Complexity in the Structure and Evolution of Language

John Bartholomew

Completed as a part of Physics Honours Course: Complex Systems 2007
Lecturer: Professor Tomaso Aste
**Introduction**

Human language is an ever evolving entity of unparalleled complexity in the natural world [KH 01]. It arises and develops as a result of the interactions between 3 complex adaptive systems - biological evolution, learning and cultural evolution [SBK 03]. These three processes have dictated that the evolution of human language has created the amazing complexity and variety of both spoken and written structure. A remarkable phenomenon is that, despite this enormous diversity of language, there exist common structural elements that permeate throughout all communication. In this paper the inherent structure of language is examined and compared to the results of earlier research and possible language evolution models are suggested for further research.

Much of the motivation for studying complex systems is to develop the ability of prediction with respect to the dynamics of real world phenomenon. However, complex system analysis can also give insight into our history and the dynamics of the processes which have formed and shaped the world in which we reside. The study of language as a complex system is driven predominantly by the latter. By examining language today – how it adapts and develops – insights into the most profound questions of human language can be gained [S 00]. The emergence of written and spoken language and its evolution are two such questions that underlie the majority of research that is conducted in linguistics.

The main focus of this paper is the widespread correlation between the frequency of language elements and the complexity of these elements. Analysis of both spoken and written language is conducted emulating the work of Zipf [Z 35], who initiated such study into linguistics. The experimental data is then compared with the theory proposed by Zipf and extensions to this theory proposed by Mandelbrot [M 53, 62] and Simon [S 55].

The subsequent section of this paper was originally to be devoted to the examination of the evolution of language via a simple Iterated Learning Model (ILM) based on the work of Kirby [KH 01]. However, the limited time frame of the research did not allow the completion of this section. Despite this, the theory and proposed direction for further study are included.
1. Inherent Structure in Language

1.1. Introduction and Theory

Because of the shared feature in all human language systems – human thought processes – the seemingly countless variations of human communication all contain some element of commonality. One particular common structural element that has been demonstrated for numerous spoken and written languages [AJW 98] has become known as Zipf’s Law $^{a}$, after the Professor of Linguistics who initiated interest in this particular area of research.

After studying a wide range of lexical data, Zipf constructed an empirical law to describe the inverse relationship between the frequency with which a speech element is used in a body of written text and the complexity of that element:

$$f_i = \frac{K}{i^\alpha}$$  \hspace{1cm} (1)

where $K$ is a constant and $f_i$ is the sample frequency of the $i^{th}$ type in a ranking according to decreasing frequency [Z 35].

Zipf studied several languages, including English, Chinese, Latin and German, and also examined several hierarchies of language structure such as phonemes, syllables, words and sentences. The expression given in (1) was found to approximate the observed frequency-complexity behaviour in all cases. Zipf’s conclusion was based on psychology: common words are easier to recall and hence are used most often because they require less mental effort - the **Principle of Least Effort** [AJW 98]. However, the quest to adequately explain this phenomenon has piqued the interest of researchers from Zipf’s era to the current day [I 03].

The empirical nature of Zipf’s Law encouraged further analysis to describe the observed complex behaviour through stochastic models. The most notable among the work of Zipf’s peers, include the work of Mandelbrot [M 53,62] and Simon [S 55] who extend Zipfian analysis in different regions via contrasting models [B 91].

Mandelbrot used a Markovlan model, generating words as a string of letters under an assumption to maximize the information transmitted per symbol [M 62]. In doing so he gave formal interpretation to Zipf’s **Principle of Least Effort** and simulated data much more closely for high frequency words. The following shows his result for the frequency of the $i^{th}$ ranked element

$$f_i = \left( \frac{N + \rho}{i + \rho} \right)^{1+\varepsilon}$$  \hspace{1cm} (II)

where $N$ is the total number of distinct words in the text and $\rho$ and $\varepsilon$ are constants. Or more familiarly

$$f_i = \frac{K}{B + i^\alpha}$$  \hspace{1cm} (III)

An interesting note on Mandelbrot’s early work [M 53] was highlighted by Miller [M 57], where he considers the probability distribution of text with respect to the word length rank. Miller derives a form analogous to Mandelbrot’s extension of Zipf’s law via a model of random processes – a monkey on a typewriter.

