Applications and Explanations of Zipf's Law

David M. W. Powers
Department of Computer Science
The Flinders University of South Australia

powers@acm.org

Abstract

Recently I have been intrigued by the reappearance of an old friend, George Kingsley Zipf, in a number of not entirely expected places. The law named for him is ubiquitous, but Zipf did not actually discover the law so much as provide a plausible explanation. Others have proposed modifications to Zipf's Law, and closer examination uncovers systematic deviations from its normative form. We demonstrate how Zipf's analysis can be extended to include some of these phenomena.

1. Introduction and Motivation

In this paper we wish to revisit Zipf's study of the relationship between rank and frequency of various linguistic and social units and constructions. The paper arises out of observations in Natural Language Learning experiments of deviations from the received version of Zipf's Law. As it may not be immediately obvious why this relationship is of significance in NLL, we very briefly mention some of the places where the relationship affects research in our field, and which we feel could usefully be further explored.

1.1 Quantitative Linguistics

The field of which Zipf was the pioneer is discovering lots of interesting empirical laws, but how much has it advanced in explanation or application (Köhler,1991)?

1.2 Statistical Learning Methods

Zipf's Law tells us how much text we have to look at and how precise our statistics have to be to achieve what level of expected error. (Finch, 1993; Powers, 1996). For example, the most frequent 150 words typically account for around half the words of a corpus, although this figure varies significantly with the size of the corpus, the size of the lexicon, the genre, register and medium of communication and the linguistic complexity of the text — and this is one of the phenomena we wish to start to examine in this paper. Zipf's Law is also closely related to the Good-Turing smoothing technique, and a better law could lead to better smoothing (Samuelsson, 1996). Note that Samuelsson showed that Zipf's Law implies a smoothing function slightly different from Good-Turing.

1.3 Semantics and Information Retrieval

Zipf's Law provides a base-line model for expected occurence of target terms and the answers to certain questions may provide considerable information about its role in the corpus (Steele,1998): What does it mean to ask if a word is significant in a corpus, beyond mere occurence or relative probability? What is the range of the semantic influence of a word in a corpus? What does the pattern of occurences contribute to our assessment of its relevance in the corpus?

1.4 Parser Evaluation

Zipf's Law provides a basis for evaluating parsers and taggers (Entwisle and Powers, 1998). Again we summarize the potential role in the form of a series of questions: How does a language model developed on one corpus transfer to another? How do we translate performance estimates on a few test corpora to estimates for the language as a whole? How do differences in register, genre and medium affect the utility of a system, and how do we compensate for these differences?

1.5 Computational Psycholinguistics

Zipf's Law provides a distributional foundation for models of the language learner's exposure to segments, words and constructs, and permits evaluation of learning models (Brent, 1997). It also provides a basis for evaluation of models of linguistic and cognitive access and storage models (Segui, Mehler, Frauenfelder and Morton, 1982). Whilst qualitative explanations and evaluations have been given on the basis of an assumption of the general relationship, a more precise account will lead to more quantitative models.

Whilst the law's qualitative or coarsely quantitative roles across these areas may seem rather fuzzy, and it stretches the imagination to see how a more precise characterization of the law could improve the performance in these applications and models, we note that the relationships, particularly from a Psycholinguistic point of view, demonstrate that the law is relevant to several aspects of our field, and that explanation and understanding of the law is an intrinsically valuable scientific objective.

David M.W. Powers (1998) **Applications and Explanations of Zipf's Law.** In D.M.W. Powers (ed.) *NeMLaP3/CoNLL98: New Methods in Language Processing and Computational Natural Language Learning,* ACL, pp 151-160.

2. Zipf's Principle of Least Effort

Zipf's major work on this subject explores a theory based on a competitive process balancing the minimization of the effort of both speaker and hearer. He uses an analogy in which words are regarded as tools, which are so constructed and arranged as to be able to achieve the communication task as efficiently as possible. Note that this culmination of his research into this relationship coincided with the publication of Shannon's proposals in information theory, and we will seek to make the connections clear shortly.

Zipf considered that the speaker had to build a continuous stream of specified products, that is an ongoing stream of utterances conveying specified meanings, in such a way as to minimize his effort as speaker consistent with effective communication to the hearer, her task being simplified as the relationship between utterances and meanings approached one to one: the work involved in producing a construction consists of the work involved in fetching the tool, which is directly in proportion to the cost of fetching the tool and includes both the mass of the tool, m, and the distance, d, that it needs to be fetched, given increasing either increases the effort required. Mass corresponds to length in Zipf's model, and distance to access time. Fequency, f, and work, w, must also be directly related, so: $w = f * m * d$, assuming direct proportionality to work in each case. Also the age of the tool (word) and number of different uses (meanings) vary directly with frequency.

2.1 Access Method

We now consider what Zipf called the "close packing" of our tools. Zipf in fact considered only one model which fitted the empirical facts, but we will consider more in order to explore to what extent the law really does correspond to optimality: What is the optimum access time for a set of N tools? In computer science, the optimum organization structures which we typically think of our hash tables and trees, with o(1) and o(logN) access times respectively. The former assumes that encoding of arbitrary length words is done in the same amount of time, and thus implies both a limit on the length of words and suboptimality of this hash coding scheme since best case and worst case are the same (in machine architecture terms, the machine uses fixed length words and is synchronous and cycle limited, and this fixed length must be at least o(logN) in order to permit full addressability). The tree access technique makes similar assumptions except that length independence may be relaxed (in machine architectures, the access would be pipelined or serialized so that length of the word and depth of storage add without increasing the order).

The theoretical o(1) and o(logN) access times which we are familiar with in computer science are however not physically sustainable (Powers,1995). Thus encoding process for hashing also takes at least o(logN) time given a limitation on the degree (fanin/fanout) of the logic elements or neurons (which turns it into a tree anyway). Worse still, as we seek to pack neurons into an n-dimensional space the speed of propogation limits our access time to $o(N^{1/n})$ and our optimal tree is not practically achievable (this can be hidden in the cycle time, which then defines an upper limit for N).

Thus our class of optimal solutions is limited to the set of $o(N^{1/n})$ solutions where n is 3, 2 or 1, which correspond to volumetric, areal and linear constraints respectively. Hence our access time or storage depth for a word of rank r, d_r is related by $w/m_r = f_r * d_r = f_r * r^{1/n}$. (Note that we simplify equations for the moment by leaving units and constants.)

