Glottometrics 4
2002

To Honor G. K. Zipf

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 4 (2002), Lüdenscheid: RAM-Verlag, 2002. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 2 (2002) **ISSN 2625-8226**

Contents

Balasubrahmanyan, V.K., Naranan, S.
Algorithmic information, complexity and Zipf´s law 1

Roelcke, Thorsten
Efficiency of communication. A new concept of language economy 27

Schroeder, Manfred
Power laws: from *Alvarez* to *Zipf* 39

Wheeler, Eric, S.
Zipf´s law and why it works everywhere 45

Debowski, Lukasz
Zipf´s law against the text size: a half-rational model 49

Kornai, András
How many words are there? 61

Montemurro, Marcelo A., Zanette, D.
New perspectives on Zip´s law in linguistics:
from single texts to large corpora phenomenology and models 87

Glottometrics 4, 2002, 1-26 *To honor G.K. Zipf*

Algorithmic information, complexity and Zipf´s law

V. K. Balasubrahmanyan
S. Naranan[1]

Abstract. Zipf's law of word frequencies for language discourses is established with statistical rigor. Data show a departure from Zipf's power law term at low frequencies. This is accounted by a modifying exponential term. Both arise naturally in a model for word frequencies based on Information Theory, algorithmic coding of a text preserving the symbol sequence, concepts from quantum statistical physics and computer science and extremum principles. The Optimum Meaning Preserving Code (OMPC) of the discourse is realized when word frequencies follow the Modified Power Law (MPL). The model predicts a variant of the MPL for the relative frequencies of a small fixed set of symbols such as letters, phonemes and grammatical words. The OMPC can be viewed as containing orderly and random parts. This leads us to a quantitative definition of complexity of a string (C) that tends to 0 for the extremes of 'all order' and 'all random' but is a maximum ($C = 1$) for a mixture of both (Gell-Mann). It is found that natural languages have maximum complexity. The uniqueness of Zipf's power law index ($\gamma = 2$) is shown to arise in four different ways, one of which depends on scale invariance characteristic of fractal structures. It is argued that random text models are unsuitable for natural languages. It is speculated that a drastic change in symbol frequency distribution starting from phrases is related to emergence of meaning and coherence of a discourse.

Keywords: Zipf´s law, information, entropy, complexity, adaptive systems, lexicon, power law

1. Introduction

We are grateful for this opportunity to pay our tributes to Professor G. K. Zipf (1902-1950), who may be considered to be one of the founders of Quantitative Linguistics as a serious multi-disciplinary subject. The recent interactions with Information Theory, Statistical Mechanics, Complex system studies and the general concepts of physics, have enriched Quantitative Linguistics, which started as a modest word counting exercise in the early twentieth century. These efforts which may appear prosaic were undertaken by the French court stenographer Estoup (1916), the famous librarian Geoffrey Dewey (1923) and the psychologist Edward L. Thorndike (1932) among others. Their main objectives were to develop systems of shorthand and a basic defining vocabulary for the English language. Later, cryptologists ─ the code breakers ─ were also involved in such efforts. Almost all these were confined to the English language.

Zipf, however, undertook word counting in several European languages, Sanskrit, Chinese and others and formulated important generalizations grouped under the name 'Zipf's law' (Zipf 1935). It would appear that Zipf had intuitively felt that words in different languages have to obey mathematical laws and significant regularities, as language is a species-specific unique characteristic of the human mind. The relationship between Chomsky's

[1] Address correspondence to: S. Naranan, 20 A/3, 2nd Cross Street, Jayaramnagar, Chennai 600041, India. E-mail: snaranan@vsnl.net

powerful formulations of "hard-wiring" of the human brain for language ability and concepts of universal grammar are seen in a latent form in Zipf's early work.

Zipf observed that word occurrence frequencies displayed a remarkable stability or universality in their relative proportions, across the entire spectrum of languages, authors and genres. This is quantified by Zipf's law for language texts:

$$p(r) = A / r \tag{1}$$

Here $p(r)$ is the frequency of occurrence of a word of rank r and A is a constant approximately equal to 0.1 N, N being the total number of words (word tokens) in the text. The most frequent word has the rank $r = 1$.

Zipf intuitively guessed that underlying this law, there might be deep reasons connected with the language faculty of mankind. He formulated his *"principle of least effort"* as the basis for Zipf's law. The human brain prefers shorter words to longer words as effort needed to use them is less; so shorter words should occur more frequently than longer words. Zipf propounded the principle as a rationale for equation (1) which is a prototype of power law distributions of the form $p(x) = A\, x^{-a}$ (a is a constant). Zipf extended his work to other areas such as demography, economics and geography establishing equation (1) as a universal law pervading several areas of behavioral science – in an elaborate, eclectic and encyclopedic work *"Human Behavior and the Principle of Least Effort – an Introduction to Human Ecology"* (Zipf 1949). It was published just a year before his untimely demise at 48 years of age.

It is worth noting that the principle of least effort is the underlying basic principle of Shannon's Information Theory (Shannon 1948; Shannon and Weaver 1949). One of the first applications of Shannon's theory was to language texts and their encoding for efficient communication. As we shall see later, the optimum minimum bit code length per symbol is the Shannon entropy. It is realized by using short bit length codes for more frequent symbols and longer bit lengths for the least frequent symbols. One can now admire Zipf's remarkable insight, since Information Theory and coding concepts were generally not well known at that time. Even today there exists no universal model to explain equation (1). There exist a whole variety of models to account for the ubiquitous power law statistical distributions, which are often referred to as 'Zipfian' in recognition of Zipf's pioneering work.

Much of Zipf's work is not easily accessible. So, the contribution of C. Prün and R. Zipf to this memorial volume is an invaluable one. It has an extensive bibliography, and little known facts and facets of Zipf's private and professional lives (Prün and Zipf 2002).

In this article we mainly review some of our recent work on Zipf's law with connections to Shannon's Information Theory, Algorithmic Information Theory (Kolmogorov complexity), statistical mechanics, classical (Boltzmann, Gibbs) and quantum; complex system studies (Gell-Mann, Bennett, Zurek), Fractals and scale invariance (Mandelbrot).

One of our primary aims was to test the "goodness of fit" of Zipf's law (equation 1) to actual linguistic data, using rigorous statistical tests. It is well recognized that Zipf himself did not attempt such tests to validate his laws.

Equation (1) is the commonly used form of Zipf's law; however it is appropriate only for words occurring with high frequency (low rank r). For words of low frequency – rarely occurring words, many of them occurring just a few times in the entire text – another version of Zipf's law is more apt:

$$W(k) = B k^{-2} \tag{2}$$

$W(k)$ is the number of *different* words (word types, lexical units) each occurring exactly k times in the text. B is a constant that depends on the size of the text (N). Although equation (1) is the dominant theme in Zipf's work, there is a reference to equation (2) too in Zipf

(1935). In any typical data set of word frequencies, both forms of the law are required for proper statistical analysis of the data (Balasubrahmanyan & Naranan 1996). The low rank, high frequency words belong to the *r-domain* (equation 1) and the low frequency, high rank words belong to the *k-domain* (equation 2). It is interesting that the two domains account for nearly equal proportions of the text. The equivalence of equations (1) and (2) might have been suspected by Zipf. Their equivalence is demonstrated by Naranan & Balasubrahmanyan (1992a). See also Naranan & Balasubrahmanyan (1998).

When compared with word frequency data both equations (1) and (2) show significant deviations at low and high values of the independent variable implying departures from pure power laws. In the *r*-domain a better description of data is given by Mandelbrot (1953, 1966)

$$p(r) = A\,(r + r_0)^{-B}\,, \tag{3}$$

r_o and B are constants. Similarly for the *k*-domain Naranan & Balasubrahmanyan (1992a, b) proposed a modification of equation (2):

$$W(k) = Be^{-\mu/k}k^{-\gamma}\,, \tag{4}$$

B, μ and γ are constants. This is referred to as the Modified Power Law (MPL) and is in very good agreement with data. $\gamma \approx 2.0$, the canonical value of the index in equation (2).

The two-domain structure arose out of the need for meaningful statistical tests of validity of Zipf's law; however it was found to have some important linguistic significance too (Balasubrahmanyan & Naranan 1996). It was noticed that the *r*-domain contained mostly grammatical words (articles, prepositions and conjunctions such as *the, of, and, or,...*) whereas the *k*-domain had largely the semantic or 'content' words or C-words which are nouns, verbs, adjectives, adverbs etc. This delineation of the strong correlation between *r*-domain and grammatical words (also called "service" or S-words) and between *k*-domain and C-words, became possible only on examining the data of Dewey (1923) which listed the actual words that occurred with different frequencies. This is a departure from the usual situation where the actual words are ignored since only their occurrence frequencies or ranks are relevant, as in equations (1) to (4).

The S-words constitute a small constant set (~ 71 in English); yet they account for 40-45% of the total number of words (text length in words, word forms). They do not change with time or from text to text. In contrast, C-words form a large variable set (a few million words typically in English) growing in time. Every text has a different subset of C-words, whose number depends on the size of the text (N). The S-words therefore resemble two other symbol sets in language: (1) the set of letters in the alphabet and (2) the set of speech sounds, the phonemes. A phoneme is a speech sound, the smallest unit of speech that distinguishes different utterances in a language. For example, the English language has a 26-letter alphabet and about 32 phonemes.

Symbol frequencies can be analyzed at the level of S-words, alphabets and phonemes, all with a fixed small number of symbols (V). The rank frequencies of equation (1) are the most appropriate: $p(r)$ (r = 1,2,3....V). There is no *k*-domain for these sets. Naranan & Balasubrahmanyan (1993) studied phoneme and alphabet rank frequency distributions and Balasubrahmanyan & Naranan (1996) examined the S-word frequencies. For all the three sets of symbols, neither equation (1) nor equation (3) conforms to data. Instead, a slightly modified version of MPL (equation 4) fits the data satisfactorily:

$$p(r) = \sum_{i=r}^{V} D\, e^{-\nu/i}\, i^{-\delta}, \tag{5}$$

D, ν, δ are constants. $p(r)$, $r = 1,2 \ldots V$ are the frequencies of the V symbols. Equation (5) is called the *Cumulative Modified Power Law* (CMPL). We present some data and statistical tests of Zipf's law and its variants mentioned above for words, S-words and phonemes in section 2.

What is the genesis of Zipf's law? Is there a viable model to account for the law and its variants [equations (1) to (5)]? There is no clear-cut 'yes' or 'no' answer. There exists a variety of models for power law distributions in general in behavioral, physical and life sciences. Some of these will equally well apply to linguistics. They can be broadly classified as stochastic models and information theoretic models. The stochastic models generally invoke randomly generated texts, which are strings of letters of an alphabet. We believe that they are not appropriate for natural language texts, which have strong correlations between different words.

Mandelbrot, in his PhD thesis tried to relate Zipf's law to Information Theory. In this pioneering work (Mandelbrot 1953), words are regarded as random statistical collections of letters and a symbol for space. Space is the symbol that delineates the words. So the model is applicable to "monkey languages", an allusion to monkeys at the typewriter to generate random texts.

Naranan & Balasubrahmanyan (1992a, b) have developed an information theoretic model for Zipf's law based on words as the basic symbols (unlike letters of alphabet). The focus of the model is on the cost of encoding a text or a string of symbols. Shannon entropy H_s is expressed as bits per symbol; it is the theoretical minimum bit length for efficient encoding, using variable code lengths for different symbols. It depends only on the relative proportions with which different symbols occur and not on the sequence in which they occur. However, a distinguishing feature of language is the sequence in which the words occur, since the sequence carries the 'meaning' of the text.

Kolmogorov's Algorithmic entropy – also called Algorithmic complexity (Kolmogorov 1960) – depends on symbol sequence. Random strings have higher algorithmic entropy than strings that exhibit some regularity, although both may have the same Shannon entropy. So Kolmogorov encoding is more relevant than Shannon entropy for language texts.

Language texts, although highly structured, with strong correlations between words, do exhibit some random features besides rule-based behavior. Naranan & Balasubrahmanyan (1992a, b; 1993, 2000) and Balasubrahmanyan & Naranan (1996, 2000) have analyzed these questions and have formulated new entropies – *degenerate entropy* H_d, *algorithmic entropy* H_a – and the *Optimum Meaning Preserving Code (OMPC),* which relate linguistic discourses to complex adaptive systems. They have also proposed a higher order entropy with features similar to the '*effective complexity'* of Gell-Mann (1994) and Bennett (1990) for complex adaptive systems. Unlike algorithmic entropy (Kolmogorov), Gell-Mann complexity is a maximum for systems (strings) with a mixture of random and regular features. Applied to language texts, effective complexity or simply *complexity* C is nearly maximal (≈ 1) for natural languages which have an optimal combination of rule-based order and randomness.

Power laws, such as Zipf's law appear as a natural consequence of self-similarity or scale invariance exemplified in fractal structures (Mandelbrot 1977, 1983). Since language texts exhibit Zipf's law on all size scales, this model is relevant for us. We briefly discuss scale invariance in section 5. It is conjectured that the universal value of the Zipf index ($\gamma = 2$) may be a consequence of a particular type of scale invariance. Information theory, per se, does not fix the value of γ, but other arguments based on Gell-Mann complexity, scale invariance etc. may account for $\gamma = 2.0$.

The problem of hierarchies in language – the words, phrases, sentences etc. – and the

'emergence' of meaning in a collection of words appears to be related to the privatization of the public (lexical) meanings of words which progressively reach an asymptotic state resembling a phase transition in physical systems (section 6). The emergence of meaning appears to be related to the uniqueness of phrases, sentences etc. in a text which shares words with a public lexicon.

2.0. Zipf´s law and its variants

The problem of subjecting word frequency data to proper statistical tests is best illustrated with an example. We consider the data of Eldridge (1911) for American Newspaper English presented in Zipf (1935). N, the total number of words (word tokens) is 43989 and V, the number of different words (word types) is 6001. Let $W(k)$ represent the number of word types occurring exactly k times and let $p(r)$ be the occurrence frequency of a word of rank r. From the table of word frequencies ($1 \leq k \leq k_m$)

$$W(1) = 2976 \qquad W(2) = 1079 \qquad W(3) = 516 \ldots$$

W decreases rapidly as k increases. For $k < 54$, almost every value of k occurs. But for $k \geq 54$ there are many gaps in k values and when a k does occur $W(k) = 1$ in almost all cases. In other words, for $k \geq 54$, $W(k) = 0$ or 1. Indeed there are only 77 words in the interval $54 \leq k \leq 4291$ ($= k_m$). Clearly $k \approx 54$ (say k_o) is the demarcating frequency for the two domains.

The set of word frequencies with $k < k_o$ and $W(k) > 1$ is called the frequency domain or the k-domain. The set with $k_o \leq k \leq k_m$ and $W(k) = 0$ or 1 is called the rank domain or the r-domain since every word in this domain can be assigned unique and different ranks. The most frequently occurring word ($k = k_m$) has rank 1 [$p(1) = k_m = 4291$]. When more than one word has the same frequency [$W(k) > 1$], the words are given consecutive ranks arbitrarily. Consider assigning ranks to words in the k-domain. For example the 516 different words each of which occurs 3 times will be assigned 516 consecutive ranks ($1946 > r \geq 1430$). Further these rank values will depend on $W(4)$, $W(5)$, $W(6)$ … So the rank values in k-domain are not statistically independent. This is the rationale for using $W(k)$ in the k-domain ($k < k_o$) and $p(r)$ in the r-domain ($k \geq k_o$) for statistical tests, such as the 'chi-squared' test for goodness of fit of a hypothetical distribution to the data.

2.1. C-word frequencies

We have confined our tests to $W(k)$ *vs.* k for $k < k_o$, i.e. the k-domain which accounts for ≈99 % of the word types. Data for 10 different texts are presented by Naranan & Balasubrahmanyan (1992b) and Balasubrahmanyan & Naranan (1996). The texts are (1) complete works of Shakespeare (2) nouns in Shakespeare's play *As You Like It* (3) nouns in Shakespeare's play *Julius Caesar* (4) American Newspaper English (5) four plays in Latin by Plautus (6) short story in Russian by Pushkin (7) colloquial Chinese (Peiping dialect) (8) *Ulysses* by James Joyce (9) nouns in *Essay on Bacon* by Macaulay and (10) a sample of English texts in Dewey (1923).

The data in almost all cases deviate from a pure power law (equation 2) for low k. However, they conform very well with the Modified Power Law (equation 4). At low k values the exponential term $e^{-\mu/k}$ ($\mu > 0$) gives $W(k)$ less than values implied by equation (2). For most texts $\gamma \approx 2.0$ and slightly in excess of 2.0 with a statistical error of ≈ 0.05. The modifying parameter μ has the range $0 < \mu < 1.3$ with varying errors but less than 0.2. B is the

normalization constant that depends on the size of the text (N). The 'chi-squared' statistic (χ^2) is within 'acceptable limits' in almost all cases. The hypothesis of MPL as a representation of word frequencies cannot be rejected. For details see Balasubrahmanyan & Naranan (1996) and references listed therein.

As mentioned already, the k-domain is dominated by lexical words (nouns, verbs, adjectives etc.) which we call "content" words or C-words. Broadly speaking, C-words relate to the vocabulary of a language. Our analysis shows that the C-words (k-domain) obey Zipf's law with $\gamma \approx 2.0$, which is a language universal with few exceptions, if any. *Consequently word frequency analysis of C-words cannot help distinguish different texts across language, author, style, genre etc.* There is one clear exception: *Complete works of Shakespeare* (N = 194667, V = 30688, k_o = 100) has $\gamma = 1.6 \pm 0.01$ and $\mu = 0.02 \pm 0.03$, which implies a nearly pure power law (since $\mu \approx 0$) with a γ clearly different from 2.0. It is interesting that Shakespeare's nouns (in *As You Like It* and *Julius Caesar*) conform to $\gamma \approx 2.35$, somewhat greater than 2.0. Shakespeare's uniqueness in the literary world is perhaps reflected in the non-canonical value $\gamma = 1.6$.

Another exception is a very unconventional "linguistic" text: the *Indus Text.* It belongs to the age of the Indus Valley Civilization (c. 2300 – 1750 BC). The writings on the seals have 417 different signs (V) in 13372 occurrences (N) (Mahadevan 1977). For our analysis we have used N = 11328, V = 415. The word frequencies yield $\gamma = 1.36 \pm 0.06$ and $\mu = 0.44 \pm 0.24$.

The Indus Text is a non-standard text with a high N/V ratio (27.3) and a low γ (1.36). In our effort to study symbol frequencies at various hierarchical levels of language, we investigated the frequency distribution of digrams in English (two letter combinations such as *th, sh, at, pt* etc.) (Naranan & Balasubrahmanyan, 1992b). From the data of Gaines (1956)

$$N = 8249 \quad V = 422 \quad \gamma = 1.35 \pm 0.07 \quad \mu = 1.70 \pm 0.29. \tag{6}$$

It is striking that these parameters, especially γ, are very similar to those for Indus Text. This suggests that most of the signs could be compound symbols as proposed by Mahadevan (1977). There is also a general agreement that the writing system "is based on syllables or something akin to them and is neither alphabetic nor logographic".

What is the basis for the MPL (equation 4)? It is not devised as a curve-fitting exercise. The MPL is derived in a model based on Information Theory and some concepts of statistical physics and computer science (Naranan & Balasubrahmanyan 1992a; Balasubrahmanyan & Naranan 1996).

2.2. Phonemes, letters and S-word frequencies

Words, spoken or written, belong to the most numerous of different categories of language symbols. At a lower level, the symbols are the letters of an alphabet in written form and phonemes in spoken form.

Phoneme frequencies in six Indian languages and the letter and phoneme frequencies in English are studied by Naranan & Balasubrahmanyan (1993) and Balasubrahmanyan & Naranan (1996). The number of phonemes varies between 20 and 40, so the rank frequency analysis (r-domain) is applicable. The rank frequencies $p(r)$, $r = 1,2,\ldots,V$ do not obey equations (1) or (3). The departures from them are very significant. It appears that when the symbol set is small and fixed, Zipf's law does not apply. However, a slightly altered version of equation (3) can accommodate all the data. This is given by equation (5) which has on the right side a cumulated sum of MPL like terms. It is called the CMPL. The constants D, v, δ are analogous to the constants B, μ, γ of the MPL.

The goodness of fit test is the Kolmogorov test; K_s is the Kolmogorov statistic. It is found that the hypothesis of CMPL cannot be rejected on the basis of the value of K_s. The parameters v and δ for 8 different data sets (7 for phonemes and one for letters) are scattered over a wide range: $-0.14 < \delta < 1.65$ and $-4.64 < v < 3.88$. Note the striking contrast with the γ and μ parameters for C-words ($\gamma \approx 2.0$ and $0 < \mu < 1.3$). For details and relevant references to data see Balasubrahmanyan & Naranan (1996).

We have already mentioned that grammatical words, which form a small set, dominate the *r*-domain of word frequencies. Since the classification of words into parts of speech tends to be fuzzy we have chosen to refer to grammatical words as service words or S-words and the semantic words as C-words. S-words are used to connect other words to form phrases, sentences etc., serving the process of building the hierarchical structure of language texts. Most S-words have no number, gender, tense etc., unlike C-words. In English there are 71 S-words, a number which has remained constant over hundreds of years.

Frequencies of 71 S-words in three different English texts are analyzed by Balasubrahmanyan & Naranan (1996). The texts are (1) complete works of Shakespeare (1564-1616), (2) Sir Arthur Conan Doyle's entire Sherlock Holmes collection (1859-1930) and (3) Dewey (1923). The rank frequencies are well fit to the CMPL (equation 5) with (v, δ) values (4.31, 2.13), (0.88, 1.72) and (1.20, 1.92) respectively for Shakespeare, Doyle and Dewey. While $\delta \approx 2.0$ in all the three (analogous to $\gamma \approx 2.0$) and $v \approx 1$ for Doyle and Dewey, v for Shakespeare is a distinctive value (4.31) like the anomalous value of $\gamma = 1.6$ for C-words of Shakespeare. So Shakespeare is 'different' from the rest in the use of C-words as well as S-words.

Can S-word frequencies be used to distinguish different texts? We are aware of at least one such attempt. Alexander Hamilton and James Madison wrote 77 essays (Federalist Papers) in 1787-1788. Of these, 51 essays were by Hamilton, 14 by Madison and 12 of unknown authorship. Mosteller and Wallace (1984) studied the frequencies of 70 function words (S-words) in the essays. They favored Madison as the author of the 12 essays of unknown authorship. Recently, a sophisticated linear programming technique has been used by Bosch & Smith (1998) for the same problem. They too favor Madison as the author.

Before concluding this section we divert briefly to another kind of language: 'the language of life'. The DNA molecule in all cells – the building blocks of life – is a linear sequence of four different bases (A,G,T,C) that carries all the genetic information for life's functions. DNA contains subsequences called genes and each gene generally codes for one protein. Proteins are also linear chains of 20 different amino acids. The genetic code prescribes how the 4-letter alphabet of genes is translated into the 20-letter alphabet of the proteins. (Actually, the alphabet has an extra symbol for 'stop' to delineate one gene from another on the DNA). The code is essentially simple. The gene sequence of bases is 'read' three letters at a time; these triplets are called codons. There are 64 possible codons such as AAG, GAT … (4x4x4 = 64). Each codon is translated into a specific amino acid, e.g. AAG means the amino acid lysine. Three of the 64 codons (UAA,UAG,UGA) mean 'stop', signaling the end of the amino acid sequence of the protein. The remaining 61 codons are assigned to 20 amino acids. The details of the translation from gene to protein are complex but not relevant for our purpose. It is a remarkable fact that the genetic code is universal for all organisms evolved over the last few billion years.

We have studied the rank frequencies of the 64 codons of genes in 20 different species. They too conform to CMPL distribution (equation 5). Just as in the case of phonemes, here too the parameters (v, δ) are scattered over a wide range among the 20 species that span a few billion years in evolutionary age ($-1 < \delta < 1$, $-4.5 < v < 0.3$). The errors in v, δ are typically $\approx$ 0.1. There is a hierarchical structure in genes too. 64 codons are classified into 26 codon sets

that are organized as 21 codon groups (20 amino acids and one 'stop' symbol). The rank frequency distributions obey CMPL at all the three levels.

There is a further interesting fact. It appears that for a given set of symbols $v/(\delta\text{-}1)$ is a constant (= b) within statistical error. b may differ from one set to another (e.g. for phonemes b = 5.7 and for 64 codons b = 3.4). If this is true then for a given set, there is only one independent variable (say δ). For details and references to data see Naranan & Balasubrahmanyan (2000) and Balasubrahmanyan & Naranan (2000).

It appears that the CMPL is a very useful representation of rank frequencies of a small invariant set of symbols (phonemes, letters, S-words, codons). It is a variant of Zipf's law in a broad sense. Its genesis lies in an extension of the same model that was proposed for the MPL distribution of C-words. We now briefly describe the model for MPL and CMPL.

3.0. Models based on algorithmic information theory

Shannon's "mathematical theory of communication" deals with efficient coding of a message for transmission. It is closely related to equilibrium statistical thermodynamics of Boltzmann and Gibbs and the physical concept of entropy that is a measure of 'disorder'. Applied to a message – a string of symbols of an alphabet – the entropy is a measure of information as the decrease of uncertainty due to a received message. For an elementary introduction relevant to linguistics see Naranan & Balasubrahmanyan (1992a) and Balasubrahmanyan & Naranan (to appear).

The Shannon entropy H_s depends only on the probabilities of occurrence of V different symbols in a string of N symbol tokens: P_i (i = 1,2,…,V), $P_i > 0$ and $\Sigma P_i = 1$.

$$H_s = -\sum_{i=1}^{V} P_i \lg P_i . \tag{7}$$

Here '*lg*' is 'logarithm to base 2'. In practical terms – for coding purposes – H_s is the minimum average number of bits (binary digits 0,1) per symbol, needed to encode the message for transmission. Shannon coding is based on assigning codes of varying lengths to symbols, depending on their probability of occurrences: short codes for the more frequent and longer codes for the less frequent symbols. An essential requirement of such a coding scheme is that the code be a *prefix code* or uniquely decipherable. In other words, as the bits of the encoded string are read in sequence there is no ambiguity in decoding the sequence of the symbols in the original message. Given the probabilities P_i (i = 1,2...V), the optimum assignment of code lengths to the V symbols is determined by an elegant algorithm (Huffman 1952).

In applying Shannon's theory to natural languages there are two important considerations. (1) H_s depends only on symbol probabilities. All permutations of the symbols in the message will have the same H_s and are therefore equivalent. However in a meaningful message – a language text – the particular sequence of symbols is important. (2) In Shannon's theory all strings of length N are equiprobable. But for language texts governed by grammatical and semantic rules this is not true. Optimal coding will depend on the particular sequence of symbols. For example a 'regular' string (101010101010…) can be coded as a short computer program *"Print '10' m times"*. But most strings will be random and the program has to print every bit in sequence.

The idea of relating 'complexity' of a string to the length of the shortest length computer program that prints the string, is due to Kolmogorov (1965). This approach was developed further as 'Algorithmic Information Theory' by Chaitin (1987) and applied by Zurek (1989) to physical systems. Algorithmic complexity applies to purely random as well as partially

ordered strings. Linguistic texts, too, have a random aspect (choice of words, topic and style) and are ordered by syntactic and semantic constraints. Language is therefore a complex system with a mix of random and orderly elements. Kolmogorov complexity or entropy or Algorithmic complexity is therefore more relevant for efficient coding of individual texts.

Here, we explore the algorithmic coding approach and show that the concept of an Optimum Meaning Preserving Code (OMPC) leads to an understanding of linguistic structure and a quantitative derivation of Zipf's law (Balasubrahmanyan & Naranan, 1996).

Let the language discourse consist of N words (word tokens) and V different words (word types). Let $W(k)$ be the number of word types that occur exactly k times.

$$\sum W(k) = V \ (vocabulary\ size) \tag{8}$$

$$\sum n(k) = N \ (discourse\ size) \tag{9}$$

$$n(k) = k\,W(k) \tag{9a}$$

The summation is over the range $1 \le k \le k_m$. As seen earlier (section 2.0) the range of k values comprises two domains: k-domain [$W(k) > 1$] and r-domain [$W(k) = 0$ or 1].

We propose the following coding procedure that incorporates specifically the sequence in which different symbols appear. It is convenient to define a '*k-word'* as a word that occurs k times in the discourse. The number of k-words is $n(k) = k\ W(k)$ (equation 9a). Every word token (or simply 'word') is uniquely identified by three numbers: (1) k, the frequency of its occurrence, (2) $W'(k)$, an integer in the interval $1 \le W'(k) \le W(k)$. $W'(k)$ is the serial number of the word in a list of $W(k)$ k-words. (3) To specify the location of the word in the sequence a third quantity is required. We arbitrarily define it as $f(k)$. The entire discourse is partitioned into k-words ($1 \le k \le k_m$) [equation (9)]. The number of bits required for algorithmic coding of a word is $lg[k*W(k)*f(k)]$. The total number of bits for the entire discourse

$$H_{AN} = \sum kW(k)\ \lg\,[k\,W(k)\,f(k)]. \tag{10}$$

The average number of bits per word is

$$H_a = H_{AN} / N. \tag{10a}$$

H_a is the algorithmic entropy. The significance of $f(k)$ will be clear soon.

To derive the optimum function $W(k)$ we invoke an extremum principle commonly used in physical sciences. A desired quantity attains a maximum/minimum value, while some related quantities are constrained to have certain given values. *It is hypothesized that the optimum W(k) is the one that minimizes H_a for given values of N and V*. N and V can be viewed as boundary conditions. Using the method of 'undetermined Lagrange multipliers' the optimum word frequency distribution can be shown to be

$$W(k) = B\ e^{-\mu/k}\ /[k\,f(k)]. \tag{11}$$

Constants B, μ arise from the two constraints N and V.

Let the set $\{\ell\} = (\ell_1, \ell_2, \ell_3, ...)$ represent the C-word frequencies (k-domain) and the set $\{g\} = (g_1, g_2, g_3, \ldots)$ represent the frequencies of S-words (r-domain). The frequencies form an increasing sequence

$$\ell_1 < \ell_2 < \ell_3 \ldots < g_1 < g_2 < g_3 \ldots$$

Let f_ℓ and f_g be the function f for C- and S-words respectively. For C-words equation (11) can be written as

$$W(k) = [Be^{-\mu/k} / kf_l(k)]\delta_{kl} \qquad [k \in \{l\}], \tag{11a}$$

δ is the Kronecker delta ($\delta_{mn} = 1$ if $m = n$, $\delta_{mn} = 0$ if $m \neq n$). Comparing with equation (4) which fits the observations (section 2.0)

$$f_l(k) = k^{\gamma-1}\ \delta_{kl}\,, \tag{12}$$

and for $\gamma = 2$

$$f_l(k) = k\ \ \delta_{kl}\ . \tag{12a}$$

For S-words equation (11) can be written as

$$W(k) = [B\, e^{-\mu/k} / k\, f_g\,(k)]\ \ \delta_{kg}\,. \tag{11b}$$

From data on S-words

$$W(k) = 1 \quad [k \in \{g\}]. \tag{4a}$$

Comparing equations (11b) and (4a)

$$f_g(k) \ = \ [B\, e^{-\mu/k} / k]\ \delta_{kg}$$

Since $k >> \mu$ for S-words

$$f_g(k) = [B/k]\ \delta_{kg}\,. \tag{12b}$$

In general the set $\{\ell\}$ will have frequencies $< k_o$ and the set $\{g\}$ have frequencies $\geq k_o$. From equations (12a) and (12b) $f_\ell(k) \propto k$ and $f_g(k) \propto 1/k$ suggesting

$$f(k) \propto k^\lambda$$

where $\lambda = 1$ for C-words and $\lambda = -1$ for S-words. Since k_o is the 'dividing' frequency between the two

$$\lambda \ \ = \ \ (k_o - k)/|\,k - k_o\,|. \tag{13}$$

3.1. Algorithmic Coding Interpretation of f_ℓ and f_g

The algorithmic entropy H_a (equation 10a) can be written as a sum of two parts H_{al} and H_{ag} corresponding to the two domains of k:

$$H_a = H_{al} + H_{ag}$$

$$H_{al} = (1/N)\sum kW(k)\ \lg\,[kW(k)\ f_l(k)] \tag{14}$$

$$H_{ag} = (1/N)\sum kW(k)\ \lg[kW(k)\ f_g(k)]. \qquad (15)$$

Substituting for $f_l(k)$ (equation 12a) in equation (14)

$$H_{al} = (1/N)\sum kW(k)\ \lg\ [k * W(k) * k] \qquad [k \in \{l\}]\ . \qquad (14a)$$

Thus, the third quantity needed for coding – $f_l(k)$ – besides k and $W'(k)$, is simply k. This can be interpreted as follows. A number k' in the interval $1 \le k' \le k$ will specify the ordinal number of occurrence of a k-word at a particular location. k' will designate a number (first, second, third etc.) among the k repetitions of the k-word. This requires $lg\ k$ bits. Thus, the coding scheme specifies the ordering of the words and in that sense preserves the 'meaning' of the discourse.

For S-words $W(k) = 1$

$$H_{ag} = (1/N)\sum k\ \lg\ [k\ f_g(k)] \qquad [k \in \{g\}]. \qquad (15a)$$

Since p is used instead of k for S-word frequencies

$$H_{ag} = (1/N)\sum p(r)\ \lg\ [p(r)\ f_g(p)] \qquad [p \in \{g\}]. \qquad (16)$$

Following the algorithmic coding scheme of equation (14a) for S-words

$$H_{ag} = (1/N)\sum p(r)\ \lg\ [r\ \ p(r)]. \qquad (16a)$$

Here, in the lg term, r the rank identifies the S-word uniquely and $p(r)$ is the ordinal number of its occurrence (1,2,3….p); together they define the particular sequence encoded. Comparing equations (16) and (16a) we identify

$$f_g(p) = r \qquad [p \in \{g\}] \qquad (17)$$

whereas equation (12b) implies

$$f_g(p) = [B/p]\ \delta_{pg}\ . \qquad (12c)$$

The two equations are consistent for

$$p(r) = B/r \qquad (1a)$$

which is the well known Zipf's law.

It can be shown that different permutations of the string of N words will have, in general, different values of H_a, thereby breaking the degeneracy (or equivalence) of Shannon entropy H_s which is the same for all permutations. In the language of statistical thermodynamics, the distinct permutations of a text can be regarded as an ensemble of *microstates* each with a different algorithmic entropy H_a. They all correspond to a single *macrostate* with Shannon entropy H_s.

We have demonstrated that algorithmic coding is optimal (minimum number of bits for coding), leading to an Optimal Meaning Preserving Code (OMPC) subject to some obvious constraints (N, V) for a word frequency distribution $W(k)$ which is Zipf's law, in both the versions [equations (1) and (2)]. The modifying term $e^{-\mu/k}$, as a departure from pure power

law, follows naturally from the constraint on V, the size of the vocabulary. The modified power law (MPL) is indeed in very good agreement with the data on word frequencies.

In the above model, the index γ is fixed at the canonical Zipf value 2.0. How do we account for instances where $\gamma \neq 2$? To make γ a constant that can assume other values besides 2, we need another constraint equation analogous to constraints on N and V [equations (8) and (9)].

3.2. Degenerate Entropy

We return to Shannon entropy of a string of N symbols, V different types of symbols with probabilities P_i (i = 1,2...V). Two different entropies are defined as follows:

(1) $\boldsymbol{H_s}$: This is the conventional Shannon entropy. H_s is also interpreted as the *lg* of the number of different 'complexions' of the string. This is known to be equivalent to equation (7). The *k!* different permutations of a *k*-word lead to identical complexions. It can be shown that

$$H_s = \lg N - (1/N) \sum n(k) \lg k. \tag{18}$$

Here the sequence of the symbols too remains unaltered by the permutation.

(2) $\boldsymbol{H_d}$: Here we consider all the different *k*-words as a single group comprising all the *W(k)* different *k*-words. It consists of *n(k)* occurrences of *k*-words and all of them are indistinguishable, i.e. all the $n(k)!$ permutations of *k*-word tokens yield identical complexions. Equation (9) is the fundamental partition of N into *n(k)'s*.

$$H_d = \lg N - (1/N) \sum n(k) \lg n(k). \tag{19}$$

The relevance of H_d for our model is the following: for obtaining a histogram of *W(k) vs. k,* i.e. sorting and counting words by their occurrence frequency, the actual identity of the *k*-words is irrelevant. We define such a discourse as a *degenerate discourse* and H_d as *degenerate entropy. W(k)* is readily obtained from *n(k)* simply as *n(k)/k.* H_d was first introduced by Naranan & Balasubrahmanyan (1992a). As already mentioned H_d is the more fundamental of the two entropies H_s and H_d. The two differ in the *lg* term and clearly $H_d < H_s$. Both the entropies were used to derive the MPL version of Zipf's law.

