Extension of Zipf's Law to Words and Phrases

Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, F. J. Smith
School Computer Science
Queen's University of Belfast
Belfast BT7 1NN, Northern Ireland
q.le@qub.ac.uk

Abstract

Zipf's law states that the frequency of word tokens in a large corpus of natural language is inversely proportional to the rank. The law is investigated for two languages English and Mandarin and for n-gram word phrases as well as for single words. The law for single words is shown to be valid only for high frequency words. However, when single word and n-gram phrases are combined together in one list and put in order of frequency the combined list follows Zipf's law accurately for all words and phrases, down to the lowest frequencies in both languages. The Zipf curves for the two languages are then almost identical.

1. Introduction

The law discovered empirically by Zipf (1949) for word tokens in a corpus states that if f is the frequency of a word in the corpus and r is the rank, then:

$$f = \frac{k}{r} \qquad (1)$$

where k is a constant for the corpus. When $log(f)$ is drawn against $log(r)$ in a graph (which is often called a Zipf curve), a straight line is obtained with a slope of –1. An example with a small corpus of 250,000 tokens is given in Figure 1. Zipf's discovery was followed by a large body of literature reviewed in a series of papers edited by Guiter and Arapov (1982). It continues to stimulate interest today (Samuelson, 1996; Montermurro, 2002; Ferrer and Solé, 2002) and, for example, it has been applied to citations Silagadze (1997) and to DNA sequences (Yonezawa & Motohasi, 1999; Li, 2001).

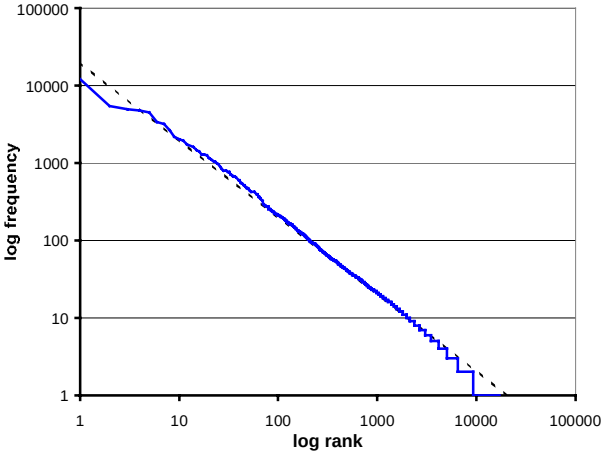

Figure 1 Zipf curve for the unigrams extracted from a 250,000 word tokens corpus

Zipf discovered the law by analysing manually the frequencies of words in the novel "Ulysses" by James Joyce. It contains a vocabulary of 29,899 different word types associated with 260,430 word tokens.

Following its discovery in 1949, several experiments aided by the appearance of the computer in the 1960's, confirmed that the law was correct for the small corpora which could be processed at that time. The slope of the curve was found to vary slightly from –1 for some corpora; also the frequencies for the highest ranked words sometimes deviated slightly from the straight line, which suggested several modifications of the law, and in particular one due to Mandelbrot (1953):

$$f = \frac{k}{(r+\alpha)^{\beta}} \qquad (2)$$

where α and β are constants for the corpus being analysed. However, generally the constants α and β were found to be only small statistical deviations from the original law by Zipf (exceptions are legal texts which have smaller β values ($\approx$0.9) showing that lawyers

use more words than other people!)(Smith & Devine, 1985).

A number of theoretical developments of Zipf's law had been derived in the 50's and 60's and have been reviewed by Fedorowicz (1982), notably those due to Mandelbrot (1954, 1957) and Booth (1967). A well-known derivation, due to Simon (1955), is based on empirically derived distribution functions. However, Simon's derivation was controversial and a correspondence in the scientific press developed between Mandelbrot and Simon on the validity of this derivation (1959-1961); the dispute was not resolved by the time Zipf curves for larger corpora were beginning to be computed.