The three dimensional solution clearly leads to the most efficient packing, with $n = 3$, but Zipf's law seems to corresponds to linear packing, with $n = 1$. Does this mean that optimality is not reached?

We answer this question in two parts: we look at information theory as a measure of efficiency, and we consider the physical constraints further. But first we look at how we move from rank to a more natural measure. The rank of a word type represents the number of word types of greater frequency — our conventional definition of the most frequent word as having rank 1 is slightly defective when ties are taken into consideration, and calling this rank 0 would lead to a more consistent definition. Thus sometimes a constant 1 needs to be added or subtracted in our formulae to allow for this. The rank associated with a particular frequency, r_f, is thus the sum of the numbers of words, n_f, of greater frequency, and r_f may thus be approximated by the integral of n_f.

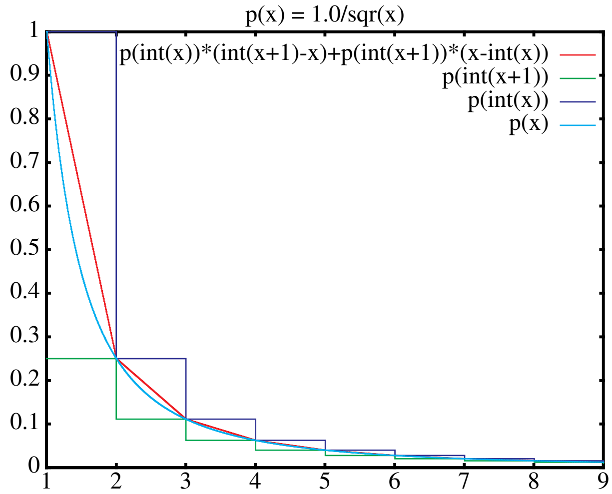

Figure 1: *Error approximating series by integral*

2.2 Error Estimates

In fact, approximating a monotonic series by an integral leads to an error which can be characterized as being slightly more than half of the first term of the series (as illustrated in Figure 1), or alternatively as representing an error of slightly under a half in the index.[1] Zipf's law for rank is thus approximated either by $r = 1/f_r + 0.64/f_r^2$ given that we use Zipfs law for number as $n_f = 1/f_r^2$ or by $r = 1/f_r$ if we use $n_f = 1/(f_r+0.43)^2$. Figure 2 illustrates the general inverse and inverse square laws, where we plot rank and number against frequency both for individual frequencies (ragged plots near gradient -1 and -2 resp.) and aggregated frequencies (step function and piecewise linear curves near gradient -1). The aggregation was by powers of two ($n' = 2^{|\log n_1|}$) as suggested by the scale. Note that we see the integrating effect not only for rank but for aggregated number.

These approximations may be used to estimate stepsize and expected error as indicated in Figure 3. The centre line is Zipf's law for rank based on the highest frequency (the word 'the' occurs 1642 times in "Alice in Wonderland") and the outermost pair of curves are based on on the highest and lowest ranks associated with the frequency 1, namely 1486 and 2620, plus or minus the maximum error. Note that the difference between these

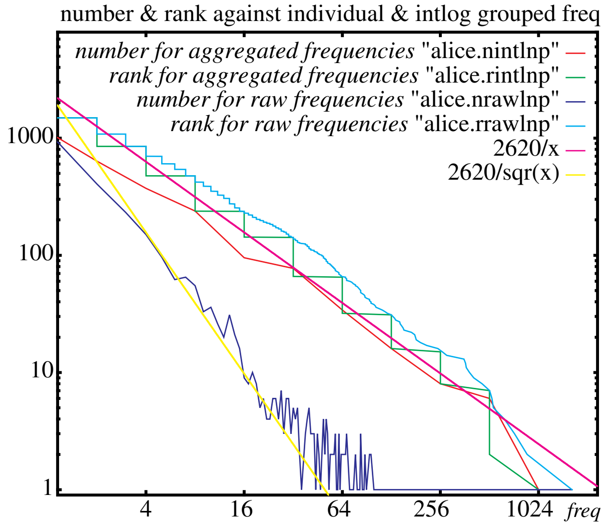

Figure 2. *Effect of aggregation of numbers to ranks*

1. This characterization of the error, closely related to a formulation due to Euler (Stanaitis, 1967), is actually considerably more accurate than that used by Zipf, and may be verified graphically from Figure 1 where the inscribed step function represents the sum whose area is underestimated by the integral of the continuous curve. The circumscribed step function represents the sum displaced by 1, corresponding to omission of the first term, f(1). The error is not only bounded above by the sum of the areas enclosed between the two stepfunctions, which is equal to the value of the first term, but it can be see to be bounded even more closely below by the chord function which excludes half this difference, e(1). Thus f(1)/2 < e(1) < f(1). Since Σn^{-2} converges to $\pi^2/6 \sim 1.64$, we have that e(1) ~ 0.64. Alternatively we can use $\Sigma(n+0.43)^{-2} \sim 1$.

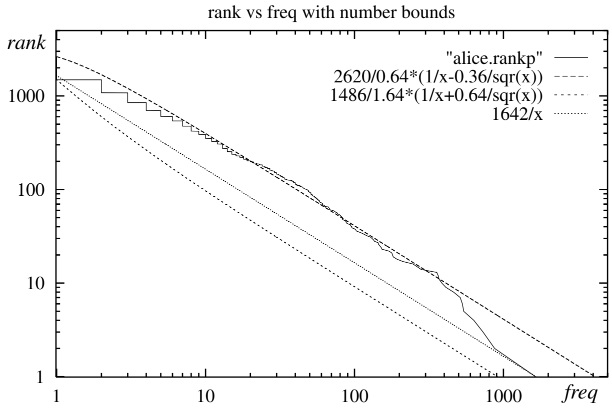

Figure 3. *Actual and expected range of rank vs freq*

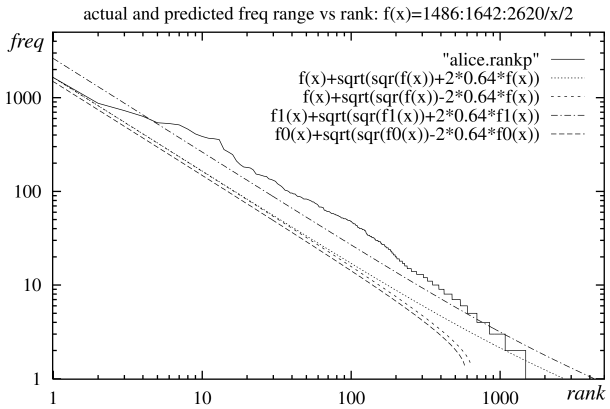

Figure 4. *Actual and expected range of freq vs rank*

ranks for frequency 1 (plus one) gives the number of words of frequency 1. This number, n_1, sets the maxima of the aggregated and unaggregated number curves ($2^{|\log n_1|}$ and n_1 resp.) that we saw in Figure 2. The number n_f at any given frequency f is represented as the stepsize, and the number of words that would have been expected to occur with this frequency is assumed to be of this order.