---

$^{a}$ As Landini [L 97] points out, Zipf actually proposed two laws – a rank-frequency law and a number-frequency law. The earlier is now commonly referred to as Zipf’s Law
The law Miller derives gives the probability of a word, \( w \), of length \( i \) - \( p(w, i) \) - as

\[
p(w, i) = \frac{p(*)}{p(L)} \left[ \frac{2(A - 1)}{A + 1} \left[ r(w, i) + \frac{A + 1}{2(A - 1)} \right] \right]^{-[1 \ln p(L)/\ln A]}
\]

(IV)

where \( r(w, i) \) is the rank order of the word with respect to length, \( p(*) \) is the probability of the monkey typing a space, \( p(L) = 1 - p(*) \) and \( A \) is the number of letters in the alphabet.

The conclusion of his paper was that the optimum solution that Mandelbrot considered was completely simulated by the random placement of spaces in between the alphabetical characters.

Simon’s contribution to linguistic analysis was contained within his more general study of a particular class of functions known as Skew Distribution Functions [S 55]. Such distributions describe certain stochastic processes, which are closely related to negative binomial or log series processes, such as the birth rate [D 53]. Simon proposed a model in which he assumes that the likelihood of the next word generated in a sequence, already existing in that sequence is proportional to the total number of times that particular word has occurred [S 55]. His assumptions give rise to a stochastic process described by the following frequency distribution

\[
f_i = A \beta(i, \rho + 1)
\]

(V)

where \( A \) and \( \rho \) are constants and \( \beta(i, \rho + 1) \) is the Beta function of \( i \) and \( \rho + 1 \) given by:

\[
\beta(i, \rho + 1) = \int_0^1 \lambda^{i-1} (1 - \lambda)^\rho d\lambda = \frac{\Gamma(i) \Gamma(\rho + 1)}{\Gamma(i + \rho + 1)}, \quad 0 < i; \ 0 < \rho < \infty
\]

(VI)

The tail of experimental data is extremely well modelled by Simon’s extension [I 03]. In this regime - \( i \to \infty \) - allowing (V) to be written as

\[
f_i \sim \Gamma(\rho + 1) \lambda^{-(\rho + 1)}
\]

(VII)

The following section describes how the theories proposed by the authors mentioned above were applied to several language data sets. Although the limited extent of the study dictated that no new insights into this area were established, the experiment demonstrated the strength of the correlation between theory and reality.
1.2. Experiment

1.2.1. Word Frequency Analysis

Imitating the work of Zipf, various language compositions were analysed with respect to the frequency of words. The programme used was provided by [R 04], which closely resembles the programme available online as detailed in [AJW 98]. The algorithm for this programme is quite simple – count the number of times each word in the composition occurs. This is achieved in the programme zipf.c by storing the words in a linked list. As suggested in [AJW 98] texts were downloaded from http://www.gutenberg.org - see References for full details.

The output of the programme zipf.c was a text document listing the rank, word and frequency of that word. This output was then transferred to an Excel spreadsheet allowing the data to be plotted as shown in Figures (1-7).

Firstly, the experimental data was plotted in terms of frequency against rank and also in log log space. From the latter a linear approximation for the data was obtained, the equation of which is displayed and the gradient gives the value of $A$. The minimum rank $i_0$, after which Zipf’s Law is obeyed, is obtained by determining the rank at which the experimental data fits the linear trendline. A Zipf’s Law function was then plotted for the particular data set using the exponent $A$ and determining the prefactor $K$ by scaling the function to equal the experimental data at $i_0$. These plots are shown in purple in Figures (1a-7a).

