The missing baselines in arguments for the optimal efficiency of languages

Fermín MOSCOSO DEL PRADO (fmoscoso@linguistics.ucsb.edu)
Department of Linguistics
University of California, Santa Barbara
Santa Barbara, CA 93106-3100 USA

Abstract

I argue that linear correlations between log word frequency,and lexical measures, cannot be taken as evidence for a "Principle of Minimum Effort". The Principle of Maximum Entropy indicates that such relations are in fact the ones most probable to be found. For such claims, one needs to compare the correlations with adequate baselines reflecting what would be expected in a purely random system. I then introduce a way of computing such baselines, and use it to show that the correlations found in a corpus are actually weaker than what one would expect to find by chance. Therefore, if an argument were to be made based on them, it would paradoxically be that language is worse for communication than what one would expect to find in a random system. More appropriately however, what these results reflect is that such correlations are not the best places to look for linguistic optimality.

Keywords: Corpus Study; Lexical Ambiguity; Principle of Maximum Entropy; Zipf's Law of Abbreviation

Introduction

Arguments about language being optimal for communication have a long tradition within the cognitive sciences, dating at least as far back as Zipf (1935). Zipf observed that, across many texts, there is an inverse correlation between a word's frequency of occurrence and its length in characters, which is now referred to as Zipf's Law of Abbreviation (ZLA). This observation led him to his "Principle of Least Effort" (Zipf, 1949): Humans prefer shorter words to refer to frequent concepts, so that the overall length of utterances will be minimized, and so will the effort required to produce them. In this form, from the speaker's (or writer's) point of view, the optimality of human language would be measured by the amount of effort required by a speaker to produce an utterance. Zipf also realized that, from the comprehender's perspective, optimality would not be so much concerned with the length of an utterance as it would with the ease with which it can be unambiguously decoded. Jointly considering both the perspective of the speaker and that of the comprehender, the structure of language would be subject to a trade-off between utterance length and degree of ambiguity. Zipf was somewhat vague with respect to how such trade-off could be measured, but his general idea is considered valid ever since.

As compelling as Zipf's arguments seem, very early on, researchers in Information Theory and Psychology noticed that they may not be as informative as Zipf thought. Both Mandelbrot (1953) and Miller (1957) realized that the negative correlation between a word's frequency of occurrence and its length in characters (i.e., ZLA) would also arise in randomly generated texts that lack any linguistic structure or communicative value whatsoever; what Mandelbrot dubbed a "typing monkeys" process, and Miller –somewhat less graphically– called an "intermittent silence" process. The validity and importance of Zipf's original *observations* on the distribution of word frequencies and word lengths is beyond doubt, as is evidenced for instance in a whole family of power-law distributions and phenomena across many unrelated fields of science being currently named in Zipf's honor (e.g., Zipf's Laws, Zipfian distributions, Zipf-Mandelbrot distribution). However, Zipf's *interpretation* that such properties reflect the optimization of human language structure is disconfirmed by the fact that those very same properties are also found in systems that are not the result of any optimization process. The properties are therefore not informative about the optimality of the process that generated them. This highlights a problem that is exhibitted by many claims on language properties that reflect some form of optimization: The lack of a non-optimal baseline against which to test whether such inferences are perhaps *non sequiturs*.

Let us consider a non-linguistic example. Suppose I put forward a theory on the processes governing the outcome obtained when throwing two particular dice. The dice themselves would be beyond my possible observations (e.g., inside a black box), but I would have access to the sum of their outcome. My theory could state that the dice are loaded so that they strongly favor a non-extreme (or optimal) outcome of three or four dots. In order to test my theory, I would collect data from many dice throws (with access only to their summed values). After obtaining a few thousand throws, if I found that the average value of the sum is seven (with some preset degree of precision), which is fully consistent with my theory. As it happens, however, seven would also be the most likely value of the sum, even if the dice were not loaded. Therefore, I could not take the evidence from the average value as support for my theory, as it would also be consistent with the *a priori* more likely theory that the dice are not loaded. As we will see below, one can objectively say that the unloaded theory is more probable *a priori* using mere probabilistic arguments (the Principle of Maximum Entropy).

