Modelling Zipfian Distributions in Language

Catriona Tullo and James R Hurford
University of Edinburgh
http://www.ling.ed.ac.uk/~jim

1 Introduction

G.K.Zipf famously discussed a number of patterns in the distributions of linguistic units, such as words and phonemes, in texts. We address several of these here, and attempt to explain their origins in terms of simple principles of language use, including, but going beyond, Zipf's own 'Principle of Least Effort'.

1.1 Rank/Frequency and Length/Frequency Correlations

The term "Zipfian distribution" refers to "a distribution of probabilities of occurrence that follows Zipf's Law". Zipf's law is an experimental law, not a theoretical one; i.e. it describes an occurrence rather than predicting it from some kind of theory. The observation that, in many natural and man-made phenomena, "The probability of occurrence of ... items starts high and tapers off. Thus, a few occur very often while many others occur rarely." The formal definition of this law is: $\mathbf{P_n = 1/n^a}$, where $\mathbf{P_n}$ is the frequency of occurrence of the **n**th ranked item and **a** is close to 1.

Applied to language, this means that the rank of a word (in terms of its frequency) is approximately inversely proportional to its actual frequency, and so produces a hyperbolic distribution. To put Zipf's Law another way: $fr = C$, where: r = the rank of a word, f = the frequency of occurrence of that word, and C = a constant (the value of which depends on the subject under consideration). Essentially this shows an inverse proportional relationship between a word's frequency and its frequency rank[1]. Zipf calls this curve the 'standard curve'. Texts from natural languages do not, of course, behave with such absolute mathematical precision. They cannot, because, for one thing, any curve representing empirical data from large texts will be a stepped graph, since many non-high-frequency words will share the same frequency.

[1]Note that this generalization is distinct from another frequency pattern also noted by Zipf, namely that $nf^2 = K$, where f is the frequency of some word, n is the number of words occurring f times in a text, and K is a constant (Zipf, 1935).

But the overall consensus is that texts match the standard curve significantly well. Li (1992:1842) writes "This distribution, also called Zipf's law, has been checked for accuracy for the standard corpus of the present-day English [Kučera & Francis, 1967] with very good results". See Miller (1951:91-95) for a concise summary of the match between actual data and the standard curve.

Zipf also studied the relationship between the frequency of occurrence of a word and its length. In *The Psycho-Biology of Language* (1935), he stated that "it seems reasonably clear that shorter words are distinctly more favoured in language than longer words." So a very few shorter words are used (spoken or written) very frequently, while others are used very rarely. Zipf did not specifically claim that the same Law which describes the connection between word rank and frequency also applies to frequency and length. He merely stated that there is a general tendency for word length to decrease as word frequency increases, "the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences", (Zipf, 1935). Nor did he expand on any possible mathematical formula to model this. It seems clear that the general length/frequency correlation is realized more messily in texts than the rank/frequency correlation. Nevertheless it is clear that there is a gross similarity between the Rank/Frequency and the Length/Frequency curves observable in linguistic texts. Both are roughly 'J-shaped', and we will refer to both distributions under the broad heading of 'Zipfian distribution'.

1.2 What a Zipfian Distribution can Explain

Several studies (Kirby, 2001; Onnis et al., 2002) propose models which explain the emergence of irregularities in language. Thus, beside the Zipfian correlations, linguists are familiar with the fact that there is also a correlation between the frequency of words and constructions and their morphological or syntactic irregularity. For instance, the most frequent verbs in a language are also those most likely to be irregular. The cited studies report computer simulations in which successive generations learn their language from the statistically biased usage of previous generations. Agents in these simulations use words or grammatical patterns with varying frequency, determined by an assumed Zipfian distribution. Taking such a Zipfian distribution as given *a priori*, the observed correlation between frequency and irregularity emerges as a stable property of the language of a simulated population.

But of course, the assumed Zipfian distributions themselves remain to be explained.

1.3 What can Explain a Zipfian Distribution?

George Miller (1965) observed that a random text would be expected to exhibit Zipfian distributions. A random text is generated by iteratively emitting random letters from a given alphabet including the space character. Maximal strings of non-space characters are counted as 'words' in such a text. Miller's idea is taken up by Li (1992) who proves that such rank/frequency and rank/length correlations are indeed to be expected in any random text of reasonably large size.

