Information content versus word length in natural language: A reply to Ferrer-i-Cancho and Moscoso del Prado Martin (2011)

Steven T. Piantadosi, Harry Tily, Edward Gibson

September 11, 2012

Abstract
Recently, Ferrer i Cancho and Moscoso del Prado Martín (2011) argued that an observed linear relationship between word length and average surprisal (Piantadosi, Tily, & Gibson, 2011) is not evidence for communicative efficiency in human language. We argue that their study of a random typing model is largely irrelevant to human language: their model critically rests on incorrect assumptions about human language, is incapable of explaining key surprisal patterns in language, and is incompatible with recent behavioral results. More generally, we argue that statistical models of a system cannot ignore what is independently known about it.

1 Introduction
One of the most famous properties of language, first studied by Zipf (1936), is that frequent words are typically also short. Zipf offered a communicative theory for this property, under which lexicons have evolved to be efficient: words which must be re-used repeatedly should be short to minimize the effort of language users. Recently, we (Piantadosi et al., 2011, henceforth, PT&G) demonstrated an improvement on Zipf’s ideas. Under a more elaborate notion of communicative efficiency, word length should depend not on frequency, but on the typical amount of information conveyed by a word. Efficient languages will convey information close to the channel capacity (Shannon, 1948) of human perceptual and cognitive systems. In this case, one should observe a linear relationship between a word’s negative log probability (surprisal) and its length, in an attempt to keep the number of bits communicated per unit time roughly constant. Such a prediction is a lexical version of a now popular idea that choices made in language production also attempt to maintain a roughly uniform rate of information transmission (Genzel & Charniak, 2002, 2003; Aylett & Turk, 2004; Jaeger, 2006; Levy & Jaeger, 2007; Jaeger, 2010). PT&G demonstrated across 10/11 languages for which corpora were readily available that information con-
tent does predict word length better than frequency, both in total correlations and partial correlations. Recently, Ferrer i Cancho and Moscoso del Prado Martín (2011, henceforth F&M) argued that the roughly linear relation observed in PT&G was not necessarily evidence of communicative efficiency. They prove that a model in which language is generated by choosing characters independently also shows a linear relationship between average information content and word length. In such a system, words are generated by randomly typing characters, and occasionally hitting the “space” character to create a word boundary. It is intuitively not surprising that such a system would show the required linear relationship, since the probability of achieving a word of a given length $l$ will decrease geometrically in $l$, so the log probability scales linearly in $l$. F&M furnish a mathematical proof for a more general version where all words must be longer than some minimum length $l_0$ and individual letters occur with arbitrary probabilities. Because random typing does not consider communicative efficiency, they argue that it is possible to achieve a linear relationship without any notion of communicative optimization. Though they do not say so explicitly, their paper implies that simple statistical models should be treated as baselines, where any properties of language that hold in them are not expected to be the result of any interesting causal processes. F&M’s random typing model is one exemplar of a long history of monkey models in psycholinguistics, so-called because they capture the statistical process of “a million monkeys typing on a million typewriters.” Such models were first articulated by Mandelbrot (1953) and Miller (1957), in an attempt to account of the power law distribution of word frequencies (Zipf, 1936).

We believe that there are several important missteps and misunderstandings in F&M’s critique. A more careful consideration of the relevant ideas in PT&G and related work demonstrates that PT&G’s results are not statistical artifacts.

Random typing cannot explain the primary finding of PT&G

First—and perhaps most importantly—F&M do not address the primary data point reported by PT&G. Our main finding was not a linear relationship. It was that a word’s average in-context surprisal predicted word length better than frequency predicts word length. This pattern held in general for several different corpora, ways of measuring word length, and ways of estimating surprisal, and in partial correlations (e.g. surprisal partialing out frequency vs. frequency partialing out surprisal). Thus, PT&G’s measure of the average amount of information conveyed by a word was a more important determinant of word length than frequency—so much so that the partial effect of frequency was near

---

1Because in their model predictability reduces to frequency, their work replicates Miller (1957) (e.g. Equation 2) and extends it to the case of unequal letter probabilities.