In the example above, the prediction used to test the hypothesized property holds trivially for the most probable outcome. One can of course design situations in which the seven sum property does not hold (e.g., by loading the sixes on dice). Still, even if it is possible to artificially design such a scenario, it is still the case that the *most probable* outcome, whether or not any optimization is at work, is that the property will hold (i.e., the dice will sum up to seven). However, even if the property were less evident than the one used in this example, testing it also on a few non-optimal baselines would enable us to see that such property does not signal the

presence of optimization.

This discussion is motivated by the recent publication of several papers making claims on the optimality of human language which suffer from the same lack of baseline problem. In what follows, I begin by summarizing four of these recent arguments for language optimality. I then introduce the Principle of Maximum Entropy, and use it to show analytically that the findings presented as evidence for communicative optimality turn out to be trivial predictions that will also be observed in the most probable non-optimized baselines. In the data section, I analyze a corpus of English to assess the strength of the effects used to argue for optimality. The results show that these effects are in fact significantly *weaker* than what one would expect to find by mere chance. In other words, if one took such correlations as evidence for optimization, one would have to conclude that human languages are actually *less optimized* than one would expect by chance. I conclude with a discussion of how Information-Theory can be used to make predictions on the optimality of human languages that do indeed survive the non-optimal baselines tests.

Some Information Theoretical Arguments for Communicative Efficiency

In a recent study, Piantadosi, Tily, and Gibson (2012) extend ZLA to the domain of lexical ambiguity. Following Zipf, they argue that short words require less effort to be produced than longer words would. Therefore, by a similar principle of economy, it would be beneficial to encode as many meanings as possible using the shorter words, and then use the redundancy present in the context to disambiguate them. They support this claim by showing that, in corpora of three languages, there is indeed a negative correlation between word ambiguity and word length when other factors (e.g., frequency) are considered.

A second prediction of Piantadosi et al. (2012) considers the fact that more frequent words have a more accessible lexical representation, as is evidenced by the fact that they elicit shorter reaction times and lower error rates in a broad range of lexical processing experiments (e.g., Oldfield & Wingfield, 1965). That a word is easy to access makes it a desirable candidate to carry many meanings (or be associated to many uninflected word lemmas) in a system that is optimized to make the production and comprehension of words as effortless as possible. Therefore, they predict a positive correlation between a word's frequency and the number of distinct meanings (or lemmas) associated with it. Their analyses of several corpora indeed find this correlation.

As convincing as these arguments might seem, I will argue below that the findings are but trivial consequences of ZLA, and do not provide support the communicative hypothesis that is put forward. I will further discuss, ZLA is itself a trivial property of most symbolic sequences, irrespective of whether they are optimized.

Piantadosi et al. (2012) also argue that, if one computes the probability of a word according to a triphone (i.e., phoneme trigram) model, those words with the highest probability correspond to those that provide a more prototypical example of the phonotactics of a language. Those words that conform better to the phonotactic constraints of the language will be easier to pronounce and recognize. Following the "least effort" argument, they predict that words with high phonotactic predictability should be associated with more meanings (or word lemmas) than words with lower phonotactic probability.

The Principle of Maximum Entropy

Before making any claims that a particular distribution of linguistic probabilities (of words, words lengths, degrees of ambiguity, etc.) constitutes evidence for language being "optimized for human communication", one should check what kind of such distributions one would expect to observe by mere chance, irrespective of the presence any hypothetical optimization process. The relevant properties of the distribution (e.g., ambiguous words being more frequent, etc.) should be found to be significantly more (or less) marked in actual linguistic data than one would expect them to be.

This raises the problem of how to determine, among the infinite possible discrete probability distributions that words could have, which ones are the most probable *a priori*. The Principle of Maximum Entropy (PME; Jaynes, 1957a, 1957b) states that, among all probability distributions satisfying a set of constraints, the most probable one will be the one that has the highest entropy. The entropy (Shannon, 1948) of a probability distribution defined over a discrete set of words $W = \{w_1, w_2, w_3, \ldots\}$ is given by

$$H(W) = -\sum_{w \in W} \mathrm{P}(W = w) \log \mathrm{P}(W = w),$$

where $P(W = w)$ denotes the probability of encountering word w in a corpus of text (i.e., its relative frequency of occurrence). In what follows, I will use the abbreviated notation $\mathrm{p}_i = \mathrm{P}(W = w_i)$. The most probable assignment of values for the p_i is the one leading to the highest value of $H(W)$, with the obvious constraint that the values of the p_i must all sum to one, so that they form an actual probability distribution.[1]