So what? Miller thought that the fact that random texts give rise to Zipf-like distributions invalidated Zipf's own 'Least Effort' explanation for the Length/frequency correlation. "It seems, therefore, that Zipf's rule can be derived from simple assumptions that do not strain one's credulity . . . , without appeal to least effort" (Miller, 1957). "Zipf's curves are merely one way to express a necessary consequence of regarding a message source as a stochastic process" (Miller, 1965). In other words, Zipf, in proposing his Least Effort Hypothesis, had not eliminated the Null Hypothesis. Miller's argument is that if the distributions to be accounted for emerge from random texts, the Null Hypothesis can account for them, and there is no need of any further explanatory mechanism, such as a Law of Least Effort. But is this a reasonable Null Hypothesis? Of course, texts in natural languages are not generated by random emission of phoneme-sized elements. They are not even generated by emission of words randomly picked from a lexicon (a zero-order approximation to a natural language – Shannon, 1948). Why, for example, could it not be an equally reasonable Null Hypothesis that all words are equiprobable? Then, of course, the Zipfian distributions would appear strikingly significant, and in need of explanation.

What strains credulity is surely Miller's idea that human language results from a stochastic monkey-and-typewriter scenario. Mathematical derivations of Zipf-like distributions from random texts deliberately ignore the fact that natural language texts are produced by intentionally communicating agents. A reasonable explanation for Zipf-like distributions should be embedded in a theory which makes realistic assumptions about the causal factors which give rise to natural language texts. Here we present such a model. Our model retains Miller's idea that the message source is a stochastic process, but situates this process in a more realistic human culturo-linguistic scenario. Our work also dovetails with models, such as Kirby's and Onnis et al.'s, which assume Zipfian distributions among their given initial conditions. Thus the present study can be seen as complementing those studies by deepening their foundations.

2 The Discourse-Triggered Meaning Choice Model

There exist two sources of meaning choice for an average speaker. The first is the environment. A speaker may react to something in their surroundings by talking about it. A commonly used example of this is the habit many people have of starting a conversation by talking about the weather. For the purposes of this discussion we assume for the moment that, from the point of view of the environment, all meanings are equally likely to be chosen by the speaker (i.e. the environment has an even frequency distribution).

Once a dialogue has begun, however, another source of word choice is available – those words used, or heard, in the preceding conversation. For example if the first speaker in a conversation (S1) begins by mentioning that a person, Fred, has gone on holiday, it is likely that S1's conversational partner, S2, will carry on the discussion by talking about one of the two topics started by S1, i.e. Fred or holidays. Any attempt to change the topic to something completely unrelated would cause a certain amount of confusion on the part of S1 and possibly a breakdown in the conversation. This means that words related to Fred, or holidays, are likely to be used most often in this discussion. So from the entire vocabulary of the language there is a small subset of words which have a much higher chance of selection.

Our hypothesis is that it is the topics or meanings used in the preceding dialogue which give the Zipfian distribution to the frequency curve of words. Some words are used more frequently than others because language users hear them more frequently than other words. So if, for some reason, a word is spoken more frequently for a while, it will then become spoken even more frequently because it has been heard more often. So, in essence, the more frequent a word is, the more frequent it will become. This should happen even if, to start with, all words are given an even frequency distribution, and speakers initially choose words at random. By chance, some words will be selected more often than others will. This inevitably leads to an uneven frequency distribution developing in the language. Our model takes a lead from Harremoës & Topsøe (2001); they write "The child gets input from different sources: the mother, the father, other children, etc. Trying to imitate their language with frequencies which are closer to Zipf's laws than the sources. As a language develops during the centuries the frequencies will converge to a hyperbolic distribution." (A conversation with Jörg Rieskamp also helped to inspire this model.)

2.1 Testing the Model by Simulation: Rank/Frequency

A computer model was devised to test the coherence of this hypothesis. A set containing the numbers 1 to 1000 was created. This corresponds to the initial word-store of the language, with each number

representing a word, so this vocabulary is created with an even frequency distribution. From this first set 1000 'words' are selected at random one at a time, and are used to create the word-store for the next generation. When a word is chosen it is copied, not moved, into the new word-store. This makes it possible for the new word-store to contain more than one instance of certain words whilst having failed to copy other words which appeared in the original set. This process is repeated, with each new word-store being selected from that of the generation before (not from the original 1000 numbers with even distribution).