We now consider algorithmic coding of a degenerate discourse on lines similar to those adopted for the coding of the standard discourse (section 3.1). Since all the *k*-words [*n(k)* of them] are equivalent, a word token is uniquely identified by its *k* value. To code the number *k*, *lg k* bits are required. Since there are *n(k) k*-word tokens, the average number of bits per symbol required is

$$H_{ad} = (1/N) \sum n(k) \lg k. \tag{20}$$

Comparing equations (20) and (18)

$$H_{ad} = \lg N - H_s. \tag{21}$$

In general $H_{ad} < H_s$; so algorithmic coding of a degenerate discourse requires fewer bits than coding the actual discourse. H_{ad} is the algorithmic degenerate entropy.

3.3. A Model for MPL with arbitrary γ

The additional constraint equation needed for obtaining a variable index γ is provided by H_{ad} (equation 20). Collecting the relevant expressions related to $n(k)$

$$\sum n(k)/k = V \tag{8a}$$

$$\sum n(k) = N \tag{9}$$

$$(1/N)\sum n(k)\ \lg k = H_{ad} \tag{20}$$

$$H_{al} = (1/N)\sum n(k)\lg[n(k)\,k] \tag{14b}$$

it is postulated that the optimum n(k) is the one that minimizes H_{al} for given values of N, V and H_{ad}. This leads to

$$n(k) = B\ e^{-\mu/k}\ k^{-(\gamma-1)}. \tag{22}$$

B, μ, γ are constants arising from the constrained quantities *N, V* and H_{ad} respectively. To obtain *W(k)* we simply divide *n(k)* by *k*:

$$W(k) = B e^{-\mu/k}\ k^{-\gamma} \tag{4}$$

which is the MPL. γ can take any value (including negative values). It is 2.0 for most texts with some notable exceptions (section 2.1). The algorithmic entropies H_{al} and H_{ad} are related to the degenerate entropy H_d through equations (19), (20) and (14b):

$$H_{al} = \lg N + H_{ad} - H_d. \tag{23}$$

Since *N* and H_{ad} are given (as constraints), it is clear that minimizing H_{al} is equivalent to maximizing H_d. Hence an alternate interpretation of the optimization procedure is the following: *for given values of N, V and H_{ad}, the degenerate entropy H_d is maximized when W(k) is given by equation (4).*

In this version of the model for MPL, both the concepts of degenerate entropy and algorithmic entropy of the degenerate discourse are used. The concept of degeneracy is borrowed from quantum statistical thermodynamics. Algorithmic entropy has connections to computer science and the optimization principle is a powerful tool for understanding laws in physical sciences.

3.4. The Cumulative Modified Power Law (CMPL)

When the symbol set is small (n) – as in phonemes, letters of alphabet, S-words or codons of DNA – each symbol has a different frequency of occurrence *p(r), r* = 1,2,3…*n*. Note that *p* is symbol frequency, instead of *k* and $p(1) = k_m$, the highest value of symbol frequency. We briefly sketch the steps involved in deriving the CMPL for *p(r)*.

To extend the model described in section 3.3 to the *r*-domain [*W(k)* = 1], a simple change of variable *p* is suggested (Naranan & Balasubrahmanyan 1993). Instead of *frequencies* (*p*'s) one considers the *intervals* between neighboring frequencies:

$$d(i) = p(i) - p(i+1) \quad [i = 1,2.....n-1] \qquad d(i) = p(n) \quad [i = n]. \tag{24}$$

The *p(r)*'s can be expressed in terms of *d(i)*'s:

$$p(r) = \sum_{i=r}^{n} d(i). \tag{25}$$

The total number of symbol tokens N is

$$N = \sum_{i=1}^{n} p(i) = \sum_{i=1}^{n} id(i) = \sum_{i=1}^{n} m(i). \tag{26}$$

N is partitioned in terms of *m(i)*'s:

$$m(i) = i\ d(i) \qquad i = 1,2,...,n\ . \tag{27}$$

Further

$$\sum_{i=1}^{n} d(i) = p(1) \qquad or \qquad \sum_{i=1}^{n} m(i)/i = p(1). \tag{28}$$

Comparing equations (26) and (28) with equations (9) and (8a) respectively, the following correspondence is seen:

$$i \leftrightarrow k \quad d(i) \leftrightarrow W(k) \quad m(i) \leftrightarrow n(k) \quad p(1) \leftrightarrow V \quad N \leftrightarrow N.$$

Following the steps similar to those outlined in section 3.3, one is led to the analog of equation (4) for optimum *d(i)* as

$$d(i) = D\ e^{-\nu/i}\ i^{-\delta}\ . \tag{29}$$

From equation (25), setting $n = V$ (the number of symbols)

$$p(r) = \sum_{i=r}^{V} D\, e^{-\nu/i}\ i^{-\delta} \tag{5}$$

which is the CMPL (section 1). D, ν, δ are constants analogous to B, μ, γ of equation (4). For details see Balasubrahmanyan & Naranan (2000). Parameter ν is determined by the highest frequency $p(1)$ which is a given constraint. So equation (5) in effect represents *p(r)*'s relative to $p(1)$. Equation (5) is shown to represent phoneme, letter and S-word frequencies adequately (section 2.2).

It is to be emphasized that whereas the MPL model was conceived to explain observed data on word frequencies, the *CMPL model is a natural extension of the model and is a prediction of rank frequencies of a small set of symbols.* The rationale for the extension is as follows. In the k-domain the k values are 'given' ($k = 1,2,3,...,k_o$) and we have to predict the *W(k)* values. In contrast, in the r-domain $W(1) = 1$ for all k and we have to predict the k values [the set $\{g\}$] or equivalently the intervals (d's) between successive k values assuming that the highest k value, k_m, is given. This leads to equation (24) and the subsequent steps automatically yield equation (5), the optimum rank frequency distribution *p(r)*.

4.0. Language as a complex adaptive system

In the introduction we mentioned the concept of 'effective complexity' of a system suggested by Gell-Mann (1994). It is different from Algorithmic complexity or Algorithmic Information Content that is low for highly ordered systems and high for disorderly or random systems. Gell-Mann proposed that effective complexity is low for both the extreme cases, but is a

maximum for a system with a mixture of order and randomness. Most interesting complex systems in real life have interacting elements of orderly and random behavior which evolve to maximum effective complexity.

Language is an excellent example of such a complex adaptive system. Rules of syntax represent order; the freedom of choice of words ─ especially the C-words ─ implies flexibility, an element of randomness allowing creativity. The predictable quality of a politician's speech makes it dull and uninteresting. It is the unpredictable diction and creative phrases of Shakespeare that make his works absorbing and fascinating. Another example is the DNA of molecular biology (section 2.2). The gene expression leading to protein synthesis is dictated by the universal genetic code and well-defined rules; yet random mutations of the bases in the DNA sequence occasionally result in harmful effects as well as beneficial ones. The latter are the driving forces for evolution of life to higher forms.

4.1. The Complexity Function

We briefly sketch below a particular *quantitative* definition of Gell-Mann's effective complexity proposed by us (Balasubrahmanyan & Naranan 1996) and slightly modified later (Balasubrahmanyan & Naranan 2000). We hereafter drop the adjective 'effective' and represent complexity by *C*. The definition is a hierarchical extension of Shannon entropy and Kolmogorov entropy.

Consider a discourse of *V* word types and *N* word tokens. Regarded as a sequence with 'meaning', its encoding is an OMPC (section 3.0). Every permutation of *N* words is a different configuration. There are *N!* configurations and the entropy is $lg\ N! \approx N\ lgN$ for large *N*. The entropy per symbol

$$H_m = \lg N \tag{30}$$

H_m is the maximum possible entropy of the string.

We define *C* as a function of two parameters α and *x*. α is the degenerate redundancy defined as

$$\alpha = 1 - (H_d / H_m) \tag{31}$$

which characterizes the macrostate. H_d is the same for all permutations of the words since it depends only on the relative probabilities of their occurrence. But each permutation – a microstate treated as an OMPC – requires H_a bits per symbol. In general each microstate has a different H_a. We proposed that an 'order' parameter *x* be defined as

$$x = (H_m - H_a) / H_d . \tag{32}$$

This is motivated partly by the observation that the numerical value of H_a lies between H_m and H_d. Since in general a string is partly ordered and partly random, phenomenologically one can regard the $N H_a$ bits of the OMPC as made of two parts:

$$NH_a = N\,x\,(H_m - H_d) + N\,(1-x)\,H_m . \tag{33}$$

The first term on the right is the orderly part and the second the random part. The 'order' results from grammar and syntax which dictate the sequence of words and is *rule based order* that is *externally* imposed. The definition of *x* (equation 32) follows from equation (33). In

contrast, the redundancy α arises from the fact that word types have a Zipfian distribution permitting variable length codes for word types and consequently data compression in encoding. This is also a type of order, but it is *internal* to a particular string.

For a measure of complexity C of a string of $N H_a$ bits we proposed

$$C = lg\ [Number\ of\ different\ complexions\ of\ the\ string]. \tag{34}$$

It is assumed that permutation of bits in each part of equation (33) gives the same configuration. This definition is analogous to the standard definition of Boltzmann and Shannon entropy.

It can be shown that C depends only on $\alpha\, y$ where $y = x/(1\text{-}x)$. The function C is given by

$$C(\alpha\, y) = \lg(1 + \alpha\, y) - [\alpha\, y / (1+\alpha\, y)] \lg(\alpha\, y). \tag{35}$$

C has the following properties: (1) C, α and x are all in the interval [0,1]. (2) C is a one-variable (x) one-parameter (α) function. (3) C has one maximum ($C = 1$) at $x = x^*$

$$x^* = 1 / (1 + \alpha) \tag{36}$$

and (4) $C \to 0$ as $x \to 0$ and as $x \to 1$. Note that *C has precisely the features required of effective complexity.*

Complexity can be interpreted as a higher order entropy defined on a binary string representing an OMPC. Substituting

$$z = \alpha\, y / (1+\alpha\, y) \tag{37}$$

equation (35) reduces to

$$C(z) = -[z \lg z + (1 - z) \lg (1 - z)]. \tag{38}$$

This is analogous to Shannon entropy H of a binary string in which '0' and '1' occur with probabilities z and $1 - z$. It is maximum when $z = 0.5$ or $\alpha\, y = 1$ (from equation 37). $\alpha\, y$ is also the ratio of the two terms 'order' and 'random' in the right hand side of equation (33). $\alpha\, y = 1$ implies that the two parts are equal when C is maximum (= 1) and the corresponding x is given by x^*.

4.2. Complexity of linguistic discourses

The complexity C depends on x and α which in turn are determined by the three entropies H_d, H_a and H_m. These entropies can be determined numerically for linguistic discourses from the data of symbol frequencies (C-words, S-words, letters etc.). The entropies x, α and C are listed in Table 1 for a variety of texts and symbols. The Shannon entropy H_s is included though it is not required in the calculation of x, α and C. Note that for S-words and phonemes H_s and H_d are equal. They differ only for C-words.

Set I is for eight texts included in the list of 10 texts for which MPL parameters were determined (section 2.1) for C-words. Set II is for 71 S-words of Dewey, Doyle and Shakespeare. Set IV contains the eight texts used for phonemes in English and six Indian languages (section 2.2). Set V is for English alphabets. For Indus Text and English digrams we had complete frequency distribution of symbols and so obtained the parameters for both the *k*-

domain and *r*-domains (Set III, section 2.0). The mean and standard deviation of the parameters are given for Sets I, II and IV.

The most striking observation we can make from the Table is the near constancy of x, α and C values in all sets except Set III. The mean values are

$$<x> = 0.55 - 0.56 \qquad <\alpha> = 0.65 - 0.68 \qquad C = 0.99 - 1.00 \tag{39}$$

with small standard deviations. It appears that the two variables characterizing external order (x) and internal order (α) cooperate to satisfy equation (36), the condition for maximal complexity ($C = 1$).

As we shall see soon, C depends on the index γ. It approaches 1 for $\gamma \approx 2.0$. It is remarkable that phonemes and C-words (Sets IV and I) have nearly identical x, α and C values although the CMPL parameters ν, δ for phonemes vary widely, unlike the near constancy of γ (≈ 2.0) for C-words. The low C values for Indus Text and digrams (k-domain) are due to the fact that for these cases $\gamma = 1.3$. We emphasize that these two texts are not in the same category as other linguistic texts. In both, the number of symbol tokens ($V \approx 400$) is too small to be regarded as vocabulary and too large to be classified as phonemes or alphabets.

We merely point out some relevant facts to help understand Table 1 without going into details. Firstly, H_s depends on size N, but H_d is independent of N. However α depends on N through the term H_m (= $lg\ N$). Secondly, $H_m - H_a$ and x seem to be independent of N. However all the parameters vary with γ. In general C too will depend on N. Equation (36) implies that for $C = 1$, x has to exceed 0.5, since $\alpha \leq 1$.

4.3. Complexity and MPL parameters B, μ and γ

In section 4.2 we calculated α, x and C for some linguistic discourses and different types of symbols. Now we repeat the same for different values of B, μ and γ which completely determine the symbol frequencies, the entropies and complexity parameters. The computations are done for (1) B = 1000, 3000, 10000 and 30000, (2) μ = 0, 1, (3) γ in the interval 1.5 to 2.5 in steps of 0.1. Note that B determines the size N and vocabulary V. For the frequency k_o we have used $k_o = B^{(1/\gamma)}$.

In Figure 1 are given the plots of α, x and C vs. γ. In each part of the figure there are four curves for different values of B. There are two important features to be noted: (1) as γ increases the dependence of x and C on B (or equivalently on N) decreases whereas that of α on B increases. (2) C has the maximum value (= 1) for $\gamma = 2.0$ to 2.2. C is lower for $\gamma < 2$ and there is a tendency for C to decrease as γ increases beyond 2.2. Thus a *Zipf distribution with $\gamma \approx 2.0$ leads to a maximally complex linguistic discourse.* The Information theoretic model based on an extremum principle outlined in section 3.3 does not determine γ. But by forging a connection between complexity and word frequency distribution, the emergence of $\gamma = 2$ is also seen as a consequence of an extremum principle.

As a general principle, an extremum is the most stable configuration and systems evolve to reach that state. We do not, however, understand the details of the dynamics involved.

For a comprehensive discussion of the interrelation of entropies and complexity variables see Balasubrahmanyan & Naranan (2000). It was shown that codon frequencies in DNA sequences of 20 different species also imply maximally complex DNA strings.

TABLE 1. SUMMARY OF ENTROPIES AND COMPLEXITY PARAMETERS

	DISCOURSE	ko	N	V	HM	HS%	HD	HA	ALPHA	X	C
SET I C-WORDS	Julius Caesar	28	2747	963	11,424	9,052	4,184	9,173	0,634	0,538	0,9840
	As You Like it	35	3367	1238	11,717	9,361	4,280	9,571	0,635	0,515	0,9730
	Macaulay	43	7363	2041	12,846	10,078	4,805	10,176	0,626	0,556	0,9890
	Chinese	41	9453	3313	13,207	10,699	4,517	10,680	0,658	0,559	0,9940
	Russian	40	15584	4699	13,928	11,284	4,643	11,184	0,667	0,591	0,9997
	Eldridge (Eng)	52	20447	5925	14,320	11,568	4,864	11,483	0,660	0,583	0,9988
	Latin	56	22931	8366	14,485	11,968	4,679	11,798	0,677	0,574	0,9985
	Shakespeare	100	194667	30688	17,571	13,485	6,223	14,381	0,646	0,513	0,9730
	# Mean				13,132	10,573	4,567	10,581	0,651	0,559	0,9910
	# St Dev				1,217	1,117	0,256	0,986	0,019	0,026	0,0098
SET II S-WORDS	Dewey		40304	71	15,299	4,748	4,748	12,622	0,690	0,563	0,9976
	Doyle		237511	71	17,858	4,919	4,919	15,205	0,725	0,539	0,9951
	Shakespeare		264246	71	18,012	4,959	4,959	15,382	0,725	0,530	0,9928
	Mean				17,056	4,875	4,875	14,403	0,713	0,544	0,995
SET III K-DOMAIN	Indus Script	43	3018	358	11,559	7,499	5,225	9,849	0,548	0,327	0,7430
	Digrams (Eng)	58	5839	394	12,511	7,907	5,683	10,658	0,546	0,326	0,7390
	Indus Script		10353	59	13,338	5,380	5,380	10,789	0,597	0,474	0,9340
R-DOMAIN SET IV PHONEMES	Digrams (Eng)		4160	36	12,022	5,056	5,056	9,798	0,579	0,440	0,8960
	English (Good)		2000	32	10,966	4,495	4,495	8,603	0,590	0,526	0,9680
	English (Ram)		8516	31	13,056	4,527	4,527	10,712	0,653	0,518	0,9780
	Hindi		9284	30	13,181	4,211	4,211	10,792	0,681	0,567	0,9977
	Telugu		9330	31	13,188	4,357	4,357	10,813	0,670	0,545	0,9910
	Tamil		9325	25	13,187	4,070	4,070	10,893	0,691	0,564	0,9977
	Kannada		9359	31	13,192	4,292	4,292	10,796	0,675	0,558	0,9954
	Malayalam		8832	32	13,109	4,293	4,293	10,716	0,672	0,557	0,9950
	Marathi		9366	30	13,193	4,199	4,199	10,795	0,682	0,571	0,9983
	@ Mean				13,158	4,278	4,278	10,788	0,675	0,554	0,9933
	@St Dev				0,054	0,143	0,143	0,062	0,012	0,018	0,007
SET V	English (Alph)		4500	26	12,136	4,130	4,130	9,867	0,660	0,549	0,9914

excluding Shakespeare, @ excluding English (Good), % For S-words and Phonemes HS = HD

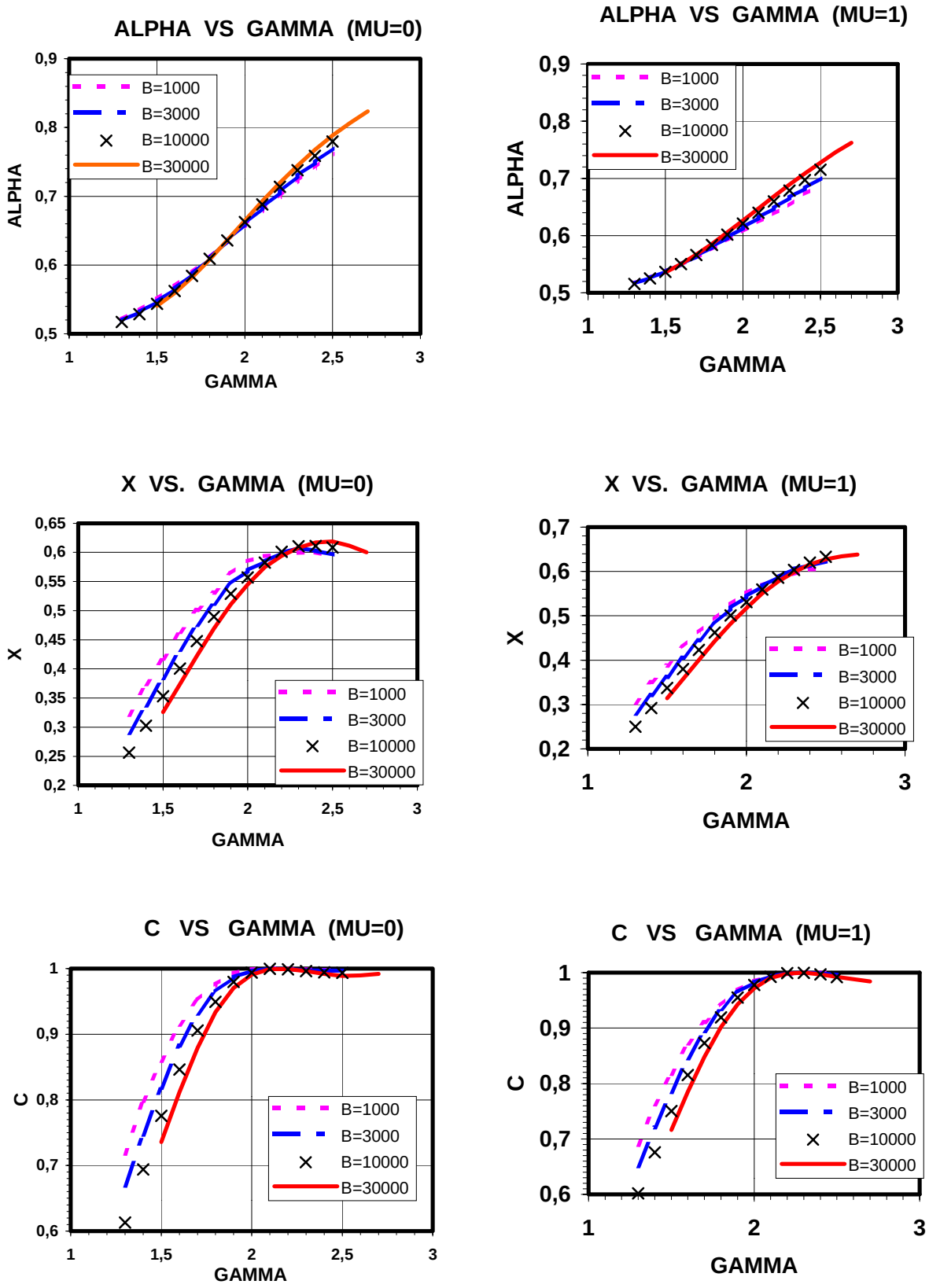

Fig. 1. Plots of α vs. γ (top), *x* vs. γ (middle) and C vs. γ (bottom). In each pair, the left figure is for μ = 0 and the right figure is for μ = 1. In each figure, there are four curves corresponding to different values of parameter B (1000, 3000, 10000, 30000).

5.0. Zipf´s law and scale invariance

Zipf's law is a quintessential example of a power law frequency distribution

$$F(x) = A\ x^{-\gamma} \tag{40}$$

with the index $\gamma \approx 2$. A is a constant. Equation (40) holds for frequency data in diverse disciplines. For a long time power laws were ignored in statistical theory because they did not conform to the requirements of a mathematically tractable theoretical framework. For an elementary review see Naranan and Balasubrahmanyan (1998) and for an in-depth treatment of the theoretical problems involved see Liebovitch & Scheurle (2000) and Sornette (2000).

Mandelbrot's work on fractal structures and geometry (Mandelbrot 1977, 1983) triggered a resurgence of interest in power laws. Fractals exhibit similar patterns on all size scales. This self-similarity or scale invariance naturally leads to power laws as statistical distributions characterizing fractals in nature.

F(x) is scale invariant if it remains unchanged when x is scaled by a factor b. It is easy to see that equation (40) has such a property. Substituting bx for x,

$$F(b\,x) = b^{-\gamma}\ F(x)\ . \tag{40a}$$

F(x) and *F(bx)* differ only by a constant multiplier $b^{-\gamma}$ that is independent of x but depends on the scale factor b. To accommodate a wide range of scales one can average over all b values. If 'strict' self-similarity is demanded, i.e. $\langle F(bx)\rangle = F(x)$, then the multiplier is 1. It can be shown that γ can take only two values 2 and 0 corresponding to $b > 1$ and $b < 1$ respectively. This implies that *there are two domains of strict self-similarity: $F(x) = A\ x^{-2}$ and $F(x) = A$.* It is indeed noteworthy that they correspond to the word frequency distributions in k-domain and r-domain respectively [$W(k) \propto k^{-2}$ and $W(k) = 1$]. Since the word frequency distribution is known to be the same for all sizes of discourse (N), it is scale invariant.

A more general scale invariant function is

$$S(x) = A(x)\ x^{-\gamma} \tag{41}$$

Here the power law is modulated by a function *A(x)* which satisfies the condition $A(bx) \propto A(x)$. A function involving trigonometric series of Weierstrass has such a property. There is some evidence that some fractal structures in nature are related to such a function.

The MPL function (equation 4) is not scale invariant since the modulating term $e^{-\mu/k}$ is not scale invariant. $e^{-\mu/k}$ and $e^{-\mu/bk}$ are not related by a constant multiplying factor. It is not unlikely that a scale invariant Weierstrass-type function *A(x)* may indeed account for the observed departures from pure power law at low k. It is also possible that scale invariance breaks down at low k as implied by the MPL. For further details see Naranan & Balasubrahmanyan (1998).

We now describe another scale invariant property of a language discourse. Consider a k-word. The average interval or number of words between consecutive occurrences of the k-word is $\lambda \approx N/k$. It can be shown that the distribution of the 'gap length' λ is invariant with respect to N only when $\gamma = 2$ (Naranan & Balasubrahmanyan 1992b). When this is true the distribution $g(\lambda)$ is a uniform distribution.

5.1. Why is $\gamma = 2$?

We have described in this article four different paths leading to the universal value $\gamma = 2$. They are recapitulated here.

(1) In section 3.0 we defined H_a, the algorithmic entropy of a meaning (sequence) preserving code of a discourse. H_a is a minimum when $\gamma = 2$ (k-domain). In the r-domain [$W(k) = 1$], $\gamma = 0$.

(2) In section 4.1 we defined a complexity measure C which is a function of the redundancy α and an order parameter x. C has its maximum possible value (= 1) when $\gamma \approx 2$. C is nearly independent of discourse size N (Figure 1).

(3) A special kind of 'scale invariance' invoking "averaging over all scales" yields $\gamma = 2$ or 0 as the only solutions (section 5.0).

(4) The distribution of gap length between two consecutive occurrences of the same word is independent of discourse size only when $\gamma = 2$.

It appears that the principles involved are general and go beyond linguistic discourses. The general principles are counting, sorting and optimum coding; it is speculated that the ubiquity of Zipf's law and the uniqueness of the index γ (= 2) in systems which are very disparate, entitles the law to be characterized as a *mathematical law* although it does not have, for example, the elegance of the Central Limit Theorem for the ubiquitous Gaussian or normal distribution (Naranan & Balasubrahmanyan 1998).

6.0. Public lexicon, private discourses and emergence of meaning

We have confined ourselves so far to frequency distributions for symbols at the lower hierarchical level of phonemes and words. The still higher level symbols are phrases, clauses, sentences etc. The distributions of the *symbol lengths* (e.g. the number of letters in a word, number of words in a phrase etc.) have been extensively studied and many regularities noted. The distribution is well fit to a long tailed *lognormal distribution* (Naranan 1992, Naranan & Balasubrahmanyan 1992a, Balasubrahmanyan & Naranan 1996 and references therein).

What do we know about the occurrence frequencies $W(k)$ of higher level symbols such as phrases, sentences etc.? The results are rather uninteresting: the frequencies tend to concentrate at a single value $k = 1$. In other words $W(1) \approx N$ and $W(k) = 0$ for $k \neq 1$. $W(k)$ behaves like Dirac's delta function [$\delta(x) = 1$ at $x = 0$ and $\delta(x) = 0$ at $x \neq 0$]. In the complex systems approach to language, this simple observation may have a profound significance.

First we begin with some data. In Table 2 are presented $W(k)$ *vs* k for phrases and sentences in a short text of 614 words organized into 256 different phrases and 45 sentences. Phrases were identified as groups of two or more words and the rest are "one-word" phrases. 248 phrases (97 % of the total) occur only once, 5 occur twice, 2 occur thrice and one phrase occurs 15 times (the last being the one-word phrase 'I'). All the 45 sentences occur only once each. This sharp peaking at $W(1)$ for higher level symbols is radically different from the frequencies of words given by Zipf's law and MPL, and of phonemes and S-words given by the CMPL. For example the ratio $W(2)/W(1) \approx 0.1 - 0.2$ for words whereas from Table 2, the ratio is 0.02 for phrases and 0 for sentences.

Complex adaptive systems exhibit self-organization and at some point in evolution have *emergent* properties, properties that were not present in the constituents of the system (Johnson 2001). When an interconnected system of relatively simple elements self-organizes to form a more adaptive, more intelligent system at a higher hierarchical level the phenomenon is called 'emergence'. In fact emergence is often considered the defining property of a complex system. In language as a complex adaptive system, the emergent properties are of context, content, coherence and meaning of a text which consists of collection of words

(morphemes) governed by rules of syntax, word formation, phonetics etc. Although 'meaning' is a very intangible concept, hard to analyze, one can grant that the whole purpose of a linguistic discourse is to communicate content or meaning. The hierarchy at which language acquires emergent properties seems to be phrases and it is at this level that *W(k)* tends to become a δ-function. Could the two be related?

Table 2. Frequencies of phrases and sentences

# OF SYMBOLS	\ FREQ *(k)*	1	2	3	15	Total
W(k)						
Phrases		248	5	2	1	256
Sentences		45	0	---	---	45

Text: "In my hands" by Irene Gupopdyke with Jennifer Armstrong
Reader's Digest, December 2000 (Indian Edition)

6.1. Privatization of language discourse

The symbols of language – letters, phonemes, syllables, words (morphemes) – can be considered as a public lexicon, a collection of symbols in a repository. New words are added periodically to a language through acquisition from other languages and invention of new words. (Shakespeare is credited with adding about 3000 new words in his works.) Similarly the rules of grammar, phonetics, sentence construction etc. are also reasonably well defined.

At higher levels of organization (e.g. phrases, sentences) there is clear qualitative departure. Complete lexicons of phrases, clauses and sentences do not exist, simply because there are too many of them and more important, they are not enumerable in principle. A speaker or writer uses the public lexicon of lower level symbols and the rules of grammar to build up a *privatized* communication. The process appears to start with phrases that severely limit the usage of words to suit context and topic. In our view, this is the starting point of emergence of meaning. At the sentence level, privatization becomes even more dominant. It is almost impossible to find the same sentence in two different books (exceptions being quotes and plagiarisms). The transition from a public domain of the symbols to a privatized communication with meaning constitutes the emergence in language.

What are the implications of this transition for the entropies, complexity parameters? What are the consequences of *W(k)* becoming a δ-function [$W(1) = N$, $W(\neq 1) = 0$]? First we note that $H_d = 0$ since all the N symbols are indistinguishable (equation 19). Next, from equations (14a), (18) and (30) $H_m = H_a = H_s$. All the entropies except H_d become equal to the maximum possible value for N symbols. Thus the discourse becomes a maximum information system according to Shannon's theory.

Complexity is defined as a function of α and x. α [$= 1- (H_d / H_m)$] becomes 1 and x = [$(H_m - H_a)/ H_d$] becomes indeterminate. For complexity $C = 1$, the necessary condition is $\alpha y = 1$. When $\alpha = 1$, then $y = 1$ or $x = 0.5$. The (α, x) values (1, 0.5) represent the extreme possible values for maximum complexity. For maximum complexity to be attained, the minimum value of x is 0.5; or *'order' has to be at least as important as 'disorder'*. Emergent properties seem to be a consequence of this double extremum (of C and x).

The transition in entropy values leading to emergence (of meaning) in linguistic discourses resembles the phase transition in physical systems. In particular the Bose-Einstein Condensation in quantum statistical mechanics is a good analog (see e.g. Venkataraman 1992). Bosons are particles which have no restrictions in occupying the same physical state (say a given energy) whereas for fermions, utmost one particle is allowed per each state. Phrases and sentences – all concentrated at a single frequency or "occupancy state" ($k = 1$) –

are like bosons. At low temperatures bosons condense to form a single entity behaving like a giant atom with strong correlations. The language text considered as a string of sentences (as symbols) is analogous to a Bose-Einstein Condensation with strong correlations among its constituents and the discourse attains meaning and a high level of coherence. Extending the analogy further, S-words with $W(k) = 0$ or 1 for all k behave like fermions.

7.0. Discussion and summary

The most widely discussed models for Zipf's law are random text models. Pseudowords, delimited by a space symbol follow Zipf-Mandelbrot law (equation 3) [Mandelbrot 1955; Miller 1957; Nicolis 1989; Li 1992]. This follows from computer simulations of random texts as well as statistical theory. Refinements, including unequal symbol probabilities and limiting the pseudoword length to a maximum value, help improve the rank frequencies to conform better with those observed for natural languages. Therefore it is considered that Zipf's law is "linguistically shallow" (Mandelbrot 1977).

We believe that the random texts – also termed as 'monkey languages' – are not comparable to natural languages for various reasons. In random texts, *all* combinations of m letters are legitimate "words" of length m and word frequencies decrease exponentially with increasing m. This is not true for natural languages in which word lengths are distributed according to the long tailed lognormal distribution and not the steeply declining exponential distribution. Further, in natural languages which have a hierarchical structure unlike the random texts, the lognormal distribution of symbol length extends to all levels of hierarchy: words, phrases, clauses, sentences etc. For further elaboration on the contrasts between natural languages and random texts see Naranan and Balasubrahmanyan (to appear).

Although Zipf did not validate his empirical laws with statistical rigor, it is now accepted that his laws do hold not only in linguistics but also in many other disciplines. We have pointed out the need to deal with two distinct domains of word frequencies ($k < k_o$ and $k \geq k_o$) in sections 1 and 2. It is found that the two-domain structure is linguistically significant. "Content" words (C-words) which are usually nouns, verbs, adjectives etc. are all in the k-domain ($k < k_o$). The grammatical words (articles, prepositions, conjunctions etc.) called the S-words, dominate the r-domain ($k \geq k_o$). Observations do exhibit a departure from pure power law at low k, which is accounted by a modifying exponential term (section 2.0). The Modified Power Law (equation 4) proposed by us is not arbitrary but follows from a model based on Shannon's Information theory, Algorithmic coding, concepts of degeneracy from quantum statistics and extremum principles commonly used in establishing physical laws (section 3.0). It is interesting to recall here the words of the famous mathematician L. Euler: "Nothing takes place in the world whose meaning is not that of some maximum or minimum". Two new kinds of entropy called the algorithmic entropy (H_a) and degenerate entropy (H_d) are introduced. The model also predicts the existence of two domains of word frequency distribution and the corresponding two versions of Zipf's law.

An extension of the model proposed for MPL leads to a prediction about the rank frequency distribution of a small fixed set of symbols (e.g. the letters of alphabet, phonemes, S-words). The distribution is a cumulated sum of MPL-like terms and is called the CMPL (section 3.4). It fits the observed distributions of letters, phonemes and S-words very well (section 2.2).

A key concept in the model is the algorithmic coding of a sequence of words, specifying the particular order of symbols in sequence. This leads to an Optimum Meaning Preserving Code – OMPC – with a minimum code length which is realized only when the word frequencies conform to the MPL.

Algorithmic coding and OMPC lead us to the concept of complexity of a string. A particular version of complexity of a system proposed by Gell-Mann – effective complexity – has the seminal property that it is low for both extremes dominated by orderly or random features. We have defined a quantitative measure of effective complexity (C) for an OMPC, that has the attributes of Gell-Mann complexity (section 4.0). C is defined as a function of an order parameter x and redundancy α which are in turn dependent on three entropies. C attains a maximum (= 1) when $x = 1/(1 + \alpha)$, a condition that is true for most linguistic texts. So OMPC implies a maximally complex linguistic discourse. Further it is found that C, which depends on γ, approaches the maximum (= 1) when $\gamma \approx 2$ (section 4.3). C can be viewed as a higher order entropy. *C quantifies the relative proportions of order and disorder in a system.* Order and disorder coexist in a complex system. To quote Niels Bohr "opposites are not contradictory but complementary". For maximum complexity to be attained it is required that 'order' exceed 'disorder' ($x \geq 0.5$). It is an oft-stressed dictum in science that quantifying a concept helps in greater understanding of the system. In this spirit, the quantitative definition of complexity of a language discourse is, we believe, a significant step. For instance, it allows comparison of texts in terms of complexity.

Zipf's law applies to all texts irrespective of their size and is therefore scale invariant. Self-similarity, scale invariance yield power laws as simplest possible probability distributions. A strict version of scale invariance extended to all size scales results in unique power law indices $\gamma = 2$ and $\gamma = 0$ (section 5.0).

Moving up from words to phrases and sentences as symbols, we ask how do the symbol frequencies behave? For phrases and sentences all 'symbols' concentrate at $k = 1$. Each symbol usually occurs only once. This limiting situation also coincides with emergence of meaning of a coherent discourse. We speculate that the phenomenon of 'emergence', an essential feature of complex systems, appears first at the level of phrases. At the level of words, the symbols constitute a public lexicon. Higher level symbols are privatized versions based on the public lexicon and the rules of grammar, used to create meaningful coherent texts. These are our first attempts to comprehend the intangible concept of meaning and its emergence in language.

Zipf was the first quantitative linguist to firmly place 'word' as the central fundamental symbol of language instead of the alphabet. He established the essential unity underlying all languages, a universality based on word as the symbol, before universal concepts in language came into vogue (e.g. universal grammar of Chomsky). His principle of least effort is a precursor to the principles of optimum coding discovered by Shannon. Zipf expanded the domain of his laws to incredibly diverse fields with emphasis on facts and figures. This expansion goes on to this day with Zipfian laws appearing in almost all areas of science.

References

Balasubrahmanyan, V.K. & Naranan, S. (1996). Quantitative linguistics and complex system studies. *Journal of Quantitative Linguistics 3*, *177-228*.

