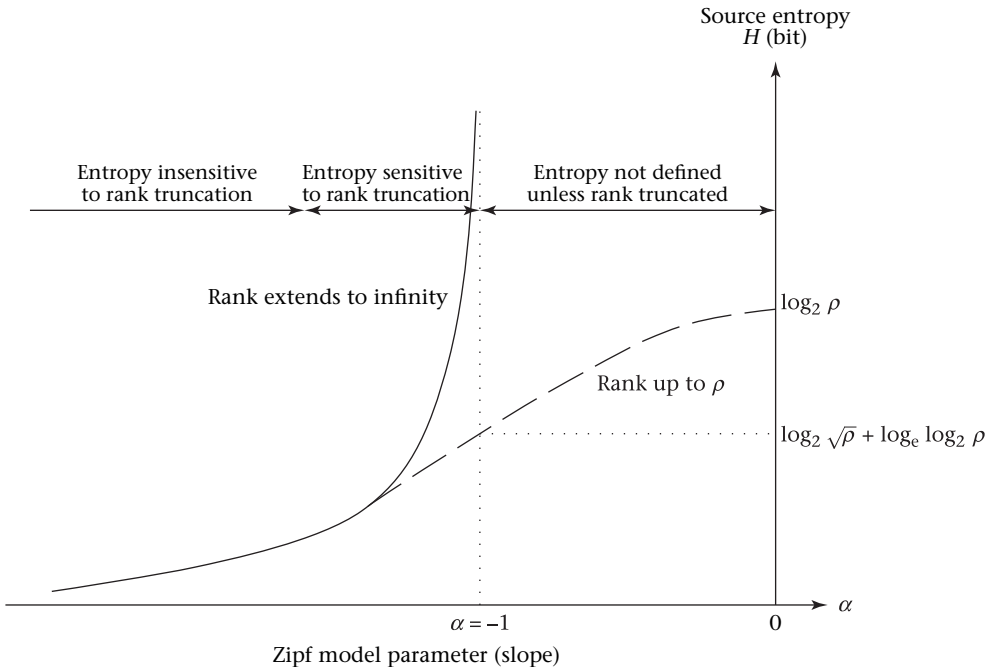

Figure 2. Source entropy H as a function of Zipf's statistic α (adapted from Mandelbrot 1952). The solid curve represents the case in which the rank extends to infinity, and the dashed curve represents the case in which the rank is truncated at a finite point, ρ.

observations, the maximum rank is necessarily limited to a finite value due to the finite observation interval. Thus 'rank truncation' can be a consequence of limited observation length, finite repertoire or both. Figure 2, adapted from Mandelbrot (1952), illustrates the two key points of his argument as summarized in the following paragraphs. This figure plots the entropy H of a Zipf's distribution source against the parameter α for the general case (solid line) and cases with rank truncation (dashed line).

The first point is that, for the general case, the entropy of a Zipf's distribution source cannot be defined for $\alpha \geq -1$. For this case, the sum of all probabilities is an infinite series $\sum_{r=1}^{\infty} cr^{\alpha}$ that diverges for $\alpha \geq -1$. We shall call a probability mass function that sums to one 'valid'. No valid mass function exists for $\alpha \geq -1$, which includes the Zipf's law case of $\alpha = -1$. Since no valid probability distribution can be defined in this region, the entropy H is also undefined for sources with $\alpha \geq -1$. For the region $\alpha < -1$, but near $\alpha = -1$, H changes steeply with α; H is much more sensitive than α to a change in the source characteristics. Two information sources with vastly different values of H will have similar values of α. Therefore, α is a poor parameter to characterize the communicative properties of information sources. This is not a desirable property for an experimental data analysis technique.