Thus when a new word-store is being built some words will be selected more than once and others will not be selected at all for the new group. This means that over time, although the actual size of the corpus of words will remain constant at 1000, the size of the vocabulary (i.e. the number of different words) will tend to decrease, as some words are lost due to not being selected.

(Note that this model falls within the 'iterated learning' paradigm as developed by Kirby in several publications (e.g. Kirby & Hurford, 2000), also called the 'E/I' (Expression/Induction) class of models in Hurford (2002).)

We evaluate the results of our simulations impressionistically as follows. A true Zipfian curve is a simple hyperbola. A simple hyperbolic graph (of the equation $rf = C$), drawn on a double logarithmic scale, is a straight diagonal line from top left to bottom right. When the data obtained from the model is plotted on the same graph as this hyperbola it is possible to compare the two lines.

This basic model was run with various additional features, as described in the following subsections.

2.1.1 Drastic Vocabulary Loss in the Basic Model

The results from running the basic model showed that the distribution of words produced was not quite 'J shaped'. There is a skewed distribution evident in the results, but it does not have the almost exponential shape of the true Zipfian distribution, as can be seen in figure 1. This shows the frequency of a word plotted against it's rank after a single run of 100 generations.

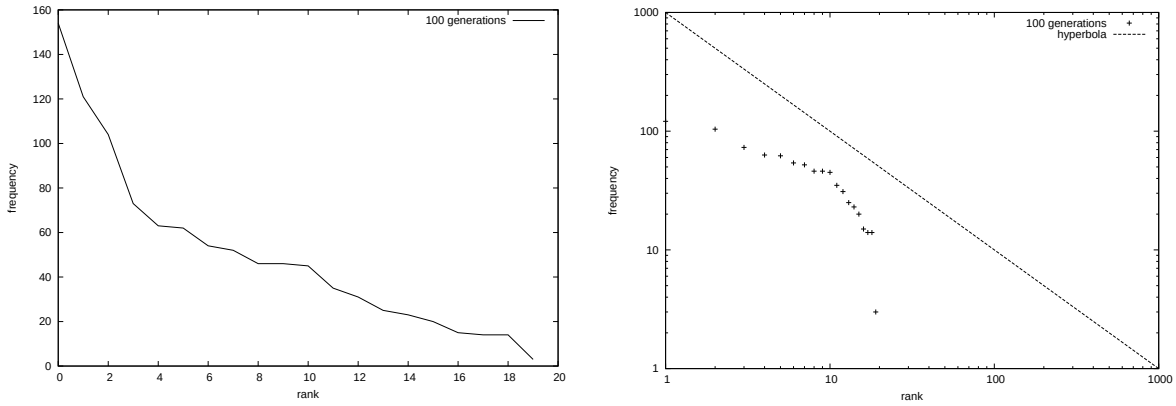

Figure 1. Left panel: The frequency of a word plotted against its rank after 100 generations. Note that there are only 19 words left in the vocabulary at this point.
Right panel: Frequency-rank distribution after 100 generations plotted on double log scale.

After 100 generations (i.e. 100 successive vocabularies) from a starting vocabulary of 1000 there are only on average only 20 different words left. This is a dramatic reduction and would have devastating effects on any real language if word-types were actually lost in this way. As the right panel of figure 1 shows, the word-stores resulting from this basic model do not have a perfectly hyperbolic, or Zipfian, distribution, although they approach it.

This different distribution could be due to the fact that in this basic model words die out completely, and there is no way of them ever being reintroduced into the language. In real languages some words are used a lot less than others, but they aren't necessarily lost from the language completely. A way of randomly re-introducing some words back into the language, or alternatively, preventing words from being lost completely, is the next step.

2.1.2 Meaning Choice Partly Triggered by the Environment

Our basic model investigated the effect of sampling previous discourse on word selection. But the environment clearly also has some influence on word choice. Speakers do not simply carry on talking about one topic for entire conversations; changes in subject matter do occur. Some of these may come from associations with the current discourse, but another source of topic is the speakers' surroundings.