Balasubrahmanyan, V.K. & Naranan, S. (2000). Information theory and algorithmic complexity: Applications to language discourses and DNA sequences as complex systems: Part II: Complexity of DNA sequences, analogy with linguistic discourses. *Journal of Quantitative Linguistics 7*, *153-183.*

Balasubrahmanyan, V.K. & Naranan, S. (to appear). Entropy, information and complexity. In: *Handbook of Quantitative Linguistics*.

Bennett, C.H. (1990). How to define complexity in physics and why? In: W.H. Zurek (Ed.), *Complexity, entropy and the physics of information 7, 137-148*. Redwood City, CA: Addison Wesley.

Bosch, R.A & Smith, J. (1998). Separating hyperplanes and the authorship of the disputed Federalist papers. *American Mathematical Monthly 105*, *601-607*.

Chaitin, G.J. (1987). *Algorithmic information theory.* Cambridge: Cambridge University Press.

Dewey, G. (1923). *Relative frequencies of English speech sounds.* Cambridge, MA: Harvard University Press.

Eldridge, R.C. (1911). *Six thousand common English words.* Buffalo: The Clements Press.

Estoup, J.B. (1916). *Gammes stenographiques. Methodes et exercises pour l'acquisition de la vitesse* (4th ed.) Paris: Institut Stenographique.

Gaines, H.F. (1956). *Cryptanalysis.* New York: Dover Publications.

Gell-Mann, M. (1994). *The quark and the jaguar, adventures in the simple and the complex.* New York: W.H. Freeman & Co.

Huffman, D.A. (1952). A method for the construction of minimum redundancy codes. *Proc. Inst. Radio Engineers 40*, *1098-1101.*

Johnson, S. (2001). *Emergence: The connected lives of ants, brains, cities and software.* New York: Scribner.

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission 1*, *3-7*.

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

Liebovitch, L.S. & Scheurle, D. (2000). Two lessons from fractals and chaos. *Complexity 5*, *34-43.*

Mahadevan, I. (1977). *The Indus Script, texts, concordance and tables.* New Delhi: Archaeological Survey of India.

Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In: W. Jackson (Ed.), *Communication theory: 486*. London: Butterworths.

Mandelbrot, B. (1955). Information networks. In E. Weber (Ed.), *Brooklyn Polytechnic Institute Symposium*: *205-221*. New York: Interscience.

Mandelbrot, B. (1966). Information theory and psycholinguistics: a theory of word frequencies. In: P.F. Lazarsfield & N.W. Henry (Eds.), *Readings in Mathematical Social Sciences: 151-168*. Cambridge, MA: MIT Press.

Mandelbrot, B. (1977). *Fractals, form, chance and dimension*. New York: W.H. Freeman.

Mandelbrot, B. (1983). *The fractal geometry of nature.* San Francisco: W.H. Freeman.

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

Mosteller, F. & Wallace, D.L. (1984). *Applied Bayesian and classical inference: The case of the Federalist papers* (2nd Ed.) Berlin, Heidelberg, New York, and Tokyo: Springer-Verlag.

Naranan, S. (1992). Statistical laws in information science, language and system of natural numbers: Some striking similarities. *Journal of Scientific and Industrial Research 51*, *736-755.*

Naranan, S. & Balasubrahmanyan, V.K. (1992a). Information theoretic models in statistical linguistics – Part I: A model for word frequencies. *Current Science 63*, *261-269.*

Naranan, S. & Balasubrahmanyan, V.K. (1992b). Information theoretic models in statistical linguistics – Part II: Word frequencies and hierarchical structure in language – statistical tests. *Current Science 63*, *297-306.*

Naranan, S. & Balasubrahmanyan, V.K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research 52*, *728-738.*

Naranan, S. & Balasubrahmanyan, V.K. (1998). Models for power law relations in linguistics and information science. *Journal of Quantitative Linguistics 5*, *35-61.*

Naranan, S. & Balasubrahmanyan, V.K. (2000). Information theory and algorithmic complexity: Applications to language discourses and DNA sequences as complex systems: Part I: Efficiency of the genetic code of DNA. *Journal of Quantitative Linguistics*, *7*, *129-152.*

Naranan, S. & Balasubrahmanyan, V.K. (to appear). Power laws in statistical linguistics and related systems. In: *Handbook of Quantitative Linguistics.*

Nicolis, G., Nicolis, C. & Nicolis, J.S. (1989) Chaotic dynamics, Markov partitions and Zipf's law. *Journal of Statistical Physics 54*, *915-924.*

Prün, C. & Zipf, R. (2002). Biographical notes on G.K. Zipf. *Glottometrics (To honor G.K. Zipf Volume 1), 3, 1-10.*

Shannon, C.E. (1948). A mathematical theory of communication. I, II. *Bell System Technical Journal, 27*, *379-423, 623-656*. Reprinted in Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*, Urbana: University of Illinois.

Sornette, D. (2000). *Critical phenomena in natural sciences: chaos, fractals, self-organization and disorder. Concepts and tools.* Heidelberg: Springer-Verlag.

Thorndike, E.L. (1932). *A teacher's Word book of 20,000 words.* New York: Teacher's College.

Venkataraman, G. (1992). *Bose and his statistics.* Hyderabad: University Press.

Zipf, G.K. (1935). *The psychobiology of language.* Boston: Houghton Mifflin Co. Reprinted (1968). Cambridge: M.I.T P*ress.*

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Reading: Addison-Wesley.

Zurek, W.H. (1989). Algorithmic randomness and physical entropy. *Physical Review A 40*, *4731-474.*

Glottometrics 4, 2002, 27-38 *To honor G.K. Zipf*

Efficiency of communication

A new concept of language economy

Thorsten Roelcke[1]

Abstract. George Kingsley Zipf is known not only as the "father" of language statistics or quantitative linguistics in general, but also as one of the first who discussed the phenomenon of linguistic economy in detail. The following discussion in linguistics and communication sciences shows a wide spread of more or less scientific grounded concepts. This great conceptual diversity however disturbs the scientific discussion further on. Hence in the following contribution a new concept of language economy that fulfils holistic (and atomistic) requirements will be shown.

Keywords: communication, economy, efficiency, effectiveness

Concepts of language economy

Language economy is certainly one of the most interesting and heavily discussed linguistic phenomena. Thereby, modern linguistic concepts of language economy are corresponding in view of a distinction between something like expense (or cost) and something like result (or profit) of linguistic communication. But further on these concepts are quite different, whereby the scientific discussion of language economy mainly focuses on the following aspects:

- definition of linguistic expense and linguistic result itself,
- relation between linguistic expense and linguistic result,
- synchronic and diachronic aspects of language economy,
- systematic and pragmatic aspects of language economy.

A view on some language economy concepts shows the following definitions of linguistic expense and linguistic result (in historical order; cf. Roelcke (to appear) and Table 1): The ease-theory by Jespersen (1922; 1941) or by Martinet (1963) grasps the linguistic expense as a linguistic system and the linguistic result as a success in communication and this way takes a reasonable large scope of linguistic economy. Language statistics developed by Zipf (1935; 1949) distinguish between the frequency of words on the one hand and their rank of frequency on the other, and this way reduce the scope of linguistic economy to the lexicon. We also find this reduction of scope to the lexicon in the terminology work in the tradition of Wüster (1931/1970) who defines the linguistic expense as shortness and comprehensibility and the linguistic result as difference and exactness of terminological items. Within the principles of conversation Grice understands the linguistic expense in a wide interpretation as way of conversation, the result as transfer of information (cf. Grice 1989). A large scope of language economy is also stated by Moser's concept of history of language and culture (Moser 1970; 1971), where the linguistic expense is defined as linguistic system and variety and the

[1] Address correspondence to: Thorsten Roelcke, Riggenbacher Weg 3, D-79872 Bernau im Schwarzwald. E-mail: roelcke@t-online.de

linguistic result as specific transfer of information. In cognitive semantics the scope of linguistic economy is reduced to the lexicon again; following Rosch (1977; 1978) the expense is constituted by semantic prototypicalization, the result, however, by cognitive categorization. It is not surprising that the scope of economy in linguistic synergetics is reasonably wide and embraces quality and quantity of linguistic items as expense and lexical or syntactic information as result (cf. Köhler 1986; 1999). In the model of textual, especially lexico-graphic condensation described by Wolski (1989) or Wiegand (1998) the linguistic expense and the linguistic result occur as microstructures of dictionaries on the one hand and as lexicographical information on the other. In a wide interpretation another model of language change connects expense with iconicity and frequency of linguistic items and result with linguistic performance (cf. Werner 1989; 1991; Ronneberger-Sibold). One of the most famous concepts of language economy today is Chomsky's minimalist program (cf. Chomsky 1995): In the minimalist program linguistic elements and principles of linguistic construction and interpretation are understood as the economic extension of language while the economic result of

Table 1
Linguistic concepts of language economy (overview)

Concept	Expense	Result	Level	Principle	Aspect
Ease theory (Martinet; Jespersen)	linguistic system	success in communication	language as a whole (system)	minimization	diachronic
Language statistics (Zipf)	rank of frequency	frequency of words	lexical level (text)	minimization	synchronic
Terminology work (Wüster)	shortnes, comprehensibility	difference, exactness	lexical level (system)	minimization	synchronic
Principles of conversation (Grice)	way of conversation	transfer of information	language as a whole (text)	minimization/ maximization	synchronic
Language and culture (Moser)	linguistic system and variety	specific transfer of information	language as a whole (system)	minimization/ maximization	diachronic
Cognitive semantics (Rosch)	semantic prototypicalization	cognitive categorization	lexical level (system)	minimization/ maximization	synchronic
Linguistic synergetics (Köhler)	quality/quantity of ling. units	lexical/syntactic information	language as a whole (system)	minimization	synchronic
Lexicography (Wiegand; et al.)	microstructure of dictionaries	lexicographical information	textual level (text)	minimization	synchronic
Language change (Werner; et al.)	iconicity and frequency	linguistic performance	language as a whole (system)	minimization	diachronic
Minimalist program (Chomsky)	ling. elements and principles	fulfilment of requirements	syntactic level (system)	minimization	synchronic
Optimality theory (Prince/Smolensky)	regulation and violation	fulfilment of requirements	phonemic level (system)	minimization	synchronic

language is connected with the fulfilment of requirements of human communication. In a further development of minimalist ideas in a phonemic context the optimality theory by Prince and Smolensky (1997) defines the linguistic expense in a narrower sense as regulation and violation of hierarchical rules and the linguistic result again as fulfilment of requirements of human communication.This short (and not complete) overview shows that the scientific concepts of linguistic expense and linguistic result both, intensionally and extensionally, are very different and concern the language either as a whole (ease theory, principles of conversation, language and culture, linguistic synergetics, and language chance) or their phonemic (optimality theory), syntactic (minimalist program), lexical (language statistics, terminology work, and cognitive semantics) or textual level (lexicography). However, these concepts not only differ in the definition of linguistic expense and linguistic result itself, but also in setting a relationship between linguistic expense and linguistic result (cf. Table 1 again): Some concepts of language economy only develop a vague relationship between expense and result (cf. Martinet 1963, 165f.). Other concepts take language economy for granted, if a minimum of expense is combined with a maximum of result (cf. Jespersen 1941: 6; Searle 1971: 50; Sperber/Wilson 1986, vii; Wurzel 2001: 384ff.). In a more scientific way of argumentation this so called *mini/max-principle* has to be rejected, because it amounts to a somehow theological or philosophical, but certainly not linguistic concept of somewhat of a creatio ex nihilo. Most linguistic concepts of language economy prefer minimization of expense with regard to a particular result as major economic principle (cf. Zipf 1935; 1949; Wüster 1931/1970; Martinet 1963: 164; Schmidt 1972: 54, 162; Ronneberger-Sibold 1980: 3; van der Elst 1984: 324; Köhler 1986; 1999; Wolski 1989; Wiegand 1998; Werner 1989; 1991; Ronneberger-Sibold 1980; Chomsky 1995; Prince/ Smolensky 1997). The terminological result of this preference in linguistics is expressions like *principle of least effort* (Zipf 1949) or *economy of effort* (Jespersen 1922, 261). A maximization of result with regard to a particular expense is known as economic principle in linguistics, too (cf. Grice 1989; Moser 1970: 9; 1971; Rosch 1977; 1978; Wurzel 1997: 305f.). But some scientists express (without sufficient explanation) scepticism about this principle (cf. Wilder/Gärtner 1997: 2; Ronneberger-Sibold 1980: 241). As to these differences some linguistic concepts of language economy prefer a synchronic point of view (*synchronic concepts:* language statistics, terminology work, rules of conversation, cognitive semantics, linguistic synergetics, lexicography, minimalist program, and optimality theory), others a diachronic point of view (*diachronic concepts:* ease theory, language and culture, and language change). And last, but not least some concepts are interested in the linguistic system (*systematic concepts*: ease theory, terminology work, language and culture, cognitive semantics, linguistic synergetics, language change, minimalist program, optimality theory), others in linguistic texts (*textual concepts*: language statistics, rules of conversation, lexicography). Unfortunately the textual concepts of language economy consider the text itself, but do ignore the producers and the recipients of the text; this way, they can only be characterized as *textual* and cannot be characterized as *pragmatic* concepts of language economy.

To sum up we can say that the concepts of language economy differ in definition and relation of linguistic expense and linguistic result. Further they are either synchronic or diachronic and either systematic or textual (but not pragmatic) concepts. This conceptual situation is not satisfactory and demands a holistic (and atomistic) model of language economy. Such a model has to consider the following points:

- holistic definition of linguistic expense and linguistic result itself,
- holistic relation of linguistic expense and linguistic result in communication,
- synergetic conjunction between synchronic and diachronic aspects,
- consideration of producer and recipient of linguistic communication.

A new concept of language economy is sketched in the following paragraphs. To fulfil the holistic demands it is based on a plain model of communication and the distinction between effectiveness and efficiency as known from economics.

Communication, effectiveness and efficiency

The sketched concept of language economy is based on a plain model of communication (cf. Roelcke 1994: 7-23; 1999: 15-31) that contains the following elements and relations (cf. figure 1): The producer of a certain text (*producer T*), the text itself (*text*) and the recipient of this text (*recipient T*); the producer not only produces the text but also controls this production while the recipient not only adopts the text, but also interprets it. In a communicative dialogue the recipient of the text may produce and control an answering text (*answer*) and appear to be a producer (*producer A*) being opposed to the interpreting recipient of the answer (*recipient A*). The producer T (recipient A) and the recipient T (producer A) have their own linguistic systems at their disposal (*system P* resp. *system R*); these linguistic systems partly coincide (*mutual system*) and this way constitute the basis of communication between producer and recipient. Both producer and recipient interact in a social and cultural context (*context P* resp. *context R*) and in a textual context (*cotext P* resp. *cotext R*) as well; the communication of producer and recipient takes place in a mutual context as well as a mutual cotext (*mutual cotext*). – The linguistic investigation of the mutual system and system P and system R resp. represents a systematic point of view. However the investigation of the text and the cotext stands for a textual, and the investigation of the context for a pragmatic interest. And the investigation of the producer and the recipient itself is the approach of cognitive sciences. The sketched concept of language economy resp. efficiency of communication takes economy or efficiency as a cognitive principle that appears in linguistic systems and in linguistic texts as well. Hence below three models of efficiency of communication are distinguished: a general one, a systematic one and a textual one.

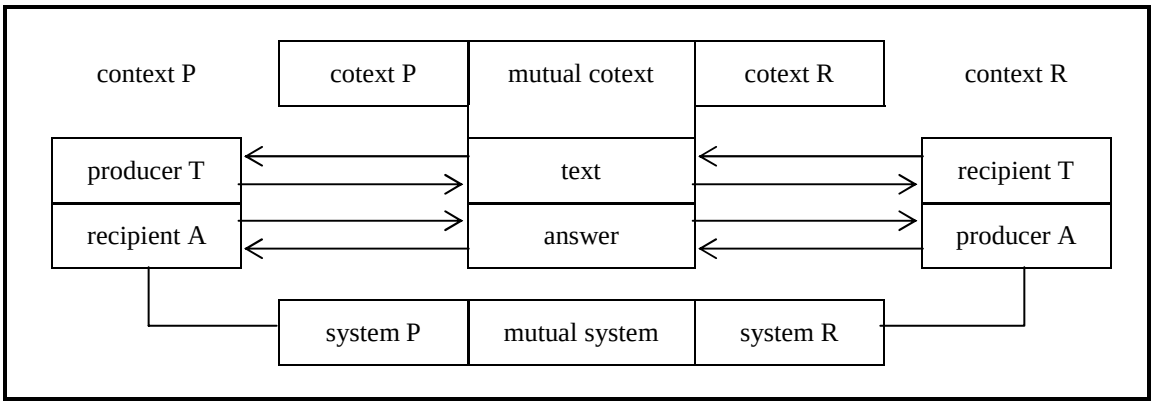

Figure 1. Model of communication (cf. Roelcke 1994, 16)

A scientific investigation of language economy has to reject the everyday concept of a vague relationship between expense and result and the pre-scientific concept of combining a minimization of expense and a maximization of result. A more appropriate approach can be seen in the economic distinction between effectiveness and efficiency of technical processes or human actions. According to this, we have to define as follows (cf. figure 2): A human action is effective, if a particular result (independent of the extent of expense) is obtained. However, if such a specific result (independent of the extent of expense again) is not

obtained, the human action in question is ineffective. A human action is efficient either, if a particular result is obtained with a minimum of expense, or, if a particular expense is combined with a maximum of result; we can call the first case *efficiency of expense*, the second *efficiency of result*. However, if a particular result is obtained without a minimum of expense, or, if a particular expense is not combined with a maximum of result, the human action in question may be effective but is inefficient anyway. These definitions ignore the faculty and the concentration of the human being itself. In an economic or technical context this idealization is possible (and allowed), in a mental or psychic one it is not. In this context the cognitive capacity of producer and recipient has to be considered, because it differs from person to person and this way determines the effectiveness and the efficiency of communicative actions.

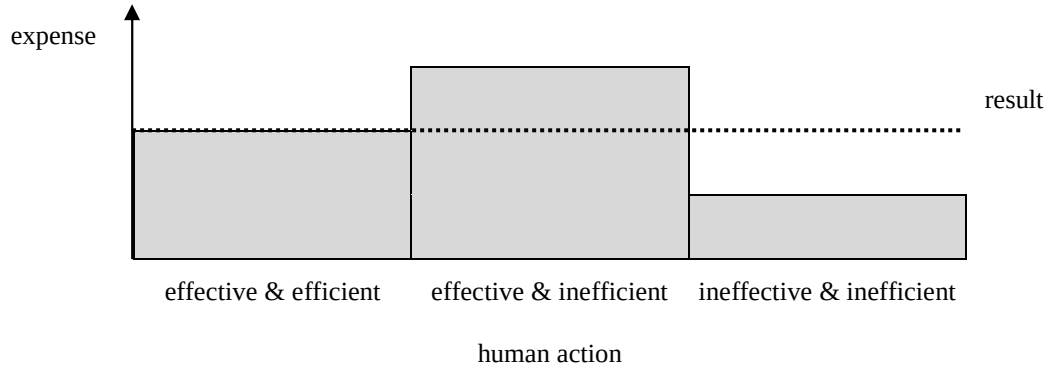

Figure 2. Effectiveness and efficiency of human actions

Efficiency of communication: elements

Efficiency of communication has to be described as a function of the following elements: intension and extension, competence and concentration, communicat and communicant, and complexity and capacity. These elements will be defined in a general, in a systematic, and in a textual way (cf. Table 2). The general terms are hyperonyms with regard to the systematic and the textual terms. The systematic and textual terms therefore are hyponyms with regard to the general terms and cohyponyms with regard to each other.

To define these elements in general first: The communicative *intension* is the cognitive result of communication and grasps its linguistic information and instruction in general; it is assumed that both, information and instruction, are variable and show various degrees. The communicative *extension* is the cognitive expense of communication and embraces its linguistic elements and relations in general; it is assumed that both elements and relations are variable and show various degrees, too. Information and instruction on the one hand and elements and relations on the other may be conceptually united to the *communicat* (term according to Roelcke 2002) of linguistic communication in general: This communicat shows a scale of degrees again meaning a variation of *complexity*, that itself depends on different degrees of information and instruction resp. elements and relations. The communicative *competence* consists of the intelligence and the instruments for linguistic communication in general; it is assumed that both, intelligence and instruments, are variable. Recently communicative *concentration* means intention and interest of linguistic communication with various degrees in general. As intension and extension, competence and concentration may be united to the *communicant* (term according to Roelcke 2002) of linguistic communication. This communicant also shows a scale of degrees which in this case means a variation of

capacity that itself depends on different degrees of intelligence and instruments resp. intention and interest.

Table 2
Elements of efficient communication (overview)

	General	Systematic	Textual
Intension	information/instruction	categorization/schematization	proposition/illocution
Extension	elements/relations	inventory/principles	words/sentences
Competence	intelligence/instruments	(ethic) nature	capability of production resp. reception
Concentration	intention/interest	(ethnic) culture	readiness of production resp. reception
Communicat	information/instruction & elements/relations	categorization/schematization & inventory/principles	proposition/illocution & words/sentences
Complexity	degree of information/ instruction & elements/ relations	degree of categorization/ schematization & inventory/ principles	degree of proposition/ illocution & words/sentences
Communicant	intelligence/instruments & intention/interest	(ethic) nature & (ethnic) culture	capability & readiness of production resp. reception
Capacity	degree of intelligence/ instruments & intention/ interest	degree of (ethic) nature & (ethnic) culture	degree of capability & readiness of production resp. reception

The systematic definitions of the elements of efficiency of communication follow the given definitions in general: From a systematic point of view the communicative intension consists of the categorization and schematization of real objects and processes; and here again it is assumed that categorization and schematization of *systematic intension* are variable and show various degrees. The *systematic extension* is constituted by the lexical inventory and the syntactic, semantic and pragmatic principles of combining lexical items to more complex signs; it is also assumed that both this inventory and these principles are variable and show various degrees. Categorization and schematization on the one hand and inventory and principles on the other are to be conceptually united to the *systematic communicat* of linguistic communication. This systematic communicat shows various degrees again, in other words a variable *systematic complexity*, that itself depends on different degrees of categorization and schematization resp. inventory and principles. The *systematic competence* is the ethic nature of human communities independent of their particular social and cultural contexts; and it is not only caused by political correctness but also by anthropologic knowledge to assume that this systematic competence is more or less constant, if the communities in question show a certain quantity of population. The *systematic concentration* is determined by the social and cultural context a certain community lives in; this way it is certainly variable and shows different degrees in differing contexts. Systematic competence and systematic concentration have to be united to a *systematic communicant* of linguistic communication. And this

systematic communicant shows different degrees, too, in other words a variation of *systematic capacity* that depends on different varieties of ethnic culture (not ethic nature) itself.

The textual definitions of the elements of efficiency of communication again follow the definitions in general and are parallel to the given systematic ones. First the *textual intension* is constituted by the propositions and illocutions connected with particular linguistic utterances and – like the systematic intension – it is variable. The *textual extension*, however, consists of the words and sentences that constitute these utterances themselves; like the systematic extension this textual extension shows various degrees, too. Both propositions and illocutions and words and sentences may be conceptually united to the *textual communicat* of linguistic communication; this textual communicat shows various degrees resp. a variable *textual complexity*, i.e. it depends on different degrees of propositions and illocutions resp. words and sentences themselves. The *textual competence* is the personal capability of production or reception of linguistic utterances and differs from person to person. Beside this textual competence there is a personal and situative differing readiness of production or reception, the *textual concentration*. The capability and the readiness of production or reception have to be united to the *textual communicant* of linguistic communication. And in the end this textual communicant also shows a variable *textual capacity*, in other words various degrees depending on different degrees of the capability and the readiness of linguistic production resp. reception themselves.

Efficiency of communication: relations

After introducing the elements needed to grasp efficiency of communication the relations between these elements have to be described. As above firstly the relations in general and then the systematic and textual relations are considered.

In general, efficiency of communication depends on the complexity of the communicat resp. its intension and extension on the one hand and the capacity of the communicant resp. its competence and concentration on the other (cf. figure 3). In case of efficient communication there is a more or less equal degree of complexity and capacity; in other words communication is efficient, if intension and extension of the communicat and competence and concentration of the communicant are in a more or less balanced relation. However, if complexity is of a higher degree than capacity, communication is neither efficient nor effective; and if capacity is of a higher degree than complexity, communication is effective but not efficient. The complexity of the communicat itself depends on the degree of intension and the degree of extension, in which a high degree of information and instructions in comparison with elements and relations causes a high degree of communicative complexity. The capacity of the communicant, however, depends on the degree of competence and the degree of concentration, in which – compared to one another – a high degree of intelligence and instruments or a high degree of intention and interest causes a high degree of communicative capacity, too. According to these considerations we come to the following statements:

- *Communicative efficiency* in general exists, if the degree of information and instruction of communicative intension and the degree of elements and relations of communicative extension on the one hand and the degree of intelligence and instruments of communicative competence and the degree of intention and interest of communicative concentration on the other are of a more or less equal degree in toto.
- *Communicative ineffectiveness* in general exists, if the degree of information and instruction of communicative intension and the degree of elements and relations of communicative extension on the one hand are higher in toto than the degree of

intelligence and instruments of communicative competence and the degree of intention and interest of communicative concentration on the other.

- *Communicative inefficiency* in general exists, if the degree of information and instruction of communicative intension and the degree of elements and relations of communicative extension on the one hand are lower in toto than the degree of intelligence and instruments of communicative competence and the degree of intention and interest of communicative concentration on the other.

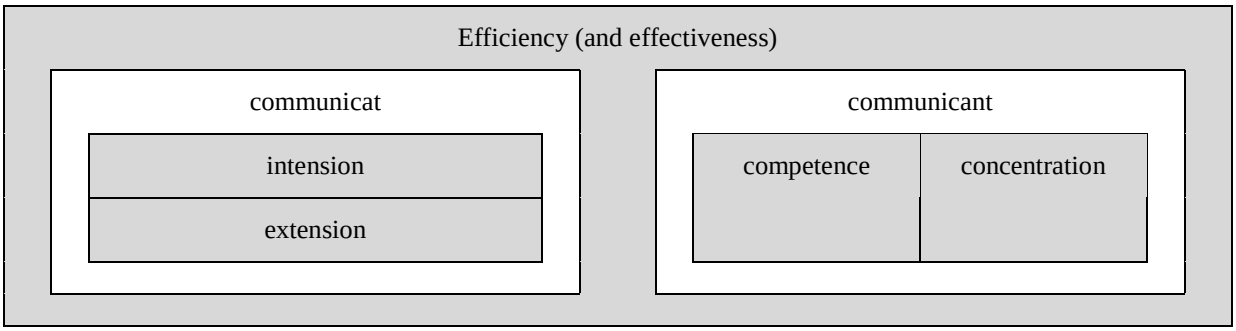

Figure 3. Intension, extension, competence, and concentration in efficiency of communication (overview)

At the level of language system efficiency of communication depends on the complexity of systematic intension and extension on the one hand and the capacity of systematic competence and concentration on the other. In case of efficient communication there is a more or less equal degree of intension and extension of systematic complexity and competence and concentration of systematic capacity. Communication is neither efficient nor effective, if systematic complexity is of a higher degree than systematic capacity; and it is effective but not efficient, if systematic capacity is of a higher degree than systematic complexity. The complexity of the systematic communicat depends on the degree of systematic intension and the degree of systematic extension, in which – compared to the degree of inventory and principles – a high degree of categorization and schematization causes a high degree of systematic complexity. The capacity of the systematic communicant depends on the degree of systematic competence and the degree of systematic concentration, in which – compared to one another – a high degree of ethic nature or a high degree of ethnic culture causes a high degree of systematic capacity. According to these considerations the following statements may be permitted:

- *Systematic efficiency* exists, if the degree of categorization and schematization of communicative intension and the degree of inventory and principles of communicative extension on the one hand and the degree of ethic nature of communicative competence and the degree of ethnic culture of communicative concentration on the other are more or less equal in toto.
- *Systematic ineffectiveness* exists, if the degree of categorization and schematization of communicative intension and the degree of inventory and principles of communicative extension on the one hand are in toto higher than the degree of ethic nature of communicative competence and the degree of ethnic culture of communicative concentration on the other.
- *Systematic inefficiency* exists, if the degree of categorization and schematization of communicative intension and the degree of inventory and principles of communicative

extension on the one hand are in toto lower than the degree of ethic nature of communicative competence and the degree of ethnic culture of communicative concentration on the other.

Finally, at the level of linguistic texts efficiency of communication depends on the complexity of textual intension and extension and the capacity of textual competence and concentration. In case of efficient communication there is also a more or less equal degree of intension and extension of textual complexity and competence and concentration of textual capacity. If textual complexity is of a higher degree than textual capacity, communication is again neither efficient nor effective; and if textual capacity is of a higher degree than textual complexity, communication is effective but not efficient. The complexity of the textual communicat depends on the degree of textual intension and the degree of textual extension, in which – compared to the degree of words and sentences – a high degree of propositions and illocutions causes a high degree of textual complexity. The capacity of the textual communicat depends on the degree of textual competence and the degree of textual concentration, in which – compared to one another – a high degree of capability of production resp. reception or a high degree of readiness of production resp. reception causes a high degree of textual capacity. According to these considerations we come to the following statements:

- *Textual efficiency* exists, if the degree of propositions and illocutions of communicative intension and the degree of words and sentences of communicative extension on the one hand and the degree of capability of production resp. reception of communicative competence and the degree of readiness of production resp. reception of communicative concentration on the other are more or less equal in toto.
- *Textual ineffectiveness* exists, if the degree of propositions and illocutions of communicative intension and the degree of words and sentences of communicative extension on the one hand are in toto higher than the degree of capability of production resp. reception of communicative competence and the degree of readiness of production resp. reception of communicative concentration on the other.
- *Textual inefficiency* exists, if the degree of propositions and illocutions of communicative intension and the degree of words and sentences of communicative extension on the one hand are in toto lower than the degree of capability of production resp. reception of communicative competence and the degree of readiness of production resp. reception of communicative concentration on the other.

Final remarks

To sum up: In this paper a new concept of language economy is sketched in four steps: 1st some linguistic concepts of language economy (for example ease theory, language statistics, terminology work, rules of conversation, language and culture, cognitive semantics, linguistic synergetics, lexicography, language change, minimalist program, and optimality theory) are discussed with regard to definitions and relations of linguistic expense and linguistic result and synchronic resp. diachronic and systematic resp. pragmatic aspects. 2nd it is shown that linguistics needs a new and holistic concept of language economy based on a plain model of communication and the economic distinction between effectiveness and efficiency of technical processes and human actions. 3rd the general, systematic and textual elements of this model of efficiency of communication are introduced (intension and extension, competence and concentration, communicat and complexity, and communicant and capacity). 4th the relationships between these elements are discussed and it is shown that efficiency of

communication exists, if the communicative complexity of intension and extension and the communicative capacity of competence and concentration are of a more or less equal degree. And in comparison with communicative capacity higher communicative complexity, however, causes communicative ineffectiveness, a lower complexity communicative inefficiency. – In future at least four desiderata are to are to be fulfilled:

- a conceptual expansion of the model with regard to a connection of systematic and textual efficiency and a separation of production and reception within linguistic communication,
- a mathematical formalization of the model to obtain scientific laws of language economy resp. efficiency of communication,
- an empirical exemplification of the model to obtain more empirical details of language economy resp. efficiency of communication,
- a practical application of the model with regard to language criticism or language didactics from an economic resp. efficient point of view.

References

Chomsky, Noam (1995). *The Minimalist Program*. Cambridge, Mass.: MIT Press.

Grice, H. Paul (1989). *Studies in the Way of Words*. Cambridge, Mass.

Jespersen, Otto (1922). *Language. Its Nature, Development, and Origin*. London: Allen & Unwin.

Jespersen, Otto (1924). *The Philosophy of Grammar*. London: Allen & Unwin.

Jespersen, Otto (1941). Efficiency in Linguistic Change. In: *Historisk-Filologiske Meddelelser*, udgivet af det Kgl. Danske Videnskabernes Selskab *27, 4, 1-90.*

Koenraads, Willy Henri August (1953). *Studien über sprachökonomische Entwicklungen im Deutschen*. Amsterdam: Meulenhoff.

Köhler, Reinhard (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik* (= Quantitative Linguistics 31). Bochum: Brockmeyer.

Köhler, Reinhard (1999). Syntactic Structures: Properties and Interrelations. *Journal of Quantitative Linguistics 6, 46-57.*

Köhler, Reinhard, Altmann, Gabriel (2000). Probability Distributions of Syntactic Units and Properties. In: *Journal of Quantitative Linguistics 7, 189-200.*

Martinet, André (1963). *Grundzüge der Allgemeinen Sprachwissenschaft.* Autorisierte, vom Verfasser durchgesehene Übersetzung aus dem Französischen von Anna Fuchs, unter Mitarbeit von Hans-Heinrich Lieb. Stuttgart: Kohlhammer [Original: Eléments de linguistique générale. Paris 1960].

Moser, Hugo (1970). Sprachliche Ökonomie im heutigen deutschen Satz. In: *Studien zur Syntax des heutigen Deutsch. Paul Grebe zum 60. Geburtstag:* 9-25. Düsseldorf: Schwann,

Moser, Hugo (1971). Typen sprachlicher Ökonomie im heutigen Deutsch. In: *Sprache und Gesellschaft. Beiträge zur soziolinguistischen Beschreibung der deutschen Gegenwartssprache.* Düsseldorf: Schwann, 89-117.

Prince, Alan, Smolensky, Paul (1993). *Optimality Theory: Constraint Interaction in Generative Grammar.* Ms. Rutgers University, New Brunswick and University of Colorado, Boulder.

Prince, Alan, Smolensky, Paul (1997). Optimality: From neutral Networks to Universal Grammar. *Science* 275, 1604-1610.

Roelcke, Thorsten (1994). *Dramatische Kommunikation. Modell und Reflexion bei Dürrenmatt, Handke, Weiss* (= Quellen und Forschungen zur Sprach- und Kulturgeschichte der germanischen Völker). Berlin, New York: de Gruyter.

Roelcke Thorsten (1999). *Fachsprachen* (= Grundlagen der Germanistik 37). Berlin: Schmidt.

Roelcke, Thorsten (2002). *Kommunikative Effizienz. Eine Modellskizze* (= Sprache – Literatur und Geschichte 23). Heidelberg: Winter.

Roelcke, Thorsten (to appear). Sprachliche Ökonomie / Kommunikative Effizienz. In: Gabriel Altmann, Reinhard Köhler, Rajmund Piotrowski (eds.), *Quantitative Linguistik / Quantitative Linguistics. Ein internationales Handbuch / An International Handbook* (= Handbücher zur Sprach- und Kommunikationswissenschaft). Hrsg. von. Berlin, New York: de Gruyter.

Ronneberger-Sibold, Elke (1980). *Sprachverwendung — Sprachsystem. Ökonomie und Wandel*. Tübingen: Niemeyer.

Ronneberger-Sibold, Elke (1997). Sprachökonomie und Wortschöpfung. In: Thomas Birkmann, Heinz Klingenberg, Damaris Nübling und Elke Ronneberger-Sibold. (eds.), *Vergleichende germanische Philologie und Skandinavistik. Festschrift für Otmar Werner: 249-261*. Tübingen: Niemeyer.

Rosch, Eleanor (1977). Human Categorization. In: Neil Warren *(ed.), Studies in Cross-Cultural Psychology, Volume I, 1-49*. London: Academic Press.

Rosch, Eleanor (1978). Principles of Categorization. In: Eleanor Rosch and Barbara B. Lloyd (eds.), *Cognition and Categorization: 27-48*. Hillsdale, N.J.: Erlbaum.

Searle, John R. (1971). What is a Speech Act? In: John R. Searle (ed.), *The Philosophy of Language: 38-52*. Oxford.

Sperber, Dan, Wilson, Deidre (1986). *Relevance. Communication and Cognition*. Oxford: Blackwell.

Werner, Otmar (1989). Sprachökonomie und Natürlichkeit im Bereich der Morphologie. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 42, 34-47.*

Werner, Otmar (1991). Sprachliches Weltbild und / oder Sprachökonomie. In: *Begegnung mit dem ‚Fremden'. Grenzen — Traditionen — Vergleiche. Akten des VIII. Internationalen Germanisten-Kongresses, Tokyo 1990: 305-315*. Hrsg. von Eijiro Iwasaki. Band 4. Hrsg. von Yoshinori Shichiji. München: Judicium.

Wiegand, Herbert Ernst (1996). Textual Condensation in Printed Dictionaries. A Theoretical Draft. *Lexikos 6, 133-158.*

Wiegand, Herbert Ernst (1998): Lexikographische Textverdichtung. Entwurf zu einer vollständigen Konzeption. In: Arne Zettersten, Viggo Hjørnager Pedersen and Jens Eric Mogensen (eds.), *Symposium on Lexicography VIII. Proceedings of the Eighth International Symposium on Lexicography, May 2-6 1996, at the University of Copenhagen: 1-35*. Tübingen: Niemeyer.