Mandelbrot’s extension was then plotted by trialling various values of $\rho$ and $\epsilon$, concentrating trials for $\epsilon$ where $1 + \epsilon \approx A$. The value of $\rho$ was determined by ceasing the trials when the value of Mandelbrot’s Extension and the value of the experimental data were equal for $i = 1$. These plots are shown in yellow in Figures (1a,b-7a,b).

Simon’s Extension was also plotted for one data set – Great Expectations – by setting $\rho = 1^\alpha$ and scaling the function such that the value of Simon’s Extension and the value of the experimental data were equal for $i = 1$. This plot is shown in light blue in Figure (1a).

1.2.2. Word Length Analysis

Inspired by Miller’s paper [M 57], randomly generated text was also investigated in terms of frequency-length analysis and probability-length analysis. ‘Books’ of randomly generated text were obtained from the programme MonkeyTypewriter.java and analysed in LengthDistribution.java, both written by J. Bartholomew. The random text programme allows the ‘monkey’ to select any letter from the English alphabet or a space with equal probability. When a space is selected a new word is begun and all these words are written to an output text file. The analysis programme counts the number of words of each length and outputs a text file containing this information. Both programmes were run for ‘books’ containing 10, 100, 250, 500, 750, 1000, 10000 and 100000 words respectively. Again, these results were transferred to Excel to allow visualisation of the data and are shown in Figures (8-11).

The output text files of MonkeyTypewriter.java were also analysed manually in terms of Miller’s theory. The analysis was performed for 2 data sets, 1000 words and 10000 word, and in the first case the number of occurrences of words of 1 or 2 letters was recorded, whilst in the latter case only the words of length 1 were considered. These were compared against (VI) with $p(*) = 1/27$, $p(L) = 26/27$, $A = 26$. The results are shown in Figures (12-13).

---

* Following the results of Simon [S 55]
1.3. Results and Analysis

1.3.1. Word Frequency Analysis

The results for this section are summarized in Table 1.

<table>
<thead>
<tr>
<th>Frequency vs Rank for Great Expectations</th>
<th>Ln(Frequency) vs Ln(Rank) for Great Expectations</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Frequency vs Rank for Great Expectations](FIG 1a)</td>
<td>![Ln(Frequency) vs Ln(Rank) for Great Expectations](FIG 1b)</td>
</tr>
<tr>
<td>![Ln(Frequency) vs Ln(Rank) for Great Expectations](y = -1.2047x + 11.041)</td>
<td>![Ln(Rank)](y = -1.2047x + 11.041)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequency vs Rank for MacBeth</th>
<th>Ln(Frequency) vs Ln(Rank) for MacBeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Frequency vs Rank for MacBeth](FIG 2a)</td>
<td>![Ln(Frequency) vs Ln(Rank) for MacBeth](FIG 2b)</td>
</tr>
<tr>
<td>![Ln(Frequency) vs Ln(Rank) for MacBeth](y = -1.0616x + 8.3)</td>
<td>![Ln(Rank)](y = -1.0616x + 8.3)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Frequency vs Rank for War of the Worlds</th>
<th>Ln(Frequency) vs Ln(Rank) for War of the Worlds</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Frequency vs Rank for War of the Worlds](FIG 3a)</td>
<td>![Ln(Frequency) vs Ln(Rank) for War of the Worlds](FIG 3b)</td>
</tr>
<tr>
<td>![Ln(Frequency) vs Ln(Rank) for War of the Worlds](y = -1.0722x + 9.2876)</td>
<td>![Ln(Rank)](y = -1.0722x + 9.2876)</td>
</tr>
<tr>
<td>Source Data</td>
<td>Form and Language</td>
</tr>
<tr>
<td>-------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>C. J. H. Dickens, <em>Great Expectations</em>, (1861)</td>
<td>Written Text English</td>
</tr>
<tr>
<td>W. Shakespeare, <em>Macbeth</em></td>
<td>Written prose English</td>
</tr>
<tr>
<td>Compiled Transcripts from ABC National The Science Show 24/03/07</td>
<td>Spoken Interviews English</td>
</tr>
</tbody>
</table>

Table 1: Summary of Zipfian Analysis
The overall results of the rank-frequency analysis showed extremely good correlations with the empirical formula proposed by Zipf and also of the derived model proposed by Mandelbrot.