2Unfortunately, F&M do not specify what they mean by communicative efficiency. For PT&G and other UID work before them, language would be efficient if it tended to communicative bits of information at the channel capacity of human cognitive systems. Under this definition, F&M’s random typing model actually could be efficient.
zero in some languages. This pattern is problematic for F&M’s critique because in random typing, frequency and surprisal are mathematically identical—a fact used in their derivations. Thus, random typing provably cannot explain the ability of surprisal to predict word length better than frequency.

**Random typing is independently known to be incorrect**

A more basic reason why F&M’s findings are not informative about real language is that real human language is not generated by anything like the type of random typing model that F&M advocate studying. Words are not chosen by sampling their component pieces. Instead, speakers know whole words—this is perhaps the most uncontroversial assumption one could make in a statistical model of language. The statistical analysis of natural language strongly favors the existence of words, and is at odds with the notion that component pieces are generated randomly. Idealized statistical models over natural sequences of characters infer not only the presence of words but the correct words themselves (Goldwater, 2006; Pearl, Goldwater, & Steyvers, 2011). Though such models were proposed as language acquisition models, they equally serve as idealized data analysis models, demonstrating that the evidence provided by statistical dependencies between characters in natural language strongly favors the existence of words. Similarly, monkey models’ assumed independence between word components is very easy to reject as a statistical hypothesis. For instance, in the tiny amount of linguistic data provided by only the abstract of F&M’s own paper, a chi-squared test reveals that neither characters ($\chi^2 = 1836.9, p < 0.001$) or syllables ($\chi^2 = 27279.2, p < 0.001$) within words are independently generated.

The existence of words as psychological units undermines F&M’s primary point, that “a linear correlation between information content and word length may simply arise internally, from the units making a word (e.g., letters) and not necessarily from the interplay between words and their context ...” If humans generate language by remembering entire words, rather than by sampling their component parts, then the psychologically real lexical system assigns no necessary relationship between word length, frequency, and information content. Indeed, it is not relevant to understanding human language how systems that randomly sample word components behave.

---

3 Howes (1968) presented a similar critique to Miller’s monkey model for deriving the Zipfian distribution of word frequencies: “If Zipf’s law indeed referred to the writings of ‘random monkeys,’ Miller’s argument would be unassailable, for the assumption he bases it upon are appropriate to the behavior of those conjectural creatures. But to justify his conclusion that people also obey Zipf’s law for the same reason, Miller must perforce establish that the same assumptions are also appropriate to human language. In fact, as we shall see, they are directly contradicted by well-known and obvious properties of languages.”

4 Excluding citations and notes in parentheses. Spaces were removed to ensure that these effects were not the result of multiple spaces not occurring sequentially. The $\chi^2$ test compared the joint distribution of letter or syllable bigrams to that which would be expected under random generation. P values were computed by simulation in the R function `chisq.test`.
Random typing is not even a good statistical model

It is still worth considering that even though the generative assumptions F&M’s model is plainly incorrect, it still may provide a useful statistical description of language. Unfortunately, a considerable amount of evidence has amassed that such models are poor statistical theories. Baayen (2001, pg106-107) analyzes the predictions of a monkey model with respect to coarse properties of lexical systems, including word frequency distributions, frequency/length relationships, and neighborhood density. He finds that while a monkey model provides qualitative trends in the right directions, its quantitative fit is not very good and is eclipsed by the fit of other models such as a Yule-Simon model. Indeed, a considerable amount of work—much of it by the first author of F&M—has detailed the ways in which the output of monkey models are unlike those found in natural languages, especially with respect to Zipf’s law (Tripp & Feitelson, 1982; Baayen, 2001; Montemurro, 2001; Ferrer i Cancho & Solé, 2002; Ferrer i Cancho & Elvevåg, 2010). Other statistical properties of random texts have been found to be divergent from real language. For instance, Bernhardtsson, Baek, and Minnhagen (2011) shows that random texts, but not natural texts, follow the statistics of Heaps’ law, a growth pattern relating types and tokens to text sample size (Heaps, 1978). Cohen, Mantegna, and Havlin (1997) compares entropy-based measures on natural and monkey model texts, with the goal of finding metrics that best distinguish these texts. Our other work has detailed ways in which lexical systems are not only non-random, but specifically structured for communicative efficiency, in terms of ambiguity (Piantadosi, Tily, & Gibson, 2012) or lexical properties such as stress (Piantadosi, Tily, & Gibson, 2009). Additionally, not all short, phonotactically possible words are used in language (Cohen, 2006), contrary to the predictions of a monkey model, but consistent with communicative theories based on entropy rate (PT&G) or possibly the avoidance of confusable code words. In short, it is clear that monkey models don’t even produce the correct detailed statistical properties of language, although they may appear qualitatively similar with respect to some coarse-grained properties.