Wilder, Chris, Gärtner, Hans-Martin (1996). Introduction. In: Chris Wilder, Hans-Martin Gärtner and Manfred Bierwisch (eds.), *The Role of Economy Principles in Linguistic Theory: 1-35*. Berlin: Akademie Verlag.

Wolski, Werner (1989). Formen der Textverdichtung im allgemeinen einsprachigen Wörterbuch. In: Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand und Ladislav Zgusta *(eds.), Wörterbücher / Dictionaries / Dictionnaires. Ein internationales Handbuch zur Lexikographie*: *956-967*.. [...]. 3 Teilbände. Berlin/New York: de Gruyter.

Wurzel, Wolfgang Ullrich (1994). *Grammatisch initiierter Wandel* (= Prinzipien des Sprachwandels 1). Unter Mitarbeit von A. und D. Bittner. Bochum: Brockmeyer.

Wurzel, Wolfgang Ullrich (1997). Natürlicher Grammatischer Wandel, ‚unsichtbare Hand´ und Sprachökonomie — Wollen wir wirklich so Grundverschiedenes? In: Thomas Birk-

mann, Heinz Klingenberg, Damaris Nübling, Elke Ronneberger-Sibold (eds.), *Vergleichende germanische Philologie und Skandinavistik. Festschrift für Otmar Werner: 295-308*. Tübingen: Niemeyer.

Wurzel, Wolfgang Ullrich (2001). Ökonomie. In: Martin Haspelmath, Ekkehard König, Wulf Oesterreicher, Wolfgang Raible (eds.),. *Language Typology and Language Universals / Sprachtypologie und sprachliche Universalien / La typologie des langues et les universeaux linguistiques. An International Handbook / Ein internationales Handbuch / Manuel international* (= Handbücher zur Sprach- und Kommunikationswissenschaft 20), *Vol.1, 384-400.* Berlin, New York: de Gruyter.

Wüster, Eugen (1931/1970). *Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik. (Die nationale Sprachnormung und ihre Verallgemeinerung).* Dritte, abermals ergänzte Auflage. Bonn: Bouvier, 1970 [1. Aufl. 1931].

Zipf, George Kingsley (1935). *The Psycho-Biology of Language. An Introduction to Dynamic Philology.* Boston: Houghton-Mifflin [2nd ed. Cambridge, Mass.: MIT Press 1968].

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Cambridge, Mass.: Addison-Wesley [2nd ed. New York: Hafner 1972].

Glottometrics 4, 2002, 39-44 *To honor G.K. Zipf*

Power laws: from *Alvarez* to *Zipf*

Manfred Schroeder[1]

This article is a reprint of a chapter from M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise* (33-38), New York: Freeman 1991, published here with the kind permission of the author and the Freeman Publishing House. The book, very popular among linguists, makes us realize that we are not alone in the universe of sciences but have joined with all other disciplines at least due to the omnipotent power law which, almost everywhere, bears the name of the linguist G.K. Zipf. At the same time it shows us how physicists look at language (G.A.).

Homogeneous power laws, like Newton's universal law of graviational attraction $F \sim r^{-2}$ abound in nature – dead and alive alike. Since homogeneous power laws, upon rescaling, remain homogeneous power laws with the same exponent (-2 in Newton's case), such laws are, by definition, self-similar. In other words, Newton's law is *true on all scaIes,* from the wavelength of light to light-years; it has no built-in scale of its own. Newton's gravitational universe, if we so wished, could be compressed or inflated at will.[2]

The same inverse square law that governs gravitation also describes the fall off of radar power with distance. This simple fact was exploited by German submarines during World War II. By measuring the increase in radar intensity, they could gauge the rate of approach of an enemy plane and dive undersea for safety before the plane could attack.

This tactic worked very well for Grand Admiral Karl Dönitz until the American physicist, Luis Alvarez (1911-1988) had a foxy vision, code-named Vixen. Alvarez suggested reducing the radar power so that it would be proportional to the *third* power of the range of the submarine. Thus, while the plane was approaching, the power incident on the unsuspecting U-boat was actually *decreasing*, giving the false impression that the radar plane was flying *away.* A grand idea indeed! (For the attacking plane, however, the *received* radar power inflected from the boat would still increase as it closed in Alvarez 1987[3]).

Another wide-ranging example of a homogeneous law is the one that connects the areas A of similar plane figures with their diameters, their perimeters, or any other of their linear dimensions l: areas are proportional to linear dimensions squared, or $A \sim l^2$. Of course, this is not true for areas an curved surfaces; the radius of curvature introduces a length scale that destroys "truth on all scales." In fact, as everyone knows, distances and areas on the surface of a sphere are limited to a maximum size, given by the radius of the sphere.

---

[1] Address correspondence to: Manfred R. Schroeder, University of Göttingen, Dept. of Physics, Buergerstrasse 42-44, D-37073 Göttingen, Germany. E-mail: mrs17@aol.com

[2] Recently, though, some doubt has been cast on the unlimited validity ot Newton's law. A still mysterious "fifth force" appears to knock on Newton's underpinnings, adding terms that introduce a natural length scale of a few hundred meters (Ander et al. 1989). At very small scales, Newton's law runs into the *Planck length* (10^{-35} m), which reminds us that eventually gravitation needs to be properly quantized and endowed with uncertainty.

[3] This scheme of Alvarez is somewhat reminiscent of Genghis Khan ("Universal Ruler") and the wily Mongol tactic perfected by the horsemen of the Golden Horde. While seeming]y galloping away from their pursuers, they would actually allow them to close in and then suddenly stand up in their stirrups, turn around in their saddles, and launch their arrows at the dumbfounded enemy.

In contrast to gravitation, interatomic forces are typically modeled as *in*homogeneous power laws with at least two different exponents. Such laws (and expontential laws, too) are not scale-free; they necessarily introduce a characteristic length, related to the size of the atoms.

Power laws also govern the power spectra of all kinds of noises, most intriguing among them the ubiquitous (but sometimes difficult to explain) 1/f noise. Thus, the noise in many semiconductor devices is not "white" (i.e., independent of frequency) and not "brown" (with a $1/f^2$ frequency dependence, like Brownian motion), but has an in-between exponent, which is why it is sometimes called *pink noise*. Pink noise is also a preferred test signal in auditory research, because it has constant power per *octave* (not per hertz) and is thus well matched to the inner ear's frequency scale.

And, as we shall see in the course of our excursion into the world of fractals, power-law exponents do not have to be integers; they can be, and often are, *fractions*.

Not surprisingly, we find homogeneous power laws not only in the inanimate world; they inhabit living nature, and particularly human perception, too. Thus, over much of the auditory amplitude range, subjective loudness L is proportional to the physical sound intensity I raised to the three-tenths power: $L \sim I^{0.3}$. This means that merely to double the loudness of a rock group of five musicians, say, we have to increase their number *tenfold***,** to 50 players of equal power output. (This minor calculation explains the resounding enamoration of popular music makers with electronic amplifiers.)

By the same token, if we want to halve the loudness of a continuous "rumble" emanating from a busy highway, we have to reduce the acoustic noise output by a factor of ten! This may sound difficult, but it is not, at least not from a purely physical point of view: tire noise – the main culprit at steady highway speeds – decreases drastically with decreasing vehicle speed. In fact, the noise intensity is approximately proportional to the *fourth* power of speed.

On the other hand, a tenfold increase in the average intensity of traffic noise caused by a tenfold increase in traffic densify can raise the rate of complaints by irate residents perhaps a *hundred* fold: one loud truck every 5 minutes may be tolerable, but one every 30 seconds could be a nightmare and would certainly make outdoor conversation nearly impossible. And what is true for trucks is just as true for low-flying aircraft.

Power laws are also ubiquitous in economics. In fact, nearly 100 years ago, the Italian economist Vilfredo Pareto (1848-1923), working in Switzerland, found that the number of people whose personal incomes exceed a large value follows a simple power law (Pareto 1896, Mandelbrot 1963a). Other instances of power laws in economics and the fallacies of trading schemes based on them are discussed by Mandelbrot (1963b,c).

One of the more surprising instances of a power law in the humanities is *Zipf's law* connecting *word rank* and *word frequency* for many natural languages. (The word with rank r is the rth word when the words of a language are listed with decreasing frequency.) This law, enunciated by George Kingsley Zipf (1902-1950), states that, to a very good approximation, relative word frequency f in a given text is inversely proportional to word rank r:

$$f(r) \approx \frac{1}{r \ln(1.78\,R)}$$

where R is the number of different words (Zipf 1949). Laws like $f(r) \sim 1/r$ are called *hyperbolic laws*. lf we assume $R = 12{,}000$, for example, we find that the relative frequencies of the highest-ranking words (*the, of, and, to,* and so on, in order of rank) are approximately 0.1, 0.05, 0.033, 0.025, and so on.

Figure 19 shows the close match between Zipf's homogeneous power law and actual data. Claude Shannon, the creator of information theory, has used Zipf's law to calculate the

entropy of a source of English text that sputters words independently with Zipf's probabilities (Shannon 1951). This entropy is given approximately by

$$H = \frac{1}{2}\log_2(2R \ln 2R) \qquad \text{bits per word.}$$

For R = 12,000, we get $H \approx$ 9 bits per word, while R = 300,000 yields an entropy of about 11.5 bits per word. Of course, this is only an upper bound, because words (though perhaps independent of actions) are not independent of each other – except in random "poetry." This interdependence of words ("redundancy") in a meaningful text, of course, reduces the entropy.

Considering that the average length of English words is about 4.5 letters, or 5.5 "characters" including one space between words, we see that the entropy of English text is roughly bounded by 2 bits per character.

Zipf's hyperbolic law, which is applicable not only to the language as such but also to individual writers, has some rather curious consequences. To wit, for a good writer with an active vocabulary of R = 100,000 words, the 10 highest-ranking words occupy 24 percent of a text, while for basic (newspaper?) English with one-tenth the vocabulary (R = 10,000), this percentage barely increases (to about 30 percent). Of course, any writer would find it difficult to avoid words like *the, of, and*, and *to*.

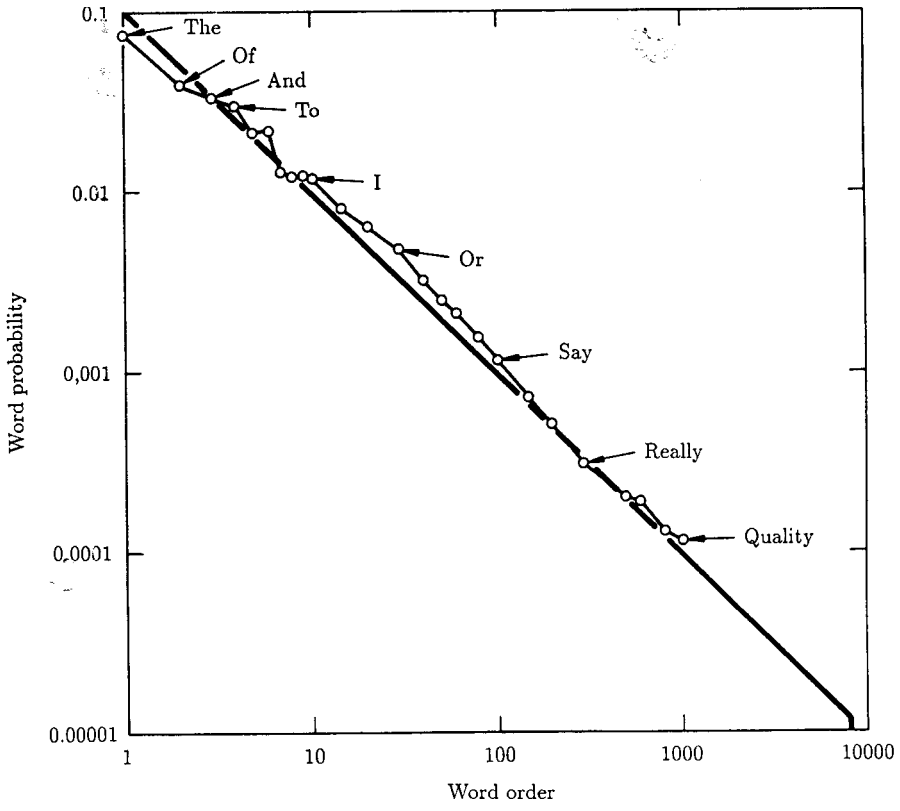

Figure 19. Word frequency as a function of word rank follows Zipf's law.

Zipf has endeavored to derive his law from *Human Behavior and the Principle of Least Effort* (the title of his 1949 treatise). But Mandelbrot, in an early effort, has shown that a monkey hitting typewriter keys *at random* will also produce a "language" obeying Zipf´s hyperbolic law (Mandelbrot 1961). So much for lexicographic *Least Effort*!

A detailed analysis shows that if the monkey´s typewriter has N equiprobable letter

keys and a space bar (with probability p_0), then his words (defined as letter sequences between spaces) have relative frequencies

$$f(r) \sim r^{-1+\log(1-p_0)/\log N}.$$

With $N = 26$ and $p_0 = 1/5$, say, the exponent of r equals -1.068, only slightly less than -1. In general, the monkey words can be modeled as a Cantor set with a fractal dimension D that equals the reciprocal of the exponent of $1/r$. In our example,

$$D = \frac{1}{1-\log(1-p_0)/\log N} \approx 0.936.$$

For a nine-letter alphabet and $p_0 = 1/10$, the exponent equals -1.048, corresponding to a Cantor "dust" with D ≈ 0.954. An arithmetic model for the (infinitely many) words of this nine-letter "language" is all the decimal fractions between 0 and 1 in which the digit 0 never occurs (not counting 0s at the end of terminating fractions).

Here are a few "three-letter" words of this language: .141, .241, .643, .442, .692, .121. Of course, .103, .707, and $.0\overline{3}$ are nonwords because they contain 0s.

Such languages do not have an average rank, but the *median* word rank of our "exemplary" language is an astonishing 1,895,761; that is, it takes the 1,895,761 most frequent words of the language to reach a total probability of one-half. (By contrast, the median word rank of English texts lies between 100 for typical media output and 500 for highly literate writers.) Thus, the monkey, while strictly clinging to Zipf's law, produces a rather wordy (and otherworldly) language.

Another, equally surprising "speech pathology" of the monkey language is the impossibility of constructing a dictionary for it, because its words form an *uncountable* Cantor set. (We would perhaps not be put off by an infinitely thick dictionary, as long as its entries could be sequentially numbered ─ but we could never countenance an uncountable compendium.)

If the monkey language has a fractal dimension, does it have any selfsimilarities? It certainly has. Multiply all words of the "decimal language" by 10 and drop the integer part (or, in general, just strike out the leftmost "letter" of each word) and you have another monkey word (most likely forming a nonsensical word sequence). In fact, the words of such languages grow on self-similar trees. Take any branch, no matter how high it is and seemingly small: it is *identical* to the entire tree.

And here we see the difference from natural languages most clearly: commonly spoken and written languages do *not* grow on self-similar trees ─ or, if we insist on hanging them from such trees (perish the thought), most branches would be dead.

Indeed, in natural languages many letter combinations are nonwords. Nevertheless, numerous English words are *homographs* (identical spellings) of words in other languages. And I do not mean such trivial cases as the uni(n)formed GENERAL, which means the same "thing" in many idioms. No, the interesting instances are "incognates" (unrelated words) such as the English word STRICKEN, which means *to knit* in German, or FALTER (a German *butterfly)* and LINKS (the German *left).* And what about such triplets as ART, which is a German word for KIND, which may mean MINOR in German, which in turn is a technical term in the theory of determinants (in either language). Finally, a fivefold string: ROT-RED-TALK-STEATITE-SOAPSTONES. Who can conceive sextuplets?

There are "letterally" hundreds of Anglo-German words like that, and I once composed a (short) German story using only English words. When I showed this story to a German-speaking Hungarian in the United States, his bored comment was "nothing but random

poetry" ─ even after repeated proddings to look at the text, with an open mind, as in visual texture (figure-background) discrimination, one of his research interests. When, half a year later, I showed the same text once more to the same Hungarian friend, this time in Germany, he read it and commented "Interessant! Interessant!!" Talk about the impact of context in human perception! (I leave it as an exercise to the linguistically inclined reader to compose a novel that makes *sense* in both German and English, or any other pair of languages in which at least the letter frequencies are not too different.)

How about the French woman who was amazed at the quantities of "soiled underwear" offered for sale in the United States when she first came upon the common come-on *Lingerie Sale?*

Sometimes a double-duty word engenders a double entendre, or rather a twofold misunderstanding. Shortly after I moved to Göttingen, the building superintendent of the physics institute, who collected my foreign parcel post from customs, went around the campus confiding that "Professor Schroeder is importing *poison* from the U.S.A. It even says so right on the packages: Gift!" Gift indeed, the German word for poison, and cognate to the English gift, because *gift* is something one gives (occasionally, anyhow), as in the surviving *Mitgift,* the bride's dowry.

When I told this tale to the (research) chemist Francis O. Schmitt of the Massachusetts Institute of Technology, he parried with the perfect misunderstanding in reverse. One of his students had once reported from a postdoctoral stay in Germany how generous indeed the indigenous chemical industry was: every other bottle in his lab was labeled GIFT! So, in certain parts of the word, better not to swallow the "presents."

Of course, not all homographs are quite so harmless. Consider *Not,* the German *emergency.* An Australian friend of mine (a linguist, no less) once found himself trapped inside a building in Austria (was the place on fire?), but every door that he approached repulsed him with a forbidding "verboten" sign saying NOTAUSGANG! ─ not exit? My increasingly frantic friend, desperately seeking *Ausgangs,* knew enough Latin and German (besides his native English) to properly decode *aus-gang* as *ex-it.* But in the heat of the emergency, he never succeeded in severing the Gordian knot: *Not* is not *not.*

References

Alvarez, L.W. (1987). *Adventures of a physicist*. New York: Basic Books.

Ander, M.E., Zumberge, M.A., Lautzenhiser, T., Parker, R.L., Aiken, C.L.V., Gorman, M.R., Nieto, M.M., Cooper, A.P.R., Ferguson, J.F., Fisher, E., McMechan, G.A., Sasagawa, G., Stevenson, J.N., Backus, G., Chave, A.D., Greer, J., Hammer, P., Hansen, B.L., Hildebrand, J.A., Kelty, J.R, Sidles, C., Wirtz, J. (1989). Test of Newton´s inverse-square law in the Greenland ice cap. *Physical Review Letters 62, 985-988.*

Mandelbrot, B.B. (1961). On the theory of word frequencies and on related Markovian models of discourse. In: Jakobson, R. (ed.), *Structure of language and its mathematical aspects: 190-219*. Providence, R.I.: American Mathematical Society.

Mandelbrot, B.B. (1963a). The stable Paretian income distribution when the apparent exponent is near zero. *International Economic Review 4, 111-115.*

Mandelbrot, B.B. (1963b). The variation of certain speculative stock prices. *J. of Business 36, 394-419.*

Mandelbrot, B.B. (1963c). New methods in statistical economics. *J. of Political Economy 71, 421-440.*

Pareto, V. (1896). *Oeuvres Complètes*. Geneva: Droz.

Shannon, C.E. (1951). Prediction and entropy of printed English. *Bell System Technical J. 30, 50-64.*

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass.: Addison-Wesley.

Glottometrics 4, 2002,45-48 *To honor G.K. Zipf*

Zipf's Law and why it works everywhere

Eric S. Wheeler[1]

Abstract. Zipf's law is a consequence of independently categorizing items, and rank ordering the categories. Therefore, it can be applied to almost anything. Testing methods on random input helps us see what is the artifact of the method rather than the property of the subject matter.

Keywords: Zipf´s law, ranking, testing methods

Since G.K. Zipf (1949) first proposed that, for a given region, the number of cities with a given population size is inversely proportional to the population size, many people have observed a similar relationship in diverse fields, and people continue to make the same discovery. Recent papers have applied the Zipf Law to cities in the US (Urzua 2000), cities in Denmark (Knudsen 2001), companies in the USA (Axtell 2001), and whistles of dolphins (McCowan et al. 1999), to cite but a few (see Simon 1968: 439ff). Linguists have frequently applied the law to words, phonemes and even the gaps between words.

1. Not seeing a common thread among these varied subjects, I once became skeptical enough (Wheeler 1979) to apply the method of one author (Králík 1977) not to his real-world subject, but to random numbers. A real-world subject may have structure that a method will reveal, but random numbers should have no pattern. The result, however, was that the Zipf law fit the table of random numbers even better than it had the real linguistic data (Figure 1).

If such a pattern happens when the subject matter is random and content-free, then the pattern is the result of the method of analyzing and presenting the data, and not a property of the subject matter at all.

Why is this so? The explanation is simple. The method used to get a Zipf law in effect says: for a given total number of items, (1) assign each item to one of a set of categories (e.g. people to the city they live in) and (2) rank order the categories by size. If the items are assigned without regard to where other items are assigned or to the number of items already in the category, it is not surprising that there are only a few categories that have many items and many categories that can have few items. Once you have taken many items from the total, there is much less left to categorize.

Think of the process of cutting random-sized pieces from a cake; the first few might be large, but after they are gone and the cake is much smaller, you are only able to cut smaller sized pieces.

In algebraic terms: a unit interval [0 1] can have at most k elements (k = 1, 2, 3…) of length $1/k$ To have many elements (k large), the length must be small (no more than $1/k$).

[1] Address correspondence to: Eric S. Wheeler, 33 Peter Street, Markham, Ontario, Canada L3P 2A5. E-mail: wheeler@wheeler-and-young.on.ca

2. Any subject where the assignment of an item to a category is essentially independent of other assignments will lead to a Zipf law distribution. Furthermore, even if the assignments are not entirely independent, there is still going to be the same effect when one categorizes items and rank orders the categories.

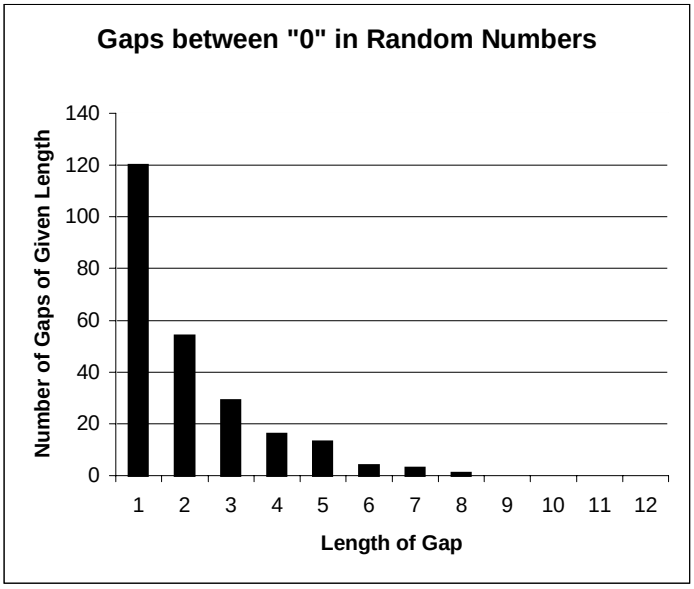

Figure 1. A reconstruction of the random number test shows that the number of gaps between a given digit in a set of random numbers varies inversely with the size of the gap.

Even if people prefer, for example, living in large cities (and therefore choose their residence place according to the size of the category it is in), one can see that there will still be few large cities and many smaller ones for a given population. A heavy bias to one end of the scale or the other does not necessarily change the overall shape of the distribution curve. It would take special circumstances to get the categories, say, to be all equal, or evenly divided into just 2 sizes.

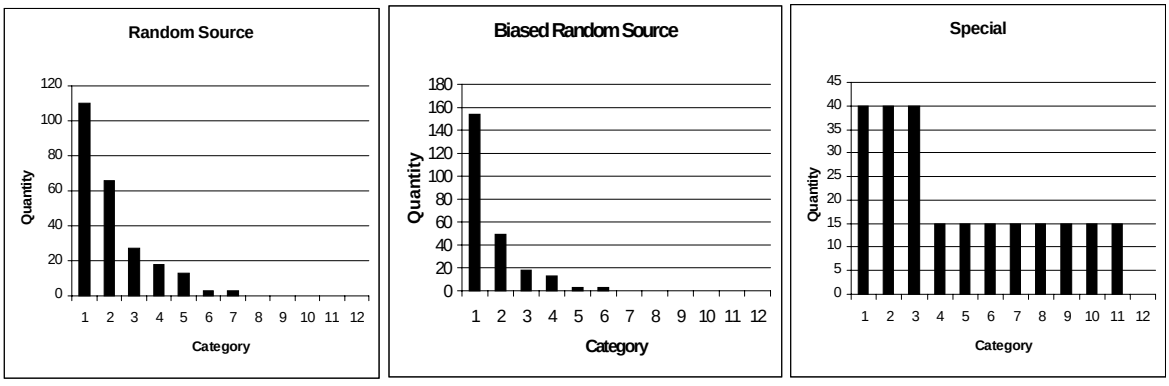

Figure 2.1. The distribution of the random source (left graph) is similar to the distribution of a source biased to category 1 (centre graph). It takes special constraints to get a distribution that clearly is not the same (right graph). (Note: Y-scales vary. Total items = 240)

3. The places where the data does *not* fit the Zipf model are the most interesting. That result is an indicator that the assignment of items to categories is not independent, and suggests the existence of some constraint or mechanism belonging to the subject area itself. Králík (1977) reported fewer than expected short gaps between successive occurrences of a given word in a text. That is consistent with our intuition that we do not like to repeat the same expression twice close together, even if we need to repeat the reference. But that data was just the part that did not fit well with the Zipf model.

4. Methods of arranging, analyzing, summarizing and presenting data all have their impact on what the data looks like when it is finally presented. It is not just the particular method used here that can have effects (although "Zipf's Law" has perhaps caught the imagination of people more so than most methods). Averages, medians, bar graphs and normal distributions (and so on) of a given set of data will shape our view of that data in a particular way. Normally, of course, we see past that shaping and are not surprised that "cities in Sweden" have an average population, or that survey results have a range of accuracy "19 times out of 20". The same should be true for the rank-ordering of categories.

Perhaps every complex method that we apply to real world data should also be benchmarked against a set of randomly generated data. If we find significant trends or relationships in the invented data, we can fairly attribute them to the methods used. There should be no pattern to find in random data. Subsequently, when the methods are applied to real world data, and there are trends or relationships beyond what is a by-product of the method, then we have something substantial. Consider, if the average height of zoo keepers at various zoos is always found to be an even number, it would be wise to try our averaging method on some randomly generated heights. Perhaps the method is rounding the heights to the nearest even number. But if we modify our method so this is not happening, and still find that the average number of eyes on the lions, and hands on the zookeepers, and legs on the monkeys are all even numbers, then it may be that we have discovered some interesting property of animals (bilateral symmetry).

5. The Zipf model is useful: dictionary writers need to know that the top 10 or 20 most frequently occurring words will account for a substantial portion of the source corpus they use, and conversely, a substantial portion of the words in the corpus will occur only once.

However, we need not be surprised to find that the Zipf model applies to many things, indeed to almost everything, and we should not overlook the places where the Zipf model does not apply, because that is where something noteworthy may be happening.

References

Axtell, R.L. (2001). Zipf distribution of US firm sizes. *Science 293, 5536, 1818-1820.*

Knudsen, T. (2001). Zipf's law for cities and beyond - The case of Denmark. *American Journal of Economics and Sociology 60, 123-146.*

Králík, Jan (1977). An application of exponential distribution law in Quantitative Linguistics. *Prague Studies in Mathematical Linguistics 5, 233-235.*

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

Simon, H.A. (1968). On judging the plausibility of theories. In: B. Van Rootselaar and J.F. Stall. (eds.), *Logic, Methodology and Philosophy of Science III: 439-459*. Amsterdam. North Holland Publishing.

Urzua, C.M. (2000). A simple and efficient test for Zipf's law. *Economic Letters 66, 257-260.*

Wheeler, Eric S. (1979). How not to count gaps. *Canadian Journal of Linguistics 24, 147-149.*

Zipf, G.K. (1949). *Human behavior and the Principle of Least Effort. An introduction to human ecology* (facsimile 1965). New York: Hafner Pub. Co.

Glottometrics 4, 2002, 49-60 *To honor G.K. Zipf*

Zipf's law against the text size: a half-rational model

Łukasz Dębowski [1]

Abstract. In this article, we consider Zipf-Mandelbrot law as applied to texts in natural languages. We present a simple model of dependence of the law on the text size, which is featured by variable power-law tail and constant ratio of the most frequent words. As a result we derive several closed formulas, which accord with empirical data qualitatively and partially quantitatively. For example, there appears to be a minimal length of literary texts equal to ≈ 159 word tokens for English.

Keywords: Zipf's Law

Introduction

For a definite object in which we can identify tokens and count them as instances of some identifiable types, Zipf's law (Zipf 1935, 1949) is a statement that the frequency $f(w)$ of all tokens belonging to given type w is roughly inversely proportional to rank $r(w)$ of the type,

$$f(w) \approx \frac{const}{r(w)}. \tag{1}$$

Rank $r(w)$ is defined as the ordinal number w on the list of all empirical types sorted in descending order according to $f(w)$.

Zipf's law forms a beautiful example of quasi-interdisciplinary empirical regularity which possesses the following features:

1. Regularity is observed in data resuming phenomena investigated in various scientific disciplines. Examples are biology (Camacho, Solé 1999), economics (Pareto 1897), linguistics (Estoup 1916), physics (Tsallis 2000). (In non-linguistic applications, rank $r(w)$ can be proportional to some simple variable describing physical size or magnitude of w, such as income in the distribution of personal incomes.)

2. Regularity inspires a multitude of half-explanations introducing assumptions which do not seem to be so universal as the regularity itself. Examples are random-typing text model (Belevitch 1956; Li 1992), effects of artificial ranking of sample taken from distribution with large variance (Günther, Levitin, Schapiro, Wagner 1996), eco-system dynamics (Camacho, Solé 1999), fractal vocabulary model (Mandelbrot 1983), new thermodynamic formalism (Denisov 1997), sampling of LNRE distributions (large number of rare events) (Khmaladze 1987; Baayen 2001), information theoretic models of language learning (Harremoës, Topsøe preprint; Dębowski 2002).

3. Regularity is roughly described by a simple mathematical formula, but domain speci-

[1]Address correspondence to: Łukasz Dębowski, Institute of Computer Science, Polish Academy of Sciences, ul. Ordona 21, 01-237 Warszawa, Poland. E-mail: ldebowsk@ipipan.waw.pl

fic investigations usually uncover many finer significant departures. Examples include initial bend and final power law (Mandelbrot 1954), better fit of tail by the inverse-square distribution of frequencies (Kornai 1999) (also independently observed by us and by M. Montemurro), another bend for very large ranks (Montemurro 2001) (also observed in economical data), parameter dependence on the size of texts (Orlov 1982), or investigated fraction of a text (Baayen 2001).

In dealing with empirical regularities such as Zipf's law, for which there is no unique, accurate and enchanting theory, one usually follows either of often disjoint ways: empirical or rational. In the empirical approach, one seeks for the formula most tightly interpolating given experimental data, even at the cost of introducing obscure additional parameters and worsening the extrapolations. The rational method consists in deriving the regularity from simpler principles or other empirical facts, even at the cost of worsening the fit of already observed data in comparison to more elaborated but obscure models.

Nonetheless, the possibility of combining empirical and rational approaches arises sometimes. For example, one can complicate a formula in concern to have additional apparently random parameters and to fit better some portion of data, but the formula with the same class of parameters plus some conditions on their variability can fit a much larger scope of data than before the modification. In fact, it is the way how many fundamental theories in natural sciences have been born.

The aim of this article is to present a modest example of combined empirical-rational approach given by a simple model of Zipf's law variability across texts of different length. We will speak only of Zipf's law applied to texts in natural languages. We are going to show that Mandelbrot's modification (2) of formula (1), introducing two unknown parameters, can be perceived as less arbitrary if we let the values of the parameters be linked with the text length according to several common-sense postulates.

The model

In quantitative linguistics, Zipf's law (1) is formulated for types w being types of words (word-forms or lemmas) encountered in some finite text. The tokens are occurrences of words at consecutive positions in the text. It is also this text against which both counts $f(w)$ and ranks $r(w)$ are computed. (In the case of words with the same count, we assign them distinct ranks.) It is important to note that language texts treat various subjects, so rank $r(w)$ of particular word w strongly depends on the particular text. The exception for this rule is a group of words constantly occupying the lowest ranks and identifiable with functional (grammatically auxiliary) words.

Mandelbrot (1954) observed that instead of formula (1), formula

$$f(r) \approx \left\lfloor \left[\frac{V+\rho}{r+\rho} \right]^{1+\varepsilon} \right\rfloor \tag{2}$$

where we abbreviate $f(r) = f(w(r))$ for $r = r(w)$ and $\lfloor x \rfloor$ is the greatest integer smaller than x, approximates statistics of words better. We have $r \in \{1,2,...,V\}$, where V is the size of vocabulary for the given text. Formula (2) fits the whole range of finite-text data better than (1), but there are some departures still (Baayen 2001). The formula contains also two new parameters to estimate: ε and ρ. (For very large texts $0 < \varepsilon << 1$ and $0 < \rho < 10$.)

It is important to note that parameters V, ρ, and ε depend quite regularly on the size N of the text, i.e. N being the number of word tokens in the text. Especially, $\varepsilon < 0$ and $\varepsilon > 0$ for $N > N_0$, where N_0 is some characteristic text length, called Zipfian size (Orlov 1982).

Orlov (1982) described this phenomenon and gave it some mathematical model in terms of interpolation formulas for random sample (urn model, or IID process) drawn from LNRE distribution. See also Khmaladze (1987), Baayen (2001) for more elaborate calculations. We had learned of the article by Orlov (1982) from an article by Sambor (1988). By the time we collected a copy of Orlov (1982), we had found out a very different heuristic model of Zipf's law variability which we introduce here.

If one considers an ensemble of texts of variable size N written in the same language, it is reasonable to assume that the same grammar is obeyed in the whole ensemble. The conservation of grammar across the ensemble may imply the stability of probability estimates for the functional words which occupy constantly the same lowest ranks. Thus

(3) $$\frac{f(r)}{N} \approx \text{const} \quad \text{for } r = 1,2,...,K$$

for the majority of texts, where K is some small natural number and N is assumed to be the length of the text in the ensemble. If $N \to \infty$, however, $f(r)/N \to$ const for all r. Thus

(4) $$\varepsilon \to \text{const} \quad \text{and} \quad \rho \to \text{const} \quad \text{for} \quad N \to \infty .$$

Postulates (3), (4) when applied to (2) can be approximated by the following set of three postulates:

1. There is such $N = N_0$ that $\varepsilon = 0$.
2. For all N, it is $f(0)/N$ = const.
3. For all N, it is $f'(0)/N$ = const.

(Formula (2) allows us to define the value $f(0)$ and the derivative $f'(0)$.)

In the further reasoning, we will assume $\rho >> 1$, despite the empirical data. Using (2), one can compute the number of tokens N in the text as

(5) $$N \approx \int_0^V f(r)dr = \int_0^V \left[\frac{V+\rho}{r+\rho}\right]^{1+\varepsilon} dr = \frac{\rho}{\varepsilon}\left[\frac{V+\rho}{\rho}\right]^{1+\varepsilon}\left[1-\left[\frac{\rho}{V+\rho}\right]^{\varepsilon}\right].$$

For $N = N_0$, let us write $V = V_0$, $\rho = \rho_0$. Combining postulates 2 and 3, one obtains $f(0)/f'(0) =$ const. Inserting (2) for any N and for $N = N_0$, and preserving terms linear in 1/ρ yields

(6) $$\rho \approx (1+\varepsilon)\rho_0 \quad \text{for } \rho >> 1.$$

Parameter N_0 can be rewritten by means of V_0 and ρ_0 as

(7) $$N_0 = \int_0^{V_0}\left[\frac{V_0+\rho_0}{r+\rho_0}\right]dr = \rho_0\left[\frac{V_0+\rho_0}{\rho_0}\right]\ln\left[\frac{V_0+\rho_0}{\rho_0}\right].$$

(Formula (5) can be applied directly for $\varepsilon \neq 0$. Formula (7) is its limit for $\varepsilon \to 0$.) Using (5), (7), postulate 2 with (2) for any $N \neq N_0$ and for $N = N_0$ gives

(8) $$\rho_0 \ln\left[\frac{\rho_0}{V_0+\rho_0}\right] = \frac{\rho}{\varepsilon}\left[\left[\frac{\rho}{V+\rho}\right]^{\varepsilon} - 1\right].$$

It is convenient to define

(9) $$\lambda = 1 - \frac{\varepsilon x}{1+\varepsilon},$$

$$x = \ln\left[\frac{V_0 + \rho_0}{\rho_0}\right]. \tag{10}$$

Then

$$1+\varepsilon = \frac{x}{x-1+\lambda}. \tag{11}$$

Equations (8) and (6) yield

$$\left[\frac{V+\rho}{\rho}\right]^{1+\varepsilon} = \left[1-\left[\frac{\varepsilon}{1+\varepsilon}\right]\ln\left[\frac{V_0+\rho_0}{\rho_0}\right]\right]^{-\frac{1+\varepsilon}{\varepsilon}} = [e(\lambda)]^x, \tag{12}$$

where function $e(\lambda)$ is defined as

$$e(\lambda) = \lambda^{1/(\lambda-1)}. \tag{13}$$

Resuming, one obtains

$$f(r) \approx [e(\lambda)]^x \left[\frac{\rho_0\left[\frac{x}{x-1+\lambda}\right]}{r+\rho_0\left[\frac{x}{x-1+\lambda}\right]}\right]^{\left[\frac{x}{x-1+\lambda}\right]}, \tag{14}$$

$$N \approx [e(\lambda)]^x \rho_0 x, \tag{15}$$

$$V \approx \left[\frac{[e(\lambda)]^{x-1+\lambda}-1}{x-1+\lambda}\right]\rho_0 x, \tag{16}$$

where V was computed from property $f(V) = 1$.