The error bound functions may easily be inverted to allow the more conventional plotting of frequency (and error estimates) as a function of rank, as in Figure 4. Here f(x) represents half the estimated frequency based on the highest frequency, and f0(x) and f1(x) represent those based directly on the upper and lower bounds on frequency 1. In this case we we use error e(1) = 0.64 but do not allow for error in the frequency 1 ranks themselves.

Note that Zipf associates the 'top-downward concavity' with 'informal colloquial speech' (1949, p82), an association which had been recognized by other researchers as early as 1936. The effect 'is not found in more formal material' and is attributed by Zipf to an expansion of the closed class vocabulary to include the personal pronouns (1949, p122). Both the phenomenon and the role of closed class words are of interest to us here.

2.3 Information Theory

Zipf's book on *Human Behaviour and the Principle of Least Effort* and Shannon's book on *The Mathematical Theory of Communication* were both published in 1949, and were developed totally independently, so it is interesting to look at how their concepts of efficiency interrelate. Interesting Crystal finds Zipf's explanations unsatisfactory and appeals to "a more conventional explanation in terms of probability theory" (1987, p87), by which he presumably means information theory, but he cites no literature in support of this claim.

Let us consider the probability distribution defined by dividing the frequency of each word by the length of our corpus, $p_r = f_r/L$ (possibly taken as a limit as our corpus increases indefinitely). An assumption that the lexicon can increase without bound is inconsistent with Zipf's Laws prediction that $p_r = C/r$ since summing over the distribution gives a non-convergent series, violating the constraint that the probabilities must sum to 1. Some prefer to hold onto this assumption and to seek a faster converging probability distribution for which the series converges to 1 (Brent, 1997). Such series include $1/r^2$, $1/r.\log^2 r$, $1/r.\log r.\log^2\log r$, ... all of which converge, whilst the series $1/r$, $1/r.\log r$, $1/r.\log r.\log\log r$, ... all fail to converge.

Interestingly, the terms of both sequences of series approach those of the series $\Sigma\, 2^{-L^*(r)}$ (where $L^*(x\text{-}1)$ is defined as $\log c + \log x + \log\log x + \ldots$) which does converge and is optimal in the sense that any monotonic decreasing distribution which satisfies our constraint must equal or exceed $L^*(x)-2k^*(x)$ infinitely often (Rissanen, 1989, p35), where $k^*(x)$ is the number of positive log terms in $L^*(x)$ excluding the constant. Note that the integrals **from** *r* (upto infinity) of the convergent series are $1/r$, $1/\log r$, $1/\log\log r$, ... whilst for the divergent series the integrals **upto** *r* are log r, log log r, ...

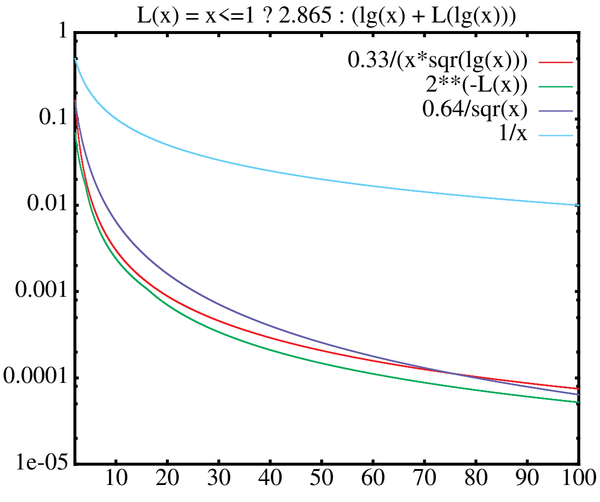

Figure 5. *Comparison of convering series with 1/x*

Now the information corresponding to probability p_r is $I_r = \log p_r$. For $p_r = 1/r$, optimal encoding of the information should take log *r* bits, and must at least specify the rank *r* which requires log *r* bits too, but a sequence of such codes could not be decoded. Adding a boolean 'finished' flag after each bit doubles the length, corresponding to squaring the probability, and allows decoding and convergence — which follows as soon as each code is a leaf in the decoding tree. Another way of delimiting is to specify a length using a more primitive scheme, then allowing minimum length encoding of the actual rank. In the extreme we specify lengths recursively till we flag we reach a length of 1, when we use our boolean flag — this corresponds to the near optimal $L^*(r)$, however for the range of lexicon size we need, one level of length encoding, $1/r.\log^2 r$, is sufficient and in Figures 5 and 6 the corresponding curves are scarcely separable. In Figure 5 we see that for the first 100 words this applies to $1/r^2$ too.

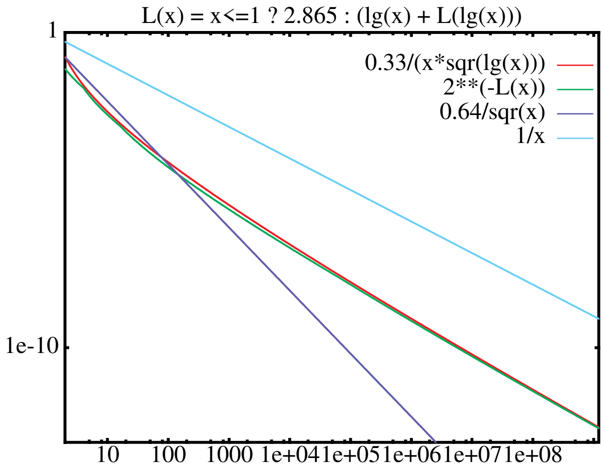

Figure 6. *Comparison of convering series with 1/x*

As discussed above, deviations from Zipf's Law are known, and the logscale which Zipf used actually hid considerable deviations for high values of either rank or frequency (and can amplify deviations for low values). We therefore now show the reciprocal of frequency against rank using a linear scale, and this in fact corresponds to a particular definition of the average interval between words. We show how this looks in Figure 7 for three different definitions of the average interval: 'interval' corresponding to dividing the corpus length by frequency (valid if imagine that the following text segments of this size have the same structure); 'initival' corresponding to treating the start of the corpus as the first reference point (valid if the interval to the first occurence is a good predictor of the interval between occurences); and 'intraval' corresponding to considering only the *f*-1 intervals between actual successive occurences. Note that the 'interval' which corresponds