To introduce this effect into the existing model a stable set containing all of the possible meanings in the language (i.e. here the numbers 1 to 1000) was preserved in the background during the runs. This represents the environment. At the beginning of a run a variable 'Environment Parameter', *E*, was

set, representing the probability of the next word to be put into the word-store coming directly from the environment, rather than from a sampling of the word-store of the previous generation. For example, where *E = 0.1*, each time a new word is selected to be added to the word-store, it has a 0.1 probability of being randomly selected from the complete environmental set of meanings, and a 0.9 probability of being randomly selected from the word-store of the previous generation, as in the basic version of the model..

It was found that as the probability of a new word being selected from the environment, rather than from the preceding conversation, was increased, the time taken for the vocabulary to decrease, and take on a Zipf-like distribution, increased. Figure 2 shows how the rank-frequency distribution after 100 generations changes with different weightings of the environment. From this it can be seen that this modification has achieved results approaching, but still somewhat far from, a hyperbolic distribution.

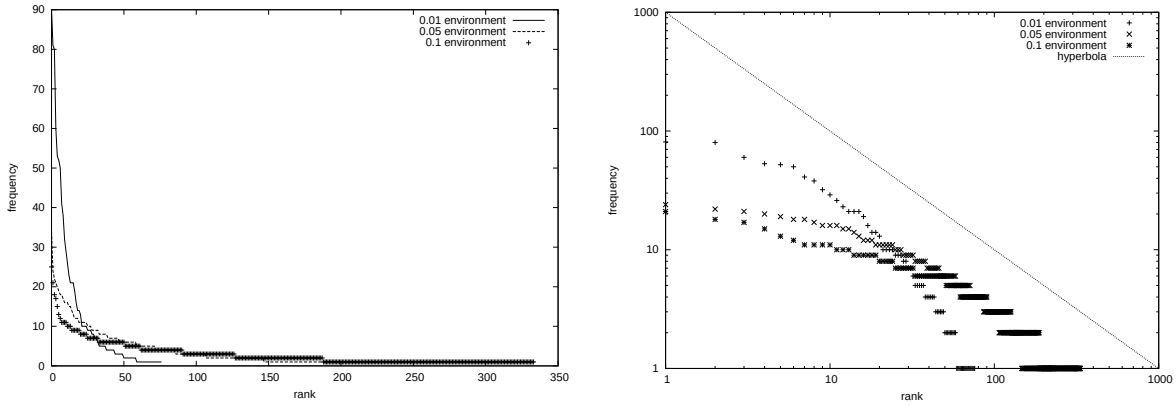

Figure 2: Rank-frequency distribution for differing environmental influence – *E* = 0.01, 0.05 and 0.1. The right panel is on a double log scale, for comparison with a hyperbolic curve (straight line).

2.1.3 Corpus Size

Thus far in this paper we have assumed that the speaker in each generation will select only 1000 words every time. The number of different word types available to the speaker initially may only be 1000, but a realistic assumption is that far more word tokens than this will actually be spoken (or put into each new word-store at each generation).

To model this the original set is left unchanged, hence the environment still contains only one instance of each word. However in every subsequent generation more than 1000 word tokens can be selected. This number is kept constant throughout every run of the model, therefore each word-store (after the original environment) will contain the same number of tokens. So although there are still only 1000

different lexical items in the language, from this, for example, 2000 word tokens can be presented at each generation. For the initial run of this, the level of the influence of the environment was 0.

It was found that increasing the corpus size again slowed the decrease in the size of the vocabulary. This is illustrated in figure 3 (left panel). Altering the corpus size causes the frequency distribution of the words to change markedly, as shown in figure 3 (right panel).