In equations (14)-(16), three parameters appear: ρ_0, x and λ. The status of ρ_0 and x is different from λ. Parameters ρ_0 and x, where $\rho_0 > 0$, $x > 1$, should be the properties of a given language, i.e. they should be constant in the whole ensemble of texts in that language. Parameter λ, where $\lambda > 0$, is a function of the size N of text and the two other parameters. Since $e(\lambda) > 1$, our model can be only applied if $N \geq \rho_0 x$.

Both text size N and vocabulary size V are monotonically decreasing functions of λ. For $\lambda \to \infty$, it is N, $V \to N_{\min} = \rho_0 x$. For $\lambda \to 0$, it is N, $V \to \infty$. Value $\lambda = 1$ corresponds to $\varepsilon = 0$ and $N_0 = \rho_0 x e^x$. We can see that for $\varepsilon < 0$ for $N < N_0$ and $\varepsilon > 0$ for $N < N_0$ actually holds for our model. (By the way, condition $x > 1$ is necessary since we need to have $f(r) > 0$ and monotonically decreasing w.r.t. r also for $\lambda \to 0$.)

If we would like the model to describe texts of any positive length, $N \geq 1$, we might like to fix $\rho_0 = 1/x$ so that the minimum of N and V be $N_{\min} = \rho_0 x = 1$. Then we have just one global parameter x left and $N_0 = e^x$. In this case, there is a simple relation between the Zipfian size N_0 and the Mandelbrot's exponent $1+\varepsilon$ for texts of asymptotically infinite length,

$$1+\varepsilon = \frac{\ln N_0}{\ln N_0 - 1}. \tag{17}$$

According to Orlov (1982), N_0 for Russian seems to range from 3000 to 20000, so we would get 1+ ε ranging between 1.143 ($x = 8.006$) and 1.112 ($x = 9.903$) respectively.

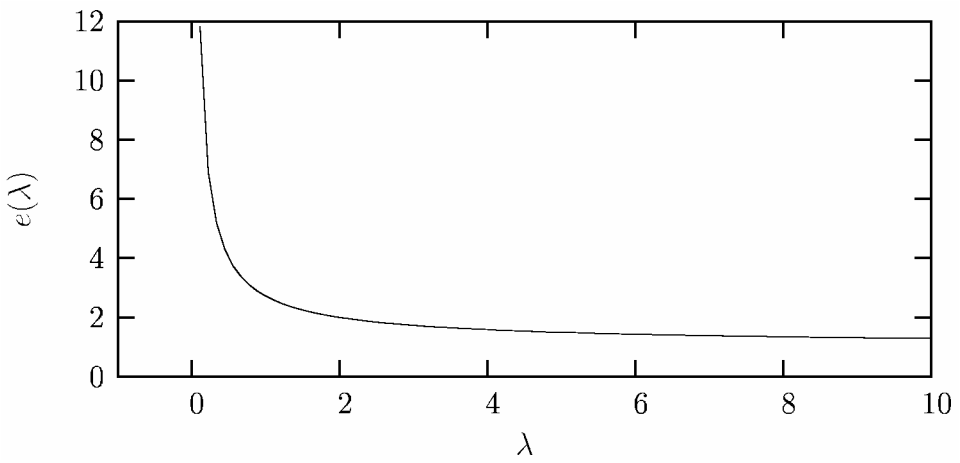

Figure 1. The plot of $e(\lambda) = \lambda^{1/(\lambda-1)}$

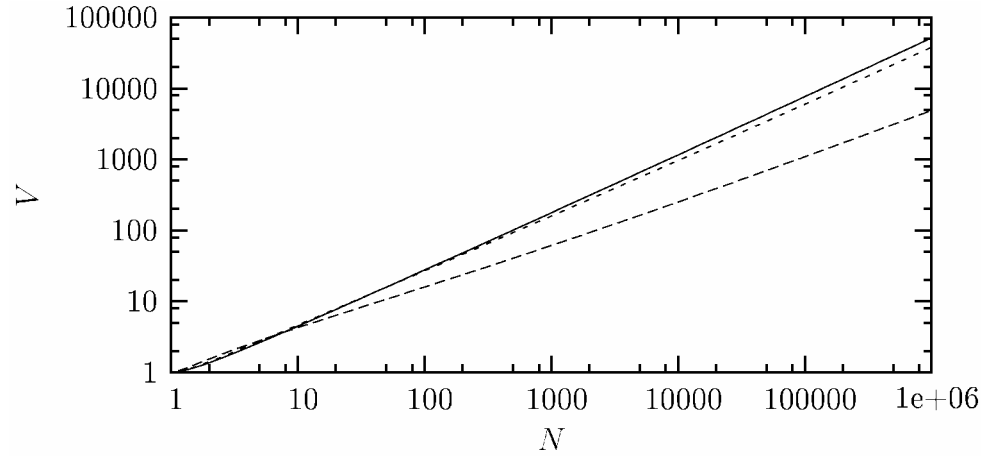

Figure 2. Plots of vocabulary size V as function of text size N for $\rho_0 = 1/x$ and x = 3.0, 8.0, 10.0 (curves increasing respectively).

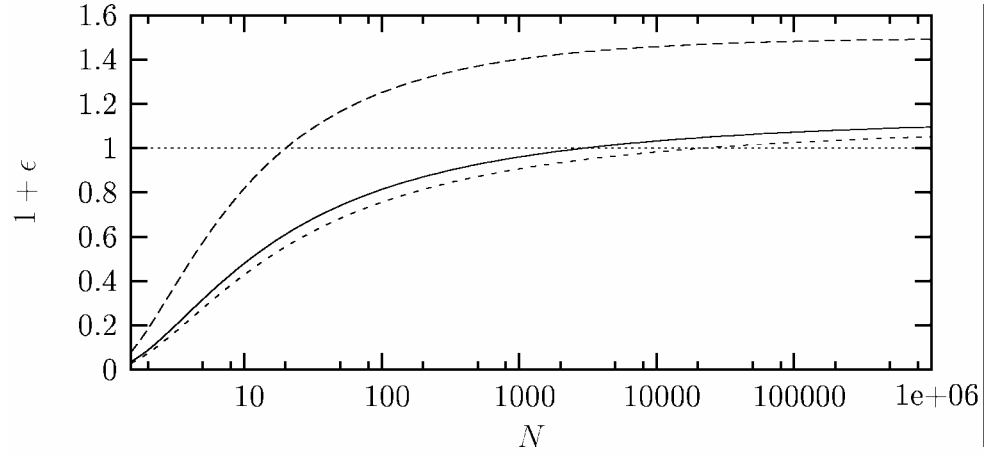

Figure 3. Plots of Mandelbrot exponent 1+ ε as function of text size N for $\rho_0 = 1/x$ and x = 3.0, 8.0, 10.0 (curves increasing respectively).

In our model, $\log V$ seems to be an almost linear function of $\log N$, but in fact, $\log V$ is a slightly concave function of $\log N$ (with $d \log V/d \log N = 0$ for the minimal point $N = V$). Anyway, since $\log V$ is an almost linear function of $\log N$, $\log V \approx a \log N + b$, we could try to approximate parameters x, ρ_0 using just linear regression for empirical data ($\log N$, $\log V$). In fact, we have

$$\frac{x-1}{x} \approx a, \qquad \log N_{\min} \approx a \log N_{\min} + b, \qquad \rho_0 x = N_{\min}, \tag{18}$$

which can be easily solved for x, $N_{\min}$, ρ_0 given a, b.

In order to compute λ as the function of N, ρ_0, and x, it is necessary to find the inverse of $e(\lambda) = \lambda^{1/(\lambda-1)}$. The inverse of $e(\lambda)$ is not a closed-form function of its argument but there is some good elementary approximation, which is presented in the appendix.

Comparison with empirical data

In order to compare our theoretical model with natural language data, we have collected a selection of texts of various sizes which we downloaded from Project Gutenberg website – http://www.promo.net/pg/index.html. The full selection is listed in Table 1. All the texts are raw English texts, in which we chose the types w to be the graphical words (word-forms) rather than their (disambiguated) lemmas. We ignored the punctuation and turned all word-forms into lower-case. In this way, all text processing and data plotting could be done automatically in several seconds on a PC using simple scripts in Perl and Gnuplot.

For the given e-text data, we have estimated the parameters of our model for two cases: (1) $\rho_0 = 1/x$ (only x is estimated), (2) ρ_0 is variable (both x and ρ_0 are estimated). The resulting values of parameters and implied characteristic constants $x/(x-1)$, $N_{\min}$, N_0 are given in Table 2. The estimation of the parameters was done using least-square fit for plot ($\log N$, $\log V$) with nonlinear theoretical curves given by chain of formulas (15), (24), (16). (Slightly better than linear regression (18).) The plot of the data including the fits is given in Figure 4.

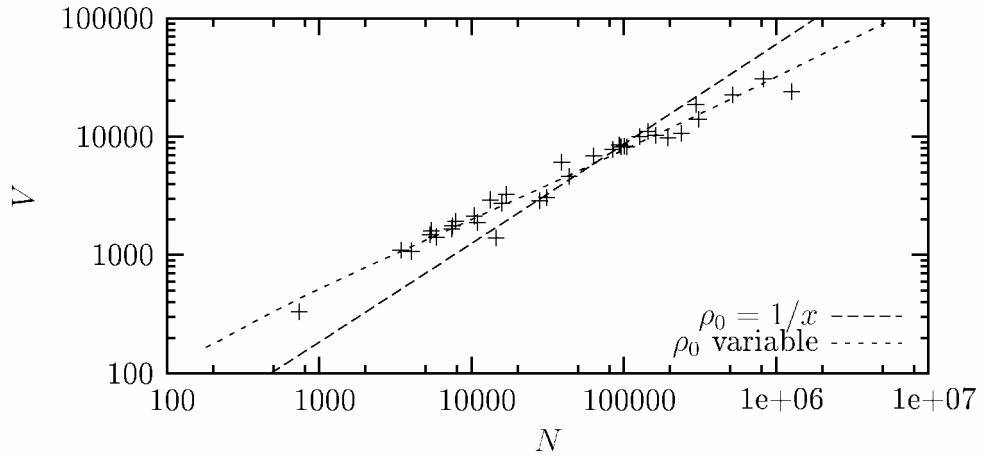

Figure 4. Plots of vocabulary size V as function of text size N for chosen e-texts.

In Figure 4, we can see that the model with variable ρ_0 accords with the data consistently better than $\rho_0 = 1/x$. For this model, the ratio of predicted and actual vocabulary size V is almost always about 1, independently of the text size N. For none of the observed data points, the ratio exceeds the range [0.5, 2] (see Figure 5).

The curious feature of the model with variable ρ_0 is that it predicts that there is a minimal text length $N_{\min} \approx 159$. Nevertheless, all e-texts considered as data were well-formed literary

texts, so this statement need not be so absurd, as it might appear i.e. for random typing texts.

Table 1
The choice of 35 e-texts from Project Gutenberg

Title	**Author**	**Text size** N	**Vocab. size** V
Peach Blossom Shangri-la	T. Yuan Ming	735	332
A Modest Proposal	J. Swift	3427	1092
On the Brain	T. H. Huxley	4017	1078
The Lake Gun	J. F. Cooper	5328	1488
Song Book of Quong Lee...	T. Burke	5440	1612
The Adventure of the Dying...	A. Conan Doyle	5857	1419
The Adventure of the Red Circle	A. Conan Doyle	7407	1668
Everybody's Business...	D. Defoe	7483	1766
Why Go to College?	A. F. Palmer	7847	1915
Dickory Cronke	D. Defoe	10426	2138
The Princess de Montpensier	Lafayette	10904	1881
Bickerstaff-Partridge Papers	J. Swift	13218	2928
The Categories	Aristotle	14488	1394
The New Atlantis	F. Bacon	15769	2750
The City of the Sun	T. Campanella	16855	3239
Alice in Wonderland	L. Carroll	27870	2868
Through the Looking-Glass	L. Carroll	31055	3059
The Battle of the Books...	J. Swift	38944	6068
Utopia	T. More	43633	4624
Around the World in 80 Days	J. Verne	63290	6853
Erewhon	S. Butler	84717	7800
Five Weeks in a Balloon	J. Verne	93252	8524
Eight Hundred Leagues...	J. Verne	95568	8210
20,000 Leagues Under the Sea	J. Verne	100598	8294
Gulliver's Travels	J. Swift	104650	8191
One of Ours	W. Cather	126621	10049
Life of William Carey	G. Smith	143849	11072
Memoirs	Comtesse du Barry	160790	10278
The Mysterious Island	J. Verne	194213	9743
The Journal to Stella	J. Swift	238787	10642
Critical & Historical Essays	Macaulay	296553	18684
The Descent of Man	C. Darwin	308171	14086
Mark Twain, A Biography	A. B. Paine	515597	22572
First Folio/35 Plays	W. Shakespeare	820016	30820
The Complete Memoirs	J. Casanova	1262287	24093

Table 2
The parameters resulting for the e-texts in two estimation schemes

	$\rho_0 = 1/x$	ρ_0 variable
x	11.477	3.1193
$x/(x-1)$	1.0954	1.4719
ρ_0	0.087132	51.108
$N_{\min}$	1	159.42
N_0	96456	3607.6

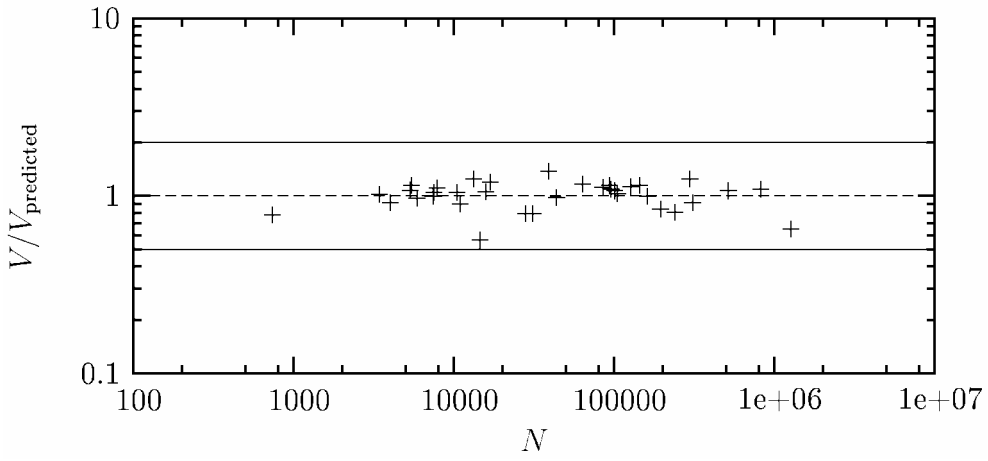

Figure 5. Plots of ratio of empirical vocabulary size V for chosen e-texts to the predicted vocabulary size $V_{\text{predicted}}$ given by our model with parameters x, ρ_0 as in the right column of Table 2 (variable ρ_0). The constant lines correspond to $V/V_{\text{predicted}} = 0.5$, 1, 2.

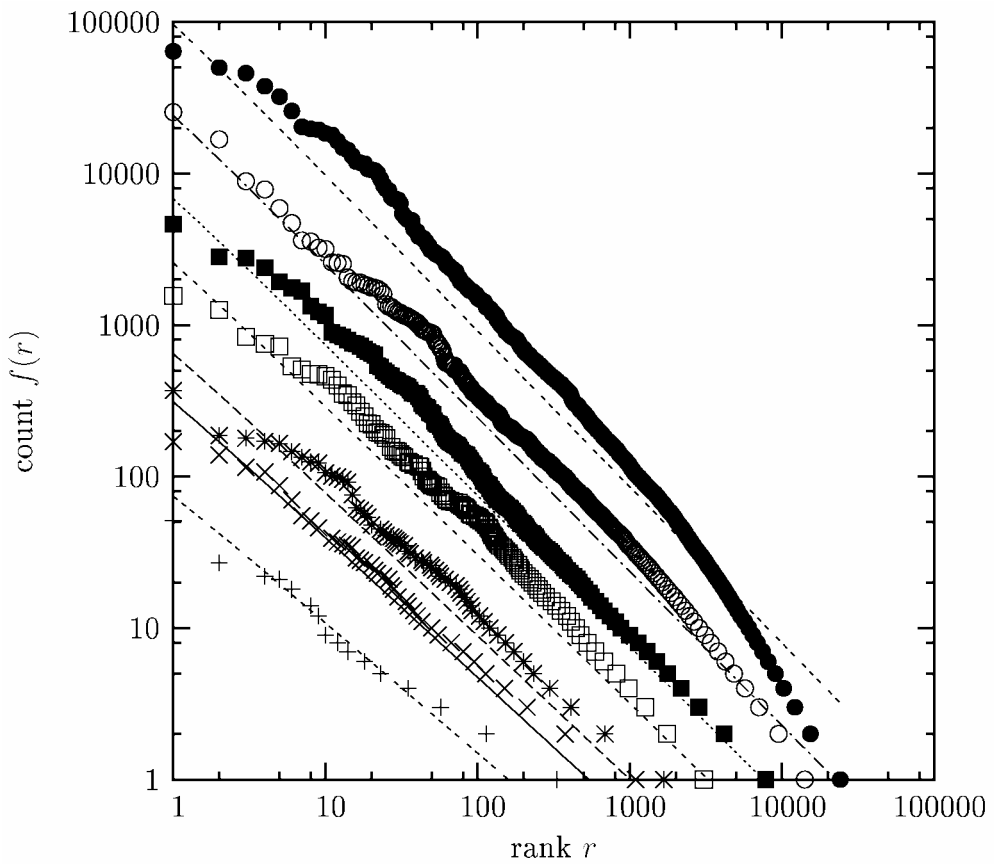

Figure 6. Plots of counts $f(r)$ against ranks r for the following texts: Peach Blossom Shangri-la, $N = 735$; A Modest Proposal, $N = 3427$; The Adventure of the Red Circle, $N = 7407$; Through the Looking-Glass, $N = 31055$; Erewhon, $N = 84717$; The Descent of Man, $N = 308171$; The Complete Memoirs, $N = 1262287$ (points growing respectively). The smooth curves stand for the count distributions predicted by our model with parameters x, ρ_0, as in the left column of Table 2 ($\rho_0 = 1/x$).

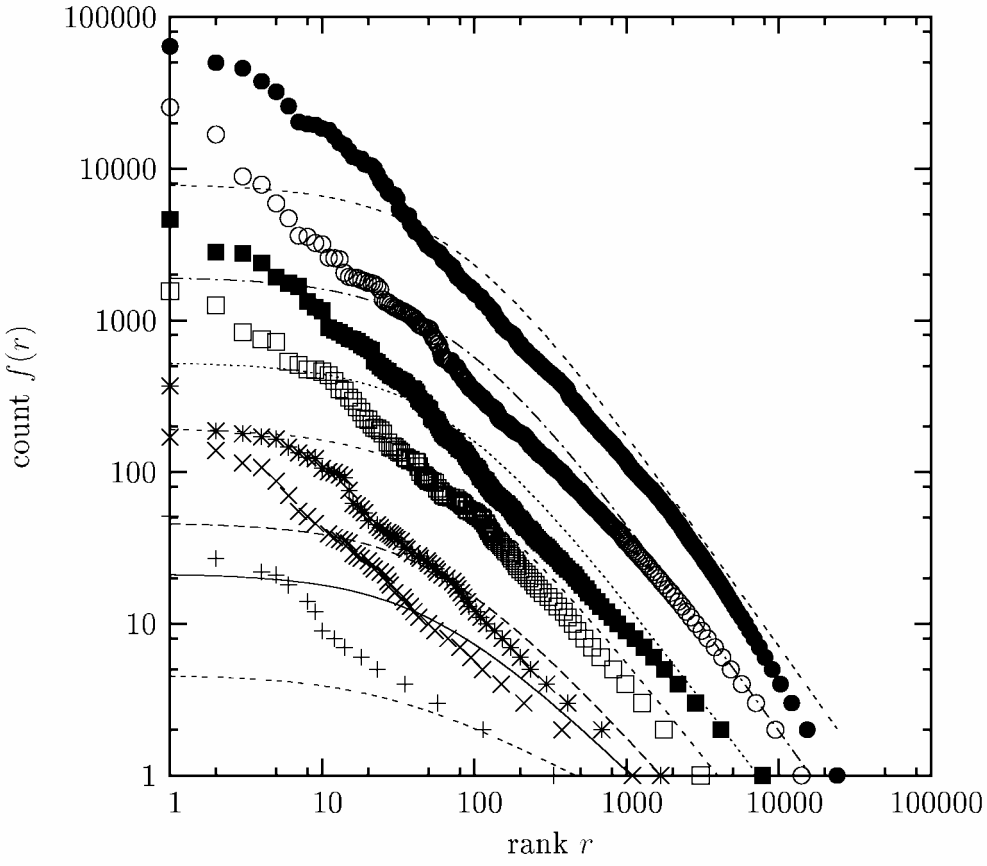

Figure 7. Plots of counts $f(r)$ against ranks r for the following texts: Peach Blossom Shangri-la, $N = 735$; A Modest Proposal, $N = 3427$; The Adventure of the Red Circle, $N = 7407$; Through the Looking-Glass, $N = 31055$; Erewhon, $N = 84717$; The Descent of Man, $N = 308171$; The Complete Memoirs, $N = 1262287$ (points growing respectively). The smooth curves stand for the count distributions predicted by our model with parameters x, ρ_0, as in the left column of Table 2 ($\rho_0 = 1/x$).

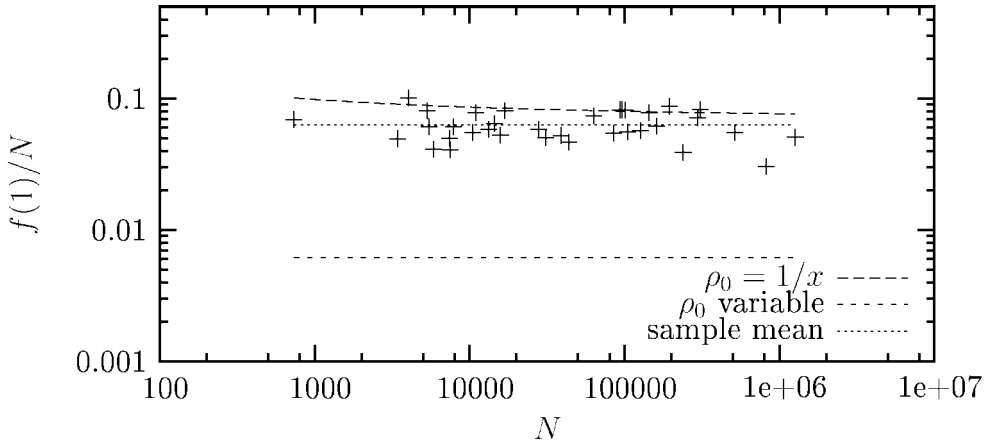

Figure 8. The relative count of the most frequent word against the predicted $f(1)/N$ and the text size N for the whole selection of e-texts.

The parameters of the theoretical model for both variable and fixed ρ_0 were estimated using (log N, log V) plot only. When we compare the rank-count distributions implied by the same parameters and the empirical rank-count distributions, we might observe greater depart-

ures. In fact, it is so. The model with $\rho_0 = 1/x$ seems to predict better the frequencies $f(r)$ for lower ranks (Figures 6, 7). Figure 8 confirms our rational assumption that $f(1)/N$ should be roughly constant across the texts of any length. Nevertheless, the model with variable ρ_0 still better reflects the variability of the power-law tail of $f(r)$ for the highest ranks r (Figure 7 as opposed to 6).

Conclusions

In this article, we have presented a very simple improvement of classical Mandelbrot-Zipf's law for natural language texts. The improvement was done not by introducing new parameters, but by letting the present ones vary with respect to the text size. The newly introduced constraint for the variability of parameters was that the relative counts of the most frequent words be constants independent of the text size.

The resulting model does not fit the data so well as much more complex LNRE models (Baayen 2001; Orlov 1982), but it still reproduces Mandelbrot´s exponent $1+\varepsilon < 1$ for text length $N < N_0$ and $1+\varepsilon > 1$ for $N > N_0$. The model also accords with empirical data qualitatively in several other aspects. Here, we have discussed theoretically the probably most complex phenomena in rank-count distribution which are still explainable by simple Mandelbrot-Zipf's formula (2).

Still, we have not checked if the quantitative departures of our model can be decreased if approximations (5), (6), assuming falsely $\rho >> 1$, were replaced by exact summations and equalities. In this case, we lose pretty closed-form formulas but maybe we could obtain better fit for self-consistent expressions.

Appendix

Approximating the inverse of $e(\lambda) = \lambda^{1/(\lambda-1)}$

The function defined by (13) is related to the definition of base e of natural logarithm. Actually,

(19) $e(1) = e$.

Function $e(\lambda)$ is an easily computable function of λ. Unfortunately the inverse is not true. Quantity λ is not a simply computable function of $e(\lambda)$. There is, however, an easily invertible and good approximation $\bar{e}(\lambda)$,

$$(20) \qquad \bar{e}(\lambda) = 1 + \frac{e-2}{\sqrt{\lambda}} + \frac{1}{\lambda}.$$

One can define the relative error of $\bar{e}(\lambda)$ as

$$(21) \qquad \bar{b}(\lambda) = \frac{\bar{e}(\lambda) - e(\lambda)}{\bar{e}(\lambda)}.$$

Function $e(\lambda)$ has the domain $\lambda \in \{0, \infty\}$. In this domain, the following substitution is convenient

$$(22) \qquad \lambda = \frac{1-u}{1+u},$$

where $u \in \{-1, 1\}$. Let $b(u) = \bar{b}(\lambda)$. Then $b(u) = 0$ for $u = -1, 0, 1$. ($b(u)$ for $u = -1,1$ is

defined by the corresponding limits.) Explicitly

$$(23) \qquad b(u) = 1 - \frac{\left[\frac{1-u}{1+u}\right]^{-1/2u}}{\sqrt{\frac{1-u}{1+u}} + (e-2) + \sqrt{\frac{1+u}{1-u}}},$$

so $b(u) = b(-u)$. Furthermore, looking at the plot of $b(u)$ (Figure 9) one can see that $0 < b(u) < 0.04$. Resuming, $\bar{e}(\lambda)$ is a good simple approximation of $\lambda^{1/(\lambda-1)}$

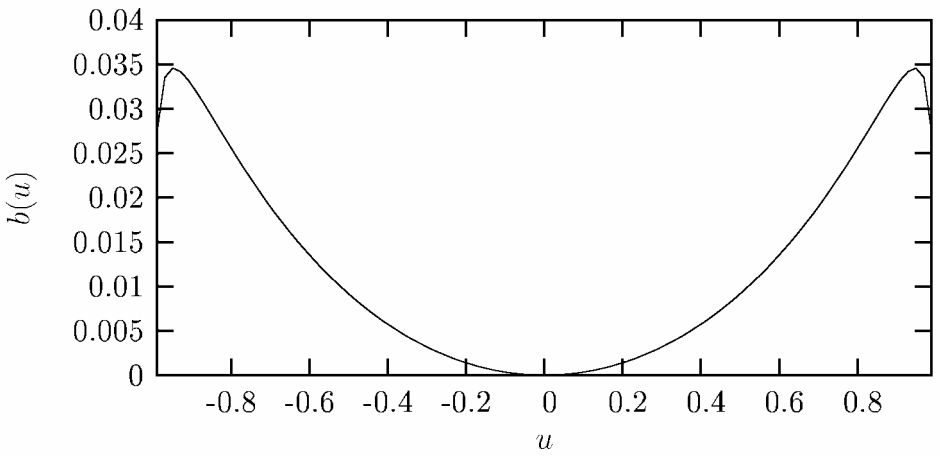

Figure 9. The plot of $b(u)$

In order to find λ for given $\bar{e}(\lambda)$, one sees that definition (20) is a quadratic equation for $1/\sqrt{\lambda}$ and it can be immediately solved,

$$(24) \qquad \lambda = \frac{4}{\left[2 - e + \sqrt{e^2 - 4e + 4\bar{e}(\lambda)}\right]^2}.$$ In formula (24), the one of two solutions was chosen which reproduces $\lambda = 1$ for $\bar{e}(\lambda) = e$.

References

Baayen, H. (2001). *Word frequency distributions*. Dordrecht: Kluwer Academic Publishers.

Belevitch, V. (1956). Théorie de l'information et statistique linguistique. *Académie royale de Belgique, Bulletin de la classe des sciences 419-436.*

Camacho, J., Solé, R.V. (1999). *Scaling and Zipf's law in ecological size spectra*. Santa Fe Institute Working Paper 99-12-076.

Dębowski, Ł. (2002). On the best theories for learningful texts. (http://www.ipipan.waw.pl/~ldebowsk)

Denisov, S. (1997). Fractal binary sequences: Tsallis thermodynamics and the Zipf's law. *Physics Letters A, 235, 447-451.*

Estoup, J.B. (1916). *Gammes sténographiques*. Paris: Institut Stenographique de France.

Günther, R., Levitin, L., Schapiro, B., Wagner, P. (1996). Zipf's law and the effect of ranking on probability distributions. *International Journal of Theoretical Physics, 15, 395-417.*

Harremoës and P. Topsøe, F. (preprint). Zipf's law, hyperbolic distributions and entropy

loss. (http://www.math.ku.dk/~topsoe/manuscripts.html)

Khmaladze, E. (1987). The statistical analysis of large number of rare events. *Technical Report MS-R8804, Dept. of Mathematical Statistics, CWI.* Amsterdam: Center for Mathematics and Computer Science.

Kornai, A. (1999). Zipf's law outside the middle range. In: *Proceedings of sixth meeting on mathematics of language: 347-356. University of Central Florida.*

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

Mandelbrot, B. (1954). Structure formelle des textes et communication. *Word, 10, 1-27.*

Mandelbrot, B. B. (1983). *The fractal geometry of nature*. New York: W. H. Freeman.

Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. (http://xxx.lanl.gov/abs/cond-mat/0104066)

Orlov, J. K. (1982). Linguostatistik: Aufstellung von Sprachnormen oder Analyse des Redeprozesses? In: J. K. Orlov, M. G. Boroda, I. Š. Nadarejšvili (eds.), *Sprache, Text, Kunst. Quantitative Analysen: 1-55.* Bochum: Studienverlag Dr. N. Brockmeyer.

Pareto, V. (1897). *Cours d'économie politique.* Lausanne, Paris: Rouge.

Sambor, J. (1988). Lingwistyka kwantytatywna — stan badań i perspektywy rozwoju. *Biuletyn Polskiego Towarzystwa Językoznawczego XLI, 47-67.*

Tsallis, C. (2000). Entropic nonextensivity: A possible measure of complexity. *Santa Fe Institute Working Paper 00-02-043.*

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley.

Glottometrics 4, 2002, 61-86 *To honor G.K. Zipf*

How many words are there?

András Kornai[1]

Abstract. The commonsensical assumption that any language has only finitely many words is shown to be false by a combination of formal and empirical arguments. Zipf's Law and related formulas are investigated and a more complex model is offered.

Keywords: Vocabulary size, Zipf's law

1. Introduction

Ask the lay person how many different words are there, and you are likely to receive a surprisingly uniform set of answers. English has more words than any other language. There are over a hundred thousand words in unabridged dictionaries. The *OED* has over three hundred thousand, but there are still a few words missing. A five year old knows five thousand words, an adult uses fifteen thousand. Shakespeare used thirty thousand, but Joyce used even more. Some primitive tribes get along with only a few hundred words. Eskimo has seventeen words for snow. Introductory linguistics courses invariably spend a great deal of energy on rebutting these and similar commonly known "facts" about languages in general and English in particular. But the basic fallacy, what we will call the *closed vocabulary* assumption, that there is a fixed number of words S in any given language, is often present even in highly technical work otherwise based on a sophisticated understanding of language. *Open vocabulary*, that words in a language form a denumerably infinite set, is a standard assumption in generative linguistics, where it is justified by pointing at productive morphological processes such as compounding and various kinds of affixation. Yet somehow the existence of such processes generally fails to impress those with more of an engineering mindset, chiefly because the recursive aspect of these processes is weak – the probability of iterated rule application decreases exponentially with the number of iterations.

In this paper we offer a new quantitative argument why vocabulary must be treated as open. We investigate vocabulary size not as an isolated number, but rather as part of the broader task of trying to estimate the frequency of words. The rest of this Introduction establishes the terminology and notation and surveys the literature. Section 2 disposes of some widely used arguments in favor of closed vocabulary by means of counterexamples and introduces the *subgeometric mean property* that will play a crucial role in the subsequent analysis of vocabulary size. Section 3 explores the regions of extremely high and extremely low frequencies, where the basic regularity governing word frequencies, Zipf's Law, is known to fail. Section 4 investigates some widely used alternatives to Zipf's Law, including the beta, lognormal, Waring, and negative binomial distributions, and shows why most of these are inferior to Zipf's Law. We offer our conclusions in Section 5.

[1] Address correspondence to: Andras Kornai, Metacarta Inc. 126 Prospect St, Cambridge, MA 02139 USA. E-mail: andras@kornai.com

1.1. Zipf's Laws

It is far from trivial to define words in spoken or signed language, but in this paper we can steer clear of these difficulties by assuming some conventional orthography or linguistic transcription system that has one to one correspondence between orthographic words (maximum non-white-space non-punctuation strings) and prosodic words. Because a large variety of transcription systems exist, no generality is lost by restricting our attention to text that has already been rendered machine readable. For the sake of concreteness we will assume that all characters are lowercased and all special characters, except for hyphen and apostrophe, are mapped on whitespace. The terminal symbols or *letters* of our alphabet are therefore $L = \{a, b, ..., z, 0, 1, ... 9, ', -\}$ and all word types are strings in L^*, though word tokens are strings over a larger alphabet including capital letters, punctuation, and special characters. Using these or similar definitions, counting the number of tokens belonging in the same type becomes a mechanical task. The results of such *word counts* can be used for a variety of purposes, such as the design of more efficient codes, typology, investigations of style, authorship, language development, and statistical language modeling in general.

Given a corpus Q of N word tokens, we find V different types, $V \leq N$. Let us denote the *absolute frequency* (number of tokens) for a type w by $F_Q(w)$, and the *relative frequency* $F_Q(w)/N$ by $f_Q(w)$. Arranging the w in order of decreasing frequency, the rth type (w_r) is said to have *rank* r, and its relative frequency $f_Q(w_r)$ will also be written f_r. As Estoup (1916) and Zipf (1935) noted, the plot of log frequencies against log ranks shows, at least in the middle range, a reasonably linear relation. Fig. 1 shows this for a single issue of an American newspaper, the *San Jose Mercury News*, or *Merc* for short.