The processing of larger corpora with 1 million words or more was facilitated by the development of PC's in the 1980's. When Zipf curves for these corpora were drawn they were found to drop below the Zipf straight line with slope of –1 at the bottom of the curve, starting for rank greater than about 5000. This is illustrated in Figure 2 which shows the Zipf curve for the Brown corpus of 1 million words of American English (Francis & Kucera, 1964).

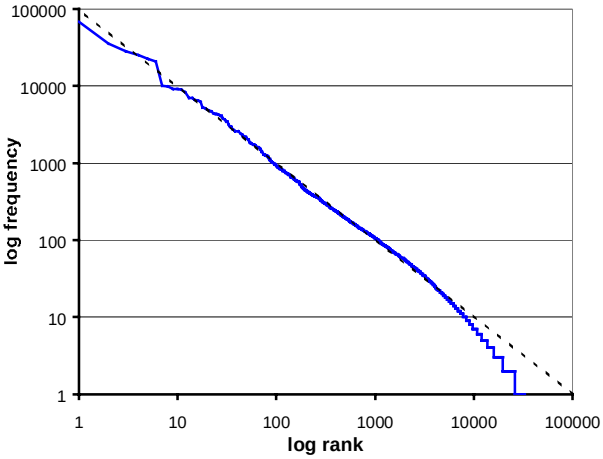

Figure 2 Zipf curve for the unigrams extracted from the 1 million words of the Brown corpus

This appeared to confirm the opinion of the opponents of Simon's derivation: the law clearly did not hold for r>5000; so it appeared that the derivation must be invalid.

2. Zipf Curves for Large Corpora

This paper is principally concerned with exploring the above invalidity of Zipf's law for large corpora in two languages, English and Mandarin. We begin with English.

English corpora

The English corpora used in our experiments are taken from the Wall Street journal (Paul & Baker, 1992) for 1987, 1988, 1989, with sizes approximately 19 million, 16 million and 6 million tokens respectively. The Zipf curves for the 3 corpora are shown in Figure 3.

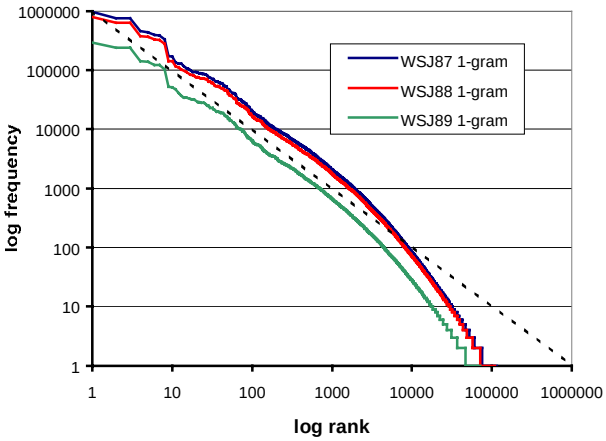

Figure 3 Zipf curves for the unigrams extracted from the 3 training corpora of the WSJ

The curves are parallel, showing similar structures and all 3 deviating from Zipf's law for larger *r*. Their separation is due to their different sizes.

Language is not made of individual words but also consists of phrases of 2, 3 and more words, usually called n-grams for n=2, 3, etc. For each value of n between 2 and 5, we computed the frequencies of all n-gram in each corpus and put them in rank order as we had done for the words. This enabled us to draw the Zipf curves for 2-grams to 5-grams which are shown along with the single word curves in Figures 4, 5 and 6 for the three corpora. These curves are similar to the first Zipf curves drawn for n-grams by Smith and Devine (1985); but these earlier curves were for a much smaller corpus.