If no additional constraints were present (i.e., any assignment of probabilities could be considered), then, for a finite set N probabilities, the maximum entropy would be the uniform distribution with $p_i = 1/N$. Of course, when one considers the probabilities of words, not all probability distributions are valid assignments. Rather, these distributions need to satisfy some basic constraints. These specific constraints can be requirements such as the existence of a mean word length or an average degree of ambiguity. Constraints of this type can be expressed as values of the means of some given functions. For instance, one such function can be the length of a word (measured in either characters, phonemes, or syllables), which maps words into natural numbers ($\ell : W \mapsto \mathbb{N}$),

[1] The general validity of the PME is demonstrated using simple combinatorics (cf., Jaynes, 2003).

and another function can the degree to which words are ambiguous, which maps words into the non-negative real numbers ($\mathcal{A} : W \mapsto \mathbb{R}^+$). The constraints would then be expressed as the existance of a mean word length ($\langle \ell(w) \rangle_W = L$) and a mean degree of ambiguity ($\langle \mathcal{A}(w) \rangle_W = A$).

The most probable distribution that satisfies the constraints would then be the solution to the maximization problem:

$$\arg\max - \sum_{w_i \in W} \mathrm{p}_i \log \mathrm{p}_i,$$

subject to

$$\begin{cases} & \sum_{w_i \in W} \mathrm{p}_i & = 1 \\ \langle \ell(w) \rangle_W = & \sum_{w \in W} \mathrm{p}_i \ell(w_i) & = L \\ \langle \mathcal{A}(w) \rangle_W = & \sum_{w \in W} \mathrm{p}_i \mathcal{A}(w_i) & = A \end{cases} \quad (1)$$

Notice that I have added one constraint to indicate that the resulting probability distribution must be normalized. The solution to such a problem is found analitically using the method of Laplace multipliers. It must have the form of a Boltzmann canonical distribution,[2]

$$\mathrm{p}_i = \mathrm{e}^{\lambda_0 + \lambda_1 \ell(w_i) + \lambda_2 \mathcal{A}(w_i)}, \quad (2)$$

where the parameters λ_0, λ_1, and λ_2 are Laplace multipliers whose value is uniquely determined by the individual values of the word lengths $\ell(w_i)$, ambiguities $\mathcal{A}(w_i)$ as well as their average values L and A.

Implications of the PME

Taking logs on both sides of Eq. 2 reveals that *a priori* – assuming the existence of a mean word length (L) and an average degree of ambiguity (A)– the most probable relation between our variables of interest is

$$\log \mathrm{p}_i = \lambda_0 + \lambda_1 \ell(w_i) + \lambda_2 \mathcal{A}(w_i). \quad (3)$$

This equation already makes important predictions. We should expect that –everything else being equal– the log probability of a word (i.e., its log frequency) should be linearly related to both its length and to its degree of ambiguity. No assumptions about language being optimized for communication are necessary to make this prediction, it just happens to be to most probable type of relation. The signs and values of the Laplace multipliers λ_1 and λ_2 will determine the strength and direction of the correlations. They therefore provide baselines for any effects whose presence is argued to reflect a form of efficiency or optimality. Without any need for efficiency, we should expect to find correlations with the strengths given by λ_1 and λ_2.

A negative value of λ_1 would indicate that ZLA is in fact the most likely relation that one should expect to find between word frequency and word length. Therefore, in order to claim that ZLA provides evidence for communicative efficiency, one should observe that the relation between log word frequency and word length is more negative than λ_1. This finding complements the previous arguments of Mandelbrot (1953), Miller (1957), or Ferrer i Cancho and Moscoso del Prado (2011) that random processes also exhibit ZLA. It provides a baseline to assess whether the ZLA observed in a real corpus is stronger than what one would have expected by mere chance.

Similarly, λ_2 indexes the relation between log word frequency and degree of ambiguity. Piantadosi et al. (2012)'s finding of a positive correlation between a word's ambiguity and its frequency of occurrence (i.e., frequent words are more ambiguous) can only be interpreted as evidence for optimality if the regression coefficient found for the degree of ambiguity is more positive than λ_2.