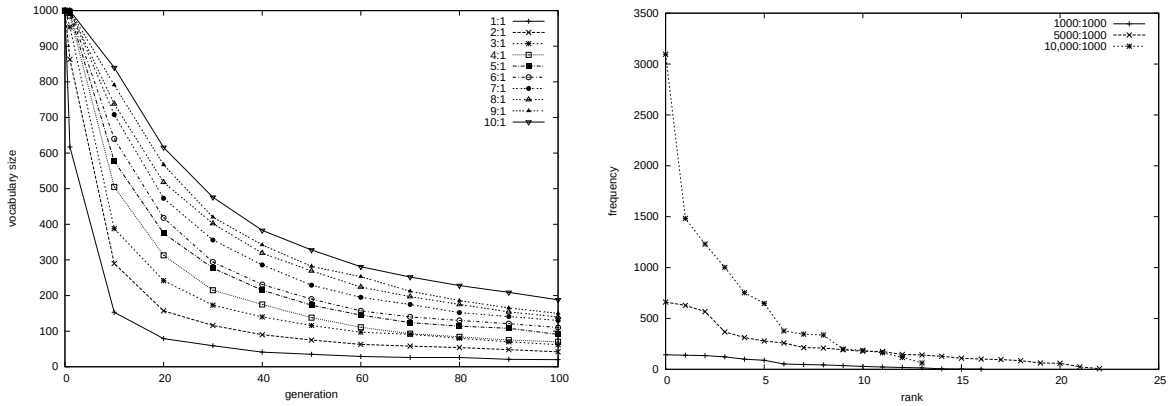

Figure 3. Results with different corpus sizes, selected from a vocabulary of 1000.
Right panel: Change in vocabulary size at each generation with differing corpus size.
Left panel: Rank-frequency distribution for corpus sizes of 1000, 5000 and 10,000.

This effect is seen because increasing the number of selections means that the initial selection from the environment contains a greater variety of words. Hence the size of the vocabulary decreases at a much lower rate than in previous runs. However, the Zipfian distribution can still emerge, albeit a little more slowly. No matter how many selections are made, some will still be selected more than others, and these words will have a greater chance of being selected in the next generation as well.

2.1.4 Combining Environmental Influence with Corpus Size

If this effect were combined with a more stable vocabulary (where words didn't die out so rapidly) a hyperbolic distribution may emerge, given that in earlier trials the environment had a positive effect on the rank-frequency distribution of a vocabulary. To this end the effect of the environment was re-introduced. The probability of a word being selected from the environment was kept small at 0.05.

It was found that allowing selections from the environment meant that the vocabulary size more or less stabilised after approximately 20 generations for each of the runs. Figure 4 (left panel) illustrates this effect (and the also the different sizes at which stabilisation occurred). Reintroducing the environmental

influence meant that the result of increasing the corpus size was even more apparent. The frequency distribution of the words continued to change. To test whether this change was towards a more Zipf-like distribution the same test was applied as before. The results were plotted on a double logarithmic scale alongside a hyperbolic line. The result of this is shown in figure 4 (right panel).

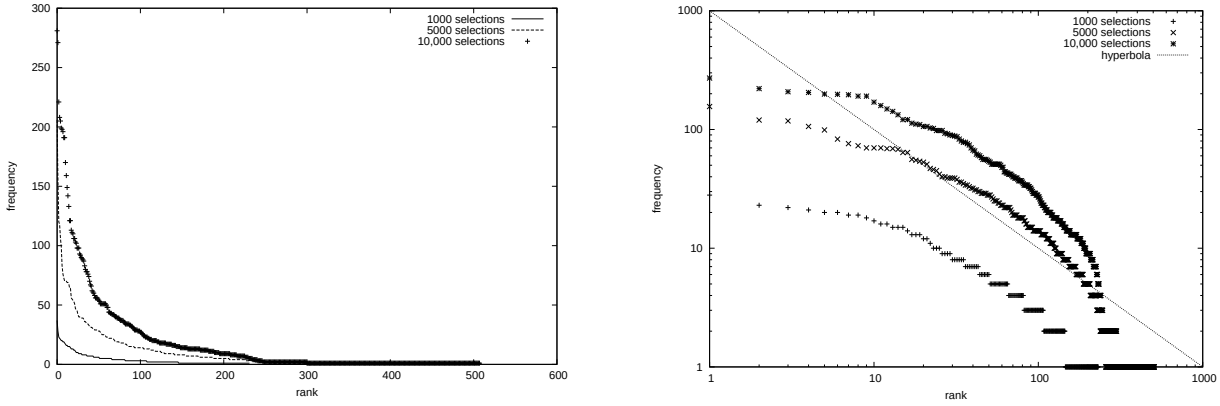

Figure 4: Combining environmental influence and varying corpus size, from a vocabulary of 1000.
Left panel: Vocabulary size against generation for corpus sizes of 1000, 5000 and 10,000 .
Right panel: Log-scale graph showing the rank-frequency distribution after 100 generations for corpus sizes of 1000, 5000 and 10,000.