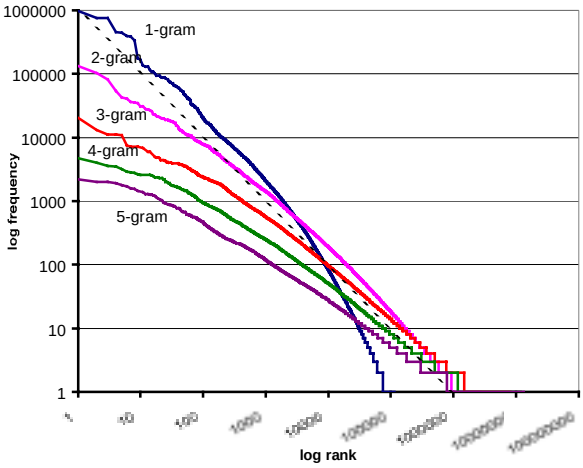

Figure 4 Zipf curves for the WSJ87 corpus

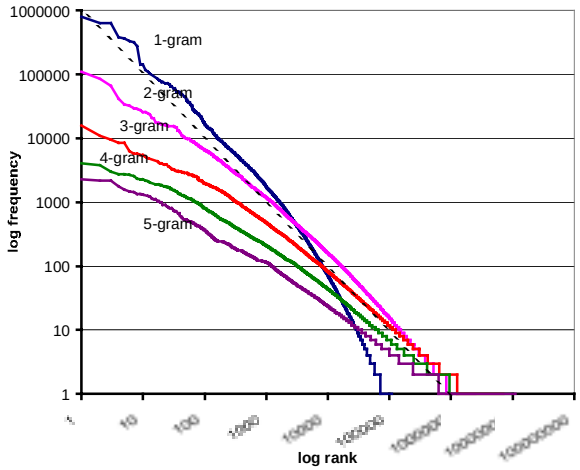

Figure 5 Zipf curves for the WSJ88 corpus

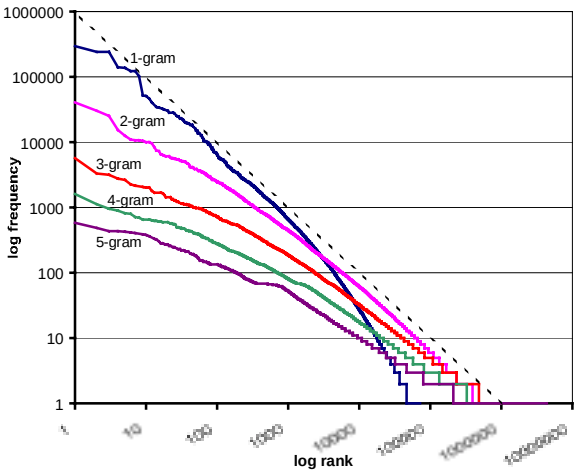

Figure 6 Zipf curves for the WSJ89 corpus

The n-gram Zipf curves approximately follow straight lines and can be represented by a single Mandelbrot form:

$$f = \frac{k}{r^{\beta}} \qquad (3)$$

where β is the magnitude of the negative slope of each line. We found the values of β for the WSJ in Table 1.

Table 1 Slopes for best-fit straight line approximations to the Zipf curves

	WSJ87	WSJ88	WSJ89	Mandarin
2-gram	0.67	0.66	0.65	0.75
3-gram	0.51	0.50	0.46	0.59
4-gram	0.42	0.42	0.39	0.53
5-gram	0.42	0.41	0.34	0.48

Note that the unigram curves crosses the bigram curves when the rank ≈ 3000 in all three cases.

The ten most common words, bigrams and trigrams in the combined WSJ corpus of 40 million words are listed in Table 2.

Mandarin corpora

The Mandarin corpus used in our experiments is the TREC Corpus. It was obtained from the People's Daily Newspaper from 01/1991 to 12/1993 and from the Xinhua News Agency for 04/1994 to 09/1995 from the Linguistic Data Consortium (http://www.ldc.upenn.edu). TREC has 19,546,872 tokens similar in size to the largest of the English corpora. The Mandarin language is a syllable-class language, in which each syllable is at the same time a word and a Chinese character. Other words, compound words, are built up by combining syllables together, similar to word n-grams in English. The most common unigrams, bigrams and trigrams are listed in Table 3.