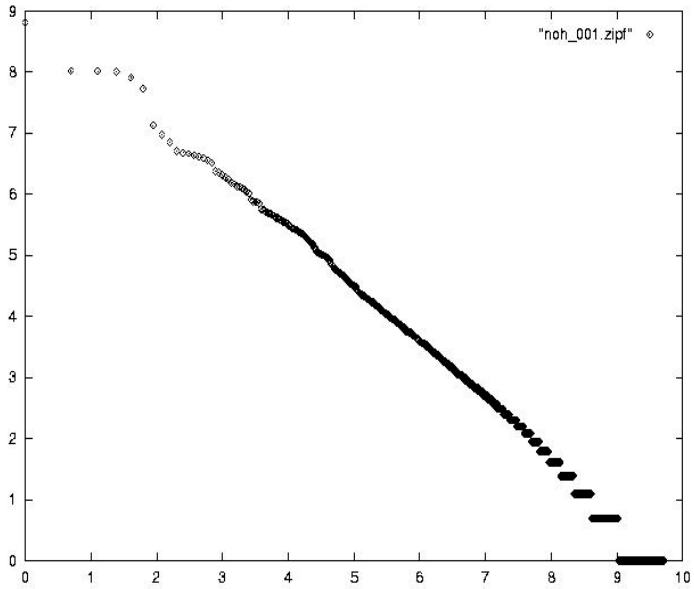

Fig. 1. Plot of log frequency as a function of log rank
for a newspaper issue (150k words)

Denoting the slope of the linear portion by $-B$, B is close to unity, slightly higher on some plots, slightly lower on others. Some authors, like Samuelsson (1996), reserve the term "Zipf's Law" to the case $B = 1$, but in this paper we use more permissive language, since part of our goal is to determine how to formulate this regularity properly. As a first approximation, Zipf's Law can be formulated as

(1) $\log(f_r) = H_N - B_N \log(r)$

where H_N is some constant (possibly dependent on our random sample Q and thus on N, but independent of r). When this formula is used to fit a Zipfian curve to frequency data, with increased corpus size not only the intercept H_N but also the slope B_N will depend on the corpus size N. We reserve the term "Zipfian" to the case where B_N tends to a constant B as N tends to infinity, but do not assume in advance $B = 1$. (1) is closely related to, but not equivalent with, another regularity, often called Zipf's Second Law. Let $V(i,N)$ be the number of types that occur i times: Zipf's Second Law is usually stated as

(2) $\log(i) = K_N - D_N \log(V(i, N))$.

1.2. Background

As readers familiar with the literature will know, the status of Zipf's Law(s) is highly contentious, and the debate surrounding it is often conducted in a spectacularly acrimonious fashion. As an example, we quote here Herdan (1966:88):

> The Zipf law is the supposedly straight line relation between occurrence frequency of words in a language and their rank, if both are plotted logarithmically. Mathematicians believe in it because they think that linguists have established it to be a linguistic law, and linguists believe in it because they, on their part, think that mathematicians have established it to be a mathematical law. [...] Rightly seen, the Zipf law is nothing but the arbitrary arrangement of words in a text sample according to their frequency of occurrence. How could such an arbitrary and rather trivial ordering of words be believed to reveal the most recondite secrets, and the basic laws, of language?

We can divide the literature on the subject in two broad categories: empirical curve fitting and model genesis. The first category is by far the more voluminous, running to several thousand scholarly papers and hundreds of monographs. Here we do not even attempt to survey this literature: *QUALICO* conference volumes and the *Journal of Quantitative Linguistics* offer a good entry point. In mathematical statistics, attempts to discern the underlying mechanism that gives rise to a given distribution are called investigations of *model genesis* a particularly successful example is the explanation why normal distribution appears so often in seemingly unrelated areas, provided by the central limit theorems. Given the sheer bulk of the literature supporting some Zipf-like regularity in domains ranging from linguistic type/token counts to the distribution of wealth, it is natural that statisticians sought, and successfully identified, different mechanisms that can give rise to (1-2) or related laws.

The first results in this direction were obtained by Yule (1924), working on a version of (2) proposed in Willis (1922) to describe the number of species that belong to the same genus. Assuming a single ancestral species, a fixed annual probability s of a mutation that produces a new species, and a smaller probability g of a mutation that produces an entirely new genus, Yule shows that over time the distribution for the number of genera $V(i,N)$ with exactly i species will tend to

(3) $$\frac{c}{i^{1+g/s}}$$

which is the same as (2) with $D_N = s/(s+g)$ and $K_N = -s \log(\zeta(1+g/s))/(s+g)$ independent of N (the Riemann ζ function enters the picture only to keep probabilities summing to one).

This is not to say that words arise from a single undifferentiated ancestor by a process of mutation, but as Zipf already noted, the most frequent words tend to be the historically older ones, which also have the highest degree of polysemy. The essential point of Yule's work is that a simple, uniform process of mutation can give rise, over time, to the characteristically non-uniform 'Zipfian' distribution: merely by being around longer, older genera have more chance to develop more species, even without the benefit of a better than average mutation rate.

The same distribution has been observed in patterns of income by Pareto (1897), and there is again a large body of empirical literature supporting Zipf's Law (known in economics as Pareto's Law). Champernowne (originally in 1936, but not fully published until 1973) offered a model where the uneven distribution emerges from a stochastic process (Champernowne 1952, 1953, 1973, see also Cox and Miller 1965) with a barrier corresponding to minimum wealth.

Zipf himself attempted to search for a genesis in terms of a "principle of least effort", but his work (Zipf 1935, 1949) was never mathematically rigorous, and was cut short by his death. A mathematically more satisfying model specifically aimed at word frequencies was proposed by Simon (1955), who derived (2) from a model of text generation based on two hypotheses: (i) new words are introduced by a small constant probability, and (ii) old words are reused with the same probability that they had in earlier text.

A very different genesis result was obtained by Mandelbrot (1952) in terms of the classic "monkeys and typewriters" scenario. Let us designate an arbitrary symbol on the typewriter as a word boundary, and define "words" as maximum strings that do not contain it. If we assume that new symbols are generated randomly, Zipf's law can be derived for $B > 1$. Remarkably, the result holds true if we move from a simple Bernoulli experiment (zero order Markov process) to higher order Markov processes.

In terms of content, though perhaps not in terms of form, the high point of the Zipfian genesis literature is the Simon-Mandelbrot debate (Mandelbrot 1959, 1961a-c; Simon 1960, 1961a,b). Simon's genesis works equally well irrespective of whether we assume closed ($B < 1$) or open ($B \geq 1$) vocabulary. For Mandelbrot, the apparent flexibility in choosing any number close to 1 is a fatal weakness in Simon's model. While we will argue for open vocabulary, and thus side with Mandelbrot for the most part, we believe his critique of Simon to be too strict in the sense that explaining too much is not as fatal a flaw as explaining nothing. Ultimately, the general acceptance of Mandelbrot's genesis as the linguistically more revealing rests not on his attempted destruction of Simon's model but rather on the fact that we see his model as more assumption-free.

2. Exponential and subexponential decay

Arguments based on counting the frequency of various words and phrases are nothing new: in the 1640s a Swedish sect was deemed heretical (relative to Lutheran orthodoxy) on the basis of larger than expected frequency of forms such as *Christ bleeding, Christ suffering, Christ crucified* found in its Sion Psalmbook. With such a long tradition, predating the foundations of modern probability theory by centuries, it should come as no surprise that a considerable number of those employing word counts still reject the standard statistical view of corpora as samples from some underlying population. In particular, Zipf himself held that collecting more data about word frequency can sometimes distort the picture, and there is an "optimum corpus size". For a modern discussion and critique of this notion see Powers (1998), and for an attempt to recast it in a contemporary statistical framework see Baayen (2001:5.2), who

traces the method to papers published in the eighties by Orlov, Chitashvili, and Khmaladze (non vidi).

The central mathematical method of this paper is to make explicit the dependence of certain model parameters on corpus size N, and let N increase without bounds. Since this method only makes sense if we assume the standard apparatus of mathematical statistics and probability theory, in 2.1 we devote some time to defending the standard view. In 2.2 we introduce a simple normalization technique that makes frequency counts for different values of N directly comparable. In 2.3 we compare the normalized distributions to exponential decay, an unrealistic, but mathematically very tractable model. We introduce the more realistic subgeometric mean property in 2.4, and the empirically observable power law of vocabulary growth in 2.5.

2.1. Corpora as samples

Suppose that the primary focus of our interest is the journalistic/nonfiction-literary style exemplified by the Merc, or that even more narrowly, our focus is just the Merc and we have no intention of generalizing our results to other newspapers, let alone other stylistic ranges. While the Merc is a finite corpus, growing currently at a rate of 60m words/year, our goal is not an exhaustive characterization of past issues, but rather predicting word frequencies for future issues as well. Therefore, the *population* we care about is an infinite one, comprising all *potential* issues written in "Merc style" and each issue is but a finite *sample* from this infinite population. The issue we wish to address is whether this population is based on a finite (closed) vocabulary, or an infinite (open) one.

It is often argued that synchronic vocabularies are by definition closed, and only in a diachronic sense can vocabulary be considered open. New words enter the language at a perceptible rate, and the Merc shows this effect as well as any other continually growing corpus. But this process is considerably slower than the temporal fluctuations in word frequency occasioned by certain geographic locations, personages, or products getting in the news, and is, at least to some extent, offset by the opposite process of words gradually falling into disuse. What we wish to demonstrate is openness in the synchronic sense, and we will not make any use of the continually growing nature of the Merc corpus in doing so. In fact, we shall argue that even historically closed corpora, such as Joyce's *Ulysses*, offer evidence of being based on an open vocabulary (see 4.1). This of course makes sense only if we are willing to go beyond the view that the words in *Ulysses* comprise the entire statistical population, and view them instead as a sample of Joyce's writing, or even more narrowly, as a sample of Joyce's writing books like *Ulysses*. It is rather unlikely that a manuscript for a sequel to *Ulysses* will some day surface, but the possibility can not be ruled out entirely, and it is only predictions about unseen material that can lend support to any model. The Merc is better for our purposes only because we can be reasonably certain that there will be future issues to check our predictions against.

The empirical foundation of probabilistic arguments is what standard textbooks like Cramér (1955) call the *stability property of frequency ratios:* for any word w, by randomly increasing the sample Q without bounds, $f_Q(w) = F_Q(w)/N$ tends to some $f(w)$ as N tends to infinity. In other words, sample frequencies must converge to a fixed constant $0 \leq f(w) \leq 1$ that is the *probability* (population frequency) of the word. In the context of using ever-increasing corpora as samples, the stability of frequency ratios has often been questioned on the basis of the following argument. If vocabulary is not closed, the pie must be cut into more and more slices as sample size is increased, and therefore the relative frequency of a word must, on average, decay.

Since word frequencies span many orders of magnitude, it is difficult to get a good feel for their rate of convergence just by looking at frequency counts. The log-log scale used in Zipf plots is already an indication of the fact that to get any kind of visible convergence, exponentially growing corpora need to be considered. Much of traditional quantitative linguistic work stays close to the Zipfian optimum corpus size of 10^4-10^5 words simply because it is based on a closed corpus such as a single book or even a short story or essay. But as soon as we go beyond the first few thousand words, relative frequencies are already in the 10^{-6} range. Such words of course rarely show up in smaller corpora, even though they are often perfectly ordinary words such as *uniform* that are familiar to all adult speakers of English. Let us therefore begin by considering an artificial example, in which samples are drawn from an underlying geometrical distribution $f(w) = 1/2^r$.

Example 1. If the rth word has probability $p_r = 2^{-r}$, in a random sample Q of size $N = 2^m$ we expect 2^{m-1} tokens of w_1, 2^{m-2} tokens of w_2, ..., 2 tokens of w_{m-1}, 1 token of w_m and one other token, most likely another copy of w_1. If this expectation is fulfilled, the frequency ratio based estimate $f_S(w_r)$ of each probability $p_r = f(w_r)$ is correct within $1/N$. Convergence is therefore limited only by the resolution offered by corpus size N, yet the number of types $V(N)$ observed in a sample of N tokens still tends to infinity with $\log_2(N)$.

Discussion. Needless to say, in an actual experiment we could hardly expect to get results this precise, just as in $2N$ tosses of a fair coin the actual value of heads is unlikely to be exactly N. Nevertheless, the mathematical expectations are as predicted, and the example shows that no argument based on the average decline of probabilities could be carried to the point of demonstrating that closed vocabulary is logically necessary. Though not necessary, closed vocabulary is still possible, and it is easy to construct examples, e.g. by using phonemes or syllables instead of words. What we will demonstrate in Theorem 1 is that closed vocabulary is logically incompatible with *observable* properties of word counts.

2.2. Normalization

We assume that population frequencies give a probability distribution over L^*, but for now remain neutral on the issue of whether the underlying vocabulary is open or closed. We also remain neutral on the rate of convergence of frequency ratios, but note that it can be seen to be rather slow, and not necessarily uniform. If rates of convergence were fast to moderate, we would expect empirical rankings based on absolute frequencies to approximate the perfect ranking based on population frequencies at a comparable rate. For example one could hope that any word that has over twice the average sample frequency $1/V(N)$ is already "rank stabilized" in the sense that increasing the sample size will not change its rank. Such hopes are, alas, not met by empirical reality: doubling the sample size can easily affect the ranking of the first 25 items even at the current computational limits of N, 10^9-10^{10} words. For example, moving from a 10m corpus of the Merc to a 20m corpus already affects the rankings of the first *four* items, changing *the, of, a, to* to *the, of, to, a*.

Since sample rank is an unreliable estimate of population rank, it is not at all obvious what Zipf's law really means: after all, if we take any set of numbers and plot them in decreasing order, the results charted on log-log scale may well be approximately linear, just as Herdan, quoted above, suggests. As a first step, we will *normalize* the data, replacing absolute rank r by relative rank $x = r/V(N)$. This way, the familiar Zipf-style plots, which were not scale invariant, are replaced by plots of function values $f(x)$ restricted to the unit square. $f(1/V(N)) = f(w_1)$ is the probability of the most frequent item, $f(1) = f(V(N)/V(N)) = 1/N$ is the probability of the least frequent item, and for technical reasons we define the values of f between $r/(V(N))$ and $(r+1)/V(N)$ to be $p(w_{r+1})$. A small sample (four articles) is plotted in this

style in Fig. 2. Since the area under the curve is $1/V(N)$, by increasing the sample size, plots of this kind get increasingly concentrated around the origin.

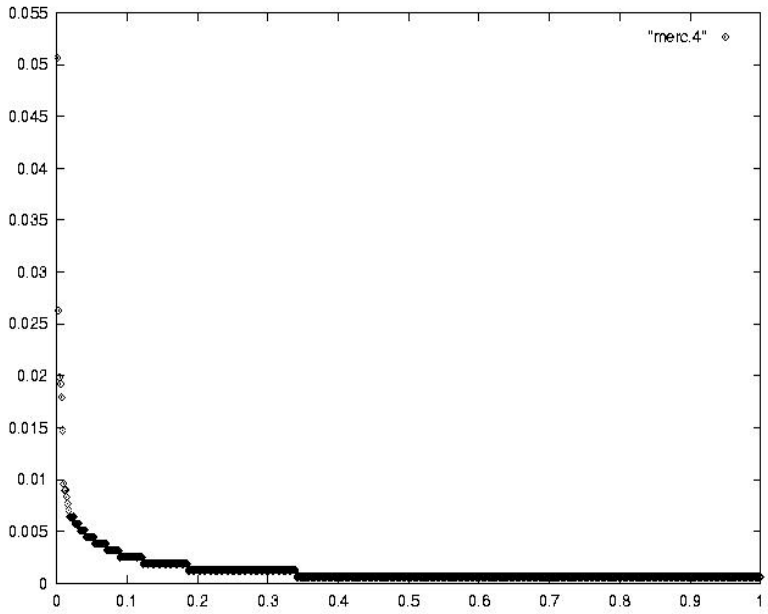

Fig. 2. Frequency as a function of normalized rank (4 articles, 1.5k words)

2.3. Exponential decay

In approximating such a curve an obvious choice would be to try *exponential decay* i.e. $f(x) \sim Ce^{-Dx}$ with some constants $C,D > 0$. However, for reasons that will shortly become apparent, no such curve provides a very good fit, and we merely use the exponential model as a tool to derive *from first principles* a lower bound for $V(N)$. We will use the following facts:

(F1) For any f obtained from a random sample Q of size N, $f(1/V(N))$ tends to p_1, the frequency of the most frequent item, as $N \to \infty$
(F2) For any f obtained from a random sample Q of size N, $f(1) = 1/N$
(F3) Word frequencies decay subexponentially (slower than $\exp(-Dx)$ for any $D > 0$).

Theorem 1. Under conditions (F1-F3) $V(N)$ grows at least as fast as $\log(N)(1-1/N)$.

Proof: $1/V(N) = \sum_{r=1}^{V(N)} f(r/V(N))/V(N)$ is a rectangular sum approximating $\int_0^1 f(x)dx$. Since $f(x)$ is subexponential, for any $g(x) = \exp(-Dx)$ that satisfies $g(1/V(N)) \geq p_1$ and $g(1) \geq 1/N$, we have $g(x) \geq f(x)$ everywhere else in the interval $[1/V(N),1]$, and therefore $1/V(N) < \int_0^1 \exp(-Dx)dx = (1 - \exp(-D))/D$. Using (F2) we compute $D = \log(N)$, and therefore $V(N) \geq \log(N)(1-1/N)$.

Discussion. Since any theorem is just as good as its premises, let us look at the conditions in some detail. (F1) is simply the axiom that sample frequencies for the single most frequent item will tend to its population frequency. Though this is not an entirely uncontroversial assumption (see 2.1), there really is no alternative: if frequency ratios can't be expected to

stabilize even for the most frequent word, there is nothing we can hope to accomplish by measuring them. On the surface (F2) may look more dubious: there is no *a priori* reason for the least frequent word in a sample to appear only once. For example, in closed vocabulary Bernoulli experiments (e.g. phoneme or grapheme counts) we would expect every item to appear at least twice as soon as the sample size is twice the inverse probability of the least frequent item. In the final analysis, (F2) rests on the massively supported empirical observation that hapaxes are present in every corpora, no matter how large (see 3.4).

It may therefore be claimed that the premises of the theorem in some sense include what we set out to prove (which is of course true of every theorem) and certainly in this light the conclusion that vocabulary *must be* open is less surprising. In fact a weaker bound can already be derived from $g(1/V(N)) \geq p_1$, knowing $g(x) = \exp(-Dx)$ and $D = \log(N)$. Since $\exp(-\log(N)/V(N)) \geq p_1$ we have $V(N) \geq \log(N)/\log(1/p_1)$, an estimate that is weakest for small p_1. Baayen (2001: 49) notes that a Turing-Good type estimate (Good 1953) can be used to approximate the rate at which the expected value of $V(N)$ changes by the left derivative $V(1,N)/N$, so that if hapaxes are always present we have $V'(N) \geq 1/N$, and by integrating both sides, $V(N) \geq \log(N)$. While the heuristic force of this simple argument is clear, it is not trivial to turn it into a rigorous proof, inasmuch as Turing-Good estimates are best viewed as Bayesian with a uniform prior over the (finite) set of types, see Nádas (1985).

2.4. The subgeometric mean property

The most novel of our assumptions is (F3), and it is also the empirically richest one. For any exponent D, exponentially decaying frequencies would satisfy the following *geometric mean property*

if r and s are arbitrary ranks, and their (weighted) arithmetic mean is t, the frequency at t is the (weighted) geometric mean of the frequencies at r and s.

What we find in frequency count data is the *subgeometric mean property*, namely that frequency observed at the arithmetic mean of ranks is systematically *lower* than frequency computed as the geometric mean, i.e. that decay is *slower* than exponential: for any $0 \leq p,q < 1$, $p + q = 1$ we find

$$f_{pr+qs} \leq f_r^p f_s^q . \tag{4}$$

In geometrical terms (4) means that $\log(f_r)$ is convex (viewed from below). This may not be strictly true for very frequent items (a concern we will address in 3.1) and will of necessity fail at some points in the low frequency range, where effects stemming from the resolution of the corpus (i.e. that the smallest gap between frequency ratios cannot be smaller than $1/N$) become noticeable. If the rth word has i tokens but the $(r+1)$th word has only $i-1$ tokens, we can be virtually certain that their theoretical probabilities (as opposed to the observed frequency ratios) differ less than by $1/N$. At such steps in the curve, we cannot expect the geometric mean property to hold: the observed frequency of the rth word, i/N, is actually higher than the frequency computed as the geometric mean of the frequency of e.g. the $(r-1)$th and $(r+1)$th words, which will be $\sqrt{i(i-1)}/N$. To protect our Theorem 1 from this effect, we could estimate the area under the curve by segregating the steps up to $\log(\log(N))$ from the rest of the curve by two-sided intervals of length N^ϵ, but we will not present the details here because $\log(N)$ is only a lower bound on vocabulary size, and as a practical matter, not a very good one.

2.5. The power law of vocabulary growth

Empirically it seems quite incontestable that $V(N)$ grows with a power of N:

(5) $$V(N) = N^{\rho}$$

where $0 < \rho < 1$ is some constant, dependent on style, authorship, and other factors, but independent of N (Herdan 1964:157 denotes this constant by C). In practice, we almost always have $\rho > 0.4$, but even for smaller positive ρ the empirical power law would still be stronger than the theoretical (logarithmic) lower bound establised in 2.3 above.

In what follows, we illustrate our main points with a corpus of some 300 issues of the *Merc* totaling some 43m words. While this is not a large corpus by contemporary standards, it is still an order of magnitude larger than the classic Brown and LOB corpora on which so much of our current ideas about word frequencies was first developed and tested, and empirical regularities observed on a corpus this size can not be dismissed lightly.

In one experiment, we repeatedly doubled the size of our sample to include 1,2,...,128 issues. The samples were selected randomly at each step so as to protect our results against arguments based on diachronic drift, and each sample was kept disjoint from the previous ones. If we plot log vocabulary size against log sample size, this experiment shows a remarkably good linear relationship (see Fig. 3), indicating that $V(N) \sim N^q$, with $q \approx 0.75$. A similar "power law" relationship has been observed in closed corpora (including some Shakespeare plays) by Turner (1997).

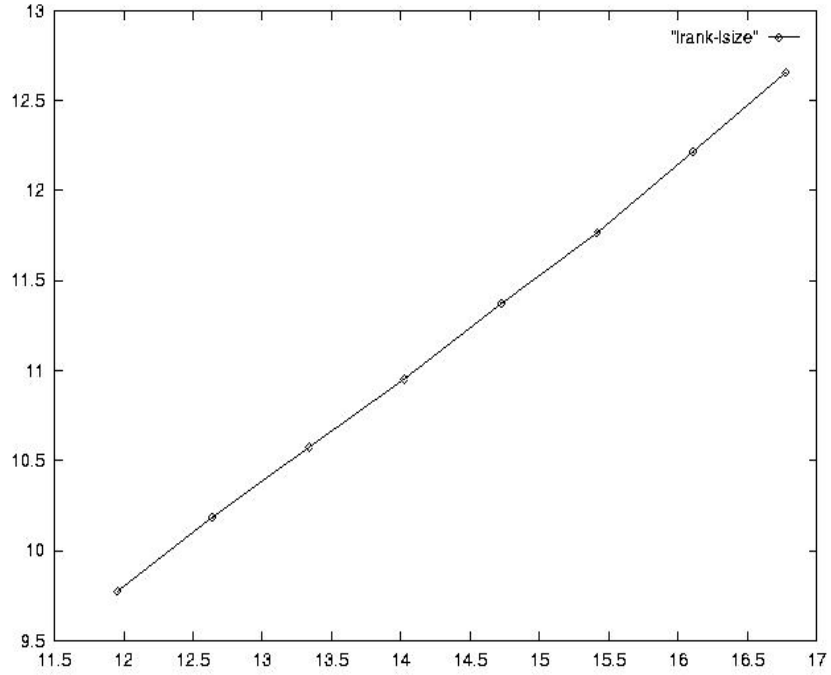

Fig. 3. Growth of vocabulary size $V(N)$ against corpus size N in the Merc on log-log scale

The assumption that a power law relates vocabulary size to sample size goes back at least to Guiraud (1954) (with $\rho = 0.5$) and Herdan (1960). The main novelty in our approach is that we need not postulate (5) as an empirical law, but will derive it as a consequence of Zipf's second law in 3.2. We should add here that (5) is just an approximate empirical law, and some

slower patterns of infinite growth such as $V(N) = N^{D/\log(\log(N))}$ would still look reasonably linear for $N < 10^{10}$ at log-log scale, and would be just as compatible with the observable data.

The lesson that we would like to take away from Theorem 1 is not the quantitative form of the relationship $V(N) \geq \log(N)(1-1/N)$, since this is a rather weak lower bound, but the qualitative fact that vocabulary size grows in an unbounded fashion when sample size is increased. Less than logarithmic growth is logically inconsistent with the characteristic properties of corpora, namely their subexponential decay and that singletons (hapaxes) are present in every corpus, no matter how large. In fact, hapaxes are not only present, they comprise a significant portion h of the word types, a matter we shall return to in 3.4.

3. Frequency extremes

Implicitly or explicitly, much of the work concerning word frequency assumes a Bernoulli-style experimental setup, in which words (tokens) are randomly drawn, with replacement, from a large urn containing all word types in fixed proportions. Though clearly not intended as a psychologically realistic model of speech or writing, it is nevertheless a very useful model, and rather than abandoning it entirely, our goal here is to refine it to fit the facts better. In particular, we follow Mandelbrot's (1961c) lead in assuming that there are *two* urns, a small one U_F for function words, and a larger one U_C for content words. The reasons why high frequency items are expected to behave differently are discussed in 3.1, where the relative sizes of the two urns are estimated by a heuristic argument. In 3.2 we argue that there need be no perceptible break between the two urns, and show how the power law (5) can be derived from (1) using only trivial facts about U_C. A more rigorous treatment of low frequency items is given in 3.3, and "ultra-low" frequency items, *hapax legomena* and *dis legomena* are discussed in 3.4.

3.1. Function words vs. content words

Starting with Herdan (1960) it is common to set aside function words in U_F, since their placement is dictated by the rules of syntax rather than by efforts to choose the semantically appropriate term. The same point can be made with respect to other Zipf-like laws. For example, in the case of city sizes, it stands to reason that the growth of a big city like New York is primarily affected by local zoning laws and ordinances, the pattern of local, state, and federal taxes, demographic and economic trends in the region, and immigration patterns: the zoning laws etc. that affect Bombay are almost entirely irrelevant to the growth of New York. But once we move to mid-sized and small population centers, the general spatial patterns of human settlement can be expected to assert themselves over the special circumstances relevant to big cities. (The issue is more complex for the largest concentrations of individual wealth, because the individuals will often diversify their holdings precisely in order to avoid disproportionate effects of laws and regulations affecting different sectors selectively. If true, this reasoning would suggest that Pareto's Law in economics suffers less from discrepancies at the high end than Zipf's Law in linguistics.) Another reason to set function words aside is that their use is subject to a great deal of idiosyncratic variation, so much so that principal component analysis on the function word counts is effective in separating different authors (Burrows:1987).

Our first task is to estimate the relative sizes of the two urns. Let $f_N(x)$ be a family of $[0,1] \rightarrow [0,1]$ functions with the following properties:

(U1) exponential decay, $f_N(x) = \exp(-D_N x)$
(U2) left limit, $f_N(1/N)$ is a constant, say $\exp(-c)$
(U3) linear area law, $\int_{1/2N}^{(V(N)+1/2)/N} f_N(x)dx = 1/N$.

To fix ideas, the f_N should be thought of as normalized frequency distributions, but the x axis is scaled by N rather than $V(N)$ as before: values of f_N for $x > V(N)/N$ are simply 0. Also, we think of the values $f_N(r/N)$ as providing the ordinate for trapezoidal sums approximating the integrals, rather than the rectangular sums used above. Since the width of the trapezoids is $1/N$ and their height sums to 1, the trapezoidal sum is $1/N$ rather than $1/V(N)$ as before.

From (U1) and (U2) we get $D_N = cN$, which for (U3) gives

$$1/N = \int_{1/2N}^{(V(N)+1/2)/N} \exp(-cNx)dx = \frac{1}{cN}\left[\exp(-c/2) - \exp(-c(V(N)+1/2))\right].$$

Since $V(N) \to \infty$ as $N \to \infty$, the last term can be neglected and we get $c = \exp(-c/2)$. Numerically, this yields $c = 0.7035$ meaning the frequency of the most frequent item is 49.4866%.

While our argument is clearly heuristic, it strongly suggests that nearly half of the tokens may come from function words i.e. the two urns are roughly the same size. An alternative to using two separate urns may be to tokenize every function word as an instance of a catchall ‘functionword' type. The standard list in Vol 3 of Knuth (1971) contains 31 words said to cover 36% of English text, the 150 most frequent used in Unix covers approximately 40% of newspaper text, and to reach 49.5% coverage on the Merc we need less than 200 words. By grouping the appropriate number of function words together we can have the probability of the dominant type approximate 49.5%.

3.2. High frequency items

From the statistical perspective the tokenization process is arbitrary. We may wish to declare *THE, The*, and *the* to be tokens of the same type, or we may wish to keep them separate. We may declare *a* and *an* to be tokens of the same type, or, if we are so inclined, we may even declare *a, an*, and *the* to be tokens of the same ‘article' type. In French we may have good reasons to tokenize *du* as *de* + *le*, in English we may keep *a* + *priori* together as a single token. Because some reorganization of the data in the tokenization step (using only finite resources) is often desirable, it should be emphasized that at the high end we cannot in general expect Zipf-like regularity, or any other regularity.

For example, Fig. 1 completely fails to show the linear pattern predicted by Zipf's law. Furthermore, it has multiple inflection points, so fitting other smooth curves is also problematic at the high end. The geometric mean property is also likely to fail for very high frequency items, but this does not affect our conclusions, since the proof can be carried through on U_C alone, either by segregating function words in U_F or by collecting them in a single functionword type that is added to U_C.

In reality, there is no clear-cut boundary between function words and content words based on rank or other observable properties. In fact, many function words like *on* are homographic to content words: for example, in *The cat is on the mat* we see the locative meaning of *on* rather than the purely prepositional one as in *go on* ‘continue'. To account for this, ideally U_C should also contain some function word homographs, albeit with different probabilities. It would require sophisticated sense disambiguation to reinterpret the frequency

counts this way, and we make no further efforts in this direction here, but note that because of this phenomenon the use of two separate urns need not result in a perceptible break in the plots, even if the functional wordsenses are governed by laws totally different from the laws governing the contentful wordsenses.

It will be evident from Table 1 below that in the Merc no such break is found, and as long as markup strings are lexed out just as punctuation, the same is true of most machine readable material. Several explanations have been put forth, including the notion that elements of a vocabulary "collaborate", but we believe that the smooth interpenetration of functional and contentful wordsenses, familiar to all practicing linguists and lexicographers, is sufficient to explain the phenomenon. Be it as it may, in the rest of 3.2 we assume the existence of some rank boundary k, ($30 < k < 200$) such that all words in $1 \leq r \leq k$ are function words and all words with $r > k$ are content words. As we shall show shortly, the actual choice of k does not affect our argument in a material way.

We assume that the function words have a total probability mass $P_k = \sum_{r=1}^{k} p_r$, (to fix ideas, take $0.3 \leq P_k \leq 0.5$) and that Zipf's law is really a statement about U_C. Normalizing for the unit square, again using $V(N)$ as our normalizing factor, sample frequencies are $f(x)$, with $k/V(N) \leq x \leq 1$. The following properties will always hold:

(D1) right limit, $f_N(1) = 1/N$
(D2) left limit, $f_N(k/V(N))$ is a constant
(D3) area under the curve, $\int_{k/V(N)}^{1} f_N(x)dx = (1 - P_k)/V(N)$.

To this we can provisionally add Zipf's law in the form given in (1), or more directly

(6) $f_N(xV(N)) = \exp(H_N - B_N \log(xV(N)))$.

Condition (D1) means $f(1) = \exp(H_N) = 1/N$ therefore $H_N = -\log(N)$. The logarithmic change in H_N corresponds to the fact that as corpus size grows, unnormalized Zipf plots shift further to the right — notice that this is independent of any assumption about the rate of vocabulary growth. In fact, if we use Zipf's law as a premise, we can state that vocabulary grows with a power of corpus size as

Theorem 2. If corpora satisfy Zipf's law, grow such that assumptions (D1-D2) above hold, and B_N tends to a fixed Zipf's constant B, vocabulary size $V(N)$ must grow with N^{ρ}, $\rho = 1/B$.

Proof. By (D1) we have Zipf's law in the form $f_N(x) = 1/Nx^{B_N}$. If $f_N(k/V(N))$ is to stay constant as N grows, $N(k/V(N))^{B_N}$ must be constant. Since k (the number of function words) is assumed to be constant, we get $\log(N) + B_N \log(k) - B_N \log(V(N))$ constant, and as B_N converges to B, $\log(N) \sim B \log(V(N))$. Therefore, $N = V(N)^B$ within a constant factor.

In our notation, $\rho = 1/B$, and as $V(N) \leq N$, we obtained as a side result that frequency distributions with $B < 1$ are sampling artifacts in the sense that larger samples from the same population will, of necessity, have a B parameter ≥ 1. Thus we find (Mandelbrot 1961c) to be completely vindicated when he writes

> Zipf's values for B are grossly underestimated, as compared with values obtained when the first few most frequent words are disregarded. As a result, Zipf finds that the observed values of B are close to 1 or even less than 1, while we find that the values of B are not less than 1 (p. 196).

We leave the special case $B = 1$ for 3.3, and conclude our investigation of high frequency items with the following remark. Condition (D3) gives, for $B > 1$, $(1 - P_k)/N^{\rho} = \int_{k/N^{\rho}}^{1} 1/(Nx^B)dx = [1 - (k/N^{\rho})^{1-B}]/N(1 - B)$. Differentiating with respect to $k = xN^{\rho}$ gives $\partial P_k / \partial k = k^{-B}$. Therefore at the boundary between content words and function words we expect $p_k \sim 1/k^B$. Looking at four function words in the Merc in the range where we would like to place the boundary, Table 1 summarizes the results.

Table 1
$B = -\log(p_k)/\log(k)$ (estimates)

Word	Rank	Frequency	B
be	30	0.0035	1.66
had	75	0.0019	1.45
other	140	0.0012	1.36
me	220	0.00051	1.41

The point here is not to compute B on the basis of estimated ranks and frequencies of a few function words, but rather to show that a smooth fit can be made at the function word boundary k. The proper procedure is to compute B on the basis of fitting the mid- (and possibly the low-) frequency data, and select a k such that the transition is smooth. As Table 1 shows, our normalization procedure is consistent with a wide range of choices for k.

3.3. Low frequency items

The fundamental empirical observation about low frequency items is also due to Zipf ─ it is sometimes referred to as his "second law" or the *number-frequency law*. Let us denote the number of singletons in a sample by $V(1,N)$, the number of types with exactly 2 tokens by $V(2,N)$ etc. Zipf's second law states that if we plot $\log(i)$ against $\log(V(i,N))$ we get a linear curve with slope close to –1/2. This is illustrated in Fig. 4 below.

Some of the literature (e.g. the web article by Landini (1997)) treats (1) and (2) as separate laws, but really the "second law", $\log(i) = K_N - D_N \log(V(i,N))$, is a straightforward consequence of the first, as Zipf already argued more heuristically.

Theorem 3. If a distribution obeys Zipf's first law with slope parameter B, it will obey Zipf's second law with slope parameter $D = B/(1+B)$.

Proof. For sample size N we have $f_N(x) = 1/Nx^B$, so the probability that an item is between i/N and $(i+1)/N$ if $i \le x^{-B} \le i+1$. Therefore we expect $V(i,N) = V(N)(i^{-\rho} - (i+1)^{-\rho})$. By Rolle's theorem, the second term is $\rho y^{-\rho-1}$ for some $i \le y \le i+1$. Therefore,

$$\log(V(i,N))/(\rho+1) = \log(V(N))/(\rho+1) - \log(\rho)/(\rho+1) - \log(y).$$

Since $\log(\rho)/(\rho+1)$ is a small constant, and $\log(y)$ can differ from $\log(i)$ by no more than $\log(2)$, rearranging the terms we get $\log(i) = \log(V(N))/(\rho+1) - \log(V(i,N))/(\rho+1)$. Since $K_N = \log(V(N))/(1+\rho)$ tends to infinity, we can use it to absorb the constant term bounded by $(\rho-1)/2 + \log(2)$.

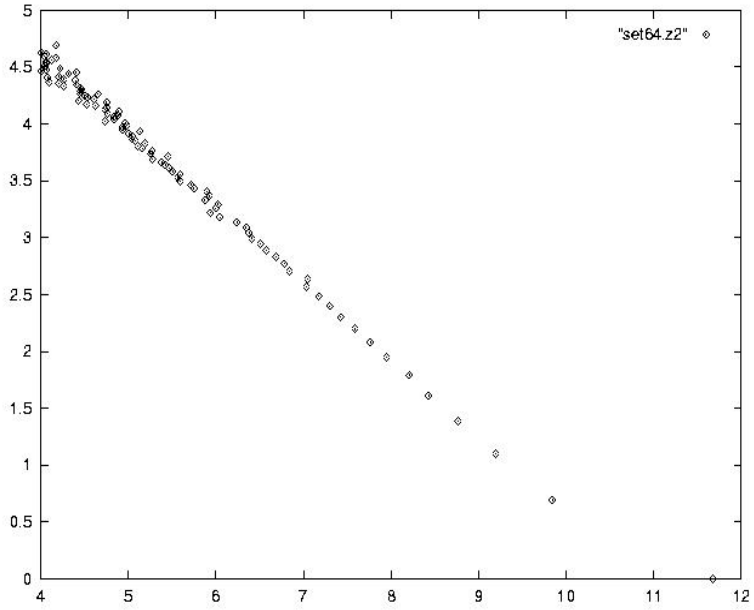

Fig. 4. Number-frequency law on the Merc (10m words)

Discussion. The normalization term K_N is necessitated by the fact that "second law" plots would otherwise show the same drift as "first law" plots. Using this term we can state the second law in a much more useful format. Since $\log(i) = \log(V(N))/(\rho+1) - \log(V(i,N))/(\rho+1)$ plus some additive constant,

(7) $$V(i,N) = mV(N)/i^{\rho+1}$$

where m is some multiplicative constant. If we wish $\sum_{i=1}^{\infty} V(i,N) = V(N)$ to hold we must choose m to be $1/\zeta(\rho+1)$, which is the reason why Zipfian distributions are sometimes referred to as zeta distributions. Since this argument assumes Zipf's second law to extend well to high frequency items, the case for using $m = 1/\zeta(\rho+1)$ is far from compelling, but it is reassuring to see that for $B \geq 1$ we always find a bound constant ($6/\pi^2$ for $B = 1$) that will make the distribution consistent.