Mandelbrot's second major point is that truncating the rank or observation interval will significantly distort the resulting entropy estimate $\hat{H}$ in the region near $\alpha = -1$, the region of greatest interest for sources close to the Zipf's law distribution. Any practical experimental or observational study is limited to a finite data record. Let ρ denote the rank of the least frequently observed word. Mandelbrot indicates two specific points of interest where the $\hat{H}$ obtained from finite observations is erroneously finite. These are: $\hat{H} = \log_2 \rho$ when $\alpha = 0$ for uniformly distributed sample points with countably many words, and $\hat{H} = \log_2 \sqrt{\rho} + \log_e \log_2 \rho$ when $\alpha = -1$ for Zipf's prescribed case, although the true source entropy H is unbounded for both cases.

Thus, even when the assumptions of Zipf's model are fully met, Mandelbrot's results indicate that the Zipf's distribution parameter α is a less valuable estimate and a less reliable representation of the source characteristics than Shannon entropy because α is insensitive to changes in the source properties. The Zipf parameter α is also a less practical statistic than Shannon entropy because α's insensitivity makes it more difficult to estimate than the entropy of a first-order model. Therefore, even if the hypothetical source follows a Zipf's distribution, the first-order model entropy estimate reflects the complexity of the source more accurately, and is easier to estimate than the Zipf's distribution parameter.

When data that are obtained through observations or experiments are analysed using any parametric model, one common technique is to find the best fit of the model parameters to the data. We discuss two problems in the following paragraphs: the possible consequences of a finite data record, and the method used to find the parameter α from a data record.

A valid empirical distribution can always be inferred from a finite data record, even when there is no valid probability distribution for the first-order information source. For example, there may be no valid and consistent mass function for the $\alpha \geq -1$ case, whereas we can still use the observed frequencies to find estimates $\hat{\alpha}$ and $\hat{c}$ that fit the finite rank-observed data in such a way that $\sum_{r=1}^{\rho} \hat{c} r^{\hat{\alpha}} = 1$. If the resulting estimate of Zipf's statistic $\hat{\alpha} \geq -1$, it means that: (1) $\hat{\alpha}$ is an inaccurate estimate of

the true α; (2) the source's vocabulary is finite; or (3) the assumption that the source follows Zipf's distribution is violated, and therefore the analysis may not be trustworthy.

Zipf's distribution model has only one free parameter. In equation (4), both α and c are parameters but c is uniquely determined from α and therefore is not a free parameter. The best fit of α for a finite set of observations must be found by adjusting α in equation (4) with $c=1/\sum_{r=1}^{\rho} r^{\alpha}$. The estimated $\hat{\alpha}$ thus obtained gives a valid Zipf's distribution model. On the other hand, if a straight line fitting on log rank versus log frequency is used with a simple regression analysis, the line model is given two free parameters, slope and intercept. That is, whereas the line-fitting procedure is allowed to fit any straight line, to fit Zipf's model one must choose the line from the limited subset of lines whose slope and intercept produce a valid Zipf's distribution satisfying the axiom of probability. Specifically, with valid models, the probabilities of all sample points must sum to one, or equivalently, the observed frequencies must sum to the sample size. It is clear that fitting Zipf's model imposes stronger restrictions on the set of allowable lines, and therefore a straight line can fit the data with a high goodness-of-fit value even if the same line fails to fit the Zipf's distribution well.

McCowan et al. (1999, Figure 1 and Table 1) used a regression that fit any line to the data, instead of constraining the line to fall within the subset of valid Zipf distributions. This erroneous application of the underconstrained line fitting may have contributed to the high R^2 values given in their Table 1. The values of R^2 given reflect the goodness-of-fit between the line obtained and the data, but the line they obtained is not a valid Zipf's distribution. Thus, their analysis is inconsistent with the very model they proposed. As shown below, the parameters of their line do not satisfy the axiom of probability, and thus their values are incorrect estimates of Zipf's distribution parameter α.