The number of syllable-types (i.e. unigrams) in the TREC corpus is only 6,300, very different from English (the WSJ87 corpus has 114,718 word types); so it is not surprising that the Zipf curve for unigrams in Mandarin in Figure 7 is very different from the Zipf curve for unigrams in English. It is similar to a previous curve for a smaller Mandarin corpus of 2,022,604 tokens by Clark, Lua and McCallum (1986). The Zipf curves for n-grams

Table 2 The 10-highest frequency unigrams, bigrams and trigrams in the WSJ corpus

Unigrams		Bigrams		Trigrams	
Frequency	Token	Frequency	Token	Frequency	Token
2057968	THE	217427	OF THE	42030	THE U. S.
973650	OF	173797	IN THE	27260	IN NINETEEN EIGHTY
940525	TO	110291	MILLION DOLLARS	24165	CENTS A SHARE
853342	A	89184	U. S.	18233	NINETEEN EIGHTY SIX
825489	AND	83799	NINETEEN EIGHTY	16786	NINETEEN EIGHTY SEVEN
711462	IN	76187	FOR THE	15316	FIVE MILLION DOLLARS
368012	THAT	72312	TO THE	14943	MILLION DOLLARS OR
362771	FOR	65565	ON THE	14517	MILLION DOLLARS IN
298646	ONE	63838	ONE HUNDRED	12327	IN NEW YORK
281190	IS	55014	THAT THE	11981	A YEAR EARLIER

Table 3 The 10-highest frequency unigrams, bigrams and trigrams in the Mandarin TREC corpus.

Unigrams			Bigrams			Trigrams		
Freq	Token	Meaning	Freq	Token	Meaning	Freq	Token	Meaning
620619	的	Of	87566	中 国	China	28037	新 华 社	New China News Agency
308826	国	State	43749	发 展	Develop/Grow	18352	期 星 期	(Of Week-)-Week*
219543	一	One	43310	经 济	Economy	18351	星 期 星	Week-(-Of Week)*
209497	中	Centre / Middle	37225	星 期	Week	13253	百 分 之	(of hundred) Percent
176905	在	In / At	36800	国 家	Country	10228	会 主 义	(So-)-cialism*
159861	和	And	29295	企 业	Enterprise	10212	社 会 主	Social-(-ism)*
143359	人	Human	28963	国 际	International	8831	社 北 京	Agency Beijing
139713	了	Perfective marker	28832	人 民	The People	8830	华 社 北	China (New) Agency Bei-(-jing)*
133696	会	Get together/Meeting/Association	28430	记 者	Journalist	7880	进 一 步	Go a step further
128805	年	Year	28402	社 会	Society	7309	委 员 会	Committee

* These phrases are incomplete/nonsense in meaning.

for the Mandarin corpus are also shown in Figure 7.

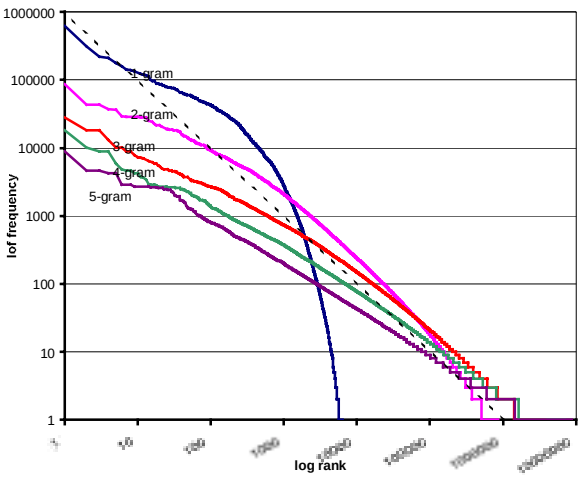

Figure 7 Zipf curves for the TREC Mandarin corpus

Except for the unigrams, the shapes of the other TREC n-gram Zipfian curves are similar to but not quite the same as those for the English corpora. In particular the bigram curve for Mandarin is more curved than the English curve because there are more compound words in Mandarin than English.