As shown in Figures (1b-7b), above some threshold rank value \( i_0 \), Zipf’s law is obeyed in all the data sets analysed. The values of \( A \) are all within \( \pm 0.2 \) of 1, which agrees with literature already published in this area [M 53, B 91, I 03]. Although there is not enough data to state with certainty that Zipf’s Law is obeyed for a wide sample of languages, this data supports such assertions made by [AJW 98] for example, as English data, both written and spoken, and written German data show adherence to the proposed law.

Mandelbrot’s Extension to Zipf’s law models the data with much greater accuracy. With one exception, the model proposed by Mandelbrot approximates the experimental data much more satisfactorily than Zipf’s Law, most apparently in the region \( 1 \leq i < i_0 \). Although in this region the values for the model are still greater than the observed trend the correlation is quite strong. The one exception was for the data set Epistola Ad Pisones, De Arte Poetica originally written in Latin and translated into English. In this case a Mandelbrot Extension plot was not able to be generated due to the fact that Zipf’s Law almost exactly approximates the experimental data in its entirety. From the original Latin, this may seem more reasonable as meaningless words such as the, and, for and alike, are not present. However, translation into English removes this explanation. Another conclusion that could be drawn is that the age of the original language determines the stronger adherence to Zipf’s law. However, further study of both the original Latin text and numerous other texts from that time period is necessary to confirm such a hypothesis.

The values for \( \rho \) and \( \epsilon \), obtained in the Mandelbrot model were such that they agree very well with the published results of [D 02], who gives that \( 0 < \epsilon << 1 \) and \( 0 < \rho < 10 \) for large texts. The two exceptions to this agreement were the text described above and the transcript of the spoken text. This failed to meet the criterion for \( \epsilon \), however, the deviation from published results is very small and hence no correlation between spoken data and this deviation is considered to exist.

Simon’s Extension predominantly models the tail of data and because of the format chosen to present the findings of this research \(^a\), the focus was on the head of the data where the different models are easily distinguished. From the one plot obtained it is apparent that Simon’s model very poorly simulates the head section of experimental data. However, not enough trials were conducted in respect to this model to confirm this hypothesis.

\(^a\) In some cases Excel lacked the precision to distinguish between the models in the low rank region
1.3.2. Word Length Analysis

The length-frequency analysis demonstrated an excellent correlation with the expected distribution for text generated under the assumptions described in 1.2.2. The data tends to an exponential distribution as the number of words in the ‘book’ increases (Figures 8-10). This is precisely what is expected. For a low number of words, random processes have a significant effect on the data and the exponential distribution is only weakly displayed. However, as the number of words increases the distribution tends towards that generated via a Poisson process, which is exactly what the random text generator is. The ‘monkey’ chooses a new symbol at a constant rate and there is a 1/27 chance that it chooses a space, beginning a new word. Hence, the expected exponent for the distribution is 1/27 = 0.037. A linear fit to the data presented in Figure (11) gives an exponent of 0.040±0.05, which adheres to the figure calculated via Poissonian statistics.
The results from the Miller analysis showed little correlation between theory and reality. The data obtained, as shown in Figures (12-13), demonstrates a slight tendency towards the distribution shape predicted by Miller [M 57], however the experiment is not well fitted by the theoretical line. It is thought likely that insufficient words were generated for the alphabet size – 26 symbols - for the proposed relationship to be demonstrated. However, further analysis using random text generated from a smaller alphabet and a study of the number of words necessary for the data to exhibit correlation to the theory would be necessary to draw decisive conclusions.