Therefore we find Mandelbrot's (1961c) criticism of $B = 1$ to be somewhat less compelling than the case he made against $B < 1$. Recall from the preceding that B is the reciprocal of the exponent ρ in the vocabulary growth formula (5). If we choose a very "rich" corpus, e.g. a table of logarithms, virtually every word will be unique, and $V(N)$ will grow faster than $N^{1-\varepsilon}$ for any $\varepsilon > 0$, so B must be 1. The following example sheds some light on the matter.

Example 2. Let $L = \{0,1,...,9)$ and our word tokens be the integers (in standard decimal notation). Further, let two tokens share the same type if their smallest prime factors are the same. Our size N corpus is constructed by N drawings from the exponential distribution that assigns frequency 2^{-i} to the number i. It is easy to see that the token frequency will be $1/(2^p - 1)$ for p prime, 0 otherwise. Therefore, our corpora will not satisfy Zipf's law, since the rank of the ith prime is i, but from the prime number theorem $p_i \sim i \log(i)$ and thus its log frequency $\sim -i \log(i)\log(2)$. However, the corpora will satisfy Zipf's second law, since, again from the prime number theorem, $V(i,N) = N/i^2(\log(N) - \log(i))$ and thus $\log(V(N))/2 - \log(V(i,N))/2 = \log(N)/2$

– log(log(N))/2 – log(N)/2 + log(i) + log(log(N) – log(i))/2, which is indeed log(i) within 1/log(N).

Example 2 shows that Theorem 3 can not be reversed without additional conditions (such as $B > 1$). A purist might object that the definition of token/type relation used in this example is weird. However, it is just an artifact of the Arabic system of numerals that the smallest prime in a number is not evident: if we used the canonical form of numbers, everything after the first prime could simply be discarded as mere punctuation. More importantly, there is a wealth of other easy to construct examples: as we shall see in Section 4, there are several standard families of distributions that can, when conditions are set up right, satisfy the second law but not the first one with any $B > 1$.

To summarize, Theorem 3 means that distributions that satisfy (1) in the the mid- and the low-frequency range will also satisfy (2) in the low-frequency range. Since the observed fit with (2) is reasonably good, there seems to be no compelling need for a separate urn in the low frequency range. This is in sharp contrast to the high-frequency range, where both theoretical considerations and empirical observations dictate the use of a separate urn.

3.4. Hapax legomena and vocabulary richness

At the extreme low end of the frequency distribution we find *hapax legomena*, types that have only one token. Though misspellings and other errors often end up as hapaxes, it is worth emphasizing that hapaxes are not some accidental contamination of corpora. In the Merc, 46m tokens fall into nearly 600k types, and more than 400k of these (69.8%) are hapaxes. To be sure, over a third of these are numbers (see 5.2), but if we remove numeral expressions from the corpus, we still have 44m tokens, 385k types, of which 218k (56.6%) are hapaxes, consistent with the observation in Baayen (1996) that in large corpora typically more than 50% of the words are hapaxes.

Using $i = 1$ in (7), Zipf's second law predicts that a non-vanishing fraction $mV(N)$ of the vocabulary will be hapaxes, and with $i = 2$ we obtain that roughly a quarter as many will be *dis legomena* (types with exactly two tokens). These predictions have massive support in the quantitative linguistics literature: for example, Herdan (1964:219) only tabulates values of the Waring distribution (see 4.4 below) for the range $0.4 \leq V(1,N)/V(N) \leq 0.6$, because this range covers all values that "are likely to arise in practical work in the area of language".

Baayen (2001:2.4), following Khmaladze (1987, non vidi), defines sequences that have $V(1,N) \rightarrow \infty$ as having a *large number of rare events* (LNRE) if $\lim_{N\rightarrow\infty} V(1,N)/V(N)$ is positive. For a sequence to behave as LNRE it is not necessary for a non-vanishing fraction of *tokens* be rare: in fact, by the power law of vocabulary growth $V(1,N)/N$ will still tend to zero, but a positive fraction h of *types* are rare. As Baayen (2001: 57) notes, word frequency distributions, even when obtained from large samples, are in the LNRE zone. This observation in fact extends to the largest corpora currently available to researchers, web indexes comprising trillions of words, where the ratio of hapaxes is even higher. Assuming $V(1,N) > hV(N)$ we can again use the Turing-Good heuristics (see 2.3 above) for $V'(N) > hV(N)/N$ which, after integration, yields the power law (5) with exponent h.

We can also turn (7) around and use the observed ratio h of hapax legomena to vocabulary size to estimate the theoretical constant m directly, the ratio of dis legomena to vocabulary size to estimate $m/2^{\rho+1}$, and so forth. On the whole we expect better estimates of m from dis legomena than from hapaxes, since the latter also serve as a grab-bag for typos, large numerals, and other marginal phenomena (see 5.2). We can include *tris legomena* and in general use $V(i,N)/V(N)$ to estimate $m/i^{\rho+1}$. Combining the observed numbers of rare words into a single least squares estimate for $2 \leq i \leq 10$, in corpora with at least a few million

words, we can actually obtain better values of the Zipf constant $B = 1/\rho$ than by direct regression of log frequency against log rank.

Clearly, any attempt to model word frequency distributions must take into account the large number of rare words observed, but the large number of hapaxes is only the tip of the iceberg as far as vocabulary growth is concerned. Tweedie and Baayen (1998) survey a range of formulas used to measure vocabulary richness, and argue that many widely used ones, such as the *type-token ratio* $V(N)/N$, fail to define a constant value. In light of the asymptotic considerations used in this paper this comes as no surprise: Guiraud's R, defined as $V(N)/\sqrt{N}$, will tend to zero or infinity if $B < 2$ or $B > 2$ respectively. Dugast's and Rubet's k, defined as $\log(V(N))/\log(\log(N))$, must tend to infinity. Aside from Herdan's C, the main measures of vocabulary richness that can be expected to converge to constant values as sample size increases without bounds are Yule's K, defined as $\sum_{r=1}^{\infty} f_r^2$, entropy, given by $\sum_{r=1}^{\infty} - f_r \log(f_r)$, and in general Good's (1953) spectral measures with $Bt > 1$.

Our results therefore cast those of Tweedie and Baayen in a slightly different light: some of the measures they investigate are truly useless (divergent or converging to the same constant independent of the Zipfian parameter B) while others are at least in principle useful, though in practice estimating them from small samples may be highly problematic. In many cases, the relationship between a purely Zipfian distribution with parameter B and a proposed measure of lexical richness such as K is given by a rather complex analytic relation (in this case, $K = \zeta(2B)/\zeta(B)$) and even this relation can be completely obscured if effects of the high-frequency function words are not controlled carefully. This important methodological point, made very explicitly in Mandelbrot's early work, is worth reiterating, especially as there are still a large number of papers (see Naranan and Balasubrahmanyan 1993 for a recent example) which treat the closed and the open vocabulary cases as analogous.

4. Alternatives to Zipf's Law

The most widely used quantitative frequency laws are (1) and (2) as proposed by Zipf. But there are many alternatives, sometimes with easily identifiable champions, but often simply as communities of practice where using a particular model is taken for granted.

4.1. Minor variations

In many cases authors simply express (1-2) using different notation but an algrebrically equivalent formula such as (3). A more interesting case is when the immediate behavior is slightly different, as in Mizutani's Expression:

$$(8) \qquad \sum_{i=1}^{s} V(i,N) = \frac{V(N)s/N}{as/N + bN}$$

with a, b constants (Mizutani 1989). Another kind of correction to (1) was suggested by Mandelbrot (1961b), who introduces an additional parameter $W > 0$ in order to guarantee that the relative frequencies define a proper probability distribution for $B > 1$:

$$(9) \qquad \log(f_r) = \log(B-1) + (B-1)\log(W) - B\log(r+W).$$

With this correction, $\sum_{r=0}^{\infty} f_r \sim (B-1)W^{B-1}\int_W^{\infty} x^{-B}dx = 1$. If W is kept constant, (7) still leaves something to be desired, inasmuch as it assigns a total probability mass of approximately $N^{(1-B)/B}$ to the region of the curve where $r > V(N)$, but at least this error tends to zero as N tends to infinity.

4.2. Beta

Many kinds of minor corrective factors would be compatible with the available empirical evidence, but not all of them show acceptable limit behavior. A case in point is the beta distribution, which Simon (1955) obtained from a model of text generation embodying two assumptions: (i) new words are introduced by a small constant probability, and (ii) old words are reused with the same probability that they had in earlier text. He gives the resulting distribution in the form

(10) $V(i,N) = AB(i,\rho+1)$

where B is the Beta function. The parameter ρ is the same as in Zipf's laws, as can be seen from comparing the estimates for $V(i,N)/V(i+1,N)$ that can be obtained from (7) and (10). In particular, the case $\rho = 1$ corresponds to Zipf's original formulation of the law as $V(i,N)/V(N) = 1/[i(i+1)]$.

But Simon's assumption (i), linear vocabulary growth, is quite problematic empirically. One example used in Simon (1955) and subsequent work is Joyce's *Ulysses*. The general claim of $V(N) = \alpha N$ is made for *Ulysses* with $\alpha \approx 0.115$. However, instead of linear vocabulary growth, in *Ulysses* we find the same power law that we have seen in the Merc (cf. Fig. 3 above). To be sure, the exponent ρ is above 0.82, while in the Merc it was 0.75, but it is still very far from 1. Leaving out the adjacent chapters *Oxen of the Sun* and *Circe* we are left with roughly three-quarters of *Ulysses*, yielding an estimate of $\alpha = 0.116$ or $\rho = 0.825$. Applying these to the two chapters left out, which have 62743 words total, we can compute the number of words in *Ulysses* as whole based on αN, which yields 31122, or based on N^{ρ}, which yields 29804. The actual number of different words is 30014, so the error of the linear estimate, 3.7%, is over five times the 0.7% error of the power law estimate.

The Shakespeare canon provides another example of a "closed" corpus that displays open vocabulary growth. Plotting $\log(n)$ against $\log(V(n,N))$ as in Figure 4 yields Figure 5 (see below). A least squares estimate of ρ at the tail end of the curve yields about 0.52, quite far from unity. If we restrict ourselves to the very tail, we obtain 0.73, and if we use (7) we get 0.76, numbers still very far from what Simon considers the range of interest, "very close to 1". To summarize, the beta distribution is not an adequate model of word frequencies, because it assumes too many words: linear vocabulary growth instead of the power law observable both on dynamically growing corpora such as the Merc and on static ones such as *Ulysses* or the Shakespeare canon (for separate power law counts on *Antony and Cleopatra* and *Richard III* see Turner 1997).

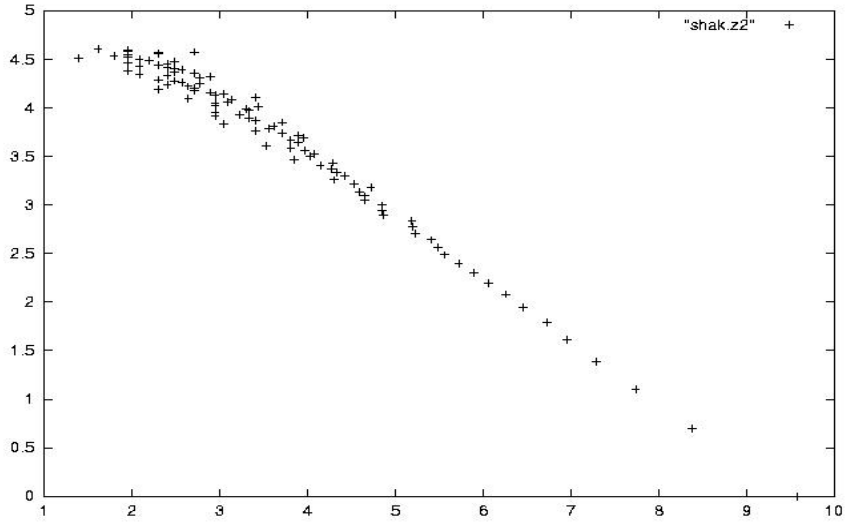

Fig. 5. Number-frequency law on the Shakespeare canon (885k words)

4.3. Lognormal

Another class of distributions that has considerable support in the literature is the *lognormal* family (Herdan 1960). As Champernowne discovered, the same genesis that leads to Pareto's law, when the assumption on minimum wealth is removed, will lead to a lognormal distribution instead. In word frequency counts, the resolution of the corpus presents a minimum barrier (everything that appears must appear at least once), but those in favor of the lognormal hypothesis could argue that this is an artifact of the counting method rather than a structural property of the data.

Theorem 1 proves that under reasonably broad conditions $V(N) \to \infty$, meaning that the average frequency, $1/V(N)$, will tend to zero as sample size increases. But if average frequency tends to zero, average log frequency will diverge. In fact, using Zipf's second law we can estimate it to be $-\log(N)$ within an additive constant R. As the following argument shows, the variance of log frequencies also diverges with $\sqrt{B \log(N)/2}$. To see this, we need to first estimate $f'_N(k/V(N))$, because the functional equation for lognormal distribution,

$$(11) \qquad f_N^2(x) = \frac{-f_N'(x)}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(\log(f_N(x)) - \mu_N)^2}{\sigma_N^2}\right)$$

contains this term. Using the difference quotient we obtain $p_{k+1} - p_k/V(N)$, and we have $V(N) = N^{\rho}$ for some constant $\rho < 1$. By Zipf's law $\log(f_N(x)) = -\log(N) - B \log(x)$. Using (D2) we get that

$$(11) \qquad \frac{1/V(N)}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{(-B\rho \log(N))^2}{\sigma_N^2}\right)$$

is constant, which can hold only if $\rho \log(N) = (1/2)\log(N)^2/\sigma_N^2$ i.e if $\sigma_N^2 = (B/2)\log(N)$.

In other words, the lognormal hypothesis does not lead to a stable limiting distribution: the means drift down with $\log(1/N)$ and the standard deviations open up with $\sqrt{\log(N)}$. This latter divergence, though theoretically more significant than the divergence of the means, is in practice barely noticeable when N grows with the current computational limits of corpus size: the predicted difference between a hundred million word corpus and a ten billion word corpus is less than 12%.

Proponents of the lognormal law could object to the above derivation, pointing out that once we assume that Zipf's first law fits the midrange $N^{\varepsilon} < i < N^{\rho-\varepsilon}$ and Zipf's second law fits the low range, surely the lognormal can not be expected to fit well. (In the high range, introducing a separate urn would benefit the lognormal approach just as much as it benefits the Zipfian.) We should, they might argue, turn the tables and see how well the Zipfian would fit, once we assume the underlying distribution to be lognormal. Because the results would be incompatible with Zipf's laws, the only means of settling the debate is to investigate which of the competing hypotheses fits the data better. Since the fit with lognormal is about as good (or about as bad, depending how we look at it) as with dzeta, there could be no compelling reason to favor one over the other.

While it is true that estimates on the divergence of the lognormal parameters are hard to quantify in the absence of an assumption that the distribution is Zipfian at least to the first approximation, for lower N the divergence is quite strong, and has been noted even by leading proponents of the lognormal hypothesis. Carroll (1967:407) has this to say:

> It will be noted that the mean and the standard deviation vary systematically as a function of sample size; that is, they appear to be biased estimates of the population value. [...] In ordinary sampling theory no bias is expected in the usual measures of central tendency applied to samples of different sizes, and the bias in estimates of the population variance is negligible for large samples and is easily adjusted for by using the number of degrees of freedom as the denominator in the calculation of variance.

Whether this bias is a barely perceptible lack of fit with the data or a gaping hole in the theoretical edifice is a matter of perspective: Herdan (1964:85) repudiates the lognormal largely because "for samples of widely different sizes from the same universe, the conventional statistics, such as the mean and the standard deviation, would not be the same". The main problem with the lognormal is the exact opposite of the problem with beta: beta distribution assumes there to be too many different words, while lognormal requires there to be too few. This problem is quite independent of the numerical details of curve-fitting: a lognormal distribution with fixed means μ and variance σ^2 predicts a fixed $V(N) = \exp(\sigma^2/2 - \mu)$, which Carroll (1967) calls the "theoretical absolute size of the vocabulary". But a fixed upper limit for $V(N)$ is incompatible with the results of our Theorem 1, and more importantly, with the empirically observable power law of vocabulary growth.

4.4. Waring

In a series of influential publications Herdan (1964, 1966) described an alternative to Zipf's law based on the Waring distribution. The basic idea is to model the classes C_i of types that have exactly i tokens: the Waring-Herdan formula asserts that the probability of a type falling in class C_i is given by

$$(12) \qquad \frac{\alpha}{\beta+\alpha}\frac{\beta}{\beta+\alpha+1}\cdots\frac{\beta+i-1}{\beta+\alpha+i}.$$

To account for the case when a token belongs to no previously seen type, the model explicitly contains the class C_0 of words not found in the sample, and assigns probability to it by the $i = 0$ case of the same formula (12). Instead of the number of visible types $V(N)$ we therefore deal with the total number of types $U(N)$ which includes the unseen words as well.

Let us first estimate the appropriate value for α and β. By Zipf's second law, we have $V(i,N)/V(i+1,N) = (1+1/i)^{\rho+1}$, with ρ is a constant < 1. From (12) we get $V(i,N)/V(i+1,N) = (\beta+\alpha+i+1)/(\beta+i)$. Therefore we need

$$1+\frac{\alpha+1}{\beta+i} \sim 1+\frac{\rho+1}{i}$$

which can work well for a range such as $2 \le i \le 200$ only if α is close to ρ and β is close to zero. With this choice, we actually obtain the correct prediction, namely that class C_0 has probability one, meaning that almost all types in the population remain unseen in any sample.

Again, proponents of the Waring distribution could object to the above derivation, especially as it assumes Zipf's second law to provide a reasonable first approximation over a broad range, while in fact the fit in this range is far from perfect. We will therefore use Herdan's own estimators for α and β, which are based on the proportion of hapaxes, $h = V(1,N)/V(N)$ and on the average class size $M = N/V(N)$ as follows:

$$\beta = \frac{1}{\frac{1}{1-h}-\frac{1}{M}-1} \qquad\qquad \beta+\alpha = \frac{\beta}{1-h}.$$

In the larger Merc samples, $1/M$ is already less than 0.01, and if we increase N without bounds, $1/M$ will tend to zero. Thus for very large samples Herdan's estimators yield $\beta = (1-h)/h$ and $\alpha = 1$. Thus we obtain a distribution with a single parameter p, which predicts that in any sample the chances of a random type being manifested as hapax legomena, dis legomena, tris legomena etc. are

$$(13) \qquad \frac{p(1-p)}{1+p}; \qquad \frac{p(1-p)}{(1+p)(1+2p)}; \qquad \frac{p(1-p)}{(1+2p)(1+3p)}; \qquad \ldots$$

These probabilities do not add up to one: the remaining probability, which is p, is assigned to unseen word types. Of course, the whole notion of equiprobable selection of types makes sense only if there is a finite number of types (closed vocabulary): what (13) predicts is that $V(N)/U(N)$, the proportion of visible types among the "theoretical absolute" number of types, is $1-p$.

One way of putting this result is that even if there are infinitely many word types, on the basis of a finite sample, with $V(N)$ manifest types, we can justify no more than $V(N)/(1-p)$ types altogether. This is not dissimilar to the logic underlying the lognormal fit, and in some sense even better, since the lognormal parameters μ and σ diverge as $N \to \infty$, but the Waring parameters α and β appear to converge to stable values 1 and $(1-h)/h$ respectively. Since h, the proportion of hapaxes among the total number of word types seen, is about 1/2, we can conclude that there are roughly as many unseen types as there are visible ones.

Herdan uses h to estimate p, but of course we could apply least squares fit in the entire range of p_i. From (13) we obtain $V(i,N)/V(i+1,N) = 1 + 2p/(1 + ip - p)$, which will be the same as the Zipfian estimate $1 + (\rho + 1)/i$ just in case $\rho = 1/B = 1$. Again, both the Waring and the ζ

distributions fit the observed numbers about equally well for small i, and both leave a lot to be desired for larger i. Therefore, the choice between the two has to be made indirectly, based in this case on the inability of the Waring distribution to support a Zipfian tail for the case of practical interest, $B > 1$.

4.5. Negative binomial

Another influential model originates with the work of Fisher (1943) on species abundance. The main relation can be formulated as

$$\frac{V(i,N)}{V(1,N)} = \frac{\Gamma(i+\alpha)}{i!\Gamma(1+\alpha)}\gamma^{i-1} \tag{14}$$

which is closely related to both (10) and (12). If (12) is to agree with (7) over a broader range, we need

$$1+\frac{\rho+1}{i} \sim \frac{i+1}{\gamma(i+\alpha)}$$

which requires γ to be close to 1 and α to be close to $-\rho$. For example, Efron and Thisted (1976) fits a negative binomial to the Shakespeare data depicted in Figure 5, obtaining $\gamma = 0.9905$, $\alpha = -0.3954$. This translates into an estimate of $\rho \approx 0.4$, much lower than the 0.73 we obtained in 4.2 based on $V(1,N)/V(2,N)$ alone, and still significantly lower than the 0.52 we obtained from the first five $V(i,N)$. Because their method gives nearly uniform weight to the whole range $1 \le i \le 40$ they consider, the discrepancy is quite visible, but for the same range our simpler fit would yield $\rho = 0.36$. Since Theorem 2 indicates that vocabulary growth is determined at the margin, we concentrate at the low end, ignoring values for $i > 5$ entirely.

In general, the negative binomial offers a considerably better fit than Zipf's second law, and in the range of interest, $-1 < \alpha < 0$, yields the same vocabulary growth formula

$$V'(N) = V(1,N)/N \tag{15}$$

as the Turing-Good method, but without assuming a uniform prior on types. As we have seen in 3.2, $V(1,N) = mV(N)$ for some constant $0 < m \le 1$. Combined with (15) we obtain, up to a constant factor, the power law (5) with exponent m. For the Shakespeare canon, $m = 0.46$, roughly halfway between the exponent predicted by (14) and the one computed from $1 \le i \le 5$.

5. Where do the words come from?

The closed vocabulary assumption, that there is a fixed number of words S in any given language, is often couched in terms of rather sophisticated statistical frameworks that assume the existence of unseen words. After discussing this issue in 5.1, we turn to the open vocabulary in 5.2, where we attempt to identify the factors fueling infinite vocabulary growth. We offer our conclusions in 5.3.

5.1. Unseen words

The lognormal and Waring distributions are examples of a broader class of hypotheses that require a distinction between the observed number of types $V(N)$ and the predicted number of types $U(N)$. In some sense, this is a very attractive distinction, for it would surely be absurd to assume that no new sample will ever contain words hitherto unseen. Also, on the basis of such hypotheses, we can obtain quantitative answers to a range of questions:

> What proportion of first names known in small villages are actually in use there? Over 80%, according to the Waring fit used in Schubert and Toma (1983).
>
> How many words could be used in children's reading materials between grades 3 and 9? 609,606, according to the lognormal fit used in Carroll (1971).
>
> How many words did Shakespeare know but never use? At least 35,000, according to the negative binomial fit used by Efron and Thisted (1976).
>
> How many words can appear in Turkish archeological texts given a sample of 7k words? Over half trillion, based on the generalized inverse Gauss-Poisson fit used in Baayen (2001:4.3.1).

It is quite conceivable that in rural Hungary first names were indeed drawn from a rather small closed set, and therefore fitting a Waring or lognormal distribution is appropriate. However, the statistics published in Carroll (1971) give no indication that children's reading materials come from a closed subset of the vocabulary, and extrapolationg by (7) suggest that it would take less than 20 times the current corpus to transcend the 609,606 types predicted by the lognormal fit.

While the third question seems to be about Shakespeare's mental lexicon, in fact Efron and Thisted (1976) posed it in a more conservative fashion: how many new words $\Delta(t)$ do we expect if a new body t times the size of the currently acknowledged Shakespearean canon was discovered? We are certainly in no position to collect 20 times the available Shakespearean corpus (for $t = 0.0004849$ see Thisted and Efron 1987) so let us for the moment pursue the issue based on the Merc.

The hope is that by fitting a Waring distribution, we can describe how many words are known by the journalists at the Merc. In the first sample, $N = 147{,}260$, we estimate $U(N)$ by $V(N)/(1 - V(1,N)/V(N))$, obtaining $U(N) = 35{,}439$. However, in an independent sample of 587k tokens, we find more than this, 38,865 types. For this sample the Waring estimate is $U(N) = 113$k. However, in an independent sample of 4.7m words we find $V(N) = 127$k and $U(N) = 280$k, but we are just as far from the elusive "theoretical absolute size" as we were before, and in an independent sample of 18.3m words indeed we find 310k different words.

What is fueling all this vocabulary growth? If indeed $V(N) \to \infty$ as sample size grows, the answer can not be that the Merc employs, say, a hundred journalists: for the total to come out infinitely large, at least some of them must already know infinitely many words. We are in no position to study the Merc contributors separately, but wherever such studies have been conducted, as on the output of Chaucer, Shakespeare, or Joyce, the same unlimited pattern of vocabulary growth can be seen. In fact, Efron and Thisted (1976) are quite explicit about the fact that their model predicts $\Delta(t) \to \infty$ as $t \to \infty$, they just somehow find this conclusion unpalatable (perhaps because it would make no sense for the case of species abundance that originally gave rise to the model).

From what we have seen so far, not only do writers have infinite vocabularies, they actually provide evidence of this in the (necessarily finite) corpus of their writings. Further,

this conclusion must hold not only for extraordinary literary geniuses but, by the pigeonhole principle, at least for some of the lesser known journalists toiling at the Merc.

5.2. Mixtures

A natural generalization of the Zipfian model is to assume that we have a *mixture* of Zipfians: instead of a single distribution with parameter B we have k independent distributions with parameters $B_1, \cdots, B_k$. In such cases, it follows from Theorem 2 that vocabulary growth will always be dominated by the smallest B_i. One application would be to separate the individual writers of multi-author corpora and inspect their contribution to the overall vocabulary separately. Another application is to look at different classes of words, such as morphologically simple vs. morphologically complex, written using digits vs. letters, etc. For example, we may assume that only 98% of the Merc data are ordinary words, and the remaining 2% are numbers. If we separate the two corpora, indeed we obtain quite different parameters: B_w = 1.65 for words and B_n = 1.31 for numbers. Assuming the power law (5), this means that for corpora with 385m words and beyond, the number of new number types will exceed the number of new word types, even though the number of new number tokens will still be only 2% of the number of new word tokens.

While it is true that the distribution of numbers in the Merc follows the same broad Zipfian patterns as the distribution of words, we certainly do not need an elaborate statistical argument to prove that the number of numbers is infinite, or that writers know infinitely many numbers. To the contrary, our goal here is to protect the conclusion, that vocabulary is open, against the counterargument "yes, but only because it includes numbers". In fact, if we remove numbers, we still find the same Zipfian pattern, and vocabulary growth still obeys the power law (5), though with a smaller exponent.

If we inspect the non-numerical hapaxes more closely, we find that other obviously infinite sources, such as proper names, foreign words, typos and eye-dialect (e.g. *Arrrrrrnnnnnold*) play a significant role. Again we need to assign these to separate mixture components, and argue that the rest still grows without bounds. Remarkably, at this point the bulk of the vocabulary growth is actually provided by productive morphological processes: about 40% of the non-numeric hapaxes are hyphenated, a clear sign of compounding, and over 7% end in the possessive suffix *'s*. Multiply suffixed forms, such as *eclectically* or *ebonizing* provide about 5%, so at 40m words the majority of non-numeric hapaxes are either monomorphemic English words such as *decade* or polymorphemic, but clearly well-formed English. To the extent that numerals freely enter into combination with nouns to form adjectives such as *958-member*, one could even argue that there is no need to treat numbers, proper names, or even foreign words as extragrammatical, but we leave this matter to the side here, as the inclusion of these classes would no doubt weaken our argument in the eyes of many.

5.3. Summary and conclusions

In this paper we answered the question posed in the title by arguing that there are an infinite number of words. We came to this conclusion not on the basis of productive morphological processes, but rather by inspecting the characteristic properties of large corpora, and deriving the open vocabulary result from these properties. Nevertheless, our results support the conclusion that the main grammatical source of infinite vocabulary growth is productive generative morphology, in particular compounding.

We inspected Zipf's law separately for the high-, mid-, and low-frequency ranges. For the high-frequency range we proposed that a separate urn, containing only a few dozen to a few hundred function words, be used, and argued that this urn will contain somewhere between 30% and 50% of the total probability mass. For the mid- and low-frequency range we noted that the frequency plot is log-covex (subgeometric mean property) and that every corpus has hapaxes. Using these properties and a simple normalization technique we proved in Theorem 1 that vocabulary size $V(N)$ tends to infinity as $N \to \infty$.

It is in the middle range that Zipf's law appears strongest, and here estimates of the Zipf constant B clearly give $B > 1$ which corresponds, as we have shown in Theorem 2, to a vocabulary growth rate $V(N) = N^{1/B}$. Theorem 3 established a simple quantitative connection between Zipf's first and second law, suggesting that there is no need to introduce a separate urn for the low range, especially as the B of this urn, were it lower than the B of the mid-frequency urn, would dominate the whole distribution for large N. If separate urns are needed at all, they should be used for numerals, typos, eye-dialect, direct quotations form other languages, and other arguably extragrammatical material that can be seen as contaminating the basic vocabulary pattern.

Altogether, there appears to be considerable empirical support for the classical Zipfian distribution with $B > 1$, both in the Merc and in standard closed corpora such as *Ulysses*. There seems to be no way, empirical or theoretical, to avoid the conclusion that vocabulary size grows approximately with a power $\rho < 1$ of N, and the most widely used competing hypotheses, in particular the beta, lognormal, and Waring distributions, are not well suited for characterizing the observed pattern of word frequencies. The negative binomial, with α negative and γ close to 1, stands out as a realistic alternative to ζ, though a satisfactory genesis, explaining why the cumulative distribution function of the Poisson parameters should follow Γ, is still lacking.

Acknowledgements

The author would like to thank Gabriel Landini, David M.W. Powers, and the anonymous reviewers for valuable suggestions and discussion of earlier drafts of this paper. In particular, the possibility of deriving a logarithmic lower bound for vocabulary size based on the Turing-Good estimates (see the end of 2.3) was called to my attention by reviewer #3.

References

Baayen, R. H. (1996). The effect of lexical specialisation on the growth curve of the vocabulary. *Computational Linguistics 22, 455-480.*

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer Academic Publishers.

Burrows, J. (1987). Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing 2, 61-70.*

Carroll, J. B. (1967). On Sampling from a lognormal model of word-frequency distribution. In: H. Kucera and W. Francis (eds.), *Computational Analysis of Present-Day American English: 406-424.* Providence, RI: Brown University Press.

Carroll, J. B. (1971). *The American Heritage word frequency book*. Boston: Houghton Mifflin.

Champernowne, D. (1952). The graduation of income distributions. *Econometrica 20, 591-615.*

Champernowne, D. (1953). A model of income distribution. *Economic Journal 63, 318-351.*

Champernowne, D. (1973). *The distribution of income*. Cambridge University Press.

Cox, D., Miller, H. (1965). *The theory of stochastic processes*. London: Methuen.

Cramér, H. (1955). *The elements of probability theory*. New York: John Wiley & Sons.

Efron, B., Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika 63, 435-448.*

Estoup, J. (1916). *Gammes Stenographiques*. Paris: Institut Stenographique de France.

Fisher, R., Corbet, A., Williams, C. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology 12, 42-58.*

Good, I. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika 40, 237-264.*

Guiraud, H. (1954). *Les charactères statistiques du vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-token mathematics*. The Hague: Mouton.

Herdan, G. (1964). *Quantitative linguistics.* London: Butterworths.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Berlin: Springer.

Khmaladze, E. (1987). The statistical analysis of large number of rare events. *Technical Report MS-R8804*, Dept of Mathematical Statistics, CWI, Amsterdam: Center for Mathematics and Computer Science.

Knuth, D. E. (1971). *The art of computer programming*. Reading MA: Addison-Wesley.

Landini, G. (1997). Zipf's laws in the Voynich Manuscript. *http://web.bham.ac.uk /G.Landini /evmt/zipf.htm*.

Mandelbrot, B. (1952). An informational theory of the structure of language based upon the theory of the statistical matching of messages and coding. In: W. Jackson (ed.), *Second Symposium on Information Theory.* London.

Mandelbrot, B. (1959). A note on a class of skew distribution functions. Analysis and critique of a paper by H.A. Simon. *Information and Control 2, 90-99.*

Mandelbrot, B. (1961a). Final note on a class of skew distribution functions: analysis and critique of a model due to Herbert A. Simon. *Information and Control 4, 198-216.*

Mandelbrot, B. (1961b). On the thory of word frequencies and on related markovian models of discourse. In: R. Jakobson (ed.), *Structure of language and its mathematical aspects: 190-219.* Providence: American Mathematical Society.

Mandelbrot, B. (1961c). Post scriptum to 'final note'. *Information and Control 4, 300-304.*

Mizutani, S. (1989). Ohno's Lexical Law: Its Data Adjustment by Linear Regression. In: S. Mizutani (ed.), *Japanese Quantitative Linguistics: 1-13.* Bochum: Brockmeyer.

Nádas, A. (1985). On Turing's formula for word probabilities. *IEEE Transactions on Acoustics, Speech, and Signal Processing 33(6), 1414-1416.*

Naranan, S., Balasubrahmanyan, V.K. (1993). Information theoretic model for frequency distribution of words and speech sounds (phonemes) in language. *Journal of Scientific and Industrial Research 52, 728-738.*

Pareto, V. (1897). *Cours d'economie politique*. Lausanne-Paris: Rouge.

Powers, D. M. (1998). Applications and explanations of Zipf's law. In: D. Powers (ed.): *NEMLAP3/CONLL98: New methods in language processing and Computational natural language learning: 151-160.*

Samuelsson, C. (1996). Relating Turing's Formula and Zipf's Law. *Proc. Fourth Workshop on Very Large Corpora*.

Schubert, A., Toma, O. (1984). Estimating the total number of first names based on occurrence frequency (In Hungarian). *Névtani Értesítő 9, 72-80.*

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Simon, H. A. (1960). Some further notes on a class of skew distribution functions. *Information and Control 3, 80-88.*

Simon, H. A. (1961a). Reply to Dr. Mandelbrot's post scriptum. *Information and Control 4, 305-308.*

Simon, H. A. (1961b). Reply to 'final note' by Benoit Mandelbrot. *Information and Control 4, 217-223.*

Thisted, R., Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika 74, 445-455.*

Turner, G. R. (1997). Relationship between vocabulary, text length and Zipf's law. *http://www.btinternet.com/g.r.turner/ZipfDoc.htm*.

Tweedie, F. J., Baayen, R.H. (1998), How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32, 323-352*.

Willis, J. (1922). *Age and area*. Cambridge University Press.

Yule, G. U. (1924). A mathematical theory of evolution. *Philosophical Transactions of the Royal Society B 213, 21ff.*

Zipf, G. K. (1935). *The psycho-biology of language; an introduction to dynamic philology*. Boston: Houghton Mifflin.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, Mass: Addison-Wesley.

Glottometrics 4, 2002, 87-99 *To honor G. K. Zipf*

New perspectives on Zipf's law in linguistics: from single texts to large corpora

Marcelo A. Montemurro[1]
Damián H. Zanette

Abstract. In this paper we revisit Zipf's law in the context of linguistics. The deviations from the original simple power law are analysed and a dynamic model for text generation is proposed whose parameters can be associated with some structural features of languages. Furthermore, for the case of large corpora a novel phenomenology is disclosed. In this case a quantitative description of all the scaling regimes is possible by considering the family of solutions of a single first order differential equation.

Keywords: Zipf´s law, text, corpus

1. Introduction

Human language evolved by natural mechanisms into an efficient system capable of coding and transmitting highly structured information (Nowak, Plotkin, Jansen 2000; Pinker 2000a,b). As a remarkable complex system it allows many levels of description across its organisational hierarchy (Aknajian et al. 1992; Van Dijk 1977; Montemurro, Zanette 2002). In this context statistical analysis stands as a valuable tool in order to reveal robust structural patterns that may have resulted from its long evolutionary history (Montemurro, Zanette 2002; Montemurro, Pury to appear; Gell-Mann 1995).