The Mandarin β-values in Table 1 are also higher than for English, on average by about 0.1, which is due to the different distribution of unigrams. For TREC, the crossing point between the unigram curve and the bigram curve is at rank: 1224, frequency: 1750, unigram: "谷", bigram: "日至". The unigram curve and the trigram curve cross each other at rank: 1920, frequency: 491, unigram: "丘", trigram: "北京九". This is very different from English.

Comparisons between the n-grams curves (n = 1 to 4) for English and Mandarin are made in Figure 8, 9, 10 and 11. The English curves are for the 3 WSJ corpora joined together making a 40 million word corpus.

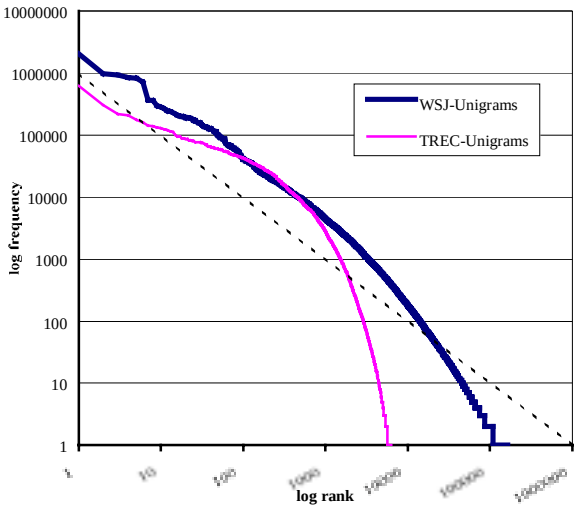

Figure 8 Zipf curve for the unigrams for the WSJ English and TREC Mandarin corpora

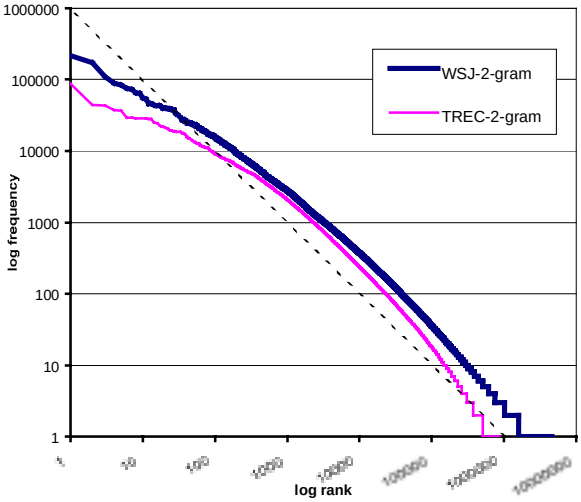

Figure 9 Zipf curve for the bigrams for the WSJ English and TREC Mandarin corpora