This clearly shows that, although the lines are closer to the hyperbole than in earlier graphs, there is still not an exact hyperbolic distribution in the frequency of words in this model. Perhaps however this is an unrealistic expectation for a model which uses only a small number of words. Even with the highest corpus size (10,000) the size of the vocabulary is still only about one third of that used by Zipf in his original study (of the distribution of words in James Joyce's *Ulysses*).

2.2 Testing the Model: Length/Frequency

The model as developed so far makes no mention of word length. At the beginning, we undertook to explain the Zipfian length/frequency correlations as well as the rank/frequency correlations. In keeping with our general approach, modelling the cultural transmission of word-stores across generations by learning, we now introduce a factor of word-length, and apply a simple implementation of Zipf's own least effort principle. Recall that Zipf made less precise statistical claims about the length-frequency corelation.

We hypothesise that the more often a word is spoken the more likely it is that at some point the

signal passing between the two individuals involved in a conversation will be degraded somehow. For example, if two people are talking in a noisy environment it is possible that a word spoken by S1 will be heard differently by S2 because of the interference of the noise in their surroundings. Of course meaning must be conserved in the real world, so the change in the word must be recognised by all speakers in the community - they must still be able to communicate and be understood by each other.

We will continue to use the model so far developed looking at frequency and rank, with some slight modifications to add a length to each word. In the current model, the initial set, or the environment, contains the numbers 1 to 1000, each of which represents one word. To this a length will be added for each word. So the initial set will consist of 1000 word-length pairs. This means that every time a word is selected to be put into the new word-store, its length is copied over with it. All words start with the same length, arbitrarily determined, at the beginning of a run.

To simulate the effect of shortening another variable was created which sets the probability of 'noise' interfering and thus the word being shortened by 1 between the source vocabulary and its destination. By this means it should occur that words which are selected more often are shortened more often and so more frequent words will become the shortest ones. (This is the same random shortening mechanism as in Kirby's (2001) model.) Obviously with this model of word shortening it is possible to have the same word with two different lengths in the same vocabulary. This is because a word may be copied over to the new vocabulary once without its length being reduced, but then copied over again and have its length reduced by one. To normalise a difference in length occurring between two instances of the same word, once a vocabulary is complete it is searched for such an event. The length of each word is then changed to match the lowest length present for that word in the vocabulary.

There are two possible procedures for analysing the data obtained from this new model. The first would be to follow the method used on the results gained from looking at the relationship between frequency and rank. This is simply to look at an individual word's frequency and its length, and plot the two against each other. This produces graphs with multiple lengths for the same frequency.

The alternative method, used by Zipf, is to group together all words with the same frequency and then take the average length for words with that frequency. Figure 5, a graph illustrating the data from an investigation of R. C. Eldridge (1911) of the English words used in four American newspapers, demonstrates this method, (Zipf, 1935). There are many fewer examples of words of higher frequencies, therefore any difference in the lengths of words of these frequencies has a much greater impact on the average length. For this reason Zipf averaged length over sets of higher frequencies, rather than simply taking the average length of words with the same frequency. This paper is attempting to recreate the results

Zipf obtained from his study of natural language. Therefore the second method described will be used to illustrate the results of the new model.

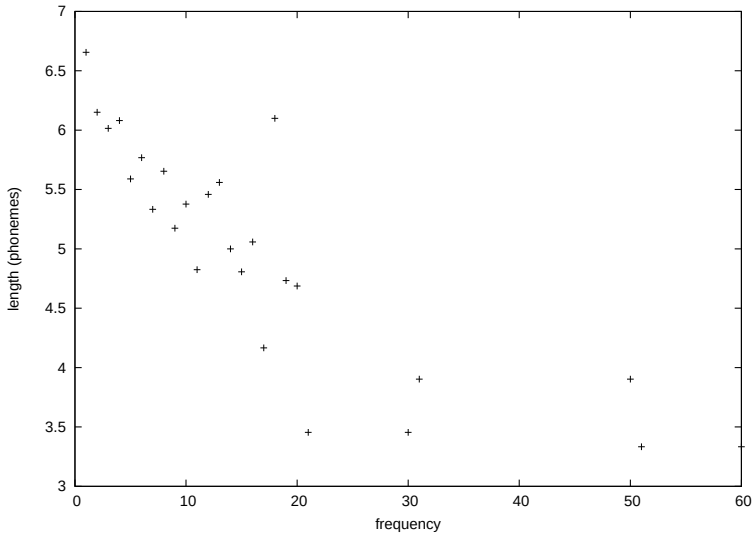

Figure 5. Eldridge's survey of frequency-length relationships of words in American newspapers.