A simple rewrite of Eq. 3 results in

$$\mathcal{A}(w_i) = \frac{\lambda_0 - \log \mathrm{p}_i + \lambda_1 \ell(w_i)}{-\lambda_2}. \quad (4)$$

This indicates that, when word frequency is kept constant or controlled for, one should also expect to find a linear relationship between a word's length and its degree of ambiguity (with a regression coefficient $-\lambda_1/\lambda_2$), as was documented by Piantadosi et al. (2012). As before, in order to accept Piantadosi and colleagues' interpretation that their finding is indicative of some form of communicative efficiency, one needs to ensure that such relation is less marked than $-\lambda_1/\lambda_2$.

Although, for reasons of space I do not detail it here, it is easy to show that a word's frequency of occurrence in a corpus is expected to be directly proportional to that word's phonotactic probability as computed from an n-phone (e.g., diphone, triphone, ...) model whose parameters were computed on that same corpus. If we denote a word's triphone-based probability as T_i, we can therefore say that $k\,\mathrm{p}_i \approx T_i$ for some value $0 < k \leq 1$ constant across all words. If we substitute on Eq. 3, we obtain

$$\log T_i - \log k \approx \lambda_0 + \lambda_1 \ell(w_i) + \lambda_2 \mathcal{A}(w_i). \quad (5)$$

Dividing both sides of Eq. 5 by $\ell(w_i)$ with a simple rearrangement results in

$$\frac{\log T_i}{\ell(w_i)} \approx \lambda_1 + \frac{\log k + \lambda_0 + \lambda_2 \mathcal{A}(w_i)}{\ell(w_i)}. \quad (6)$$

Therefore, when the degree of ambiguity is controlled for, a word's log triphone (or diphone, ...) probability normalized by its length, is expected to be non-linearly related to word length itself (i.e., linearly related to the reciprocal of word length). One would therefore expect to find the non-linearities that Piantadosi et al. (2012) found. Hence, such non-linear relation –by itself– cannot be interpreted to be the product of an optimization process, contrary to what was argued by Piantadosi and his colleagues.

In summary, I have shown that the linear relationships between log word frequency, word length, and word ambiguity –by themselves– do not warrant an interpretation that language is optimized for communicative efficient. In the following section, I will estimate the values of the parameters

[2] The general form of the solution is due to L. E. Boltzmann. For a sketch ot the derivation, see, e.g., Moscoso del Prado (2011).

λ_0, λ_1, and λ_2 from corpus data, and I will assess whether or not they provide evidence for (or against) any sort of optimization.

Corpus Study

In order to test whether the values of the Laplace multipliers (λ_1, λ_2) provide support for the hypothesis that language is optimized for communicative efficiency, I selected the 29,025 most frequent English content words (adjectives, adverbs, nouns, and verbs) present in WordNet (Miller, 1995).[3] The selection was done using their surface word spoken frequency according the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995), from where the corresponding word length in phonemes was obtained.[4] For each word I counted its number of distinct senses (i.e., 'synsets') listed in WordNet (Miller, 1995). The log of the number of senses was taken as the measure of a word's ambiguity ($\mathcal{A}(w_i) = \log N_i$, where N_i is the number of distinct senses of w_i).[5]

I normalized the word frequency counts into relative frequencies adding up to one, and estimated the mean word length as the weighted average

$$L = \langle \ell(w_i) \rangle_W = \sum_{w_i \in W} \mathrm{p}_i \, \ell(w_i),$$

with p_i being the corpus based relative frequency of w_i. Similarly, the average degree of ambiguity was estimated as

$$A = \langle \mathcal{A}(w_i) \rangle_W = \sum_{w_i \in W} \mathrm{p}_i \, \mathcal{A}(w_i) = \sum_{w_i \in W} \mathrm{p}_i \log N_i.$$

Using these estimates of L and A, and the individual values of $\ell(w_i)$ and $\mathcal{A}(w_i)$, the values of the multipliers λ_0, λ_1, and λ_2 were estimated by nonlinear maximization (using a Newton-type algorithm) of

$$H(W) = -\lambda_0 - \lambda_1 L - \lambda_2 A$$

subject to the constraints of Eq. 1. In order to keep the results comparable to those of Piantadosi and colleagues, additional parameters were added to separate the different grammatical categories.