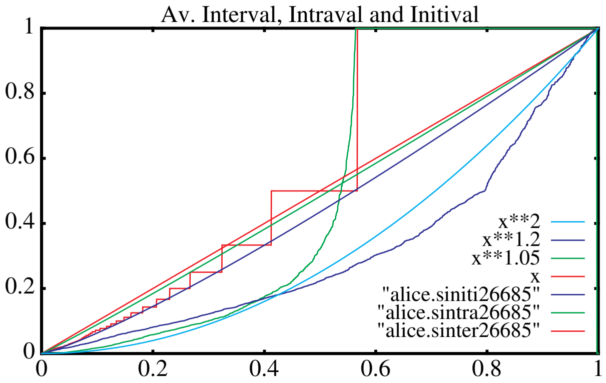

Figure 7. *Comparison of interval definitions and laws*

directly to a scaled reciprocal of the rank, is guaranteed to have a step shape as many words have the same frequency. This discretization is avoided by the other definitions. In Figure 7 we show these as a normalized intervals (divided by the corpus size, 26685) against normalized rank (divided by the lexicon size, 2620) along with lines corresponding to Zipf's Law, Mandelbrot's Modification (proposing an exponent of 1.05), the best fit power for our small test corpus (1.2) and the quadratic model of Brent (1997).

There are many details of Zipf's theory which we are unable to go into in the confines of this paper, but at this point we need to note two things. First, that Zipf claimed that claimed the reciprocal law applies only to an optimum sample size corresponding to a single cycle for the least frequent word, such that the maximum frequency and the maximum rank were equal to the intercepts of the line of best fit of gradient -1 (least squares in log scale). Of course as the corpus grows, new words of frequency 1 enter, so his principle is to select the size which gives gradient closest to -1 across a large number of samples of the corpus (which should be consistent as to genre, register, age of speaker, etc.) For most published literature this should correspond to around 10000 words (thus this is the size of the usual active lexicon) while for children around the time of starting school it is around 2000 words. Our corpus is around 2.5 times his optimum for literature, and as we are using Alice in Wonderland which is supposedly a children's book about a child, and is of informal character, perhaps he would probably have suggested using even smaller sample sizes. However, we are not content to characterize samples of an optimum size, and we are aiming to determine how the law should be adjusted to take into account sample size.

The second deviation from Zipf's practice is implicit in the mechanism for determining the optimum sample size as just explained: The correct line to draw is a line of best fit, minimizing the least squares error in the prediction of the log of frequency from the log of rank. This means are

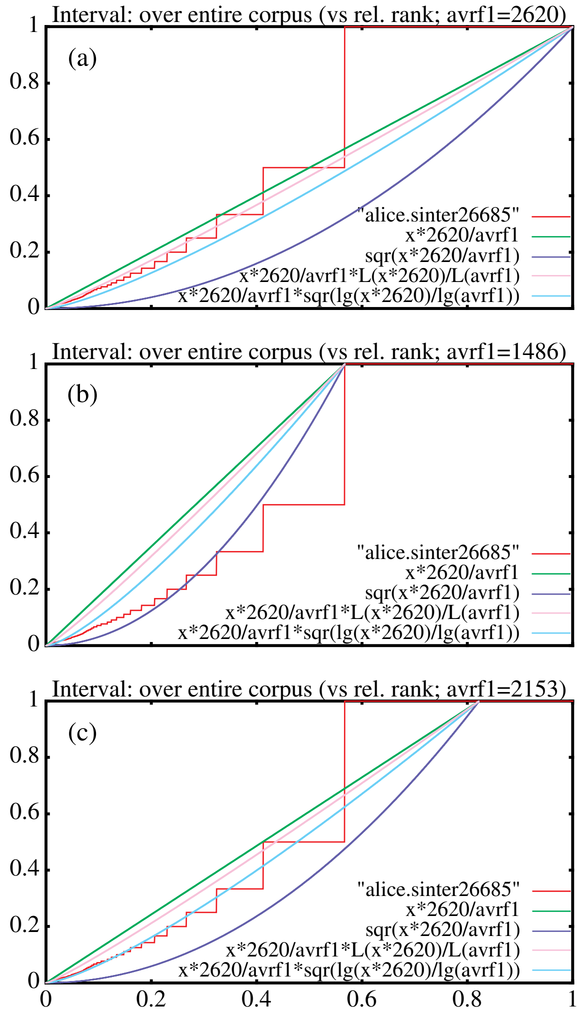

Figure 8. *Comparison of top step cut points*

line should pass through the middle of the steps in Figure 7 (as in Figure 2). This may also be viewed in another way. The midpoint of a step may be viewed as showing close to its correct frequency and rank, whilst words which should not occur an integral number of times in a sample of the selected size will be rounded to an integral frequency. They will necessarily occur more or less than the expected number of times. This problem does not apply to our alternate definitions of average intervals. Thus for 'intervals' we should fit the midpoint of the intervals, and in particular the midpoint of the frequency 1 step, for 'initivals' we have the full range of the corpus available and should fit the high point of the frequency 1 step, and for 'intravals' we should fit the low point since frequency 1 does not define an intraval and the value is arbitrarily set to a limit of 1.

Figures 8 and 9 show fits to these different points on the top step, and allow comparison with the convergent information theoretic functions we have discussed. In Figure 8c there is an evident log-squared bias.

From Figure 9 it will be observed that our new definitions of average interval have a totally different characteristic from the old. Not only have we eliminated the ‘steps’, but the resulting functions are clearly far from linear, and from the slowly convergent log-based series. Clearly for words that occur locally rather than globally in the corpus, these methods progressively add a bias relating to the location of a cluster of occurences — and better reflect the frequenct of clusters. This is something else Zipf has considered: for words of a particular average frequency (and thus interval), the number of intervals of a particular size also varies inversely with that interval size (1949, p42). In Zipf’s model this results from spreading the workload away from costly words.