2.2.1 Length/Frequency Results with Basic Settings

The first runs of the new model will use 'basic' settings: no environmental influence and a corpus size of 1000. This will give a basic shape for the graph of length against frequency. Parameters will then be altered to include environment and Kdifferent corpus sizes. We hypothesize that the relationship between frequency and length will require much the same conditions as were needed to gain a Zipf-like rank/frequency correlation in the previous section. This is due to the unavoidable link between these two relationships; if the frequency distribution is not like that of a natural language then the reported frequency-length relationship may not occur.

In this situation the rank-frequency distribution is much like it was before in the first few runs of the very first model. This meant that at the end of a run of 100 generations there were very few words left in the language (somewhere between 15 and 25 usually). This obviously has an effect on the frequency-length distribution. As so many words have been lost from the vocabulary it is very difficult to find a pattern on a graph of frequency versus length with so few points.

2.2.2 Adding Environmental Influence Again

To combat this loss of diversity the environment parameter was adjusted to try and stabilise the size of the vocabulary. This prevents words being lost so quickly and may allow the pattern of word shortening

to emerge as a result. The environment parameter was set to a relatively low level at 0.05. This was to ensure that there would be some stabilisation of the vocabulary size, whilst still allowing the vocabulary to develop to the Zipf-like shape.

The results graphed below show that the inverse relationship between frequency and length is beginning to emerge, but still not to the extent reported by Zipf. It may be that the vocabulary size, although somewhat stabilised, is not large enough for the length-frequency relationship to appear. As noise is increased to 0.5 the inverse shape of the graph begins to show through as the range of lengths at the 100th generation grows, as shown in figures 6.

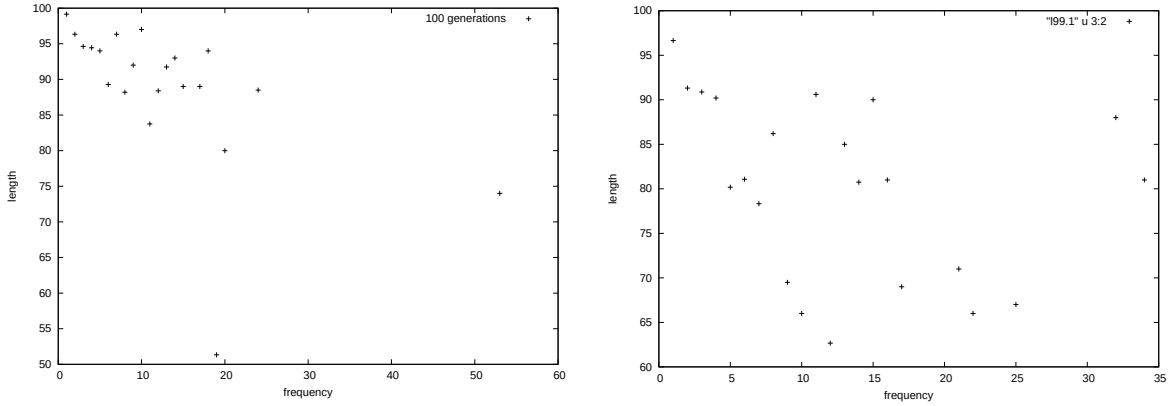

Figure 6: Frequency-length distribution after 100 generations, with environmental influence set at 0.05: **Left panel:** Noise = 0.1. **Right panel:** Noise = 0.5.

2.2.3 Varying Corpus Size Again

Corpus size is the parameter which had the greatest effect on the rank-frequency distribution. Therefore once this relationship is as Zipf found it to be, the inverse relationship between frequency and length should emerge. Runs were completed with 5000 and 10,000 selections to ascertain what effect increasing numbers of selections would have on the frequency-length distribution.

From the graphs below it can be seen that increasing the corpus size has produced a distribution closer to that Zipf obtained. Figure 7 shows the frequency-length relationship after 100 generations for corpus sizes of 5000 and 10,000. Less frequent words can be seen to have a greater length, although more frequent words tend to have a more variable length.