The values of the Laplace multipliers were estimated to be $\lambda_0 = -10.37$, $\lambda_1 = -.14$, and $\lambda_2 = 1.07$. As I discussed in the previous section, that $\lambda_1 < 0$ indicates that ZLA (a negative correlation between word length and word frequency) is the most likely relationship between these two variables, whether or not any optimization is at work. Similarly, $\lambda_2 > 0$ implies that we should expect *a priori* a positive correlation (all other factors equal) between a word's frequency and its degree of ambiguity. As suspected –by itself– the positive relation between frequency and ambiguity does not warrant the interpretation of optimization, it is rather what one should expect, contrary to Piantadosi et al. (2012).

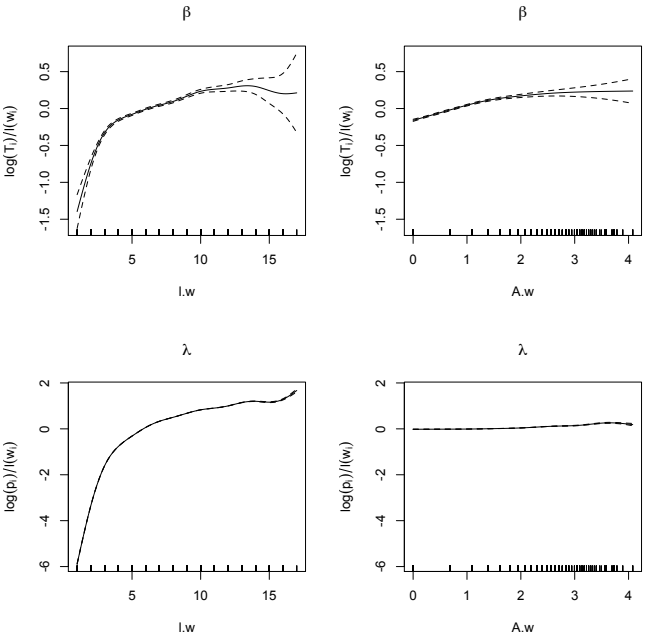

Figure 1: Non-linear effects of word length (left panels) and degree of ambiguity (right panels) on the length-normalized log triphone probability (top panels) and the length-normalized log *a priori* frequency (bottom panels).

To see the actual values of these correlations in the corpus itself, I performed a linear regression predicting a word's log probability from its length and its degree of ambiguity (once more, I also included additional parameters to separate the grammatical categories). As Piantadosi et al. (2012), I found significant effects of both word length and degree of ambiguity (both with $\mathrm{p} < .0001$). Interestingly, the estimated coefficient for word length ($\beta = -.05 \pm .003$) constitutes a much weaker effect than the one we would have expected by chance ($\lambda_1 = -.14$). This means that the supposed optimization from ZLA is actually *weaker* than what one should have expected, not supporting any optimization. In a similar vein, the coefficient estimated for the effect of ambiguity ($\beta = .84 \pm .01$) is also a *weaker* effect than the chance level ($\lambda_2 = 1.07$). Again, it seems that the negative relation between frequency and ambiguity that was claimed by Piantadosi and his colleagues to reflect optimization, is actually significantly weaker than the expected chance level. This illustrates the importance of having meaningful baselines before interpreting lexical statistics.

As discussed in the previous section the ratio $-\lambda_1/\lambda_2 = .13$

[3] The same effects reported here were also replicated for other corpora of English and French. These additional analyses are not reported here for brevity reasons.

[4] Piantadosi et al. (2012) report effects on length in syllables. I use phoneme-based lengths instead as these are more sensitive, but I also replicated the same effects using syllable-based lengths. Conversely, Piantadosi and colleagues also report that their effects also held when measuring length in phonemes.

[5] The log number of senses provides a better approximation to the psychologically relevant magnitude than does the raw count (cf., Moscoso del Prado, 2007). Note however that doing the calculation on raw counts of word senses did not result on different results.

indexes the strength of the correlation between word length and word ambiguity (after controlling for word frequency) that we should expect by chance. Indeed, we should therefore expect by chance a positive correlation between a word's length and its degree of ambiguity, irrespective of any optimization process. The strength of this relationship in the data is given by the ratio between the corresponding regression coefficients $\beta[\text{length}]/\beta[\text{ambiguity}] = .05$, which is once more weaker than what we would have expected by chance.