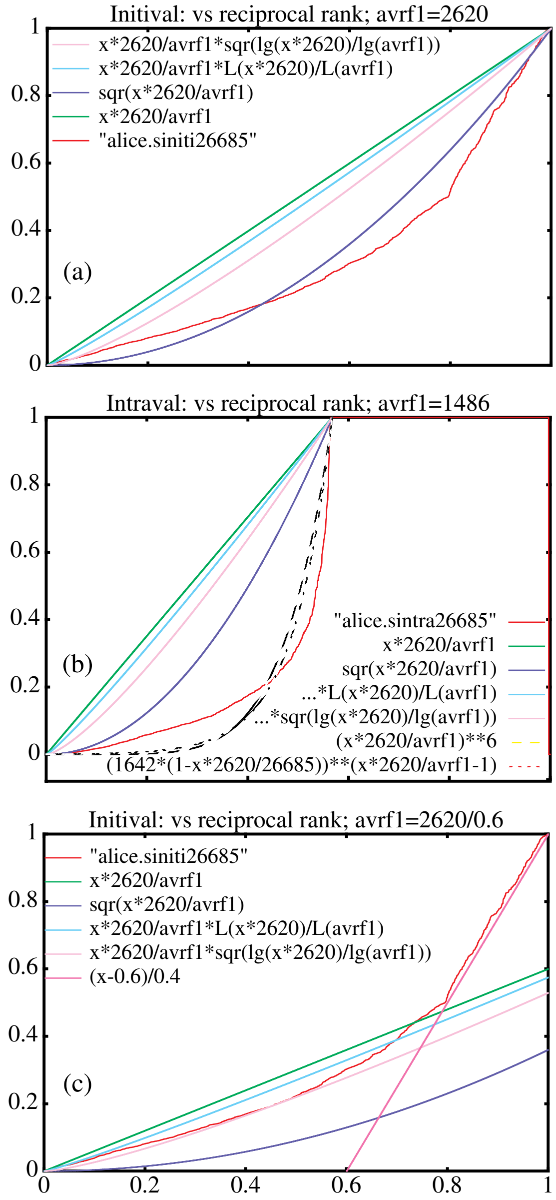

Figure 9. *Comparison of interval definitions*

Whilst it is possible to fit higher order polynomials and exponentials to to these curves, the fits are not good. In Figure 9c we have fitted a 6th order polynomial which is indistinguishable in the range from an exponential formulated in terms of the probability that a word does not occur. We do not see this as something where an explanation as a single distribution is appropriate, and later we view these as a joint distribution for open and closed class words, and in Figure 9c we show that the bends in the curves look more like transitions between two distinct distributions obeying different parameterizations of some form of Zipf’s law.

2.4 Psychological Predictions

If we believe Zipf’s law in its standard form, and scale frequencies to probabilities for a finite lexicon, then information theory suggest that the length of words should look like the log of frequency, and the access time for words should follow the log of word frequency. Assuming that the lexicon is unbounded, then information theory suggest that the length of words should be L(x) or, less optimally, log x + 2 log log x.

Zipf went further and predicted that the older words would be the more frequent, both in an etymological and a psycholinguistic sense, and performed experiments to demonstrate the law in relation to the etymology of English, as well as performing some analyses of children’s speech which were also consistent with his model. However, his experiments on length did not quantiatively demonstrate what relationship was achieved, and he was expecting a negative power relationship again. Moreover, he did not perform any experiments to check the validity that access time would reflect an inverse relationship, and expected that length (m_r) and access time (d_r) would be proportional to $r^{0.5}$.

Studies of latencies in various linguistics task have, however, been extensively studied by psychologists, and although the interpretation of the results is controversial, and the results are more qualitative than quantitative, considerable evidence exists to support a logarithmic access time, and have been the basis for one of the most influential models of word recognition, the Logogen model (Morton, 1969). There is also Event-Related Potential evidence from EEG studies, but these results are even less precise and we ignore them here (although we have undertaken some ERP experiments ourselves and hope to further elucidate certain factors in this way).

Looking more closely at the experimental data, we find, just as with the frequency data, that there are strong contextual effects (Becker, 1979) which tend to be additive, particularly for lew-frequency words. An additional confusion factor is that subjective measures of familiarity which actually can better predict access time than more objective frequency measures (Gernsbacher, 1984). As Zipf also knew, the number of distinct also plays a role, and Zipf himself found that the number of meanings decreased with the square root of frequency (1949, p75). Other reported confusion factors include concreteness, level of education, age, age of acquisition and word length. Also there is a correlation between word frequency and the signal to noise ratio which may be tolerated by a word, as well as the fixation time in reading a word. The mode of presentation and the method of testing may also influence the relationships found, as can even such factors as stress pattern and syntactic role. Thus the role of frequency as a primary determiner of access time is highly controversial although the relationship itself is well accepted (Balota & Chumbley, 1984 & 1990; Monsell, Doyle & Haggard, 1989).

As we can see from Figure 10, the length data does support a logarithmic relationship, notwithstanding that a log-squared bias was observed in Figure 8. However the square root of the logarithm is best for this data. Nevertheless the information theoretic optimum is approximated for this corpus and we would further predict a similar function for the access time for words.

Although the age of acquisition and length both show stronger correlation with latency than frequency in naming tasks, and this has been cited as evidence against word frequency having a significant effect on access time (Morrison, Ellis and Quinlan, 1992), this observation supports the predictions we made on the basis of Zipf's Law and Information Theory. Even if the age of acquisition and phoneme length fully determine access time, the fact of correlation between frequency and latency is not disturbed, and we can hypothesize that high frequency leads to early and frequent exposure to, and thus learning of, a word; and furthermore, that the early learning, in combination with constant refreshment, maintains the word at a relatively greater level of accessibility than less frequent words. Similarly, we predicted a strong correlation between length and latency