One might argue that the difference between the unconstrained fit obtained by McCowan et al. (1999) and a valid Zipf distribution is not large, and thus should not change the goodness-of-fit substantially. To investigate this possibility, we used the values given in McCowan et al.'s (1999) Table 1 to sum the frequencies of all events for the model fit they obtained. If the model is close to a valid Zipf distribution, the sum of the model frequencies should be close to the observed sample size. The sums of the frequencies we obtained from the model are 289.8 (Adults), 85.44 (<1 month), 336.3 (2–8 months) and 191.5 (9–12 months). These numbers differ considerably from the actual sample sizes of 600, 53, 424 and 293, respectively, given in their Table 1. This means that McCowan et al.'s (1999) model failed to preserve the sample size. The same consistency check may be framed in terms of the axiom of probability, computing $\sum_{r=1}^{\rho} p(r)$ for the probability distributions derived from the slopes, intercepts and number of total whistles in Table 1 (where ρ is McCowan et al.'s N, whistle types). If the McCowan et al.'s (1999) model fit is consistent with the data observed, these probabilities should sum to one. Instead, we found that the probabilities implicit in their model fit sum to 0.4831 (Adults), 1.612 (<1 month), 0.7932 (2–8 months) and 0.6536 (9–12 months). None of these values are close to the value of one required for a valid distribution. Therefore, we conclude that McCowan et al.'s R^2 values indicate an excellent fit to a model that is not internally consistent with the data they observed, since the model does not match the actual observed sample size, nor does it satisfy the axiom of probability. A fit constrained to a valid Zipf's model would necessarily result in a poorer goodness-of-fit measure.

In view of these issues, the question arises as to whether Zipf's law is still useful as a statistical comparative tool. Our conclusion is that (1) the Zipf's distribution model is not an effective way to analyse unknown information sources, even when we know that the source statistics closely follow this distribution; (2) Zipf's law analysis cannot reliably discriminate between languages and stochastic processes devoid of semantic or communicative content. Studies that have depended on Zipf's law as a language detector or to measure communication capacity should develop alternative techniques.

As a final note, we point out that McCowan et al. underestimate the number of probabilities required for higher-order Markov models. The correct number of probabilities to be estimated for n call types taken r at a time is n^r and not $n!/r!(n-r)!$ as they report (McCowan et al. 1999, page 415). Their expression is only valid under the strong (and unstated) assumptions that no call is repeated within the group of r calls, and that the temporal order of the calls within the group is irrelevant. Previous work using Markov models has assumed that calls are repeated and that the ordering of calls within a sequence is important (e.g. Hailman et al. 1995; Gentner & Hulse 1998).

We thank Vincent Janik, Jack Hailman, Ramon Ferrer i Cancho and an anonymous referee for helpful suggestions on this manuscript. This work was performed while R.S. was at the Department of Electrical and Computer Engineering and Center for Marine Science and Technology of the University of Massachusetts Dartmouth, and while visiting at the Woods Hole Oceanographic Institution. J.R.B. and R.S. acknowledge the support of NSF Ocean Sciences CAREER award 9733391, and P.L.T. acknowledges the support of ONR grant N00014-97-1-1031. While preparing the final revision of the manuscript, R.S. was an MIT Rosenblith Fellow and a Howard Hughes Medical Institute Predoctoral Fellow. This is contribution number 99-1101 from the University of Massachusetts Dartmouth Center for Marine Science and Technology and 10092 from Woods Hole Oceanographic Institution.

References

Bonhoeffer, S., Hertz, A. V. M., Boerlijst, M. C., Nee, S., Nowak, M. A. & May, R. M. 1996. No signs of hidden language in noncoding DNA. *Physical Review Letters*, **76**, 1977.

Bradbury, J. W. & Vehrencamp, S. L. 1998. *Principles of Animal Communication*. New York: Sinauer.

Cover, T. M. & Thomas, J. A. 1991. *Elements of Information Theory*. New York: J. Wiley.

Damashek, M. 1995. Gauging similarity with *n*-Grams: language independent categorization of text. *Science*, **267**, 843–848.