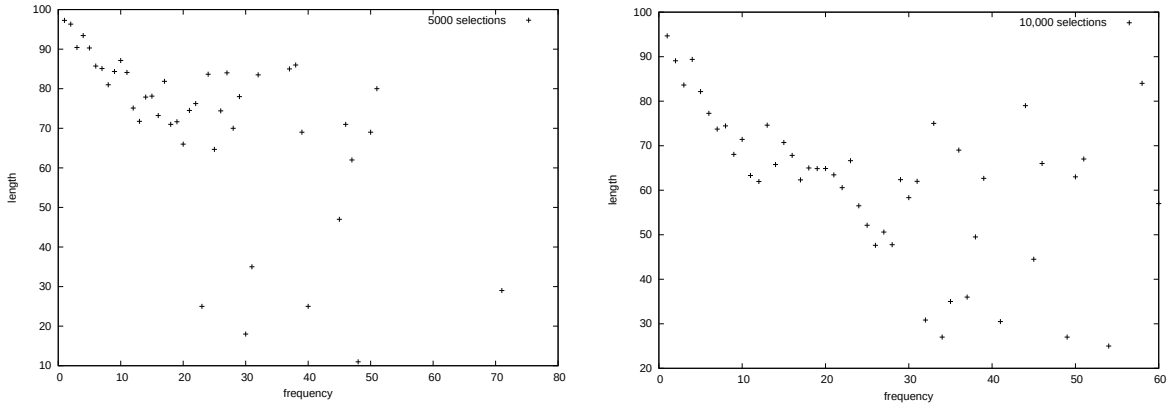

Figure 7: Frequency-length distribution after 100 generations, $E = 0.05$, noise = 0.1.
Left panel: Corpus size = 5000. **Right panel:** Corpus size = 10,000.

Eyeballing these results, the pattern in the right-hand panel is quite similar to Eldridges's data presented in figure 5.

3 Conclusion

Our simulations model the following factors in the transmission of vocabularies across generations:

- Storage of information on frequency of words heard,
- A major influence of stored word-frequency on production,
- A mild influence of non-discourse-related factors ('environment') on word-choice,
- Large corpus size relative to vocabulary size (token/type ratio),
- Noise affecting random shrtening of words.

Our results show interesting similarities with the rank/frequency and length/frequency distributions described by Zipf.

4 Bibliography

Eldridge, R. C. (1911) *Six Thousand Common English Words*, Buffalo: The Clement Press.

Harremoës, P. and Topsøe, F. (2001) "Maximum Entropy Fundamentals", *Entropy*, 3:191-226.

Hurford, James R. (2002) "Expression/induction models of language evolution: dimensions and issues".

In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, edited by Ted Briscoe, Cambridge University Press. pp.301-344.

Kirby, Simon, (2001) “Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity”, *IEEE Transactions on Evolutionary Computation*, 5(2):102-110.

Kirby, Simon and Hurford, James R (2001) “The Emergence of Linguistic Structure: an Overview of the Iterated Learning Model”, in Parisi, Domenico and Cangelosi, Angelo, Eds. *Computational Approaches to the Evolution of Language and Communication.* Springer Verlag, Berlin.

Kučera, H., and W. Nelson Francis (1967) *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island.

Li, W. (1992), “Random texts exhibit Zipf’s-law-like word frequency distribution”, *IEEE Transactions on Information Theory*, 38(6):1842-1845

Miller, George A., (1951) *Language and Communication*, McGraw -Hill, New York.

Miller, George A., (1957) “Some effects of intermittent silence”, *American Journal of Psychology*, 70, pp.311-314.

Miller, George A., (1965) Introduction to republication of Zipf (1935), MIT Press, Cambridge, MA.

Onnis, Luca, Matthew Roberts, and Nick Chater (2002) “Acquisition and evolution of natural languages: Two puzzles for the price of one”, paper given at the Fifth International Conference on the Evolution of Language, Harvard.

Shannon, Claude E., (1948) “A mathematical theory of communication”, *Bell Systems Technical Journal*, 27:379-423, 623-656.

Zipf, G. K., (1935) *The Psycho-Biology of Language*, Houghton Mifflin, Boston.

Zipf, G. K. (1949) *Human Behaviour and The Principle of Least Effort*, Addison-Wesley, Cambridge, MA.