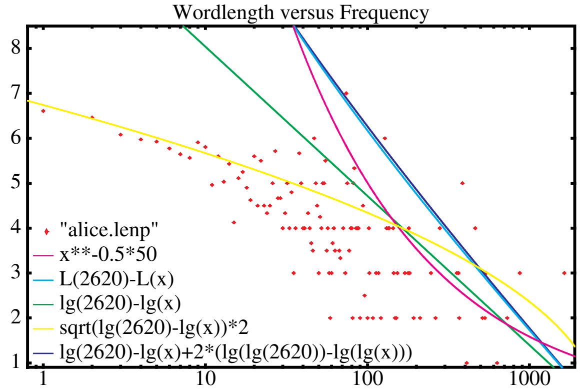

Figure 10. *Wordlength versus Frequency*

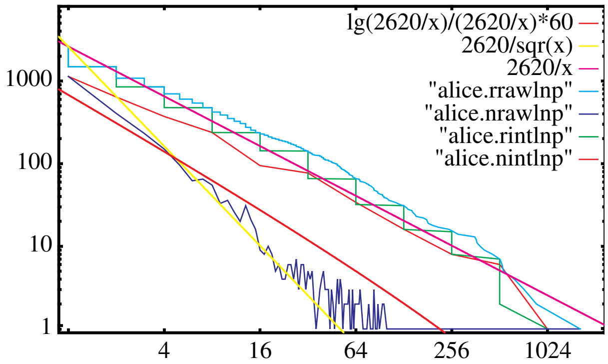

Figure 11. *Approximation to numbers given frequency*

on the basis that the logarithmic relationship is required for optimality in each case. The results do however seem to contradict Zipf's prediction that length, m, and access time, d, would be proportional to $f^{-0.5}$, although it must be emphasized that the correlations are far from perfect and the precise trends cannot be distinguished to any great degree of accuracy. Our own data in Firure 10 suggests that the relation is actually $\log^{0.5} f$, and the curve for $f^{-0.5}$ moves right away from the data at both extremes. However we have no accurate information for access time and we will simply note that both length and access time have a generally logarithmic relationship with frequency, and that $L(1/f)$ is also a better fit than $f^{-0.5}$.

Nonetheless, this has implications for the principle of least effort in that the Zipf's relationship for work,, $w = f * m * d = f / \log^2 f = 1/r.\log^2 r$, is not constant, and indeed Information Theory says we should actually do less work for more frequent items. Moreover, this relationship for work now obeys a law consistent with near optimality in an unbounded lexicon model. We could moreover replace Zipf's law by $f = 1/r.\log^2 r$, consistent with the improved empirical fit of Figure 8c.

This gives us a new relation for the number of words around each frequency. For Zipf's Law, $f = c/r$ gave us $n = c/f^2 = r/f$. For our new version, $f = c/r.\log^2 r$ gives us $n = \log r/bf$ (where our new constant, b, depends on the base of our log and is given by $b = 2 \log e$). Thus for Zipf's Law, the number expected for each frequency is the corresponding fraction of the number of words with higher frequency, but our new numbers grow more slowly, the frequency specifying a fraction of the log of the number of words with higher frequency. Substituing an overestimate for the rank using Zipf's Law, in Figure 11 we approximate the number at frequency f by $n = \log(c/f)/bf$ (which is a bit steeper than the correct inverse would be) and is at least as good a fit as c/f^2 and indeed better reflects the distribution of the sparse ranks where the expected number of words for a frequency is less than one. The corresponding optimal length function, again with rank overestimated using Zipf's law (and hence also too steep), is shown in Figure 10.

From the perspective of our own research in Natural Language Learning, the most significant results from these psychological explorations of access time are those which suggest that open class words show a more significant effect than closed class words, and have distinct roles and mechanisms in the early stages of morphological processing (Segui, Mehler, Frauenfelder and Morton, 1982; Matthei and Kean, 1989).

3. Corpus Characteristics and Sample Size

One of Zipf's claims was that a given text had particular characteristics which included a characteristic optimum sample size and lexicon size. He chose the optimum sample size to be the one with the best fit to Zipf's Law, and this implicitly specified a lexicon. However, as the sample is increased above that point, new word types continue to enter the lexicon. He furthermore notes that informal colloquial speech gets a hump in the first 150 words, one which has generally been associated with increased usage of the personal pronouns. We further noted that there is evidence (Matthei & Kean, 1989; Segui et al., 1982) that open and closed class words are treated differently, and it could also be assumed that there is a primary subject, and hence lexicon, for any specific work, as well as secondary or incidental topics. As an average of independent topics, we might expect the law to re-emerge, but the closed class (including generics) cannot increase in size as the lexicon does.

One of our motivations for undertaking this study was the observation during our language learning experiments that as corpus size increased, Zipf's Law tended to be increasingly invalidated, with the curvature increasing consistent with a move to a continually higher exponents. While it would be predicted for the highest ranks of each step to exhibit a quadratic component, due to the size of the step reflecting the number of words with that frequency, the tendency affected the lowest and median ranks as well.

To investigate this analytically, we assume that we have two samples each of which has a lexicon of size *N* which includes a common vocabulary of *pN*, and that each sample obeys Zipf's Law. We further assume that the common vocabulary includes the closed class words, and more generally the most frequent words in the language and relating to the topic of the corpus. If the *pN* words were the most frequent words of the individual samples, and were exhibiting their characteristic frequencies, they would exhibit exactly double the frequency for the same rank, thus retaining the same slope in a log-log plot. The converse is true for the remainder of the lexicon: the words will all the same frequency as in the smaller samples, but their ranks will have increased in proportion to their distance above the

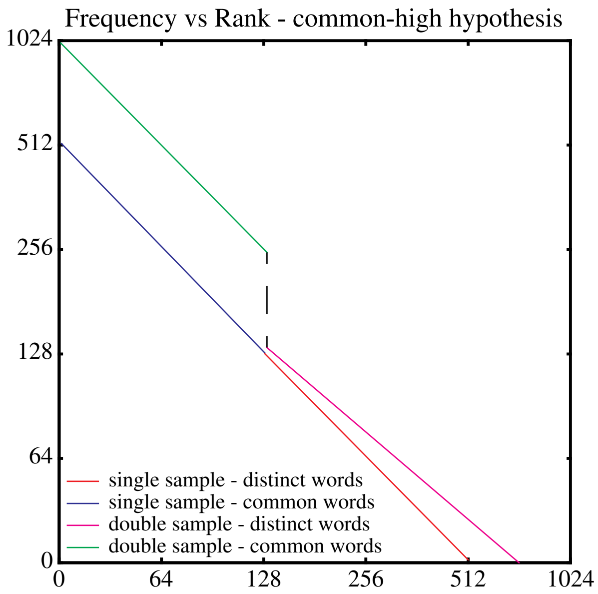

Figure 12. *Model with high freq. common vocabulary*

common set. These word types will thus exhibit a reduced slope, and there will be a jump and a sudden discontinuity in slope between the common and the distinct words as in Figure 12 In general, whereever new (or displaced) vocabulary enters the picture, there will be reduced slope — rank will increase without much change in frequency. Similarly, where disused vocabulary is displaced, rank will decrease without much change in frequency, giving rise to increased slope. In fact there is evidence of a jump and discontinuity which occur at a rank logarithmic in the size of the lexicon.