2. Evolution in Language

2.1. Introduction and Theory

Although common structures that appear in language are fascinating in themselves, research into the complexity of language is directed to deeper questions such as the emergence and evolution of language. These fundamental processes of this dynamic adaptive system may prove to be the crucial element to discovering the very origins of written and spoken human communication. However, in addition to the desire for greater understanding of our past, the motivation to provide technology for the future also fuels interest into studying language as a complex system. The commercial interest in an artificial intelligence capable of demonstrating human level language development is only limited by one’s imagination concerning the applications of such an advance. Hence, there is a fascinating interweaving of present, past and future in the study of the evolution of language.

Language as a complex dynamic system can be studied in terms of the emergence of some innate linguistic ability from a biological evolution perspective. That is, the study of how language in its spoken form originated. However, more commonly, and the focus of this brief section, is research at the level of cultural evolution of specific languages and their syntax and semantics. The ideas presented in this section closely follow the model put forward by Kirby [KH 01] who studied the emergence of compositionality as a fundamental structural property of language arising from cultural evolution.

Kirby presents an ILM based on the following algorithm [KH 01, K 06]. A parent teaches a child meaning-language pairs where the meaning is created from a finite meaning space, for example {Bob, Alice, Jack, Knows, Admires} and language is generated from a finite language space {a, b, c, d}. Thus, an example of a meaning-language pair could be – Bob admires Jack, with language representation $abddc$. The child incorporates these meaning-language pairs and then compresses the information by replacing frequent and long substrings with non-terminals from the space {S, T, U, V, W} for example. The final process is the child’s generalisation of the language where two non-terminals used in the same context are equated. This process is repeated for thousands of generations.

Further details on ILM and computer based language evolution models can be found in the following references [KH 01, K 06, SBK 03].

2.2. Proposed Direction for Future Study

As indicated in the introduction to this paper, time constraints prevented the completion of an ILM. The algorithm was written to mimic that outlined in 2.1 but instead of generalising grammar, the language evolved on the basis of a fitness defined by Zipfian analysis. That is, the most common meaning passed to the child would have its language representation set to the shortest language element passed with the meaning.
It was hoped that the shorter meanings such as *Bob admires Jack* would eventually be represented by short phrases or words and statements such as *Jack knows Alice knows Bob knows Alice admires Jack* would be represented by longer phrases or words.

In summary, the second section was originally designed to mimic language evolution in a Zipfian world and it would have provided an interesting insight into the development of language.

3. Conclusions

The research outlined in this paper, although limited in extent, provided some interesting results concerning the structure of language and the possible applications of this information. The strong adherence of a wide variety of language data sets to Zipf’s Law and Extensions to this law proposed by Mandelbrot have been demonstrated. The relationship between such Zipfian analysis and the structure of randomly generated text was also briefly explored, although no firm conclusions can be drawn from the small sample data collected. The possibility of applying such structures to language evolution models was also suggested and further study opportunities may provide further insight into the adaptation of language in a Zipfian world.
References

Texts Analysed

- C. J. H. Dickens, *Great Expectations*, (1861)
  Available online at http://www.gutenberg.org/dirs/etext98/grexp10a.txt
- W. Shakespeare, *Macbeth*
  Available online at http://www.gutenberg.org/dirs/etext97/1ws3410.txt
- H. G. Wells, *War of the Worlds*, (1898)
  Available online at http://www.gutenberg.org/files/36/36.txt
- F. W. Nietzsche, *Menschliches, Allzumenschliches*, (1877)
  Available online at http://www.gutenberg.org/dirs/etext05/7msch10.txt
  Available online at http://www.gutenberg.org/dirs/etext05/7artp10.txt
  Available online at http://www.gutenberg.org/files/61/61.txt

Available online at http://alpha2.infim.ro/~ltpd/Zipf_Law.html
Available online at http://web.bham.ac.uk/G.Landini/evmt/zipf.htm
Available online at http://www.eee.bham.ac.uk/russelm/ee3j2%5Czipf.c