Feller, W. 1968. *An Introduction to Probability Theory and Its Applications.* Vol. I. 2nd edn. New York: J. Wiley.

Ferrer i Cancho, R. & Solé, R. V. 2002. Zip's law and random texts. *Advances in Complex Systems*, **5**, 1–6.

Ferrer i Cancho, R. & Solé, R. V. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences, U.S.A.*, **100**, 788–791.

Flam, F. 1994. Hints of a language in junk DNA. *Science*, **266**, 1320.

Gentner, T. Q. & Hulse, S. H. 1998. Perceptual mechanisms for individual recognition in European starlings (*Sturnus vulgaris*). *Animal Behaviour*, **56**, 579–594.

Hailman, J. P., Ficken, M. S. & Ficken, R. W. 1995. The 'chick-a-dee' calls of *Parus atricapillus*: a recombinant system of animal communication compared with written English. *Semiotica*, **56**, 191–224.

Israeloff, N. E., Kagalenko, M. & Chan, K. 1996. Can Zipf distinguish language from noise in noncoding DNA? *Physical Review Letters*, **76**, 1976.

Li, W. 1992. Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, **38**, 1842–1845.

Loève, M. 1963. *Probability Theory*. 3rd edn. Princeton, New Jersey: Van Nostrand.

McCowan, B., Hanser, S. F. & Doyle, L. R. 1999. Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour*, **57**, 409–419.

MacKay, D. M. 1972. Formal analysis of communicative processes. In: *Non-verbal Communication* (Ed. by R. A. Hinde), pp. 3–25. Cambridge: Cambridge University Press.

Mandelbrot, B. 1952. Contribution a la théorie mathématique des jeux de communication. *Publications de l'institut de statistique de l'université de Paris*, **2**, 1–124.

Mandelbrot, B. 1982. *The Fractal Geometry of Nature*. San Francisco: W. H. Freeman.

Mantegna, R. N., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Peng, C.-K., Simons, M. & Stanley, H. E. 1994. Linguistic features of noncoding DNA sequences. *Physical Review Letters*, **73**, 3169–3172.

Martindale, C. & Konopka, A. K. 1996. Oligonucleotide frequencies in DNA follow a Yule distribution. *Computers & Chemistry*, **20**, 35–38.

Miller, G. A. 1954. Communication. *Annual Review of Psychology*, **5**, 401–420.

Miller, G. A. 1957. Some effects of intermittent silence. *American Journal of Psychology*, **70**, 311–314.

Miller, G. A. & Chomsky, N. 1963. Finitary models of language users. In: *Handbook of Mathematical Psychology* (Ed. by R. D. Luce, R. R. Bush & E. Galanter), pp. 419–492. New York: J. Wiley.

Papoulis, A. 1984. *Probability, Random Variables, and Stochastic Processes*. 2nd edn. New York: McGraw-Hill.

Rapoport, A. 1982. Zipf's law re-visited. *Quantitative Linguistics*, **16**, 1–28.

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

Shannon, C. E. 1951. Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, 50–64.

Slater, P. J. B. 1973. Describing sequences of behavior. In: *Perspectives in Ethology*. Vol. I (Ed. by P. P. G. Bateson & P. H. Klopfer), pp. 131–153. New York: Plenum.

Troll, G. & Graben, P. beim 1998. Zipf's law is not a consequence of the central limit theorem. *Physical Review E*, **57**, 1347–1355.

Tyack, P. L. 1999. Communication and cognition. In: *Biology of Marine Mammals* (Ed. by J. E. Reynolds III & S. A. Rommel), pp. 287–323. Washington, D.C.: Smithsonian Institution Press.

Visweswariah, K., Kulkarni, S. R. & Verdú, S. 1998. Source codes as random number generators. *IEEE Transactions on Information Theory*, **44**, 462–471.

Voss, R. F. 1996. Comment on 'Linguistic features of noncoding DNA sequences'. *Physical Review Letters*, **76**, 1978–1981.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.