In another model, we can imagine replacing a pair of equally likely synonyms by one member of the pair, and note that this will double it's frequency and halve its rank. This is consistent with Zipf's Law, although in between the old and new positions, ranks will increase without an increase in frequency, and after the old position, ranks will decrease without a decrease in frequency. This will produce the kind of bulge Zipf identified with informal text. Shifting words can thus cause discontinuities too.

At this stage, it may be worth saying a few words about the corpus used throughout this paper, *Alice's Adventures in Wonderland* (Carroll,1865) is an edited collection of children's stories, originally delivered verbally, and culminating at a Picnic in 1862 when Alice going down the rabbit hole provided the framework and cast the spell which eventually led to publication. As a series of adventures, there are some characters in common, some of whom recur, but entire vocabularies are limited to a single chapter. We used the Millennium Fulcrum Edition 2.9, available through the Gutenberg Project, which is significant since this attribution occurs at the beginning of the book and affects our analysis.

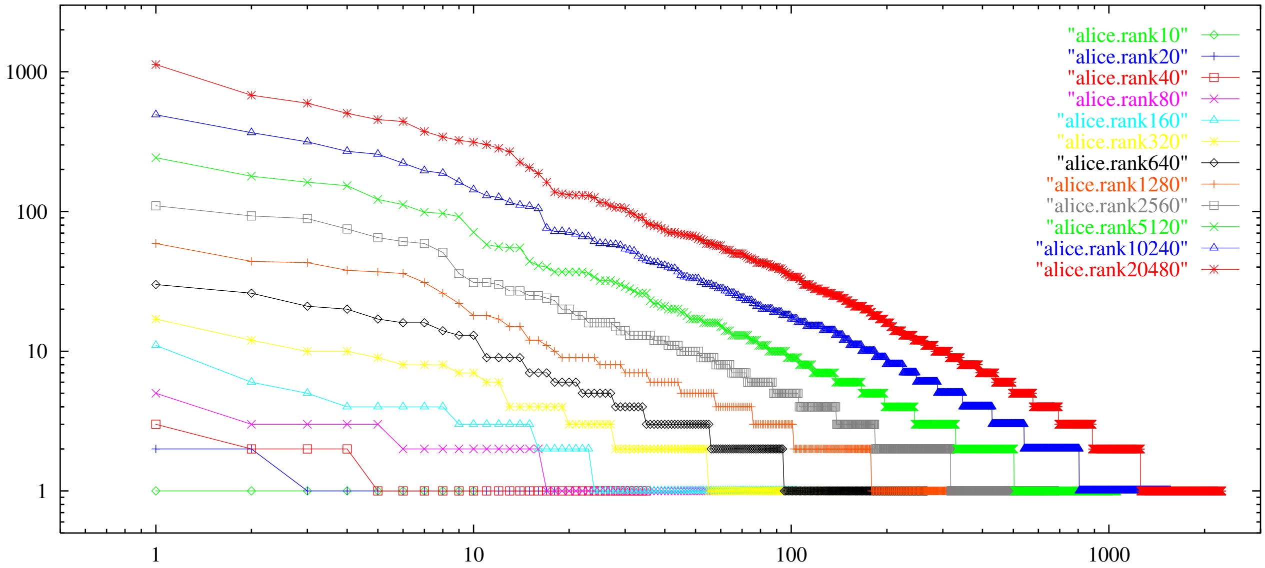

Figure 13. *Frequency against Rank as corpus doubles in size — Alice's Adventures in Wonderland*

Having made the above predictions about the shape of the curves under both extreme conditions and incremental change, we then proceeded to an analysis of Alice produced by taking successive prefixes of the book. These are shown in Figure 13, where each prefix is twice the size of its predecessor. Note the discontinuities that start in our first doubling, where closed class words start to sift above above the non-recurring words of the title page, some of which only occur once in the whole volume. This shift starts in the first order of ranks and is visible well into the second order of ranks. The dips which appear and disappear around rank 10 in the larger segments are due to the competition between the words 'said', 'in' and 'i', and the name 'Alice' as the story alternately focuses on her involvement and scenes involving other characters, and changes its balance between narrative, soliloquy, and reported speech. The beginning of the second order ranks marks the transition between closed class words and focal words. The first 20 words are: *the, and, to, a, she, it, of said, I, Alice, in, you, was, that, as, her, at, on, all, with*, and the remainder of the first 100 are all closed class words (plus narrative devices like *think* and *looked*) or characters — with the single exception of the word *head* (which is closely linked to the one character who wasn't in danger of losing hers).

Thus in this range we see a number of discontinuities as words move into and out of the focal range, and two different slopes corresonding to the closed class and focal words, and the open class words.

We have carried out a similar and more extensive study on the Bible in four languages. It also has the character of sequences of stories focusing on different people and events, and is also largely and edited version of verbal accounts. Using multiple versions retains the thematic biases, so that if the artifacts we are observing are primarily thematic, we should see similar artificacts in similar places. We should separately be able to see the effects of language and translation style. Although some translations are targeted at a more popular level and use less technical vocabularly, we have selected four traditional translations three of which use reasonably contemporary language (the versions are KJV, RSV, Louis Segond, Elberfelder).

Whereas Alice allowed us to double only 11 times, the Bible allows us to double 16 times. We have however for consistency and convenience kept with the smaller Alice corpus for the graphs shown here (the equivalent of Figure 13 for each version of the bible is about 1Meg of PostScript). In Figure 14 we show the results for French for the last seven samples, this being the one where the discontinuities were least pronounce. In each case there were one or two deep drops around rank 10, followed by a steepening of slope.