In order to assess the non-linear effects of a word's phonotactic probability, I trained a triphone model using the Brown corpus of English (Kucera & Francis, 1967) after transcribing all the words into the phonemic forms using the CMU Pronouncing Dictionary.[6] I used this triphone model to estimate the phonotactic probability of each word (T_i). For each the length-normalized log trigram probabilities ($\log T_i/\ell(w_i)$) and the length-normalized log *a priori* probabilities estimated using the λ ($\log p_i/\ell(w_i)$), I fitted a generalized additive model with a linear predictor for log word frequency (estimated for the corpus or *a priori*) and penalized spline smoothers terms for word length and degree of ambiguity. Fig. 1 plots the estimated curves. As predicted, the shape and strength of the non-linear relations is basically the shame for the actual triphone probabilities (top-panels), as it is for the word probabilities that would be predicted *a priori* (bottom panels). Once more, the shape of the relation does not warrant the interpretation of optimization.

The values of the Laplace multipliers can be used with Eq. 3 to compute what should be the *a priori* distribution of words, considering only their length and degrees of ambiguity. The log relative frequencies predicted by the method exhibit a remarkably strong correlation with the relative log frequencies actually observed ($r = .45, t[29004] = 86.41, p < .0001$). This suggest that the frequency distribution of words is not that different from the distribution one would expect to find by chance. In other words, it does not appear to reflect much specific optimization.

It could be argued that, by using the actual values of word length and word ambiguity as estimated from the corpus, I am covertly exploiting the possible correlations that are already present in their distributions, even before considering word frequency. To control for this possible confound, I used a Jackknife technique. New values of word lengths and ambiguity were randomly assigned to each word by randomly sampling (with replacement) from the original distributions. In this way, one obtains distributions of word length and ambiguity which are fully uncorrelated, but retain their original distributional shapes. I repeated this process two hundred times, re-estimating the values of the Laplace multipliers in each case. Fig. 2 compares the original λ estimates (blue dots), the distribution of λ values obtained in the resampling (box and whiskers plots), and the β parameters of the regressions on the actual corpus (red crosses). As it can be seen, even after fully decoupling word length and ambiguity, the expected effects are much stronger than those observed in the corpus. In short, the effects that allegedly reflect optimization of language for communicative efficiency are actually much weaker than we should have expected them to be.

[6] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Conclusion

These results do not question either ZLA or Piantadosi et al. (2012)'s effects, rather the effects themselves are indeed replicated here. What is questioned is the interpretation of such effects as evidence for optimization. I have shown that – by themselves– the linear relationships between log word frequency, word length, and degree of ambiguity, do not warrant the interpretation that language is optimized for communicative efficiency. The shape and direction of the effects reported in Piantadosi et al. (2012), as well as ZLA, are precisely what one would expect to obtain by chance (i.e., by a random assignment of probabilities). Furthermore, if anything, I find that the effects present in a corpus are actually *weaker* than what one would obtain by chance. Following the classical "Principle of Least Effort" interpretation is therefore not warranted by this type of correlations.

These findings complement previous studies showing that mere random typing processes (e.g., Mandelbrot, 1953; Miller, 1957; Ferrer i Cancho & Moscoso del Prado, 2011) also exhibit ZLA. More generally, the most likely observation is that a word's log probability of occurrence is linearly related with any property of the word for which a mean value exists. Therefore, in order to claim that observations of this kind are indicative of any type of process (optimization or otherwise) giving rise to the word frequency measures are not warranted, unless the effects are explicitly found to be significantly the stronger than the effects one should find by chance. In other words, such effects are meaningless unless compared to random baselines.

Notice that the same conclusions would be reached if, instead of unigram word frequencies, I had considered the *a priori* distribution of word bigrams or trigrams. It would merely be a question of applying the PME to the whole matrix of *n*-grams and we would expect to obtain the same type of linear relations between log *n*-gram frequencies and lexical measures. Thus, similar arguments as those expressed in Piantadosi, Tily, and Gibson (2011), would suffer from exactly the same problems I discussed above.