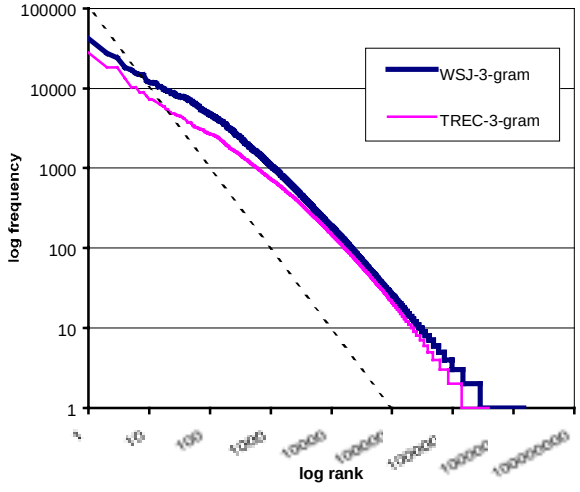

Figure 10 Zipf curve for the trigrams for WSJ English and TREC Mandarin corpora

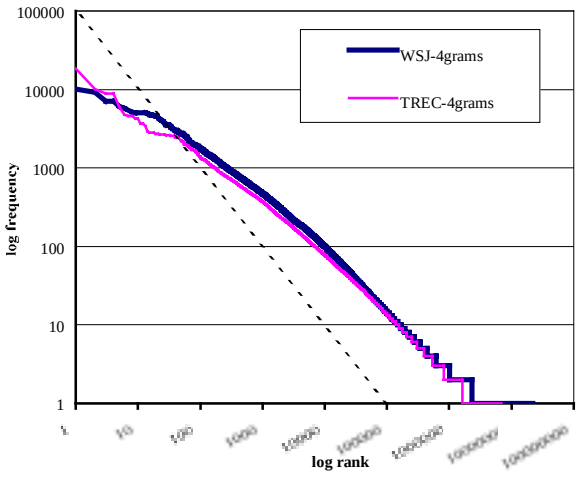

Figure 11 Zipf curve for the 4-grams for the WSJ English and TREC Mandarin corpora

3. Combined n-grams

The derivation of Zipf's law by Simon was based solely on single words and it failed for English when the number of word types was greater than about 5000 words. In Mandarin it failed almost immediately for unigrams because of the limited number of characters. However it might not have failed if the Mandarin compound words in the bigram, trigram and higher n-gram statistics had been included; this suggested that the n-gram and unigram curves should be combined. Perhaps the same may be true for English. So we should combine the English curves also.

This can be justified in another way. In a critical part of his derivation Simon gives an initial probability to a new word found in the corpus as it introduces some new meaning not expressed by previous words. However, as the number of words increases new ideas are frequently expressed not in single words, but in multi-word phrases or compound words. This was left out of Simon's derivation. If he had included it, the law he derived would have included phrases as well as words. So perhaps Zipf's law should include words and phrases.

We therefore put all unigram and n-gram together with their frequencies into one large file, sorted on frequency and put in rank order as previously. The resulting Zipf curve for the combined curves for both English and Mandarin are shown in Figure 12.

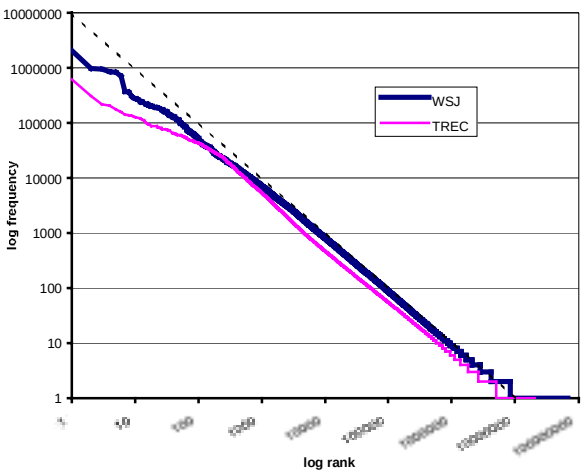

Figure 12 Combined Zipf curves for the English WSJ and the TREC Mandarin corpora

This shows that the n-grams ($n \geq 2$) exactly make up for the deviation of the two very different unigram curves from Zipf's law and the combined curves for both languages are straight lines with slopes close to –1 for all ranks>100. This result appears to vindicate Simon's derivation. However, whether Simon's derivation is entirely valid or not, the results in Figure 12 are a new confirmation of Zipf's original law in an extended form. This remarkable result has been found to be valid for 3 other natural languages: Irish, Latin and Vietnamese, in preliminary experiments.