4. Conclusions

At this stage, only tentative conclusions can be made from this preliminary studies, although further investigations are being undertaken using larger corpora in multiple languages. Zipf's theory requires effort to be constant independent of frequency, however Information Theory and Psychological experiments both indicate that this ought not to be the case, and that it in fact decreases in a way consistent with an optimal strategy for an unbounded lexicon. We have not been able to establish the validity of an optimum sample size for a particular corpus, genre or lexicon, but observe that new words tend to enter faster than they repeat, as evidenced by the fact that the number of words of frequency 1 tends to increase

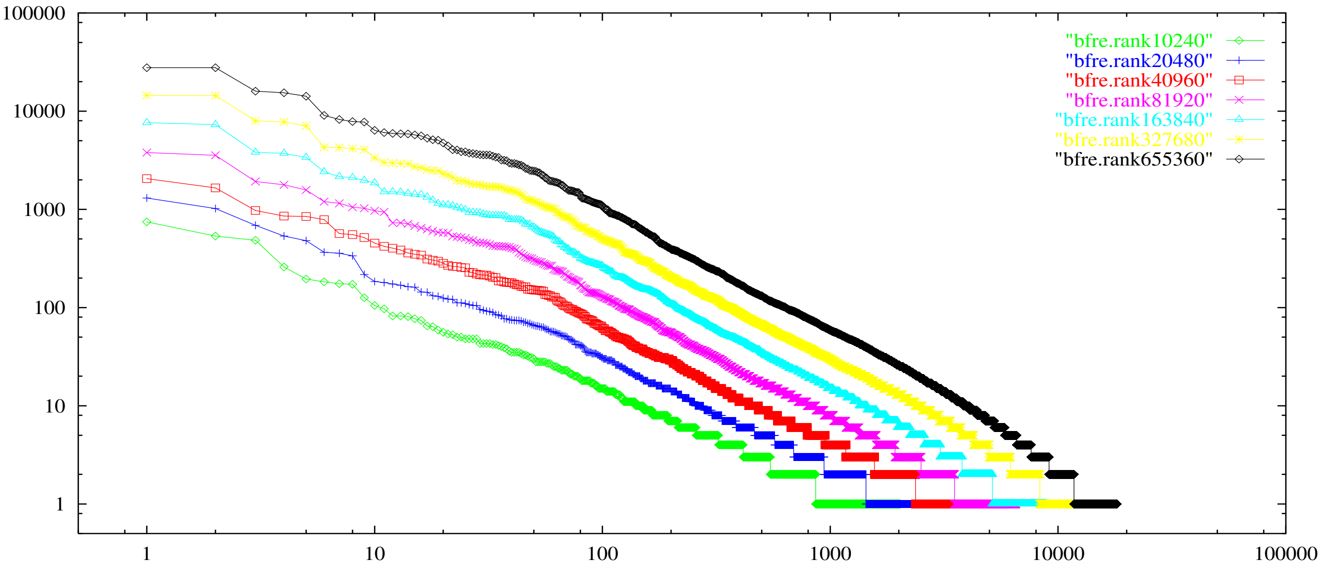

Figure 14. *Frequency against Rank as corpus doubles in size — French Bible (Louis Segond)*

as the size of sample increase. Given that language is productive, and an unbounded lexicon model has been indicated (or at least possible) in each of our experiments, this trend may well continue indefinitely, although it does seem to slow as the sample is increased (even though we increase by doubling).

5. Acknowledgements

All plots were made with GnuPlot and reformatted in FrameMaker. Aspects of this work were undertaken while a guest researcher at ENSSAT in Lannion, France and at the University of Antwerp, Belgium, with support from IRISA and CLIF respectively.

6. References

Balota D.A. and Chumbley, J.I. (1984) Are Lexical Decisions a Good Measure of Lexical Access? The Role of Word Frequency in the Neglected Decision Stage, *J.Exp. Psych: Hum.Perc. & Perf.* **10**#3:340-357

Balota D.A. and Chumbley, J.I. (1990) Where are the Effects of Frequency in Visual Word Recognition Tasks? *J. Exp. Psych: General* **119**#2:231-237

Becker, C.A. (1979) Semantic Context and Word Frequency Effects in Visual Word Recognition, *J. Exp. Psychology: Human Perc. & Perf.* **5**#2:252-259

Brent, M.R. (1997). Toward a Unified Model of Lexical Acquisition and Lexical Access. *Journal of Psycholinguistic Research* **26**:363-375.

Carroll, L. (1865). Alice's Adventures in Wonderland. The Millennium Fulcrum Edition 2.9, Gutenberg Project.

Crystal, D. (1987). *The Cambridge Encycolpedia of Language*, CUP

Entwisle, J. and Powers, D.M.W. (1998). The Present Use of Statistics in the Evaluation of NLP Parsers. *submitted*

Finch, S. (1993) *Finding Structure in Language.* Ph.D Dissertation, U. Edinburgh

Gernsbacher, M.A. (1984) Resolving 20 years of Inconsistent Interactions between Lexical Familiarity and Orthography, Concreteness, and Polysemy. *J. Exp. Psych: General* **113**#2:256-281

Köhler, R. and Rieger, B.B., eds (1991) *Contributions to Quantitative Linguistics*, Kluwer

Matthei, E.H. and Kean, M-L. (1989) Postaccess Processes in the Open vs Closed Class Distinction. *Brain and Language* **36**: 163-180

Monsell, S. Doyle, M.C., and Haggard, P.N. (1989) Effects of Frequency on Visual Word Recognition Tasks: Where are they? *J. Exp. Psych: General* **118**#1:43-71

Morton, J. (1969) Interaction of information in word recognition. *Psychological Review* **76**:165-178

Powers, D.M.W. (1995). Parallel Unification: Practical Complexity, *Australasian Computer Architecture Workshop,* Adelaide

Powers, D.M.W. (1996). Learning and Application of Differential Grammars. *CoNLL97: ACL/SigNLL Workshop on Computational Natural Language Learning*, Madrid

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* Singapore:World Scientific

Samuelsson, C. (1996). *Relating Turing's Formula and Zipf's Law*, WVLC'96

Segui, J., Mehler, J., Frauenfelder, U. and Morton, J. (1982). The word frequency effect and lexical access, Neuropsychologia 20:6 615-627

Shannon, C.E. and Weaver, W. (1949) *The Mathematical Theory of Communication.* Urbana: U. Illinois Press

Stanaitis, O.E. (1967). *An introduction to sequences, series, and improper integrals*. San Franc:Holden-Day

Steele, R. and Powers, D.M.W. (1998) Evolution and Evaluation of Document Retrieval Queries. *submitted*

Zipf, G.K. (1949) *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* AW