By this I do not intend to claim that language is not optimized for human communication. Rather the opposite, I am strongly convinced that this is indeed the case. However, pure correlational values between lexical measures (or for that matter *n*-gram measures) are not sufficient evidence to support such claims. Of course, there is a certain commonsensical aspect to the claim that human language is optimal for communication: It would be difficult to find a cognitive scientist who disagrees with such a statement. However, claims on optimization of human language should rely on specific mechanisms by which the optimization takes place, together with explicit mathematical (e.g., variational) de-

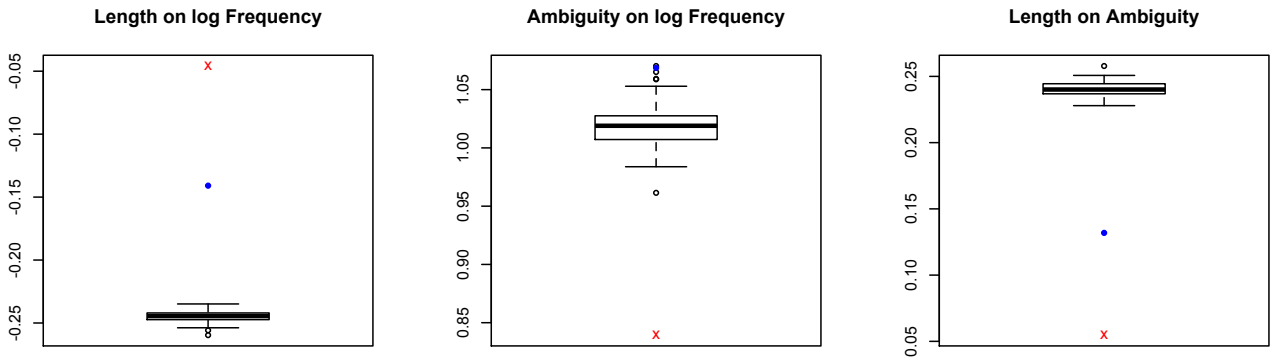

Figure 2: Estimated relations between log word frequency, word length, and degree of ambiguity. The red crosses indicate the magnitude of the effects observed in the corpora. The blue dots plot the magnitude that we should expect to observe *a priori*. The box and whisker plots plot the distribution of the *a priori* predictions once word length and ambiguity have been decoupled using Jackknife. The leftmost and middle panels respectively plot the effects of word length and ambiguity of log word frequency. The rightmost panel plots the direct relation between word length and word ambiguity. Notice the different vertical scales between the panels.

scriptions of how such an optimization proceeds, as exemplified by some recent studies (e.g., Ferrer i Cancho & Solé, 2003; Ferrer i Cancho & Díaz-Guilera, 2007).

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM)*. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Ferrer i Cancho, R., & Díaz-Guilera, A. (2007). The global minima of the communicative energy of natural communication systems. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*, P06009.

Ferrer i Cancho, R., & Moscoso del Prado, F. (2011). Information content versus word length in random typing. *Journal of Statistical Mechanics: Theory and Experiment*, *2011*, L12002.

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 of the United States of America*, *100*, 788 –791.

Jaynes, E. T. (1957a). Information theory and statistical mechanics. *Physical Review*, *106*, 620–630.

Jaynes, E. T. (1957b). Information theory and statistical mechanics II. *Physical Review*, *108*, 171–190.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In W. Jackson (Ed.), *Communication Theory* (p. 503512). New York, NY: Academic Press.

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

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41.

Moscoso del Prado, F. (2007). Co-occurrence and the effect of inflectional paradigms. *Lingue e Linguaggio*, *2*, 247–263.

Moscoso del Prado, F. (2011). Macroscopic thermodynamics of reaction times. *Journal of Mathematical Psychology*, *55*, 302–319.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology*, *17*, 273–281.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the USA*, *108*.

Piantadosi, S. T., Tily, H. J., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*, 280–291.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*, 379—423, 623–656.

Zipf, G. K. (1935). *The Psychobiology of Language; an Introduction to Dynamic Philology*. Boston, MA: Houghton-Mifflin.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Reading, MA: Addison-Wesley.