George K. Zipf (1949, 1965) noted the manifestation of several robust power-law distributions arising in different realms of human activity. Among them, one of the most striking was undoubtedly the one referring to the distribution of word frequencies in human languages. Zipf found that for many texts there is a simple approximate mathematical relation between the rank of a word in a list of all the words used in the text ordered by decreasing frequency and its frequency. If we define the normalised frequency of a word as $f(r) = n(r)/T$, where $n(r)$ is the number of occurrences of the word at position r in the frequency-ordered list and T is the total number of words in the text, then Zipf's law reads:

$$f(r) \propto r^{-z}. \tag{1}$$

In the original formulation of this empirical law, the exponent z was taken to be exactly 1. If instead, the exponent is assumed as a parameter and fitted to empirical data it may take on values slightly different from unity. However, the most striking feature of Zipf's law in not merely its simplicity, but its ubiquitous validity corroborated over a large number of human languages.

[1] Address correspondence to: Marcelo A. Montemurro, International Centre for Theoretical Physics, 34014 Trieste, Italy. E-mail: mmontemu@ictp.trieste.it. Or to Damián H. Zanette, Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche and Instituto Balseiro, 8400 Bariloche, Río Negro, Argentina

Based on some simplified arguments on the structure of language, Benoit Mandelbrot (1983) proposed the following generalisation of the original law: $f(r) = A/(1 + Cr)^z$, where C is a second parameter that needs to be adjusted to fit the data. Ever since the discovery of the law there have been opposite views regarding its origin and significance. It has been shown that this form of the law is also obeyed by random processes that can be mapped onto texts (Li 1992), hence ruling out any sufficient character for linguistic depth inherent to the Zipf-Mandelbrot law. Nevertheless, it has been argued that it is possible to discriminate between human writings and stochastic versions of texts precisely by looking at statistical properties of words that fall beyond the scope where Mandelbrot's generalisation holds (Cohen, Mantegna, Havlin 1997). This last observation draws the attention towards the deviations from Zipf's law observed in empirical data as possible carriers of linguistic content.

In this paper we shall analyse the rank-frequency distributions obtained from real text sources at two different ranges of text sizes. The first one comprises unitary texts for which there is a practical limitation in length. For them, we shall present a dynamical model that explains the empirical behaviour of words frequencies as a consequence of processes acting at two different scales: a global memory driven by context that is essentially related to the interplay between multiplicative and additive dynamics in word selection and a local grammar-dependent effect associated with the combinatorial growth of word forms due to the inflective character of language. In this way we set up a simple dynamical model that is able to reproduce realistic Zipf's distributions and, additionally, has a linguistic interpretation. In the second part, we shall scale up orders of magnitude in text size and analyse the complex phenomenology that unfolds when analysing word frequencies in large-corpora. For this case of mixed text sources we present a complete mathematical framework that accounts for all the observed regimes by considering the family of solutions of a simple first order differential equation.

2. Rank-frequency distribution in unitary texts

In its original form Zipf's law predicts a rank-frequency distribution of the form of eq. (1). However, empirical data from text sources exhibit a more complex rank-frequency distribution. In the three uppermost curves of Figure 1 we show the rank-frequency distribution obtained for three literary works written in English, Spanish and Latin respectively, from top down. In all three cases the qualitative features of the distributions are similar; however, some quantitative differences are worth mentioning. After a short transient corresponding to the most frequent words the distribution behaves like a power law for a range of ranks that for these texts spans little less than two decades. Immediately following the power law regime the distributions starts to fall slightly faster as can be seen by comparing with the pure power laws included in the graph as dot lines. For each curve we have also included the value of the exponent z measured in the power law regime. Note that it deviates appreciably from unity in the English and Latin texts.

Since the discovery of Zipf's law much effort has been devoted to understand its origin and linguistic relevance. In that direction Mandelbrot developed a model based on branching Markov processes that lead to a generalisation of the original law that fitted much better actual data in the high frequency domain. According to Mandelbrot's model Zipf's law was devoid of any linguistic content and its validity was merely a consequence of features shared by very general stochastic processes (Mandelbrot 1983).

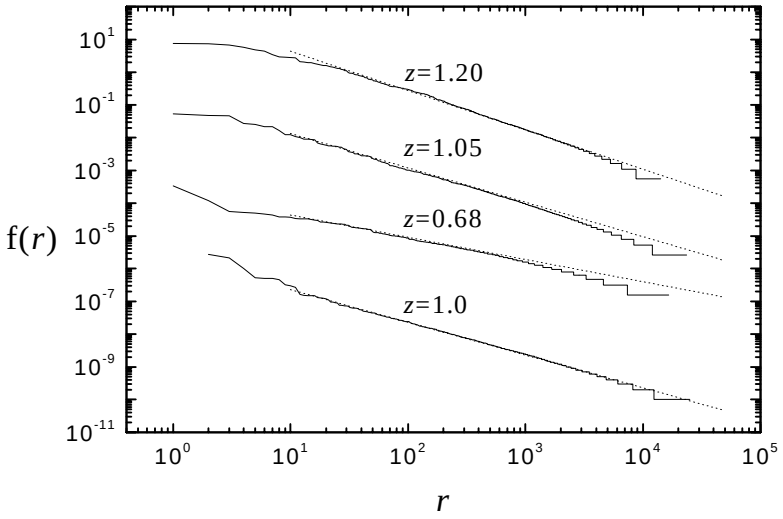

Figure 1. The three uppermost curves are the rank-frequency distributions obtained for David Copperfield by Charles Dickens, Don Quijote by Miguel de Cervantes and Aeneid by Virgil. The fourth distribution represents a realisation of the Simon model. The dotted lines are pure power laws that approximate the distribution in a restricted range of ranks, and the value of the exponents *z* are given for each curve. The curves were displaced in the vertical scale in order to avoid superposition.

Another important step towards understanding the origin of Zipf's law was taken by Herbert Simon. He proposed a mathematical model that describes the generation of a text as a multiplicative process (Simon 1955) that leads asymptotically to Zipf's law. The model sets up a multiplicative process that mimics the dynamics of text generation by specifying how words are added to the text as a function of time. A basic description of the Simon model is as follows. Suppose that at each time step t a new word is added to the text beginning with just one word at $t = 1$. Hence, at any time the length of the text is t. The Simon model says how the new word must be chosen between two alternative mechanisms that try to capture essential features of the actual process of text generation. With a fixed probability α, a new word not present in the text is added at $t+1$ or, with the complementary probability, the new word is taken among the previous t words at random. It is important to note that this dynamic process establishes a strong competition among different word types since the probability of repeating words that have already appeared is proportional to their number of previous occurrences, thereby establishing a multiplicative stochastic dynamics for the use of words. We have included in Figure 1 the results of a simulation of the Simon model, and its comparison with the distribution obtained from real texts reveals a qualitative agreement.

Up to this moment there was no clear known dynamic process based on mechanisms operating at the level of words that can lead to realistic Zipf's distributions. In the next subsection we shall present a modification of the Simon model for Zipf's law that will account for both the differences in slope and the behaviour of infrequent words.

2.1 A dynamic model for text generation

In the Simon model new words are introduced at a constant rate α, such that the vocabulary

size obeys the following simple relation on average

(2) $$V_t = \alpha\ t.$$

A more realistic approach suggests the following modification for the vocabulary growth law

(3) $$V_t = \alpha t^{\nu}$$

with the exponent $\nu < 1$ as a new independent parameter of the model. In this case the growth of the vocabulary size is sublinear and it better approximate the situation observed in real texts (Turner http). Assuming eq. (3) as the growth law for the vocabulary size then the rate per unit time at which new words are introduced will be given by

(4) $$\frac{dV_t}{dt} = \alpha\nu\ t^{\nu-1}.$$

This suggests the first modification of Simon's model. Instead of using the constant rate α as the probability of inclusion of new words, a 'time' dependent probability $\alpha\ t^{\nu-1}$ can be used - (note that we have redefined the prefactor $\alpha\ \nu$ simply as α for brevity in notation. The exponent ν has a clear interpretation since it is related to the rate at which new words are introduced in the text. In a first approximation we can distinguish two main factors that can affect the value of ν. One depends on author and style and can explain small differences of vocabulary growth varying different authors and styles. However, as we shall see, a stronger constraint on ν is imposed by the linguistic structure of the particular language used in writing the text: highly inflected languages like Latin require a higher value of ν than poorly inflected languages such as English. This is reasonable since we are considering two words to be different if they are spelled different. Therefore, in a text written in a highly inflected language word forms will proliferate significantly faster than in languages with few inflected forms.

One of the most important aspects of the Simon model besides the fact that it reproduces Zipf's law in its simplest form, is that it translates neatly the context dependent dynamics of text generation into a multiplicative process. In other words, the chances for a word to be repeated is proportional to the number of its previous occurrences, thus incorporating a memory effect in the process that has global scope. Nevertheless, at this stage another important modification to the original Simon model can be incorporated. At a given moment there will be V different words in the text with n_i (i = 1,...,V) representatives of each. The original dynamics of the model implies that when a word has to be taken from the text written so far the probability that the repeated word belongs to a given type i (i = 1,...,V) is proportional to n_i. As we said above, that multiplicative dynamics models context dependent effects in word occurrences. However, newly introduced words have not a clearly defined context and their dynamics has to be treated in a slightly different way. In this sense, another modification of the model incorporates a word-dependent threshold δ_i, in such a way that the probability of a new occurrence of a word of type i is proportional to $\mathrm{Max}\{n_i, \delta_i\}$. The choice of this set of thresholds is to a great extent arbitrary; however, for simplicity we have inclined for an exponential distribution for which we just have to specify one parameter, namely the mean δ. The dynamic effect of this new addition to the model is that words that have been recently introduced, for which $n_i < \delta_i$, have a slightly higher competition advantage and their reappearance in the text is favoured. Since this parameter has no influence on words for which $n_i > \delta_i$, we expect that the power law region of the distribution, where $n_i >> \delta$, will not be affected by δ_i

In summary the complete model has the following algorithmic dynamics for text generation:

The probability that the word added at time *t* is a new word not present in the text is

(5) $P_t = \alpha t^{\nu-1}$

The probability that the word added at time *t* is chosen among ones already present in the text is

(6) $1 - P_t$

In this last case, the particular word to be repeated is taken from the text with probability proportional to

(7) $Max(n_i, \delta_i)$.

In Figure 2 we show three realisations of the model that fit quite well the empirical data already shown in Figure 1. We can appreciate that the slope of the distribution in the power law region is reproduced accurately for the three literary works. Moreover, the slight concavity of the distribution towards the low frequency regime is also in close agreement with the actual linguistic data. In Table 1 we summarise the value of the different parameters used to fit the distributions.

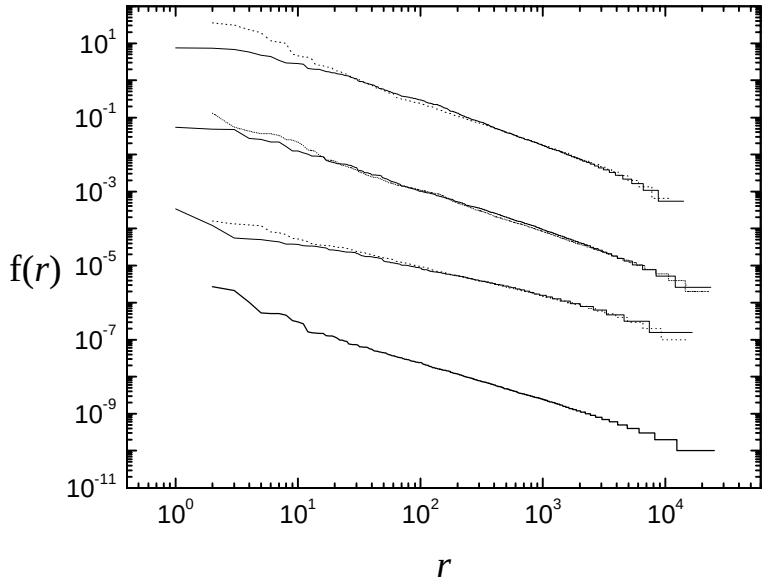

Figure 2. As in Figure 1 the three uppermost curves are the rank-frequency distributions obtained for David Copperfield by Charles Dickens, Don Quijote by Miguel de Cervantes and The Aeneid by Virgil. The fourth distribution represents a realisation of the Simon model. In this case the dotted lines are realisations of the model presented in this paper with the parameters shown in Table 1.

The same analysis was performed using other literary works obtaining consistent results. In particular, the model reproduces Zipf's distributions of texts written in languages with many inflexions, like Latin or Russian for instance, with higher values of *v* than those required to fit Zipf's plots obtained from texts written in less inflective languages such as English or

Spanish. In all cases the faster decay exhibited by the rank-frequency distributions at high ranks was obtained by choosing the parameter δ between 2 and 4. We conclude that linguistically sensible modifications of the Simon model eliminate the slight deviations between the results of the original model and actual Zipf distributions. This gives additional support to the interpretation of Zipf's law as a linguistically significant feature of written texts.

Table 1
Model parameters that reproduce the actual rank-frequency distribution for the three literary works used as test applications of the model presented in this paper.

Source	α	ν	δ
David Copperfield	0.03	0.85	3
Don Quijote	0.05	0.9	2
Aeneidos	0.25	0.95	3

3. Rank-frequency distribution in large corpora

One relevant question concerns the behaviour of word frequencies as the size of the text is increased. Individual texts, with few exceptions, contain less than one million words; therefore, the only possibility to study the form of the rank-frequency distribution beyond that limit implies the mixing of different text sources. As we shall see, a complex new phenomenology will emerge when analysing corpora made up of many individual texts, showing differences according to whether the texts were written by the same author or by different ones.

Although the original form of Zipf's law and its subsequent generalisation by Mandelbrot capture the overall aspects of the rank-frequency distribution, they represent drastic simplifications of the empirical scenario as we have already seen for the case of individual texts. This discrepancy becomes even more apparent for the case of large text samples for which a very rich phenomenology is disclosed.

The generalised expression proposed by Mandelbrot describes very well the statistical behaviour of words in a range from the highest frequencies (low ranks) to Zipf's regime, characterised by the power law region of the distribution. However, the precise mathematical form describing the behaviour of very low frequency words cannot be correctly assessed from individual text samples, since much longer texts should be required to resolve the statistical behaviour of uncommon words.

In Figure 3 we show Zipf's graphs obtained for four large corpora made up of individual texts of four different authors (in the Figure T stands for the number of tokens and V for the vocabulary size). The dots in the graphs represent local averages on windows of constant width in the logarithmic scale. It is apparent that there are three clearly identifiable regimes in the rank-frequency distribution. The first one corresponds to very frequent words and shows variations for each of the corpora considered. The second one is the Zipfian regime where all the curves show a similar power law behaviour from $r \sim 10$ to $r \sim 2{,}000$-$3{,}000$. At that point a third regime appears that is characterised by a faster decay. It is remarkable that all the curves start to deviate from the power law behaviour at approximately the same value of the rank. This suggests that regardless of the different sizes of the texts considered, the vocabularies can be divided into two parts of distinct nature (Ferrer, Solé 2001a,b): one of basic usage whose overall linguistic structure leads to the Zipf-Mandelbrot law, and a second part containing more specific words with a less flexible syntactic function.

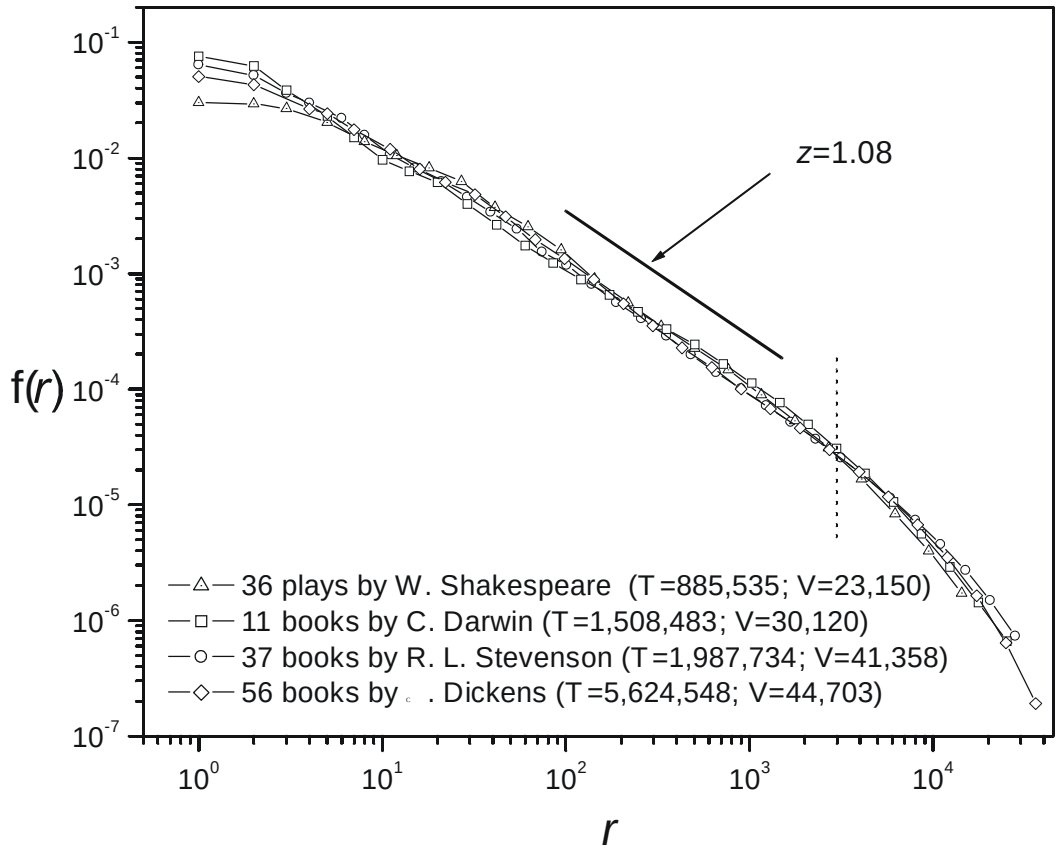

Figure 3. Rank-frequency distribution of words for four large text samples. The vertical dash line is placed approximately where Zipf's law ceases to hold.

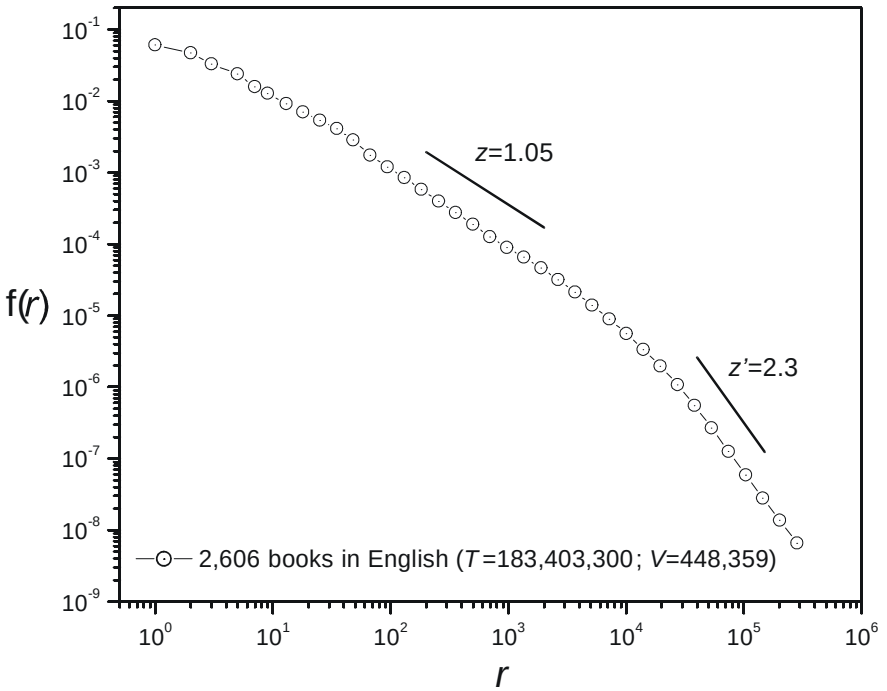

Figure 4. Zipf's plot for a large corpus comprising 2,606 books in English, mostly literary works and some essays. The straight lines in the logarithmic graph show pure power laws as a visual aid.

It is clear that in order to analyse even larger texts we will be compelled to mix works from different authors and styles. In Figure 4 we show the rank-frequency distribution of words in a very large corpus made up of 2,606 books written in English, comprising nearly 1.2GB of ASCII data. The total number of tokens in this case rose to 183,403,300 with a vocabulary size of 448,359 different words. It is remarkable that the point at which the departure from Zipf's law takes place has just moved to $r \sim 6{,}000$ despite the increase in sample size of nearly two orders of magnitude. The most interesting aspect of the curve is given by the emergence of a second power law regime for low frequency words characterised by an exponent that is greater than the one in the first Zipf regime.

In the next section a phenomenological generalisation of Zipf-Mandelbrot law will be presented that will allow unified mathematical description of all the scaling regimes discussed so far.

3.1. Phenomenological generalisation of Zipf's law in large corpora

Zipf's law gained much of its fame because it was a very simple mathematical expression that successfully described a wide range of empirical observations appearing in different scientific contexts. In this paper we have concentrated on the particular nuances in the rank-frequency distribution of words in literary texts, and found that the empirical distributions show a complex behaviour that, nevertheless, emerges as a systematic feature of real distributions. In this section, we shall show that it is possible to gather in a simple mathematical expression all the diverse regimes observed experimentally.

As a first step in that direction we start from the observation that the Zipf-Mandelbrot law satisfies the following first order differential equation:

$$\frac{df}{dr} = -\lambda f^{q}. \tag{8}$$

The solutions to eq. (8) asymptotically take the form of pure power laws with decay exponent $1/(q\text{-}1)$. Moreover, it is possible to generalise this expression in order to include a crossover to a second regime, as follows:

$$\frac{df}{dr} = -\mu f^{q'} - (\lambda - \mu) f^{q} \tag{9}$$

where we have added a new parameter and a new exponent. In the case $1 < q' < q$, and $\mu \neq 0$ the effect of the new additions is to allow the presence of two global regimes characterised by the dominance of either exponent depending on the particular value of f.

We can distinguish three qualitatively different cases in the solutions of eq. (9) according to the values assumed by the parameters. The first case corresponds to the recovery of Zipf-Mandelbrot law by taking $q' = q > 1$ or, which has the same effect, $\mu = 0$ and $q > 1$. In this case the solution of eq. (9) is

$$f(r) = \frac{1}{\left[1 + (q-1)\lambda r\right]^{\frac{1}{q-1}}} \tag{10}$$

where we chose $f(0) = 1$. This expression coincides with Zipf-Mandelbrot law after renaming $C = (q\text{-}1)\lambda$, and $z = 1/(q\text{-}1)$.

As mentioned above, Zipf-Mandelbrot law fits approximately the distribution of words for single texts, but fails when large volumes of data are analysed. However, we shall show that the solutions of eq. (9) describe accurately the rank-frequency distribution of words in all possible situations encountered. If we now take $q' = 1$ and $q > 1$, we obtain

$$f(r) = \frac{1}{\left[1 - \frac{\lambda}{\mu} + \frac{\lambda}{\mu} e^{(q-1)\mu r}\right]^{\frac{1}{q-1}}} . \tag{11}$$

This expression shows a very interesting behaviour for $\mu << \lambda$, since for small values of r it reduces to eq. (10) and then for larger values of r it undergoes a crossover to an exponential decay. Equation (11) describes with great accuracy the rank-frequency distribution obtained for large corpora for individual authors. As an example, in Figure 5 we can see the excellent fit obtained with eq. (11) for one of the large text samples already used in Figure 4. Whereas Zipf-Mandelbrot law would have only fitted a small fraction of the total vocabulary present in these corpora, eq. (11) captures the behaviour of the rank-frequency distribution along the whole range of the rank variable.

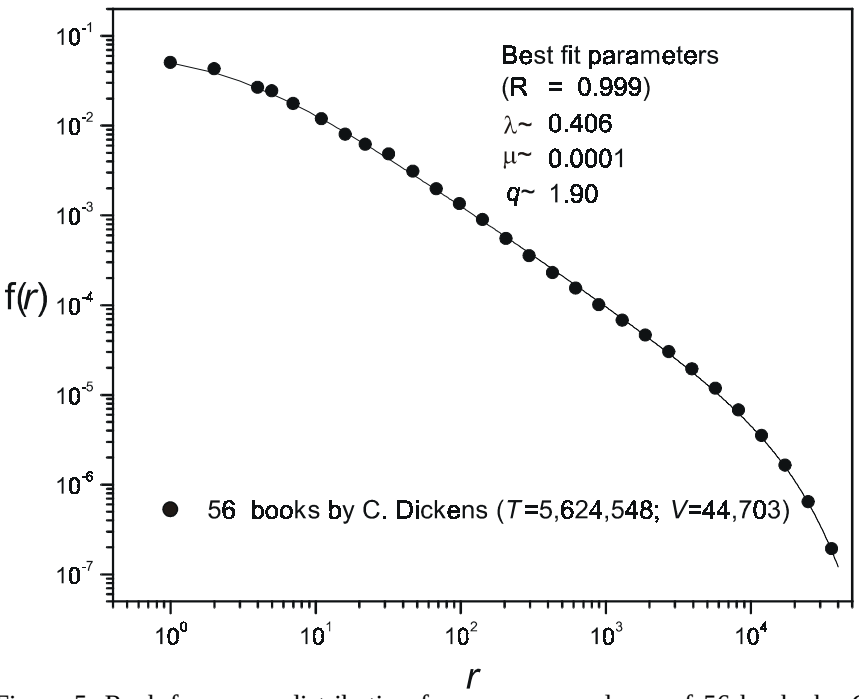

Figure 5. Rank-frequency distribution for a corpus made up of 56 books by Charles Dickens (circles) together with a fit (full line) by eq. (11).

Finally, for the more general situation $1 < q' < q$, the solution of eq. (9) can also be solved explicitly yielding a very involved expression in term of hypergeometric functions (see Montemurro 2001 for the mathematical details). However, it is possible to derive a much simpler relation for the probability density function $p_f(f)$, which is simply the normalised histogram of all the frequencies f. The value of the rank for a word with a normalised frequency f, can be written in the following way :

$$r(f) = V \int_f^\infty p_f(f')\, df' . \tag{12}$$

This can be seen by noting that in the continuous approximation $Vp_f(f')df'$ gives the number of different words that appear with normalised frequency between f' and $f' + df'$. Thus, the corresponding position in the rank list of a word with frequency f equals the total number of different words that have frequencies greater than or equal to f. Based on this we can write

$$p_f(f) \propto -\left(\frac{df}{dr}\right)^{-1} . \tag{13}$$

Consequently, in the probability density representation, the solution is

$$p_f(f) \propto \frac{1}{\mu f^{q'} + (\lambda - \mu) f^{q}} . \tag{14}$$

This result is particularly interesting in view of the mathematical simplicity of eq. (14) and in fact may be taken as the most general form of the generalised Zipf's law in the probability density representation.

Figure 6 shows the probability density function $p_n(n) = p_f(f')/T$ against the number of occurrences n (recall that $f = n/T$ as defined above) for the corpus of 2,606 books, together with the fit obtained with eq. (14). Figure 7 shows the frequency-rank distribution for the same corpus compared with a plot of the explicit solution of eq. (9) using the best fitting parameters presented in Figure 6 . In this case the parameter that controls the second power law regime takes the value q'=1.32, indicating an asymptotic decay exponent close to -3. In Figure 7 the range of the rank variable was extended in order to make evident the whole transient between the two power law regimes.

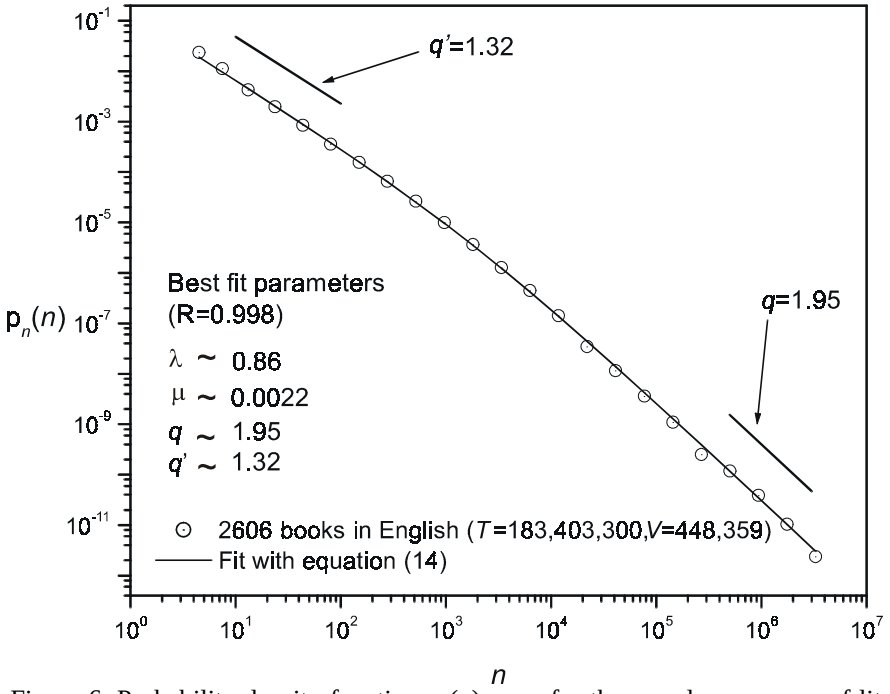

Figure 6. Probability density function $p_n(n)$ vs. n for the very large corpus of literary English and the best fit obtained with equation (14). The straight lines show pure power laws that correspond to the asymptotic forms of equation (14).

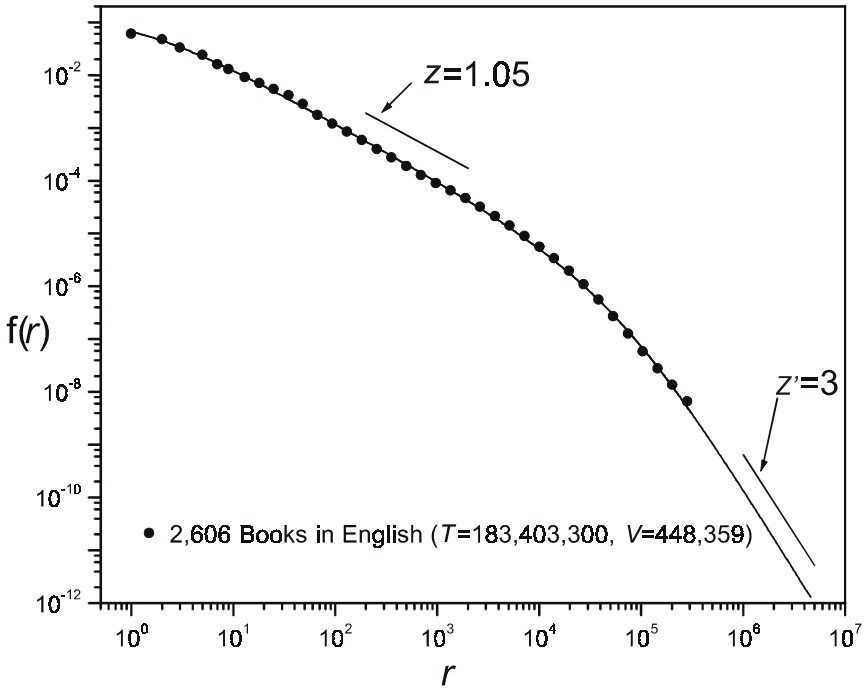

Figure 7. Actual data from the corpus of 2,606 books in English together with a plot of the explicit solution of differential equation (9) with the fitting parameters shown in Figure 6. In the Figure *z=1/(q-1)* and *z'=1/(q'-1)*.

Finally in Table 2 a summary collecting the value of the two main exponents in the distribution is presented for the text corpora analysed in this paper with the addition of a set of seventy books written in classical Latin. In the table we can see clearly that single author corpora systematically yield $q' = 1$, which means exponentially decaying tails in the rank-frequency plot. However, when different authors are combined, the final regime becomes a power law, and consequently $q' > 1$. The understanding of the full implications of these observations requires further interdisciplinary research.

Table 2
Values of the exponents q and q' for different large corpora.

Source	**q**	**q'**
36 plays by W. Shakespeare	1.89	1.00
11 books by C. Darwin	1.90	1.00
56 books by C. Dickens	1.89	1.00
37 books by R. L. Stevenson	1.91	1.00
2606 books in English	1.95	1.32
70 books in classical Latin	2.05	1.23

4. Conclusion

In this paper we presented a new dynamical model for text generation that leads to realistic word-frequency distributions. The model is based on the original one proposed by Herbert Simon and incorporates two key additions. The first one is a new parameter, *v*, that controls

the growth of the vocabulary size as the text is written. We have corroborated that real distributions of highly inflected languages are reproduced using higher values of v than those required to fit rank-frequency distributions of texts written in poorly inflected languages. Consequently, this parameter can be interpreted as a coarse-grained manifestation of the local structure of language as given by grammar. The second addition refers to a word dependent threshold that accounts for the initial context-dependent dynamics of words that have just been introduced in the text. The new parameter δ controls a global aspect of language in the sense that it regulates the long term memory effects associated with context. Furthermore, since it has the direct effect of favouring the repetition of newly introduced words, it contributes to depopulating the low frequency region of the distributions, thus leading to a faster decay for high rank words. The original Simon model was a breakthrough in the sense that it revealed a basic mechanism leading to power law behaviour that agreed approximately with a wide range of empirical observations. However, its extreme simplicity together with the remarkable diversity of the phenomena it described is a clear indication that the model would only capture the most basic underlying mechanisms. The complete understanding of particular distributions like those discussed by Zipf in linguistics or the surname distributions in society (Manrubia, Zanette 2002), for instance, require a deeper quest into the particular dynamic processes acting at different levels. We have also presented a generalisation of Zipf-Mandelbrot law for words for the case of large mixed corpora. After analysing the rank-frequency distributions of words in large text samples we found strong systematic deviations to Zipf-Mandelbrot law that emerged as robust statistical features. These statistical regularities make a complex scenario of different regimes in the distribution, and they could be all encompassed quantitatively by considering the set of solutions of a first order differential equation (9). Further investigation is required in order to propose plausible microscopic mechanisms for the emergence of the complex phenomenology described.

4. Acknowledgements

The authors wish to thank Constantino Tsallis for insightful suggestions. The texts analysed in this work were taken from the *Gutenberg Project* in the Internet http://promo.net/pg/.

References

Aknajian A., Demers, R.A., Farmer, A.K., Harnish R.M. (1992). *Linguistics. An introduction to language and communication:* Cambridge, Mass: MIT Press.

Cohen, A., Mantegna, R.N., Havlin S. (1997). Can Zipf analysis and entropy distinguish between artificial and natural language texts? *Fractals 5, 95-104.*

Ferrer, R., Solé, R.V. (2001a). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics 8, 165-174.*

Ferrer, R., Solé, R.V. (2001b). The small-world of human language. *Proceedings of the Royal Society (London) B 268, 2261-2266.*

Gell-Mann M. (1995). *The Quark and the Jaguar: Adventures in the Simple and the Complex.* New York: Freeman.

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

Mandelbrot B. (1983). *The fractal structure of nature.* New York: Freeman.

Manrubia, S.C., Zanette, D.H. (2002). At the boundary between biological and cultural evolution: the origin of surname distributions. *Journal of Theoretical Biology 216, 461-477.*

Montemurro, M.A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A, 300, 567-578*

Montemurro, M.A., Pury, P.A. (to appear). Long-range fractal correlations in literary corpora. Aaccepted for publication in *Fractals*, in press.

Montemurro, M.A., Zanette, D.H. (2002). Entropic analysis of the role of words in literary texts.' *Advances in Complex Systems 5, No. 1, 7-17.*

Nowak, M.A., Plotkin, J.B., Jansen, V.A.A. (2000). The evolution of syntactic communication. *Nature 404, 495.*

Pinker S. (2000a). *Words and rules.* New York: Harper Collins.

Pinker S. (2000b). *The language instinct.* New York: Harper Collins.

Simon H. (1955). On a class of skew distribution functions. *Biometrika 42, 425-440.*

Turner, G. (http). Relation between vocabulary, text length and Zipf's law. *http://www.btinternet.com/g.r.turner/ZipfDoc.htm*

Van Dijk T.A. (1977). *Text and Context: Explorations in the Semantics and Pragmatics of Discourse.* London: Longman.

Zipf G. K. (1949). *Human behavior and the principle of least effort.* Cambridge, Mass.: Addison-Wesley (1949).

Zipf G. K. (1965). *The psycho-biology of language. An introduction to dynamic philology.* Cambridge Mass: MIT Press.