The irrelevance of untrue causal processes

The implication of F&M (and before them, Miller and Chomsky (1963)) is that even though monkey models are implausible descriptions of the generative process of language and lousy statistical theories, they still provide a null hypothesis which should be considered in the course of scientific theorizing. Properties of language that are also exhibited by monkey models should be looked on cautiously, as phenomena which likely have a trivial and uninteresting cause.

In contrast, we believe that it is a fallacy to think that the fact a monkey model exhibits a linguistic property should cast any doubt on alternative theories, such as those proposed by PT&G and Zipf. Causal processes which are independently known to be incorrect—like the random generation of monkey models—cannot cast doubt on alternative explanations that are plausible.
The hypothesis of random typing—and all models like them—have already been disproven by other sources of evidence.

We find this point interesting because it raises a difficult issue for modeling. All models are incorrect in that they do not exactly mirror the “real” process happening in the world. But there is an important difference between most models and F&M’s: most models do or should attempt to make their core, key assumption things that are at least approximately like those in the real world. Changing “good” models to make them more realistic should not break their core predictions and properties. This is what makes them useful to science—even though they will later be eclipsed by more sophisticated theories, they at least approximately capture some aspect of the real world. But F&M’s model is different: even slight improvements to make their model better reflect what we know about real language—for instance introducing dependencies between characters—no longer necessarily demonstrates the key properties. The key aspects of F&M’s model make it critically unlike people in terms of representation, processing, and knowledge of language.

Independent evidence for PT&G’s optimization process

Aside from discussions of the plausibility of various models, there are good independent reasons for rejecting F&M’s assertion that the relationship observed by PT&G is a statistical artifact. PT&G posited that the observed relationships might result from lexicalization of phonetic reduction (e.g. Lieberman, 1963). It is well-known that speakers shorten or reduce syllables in predictable locations and PT&G’s findings plausibly result from these speech production factors being integrated into lexical representations. If a word is used in predictable locations, it will be reduced, and eventually might be learned to be its shorter form, giving a relationship between word length and predictability.

Second, there are independent behavioral studies showing that speakers actually do prefer short forms of words in predictable contexts, exactly as PT&G’s theory would predict. Mahowald, Fedorenko, Piantadosi, and Gibson (in press) gave people a choice between two synonymous pairs (e.g. “chimp” / “chimpanzee”) in either predictive or non-predictive contexts. They found that people preferred the the short form when the word was predictable and the long form was it was not. These kind of behavioral tendencies have also been examined in corpus research by Frank and Jaeger (2008), who showed that contracted forms (“do not” / “don’t”) occur more frequently in predictive contexts. Such behavior is predicted by PT&G, but not explainable under F&M’s kind of theories.

Conclusion

To summarize, monkey models embody an interestingly incorrect approach to understanding natural language. In general, one is not free to study any conceivable statistical process and conclude that it is relevant for understanding how real language works. Candidates for relevance must not ignore facts about linguistic systems. Since words are actively chosen by language users to convey
a meaning, there is no point to studying models for which the uttered word is generated according to some statistical properties of the wordform itself—that is the wrong causal direction. Indeed, the key causal process at work in monkey models—words are produced by randomly generating word parts and sometimes generating a word boundary—is plainly incorrect in real language. As such, results about monkey models only apply to systems which are critically unlike humans in terms of the structure of language, knowledge of words, statistical dependencies between words, and the transmission of meaningful information. Demonstrating that such processes can also give rise to linear relationships, power law distributions, or other statistical properties does not cast doubt on more realistic causal processes.

References