References

Booth, A. D. (1967) "*A Law of Occurrences for Words of Low Frequency*". Inform. & Control Vol. 10, No. 4, pp 386-393. April.

Clark, J. L., Lua, K. T. & McCallum, J. (1986). "*Using Zipf's Law to Analyse the Rank Frequency Distribution of Elements in Chinese Text*". In Proc. Int. Conf. on Chinese Computing, pp. 321-324. August, Singapore.

Fedorowicz, J. (1982) *"A Zipfian Model of an Automatic Bibliographic System: an Application to MEDLINE*", Journal of American Society of Information Science, Vol. 33, pp 223-232.

Ferrer Cancho, R. & Solé, R. V. (2002) *"Two Regimes in the Frequency of Words and the Origin of Complex Lexicons"* To appear in Journal of Quantitative Linguistics.

Francis, W. N. & Kucera, H. (1964). "*Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers*" Department of Linguistics, Brown University, Providence, Rhode Island

Guiter H. & Arapov M., editors. (1982) "*Studies on Zipf 's Law*". Brochmeyer, Bochum.

Li, W. (2001) "*Zipf's Law in Importance of Genes for Cancer Classification Using Microarray Data*" Lab of Statistical Genetics, Rockefeller University, NY.

Mandelbrot, B. (1953). "*An Information Theory of the Statistical Structure of Language"*. Communication Theory, ed. By Willis Jackson, pp 486-502. New York: Academic Press.

Mandelbrot, B. (1954) "*Simple Games of Strategy Occurring in Communication through Natural Languages"*. Transactions of the IRE Professional Group on Information Theory, 3, 124-137

Mandelbrot, B. (1957) "*A probabilistic Union Model for Partial and temporal Corruption of Speech"*. Automatic Speech Recognition and Understanding Workshop. Keystone, Colorado, December.

Mandelbrot, B. (1959) "*A note on a class of skew distribution function analysis and critique of a paper by H.A. Simon"*, Inform. & Control, Vol. 2, pp 90-99.

Mandelbrot, B. (1961) "*Final note on a class of skew distribution functions: analysis and critique of a model due to H.A. Simon*", Inform. & Control, Vol. 4, pp 198-216.

Mandelbrot, B. B. (1961) "*Post Scriptum to 'final note*'", Inform. & Control, Vol. 4, pp 300-304.

Montemurro, M. (2002) "*Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics*". To appear in Physica A.

Paul, D. B. & Baker, J.M. (1992) "*The Design for the Wall Street Journal-based CSR Corpus*", Proc. ICSLP 92, pp 899-902, November.

Samuelson, C. (1996). "*Relating Turing's Formula and Zipf's Law*". Proceedings of the 4th Workshop on Very Large Corpora, Copenhagen, Denmark.

Silagadze, Z. K. (1997) "*Citations and the Zipf-Mandelbrot Law*". Complex Systems, Vol. 11, No. 6, pp 487-499.

Simon, H. A. (1955) "*On a Class of Skew Distribution Functions*", Biometrika, Vol. 42, pp 425-440.

Simon, H. A. (1960) "*Some Further Notes on a Class of Skew Distribution Functions*", Inform. & Control, Vol. 3, pp 80-88.

Simon, H. A. (1961) "*Reply to Dr. Mandelbrot's post Scriptum*" Inform. & Control, Vol. 4, pp 305-308.

Simon, H. A. (1961) "*Reply to 'final note' by Benoit Mandelbrot*", Inform. & Control, Vol. 4, pp 217-223.

Smith, F. J. & Devine, K. (1985) "*Storing and Retrieving Word Phrases*" Information Processing & Management, Vol. 21, No. 3, pp 215-224.

Yonezawa, Y. & Motohasi, H. (1999) "Zipf-Scaling Description in the DNA Sequence" 10th Workshop on Genome Informatics. Japan.

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