Zipf and Type-Token rules for the English, Spanish, Irish and Latin languages

Le Quan Ha, Darryl W Stewart, Philip Hanna and Francis J Smith

School of Electronics, Electrical Engineering and Computer Science

Queen's University Belfast
Belfast BT7 1NN, Northern Ireland
q.le@qub.ac.uk

Abstract

The Zipf curves of log of frequency against log of rank for a large English corpus of 500 million word tokens, 689,000 word types and for a large Spanish corpus of 16 million word tokens, 139,000 word types are shown to have the usual slope close to –1 for rank less than 5,000, but then for a higher rank they turn to give a slope close to –2. This is apparently mainly due to foreign words and place names. Other Zipf curves for highly-inflected Indo-European languages, Irish and ancient Latin, are also given. Because of the larger number of word types per lemma, they remain flatter than the English curve maintaining a slope of –1 until turning points of about ranks 30,000 for Irish and 10,000 for Latin. A formula which calculates the number of tokens given the number of types is derived in terms of the rank at the turning point, 5,000 for both English and Spanish, 30,000 for Irish and 10,000 for Latin.

1 Introduction

Zipf's law, discovered empirically by Zipf (1949) for word tokens in an English 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 called a Zipf curve), a straight line is obtained with a slope of –1. 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.

Zipf's discovery was followed by a large body of literature, reviewed in a series of papers edited by Guiter and Arapov (1982). Notable among these are papers by Mandelbrot (1953, 1954, 1959, 1961), Miller (1954, 1957, 1958), Simon (1955, 1960, 1961), Sichel (1975, 1986), Carroll (1967, 1969), Chitashvili (1983, 1989) and Orlov (1983).

It continues to stimulate interest up to today Samuelson (1996); Baayen (1991, 2001); Evert (2004); Hatzigeorgiu, Mikros and Carayannis (2001); Montermurro (2001); Ferrer and Solé (2002) and, for example, it has been recently applied to citations Silagadze (1997), to biological

species-abundance Sichel (1997) and to DNA sequences Yonezawa and Motohasi (1999); Li (2001).

Following its discovery in 1949, several experiments aided by the appearance of the computer in the 1960's, confirmed that the law was correct. 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 from the straight line, which suggested several modifications of the law, and in particular one derived theoretically by Mandelbrot (1953) with the form

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

where α and β are constants for the corpus being analysed. However, generally the constants α and β were found to be only small varying deviations from the original law by Zipf.

A number of theoretical explanations for Zipf's law had been derived, many reviewed by Fedorowicz (1982); notably are those due to Mandelbrot (1954, 1957), Miller (1954, 1958), Simon (1955), Booth (1967), and Sichel (1975, 1986).

The processing of larger English corpora with 1 million words - the Brown corpus of American English (Francis and Kucera 1964) - was facilitated. 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.

At this laboratory (Ha et al. 2003) experiments with the large English Wall Street Journal corpus (Paul and Baker 1992) of 42 million words, have shown that the Zipf curve drops rapidly below a straight line for ranks higher than about 5,000 (see Figure 1).

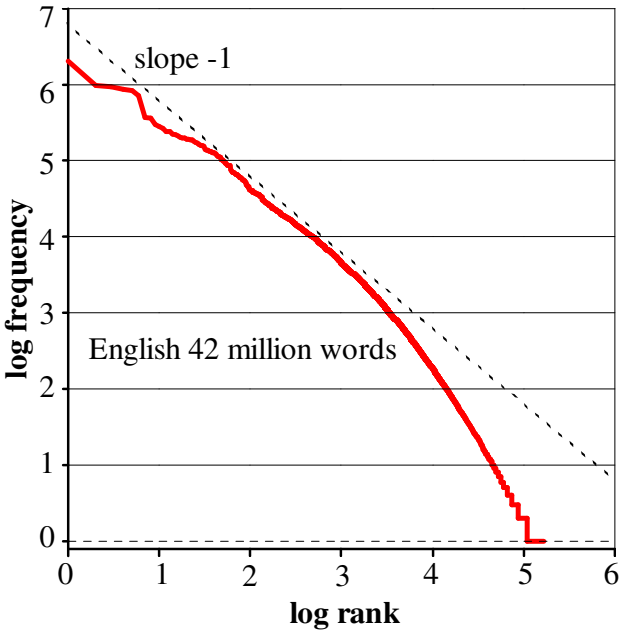

Figure 1. Zipf curves for Wall Street Journal.

2 Zipf's law for English, Spanish, Irish and Latin corpora

The English corpus used in our experiments is the North American News Text corpus from the Linguistic Data Consortium[1], size 489 million tokens with 689,028 types, including Los Angeles Times & Washington Post for May 1994 - August 1997 of 71,038,342 tokens and 253,774 types, New York Times News Syndicate for July 1994 - December 1996 of 249,365,501 tokens and 461,068 types, Reuters News Service (General of 89,884,598 tokens and 258,928 types & Financial of 25,173,907 tokens and 122,555 types) for April 1994 - December 1996 and Wall Street Journal for July 1994 - December 1996 of 54,308,157 tokens and 198,317 types.

Recently, we have developed a new large Spanish corpus named NLPSR 1.0 from SiSpain: www.sispain.org, with a material permit given by its owner, Prof Barrio. Our Spanish corpus contains 16 million tokens with 139,757 types, including only a part of Barrio's sources. Hence, we employ this Spanish corpus for our experiment in this paper.

The Irish and Latin are highly-inflected Indo-European languages. Both the beginning and end of words are regularly inflected; so it is very different from English. The Irish corpus used in our experiments is taken from a corpus of 17th and 18th century Irish from the Royal Irish Academy[2] with sizes 7,122,537 tokens with 449,968 types and the Latin corpus invented by Harvey, Devine and Smith (1994) includes 2,244,444 tokens with 147,442 types.

For pre-processing of the corpora in all languages, all numbers were replaced by the symbol *#NO* and punctuation marks were excluded. The characters *"=", "#", "~", "<", ">", "|", "+", "-", "^", "*", "@", "/"* and *"\"*, etc. were also ignored. Especially in our recent NLPSR 1.0 Spanish corpus, all the commas are replaced by <COMA>. Typographical errors, if any, appear in the hapax-legomenon (types which occur once only).

The identical curves for the large English and Spanish corpora in Figure 2 confirms the observation which we obtained previously with the Wall Street Journal corpus that the Zipf curve falls below the straight-line Zipf curve starting at about rank 5,000. For high ranks, the curve continues to bend downwards until the slope is close to –2. It then straightens out and maintains this slope to the final rank of 689,028 for English and 139,757 for Spanish (although there is an indication that it is just beginning to fall off with a slightly smaller slope of about –2.2). A similar two-slope behaviour has previously been observed by Ferrer and Solé (2002).

---

[1] http://www.ldc.upenn.edu/
[2] http://www.ria.ie/

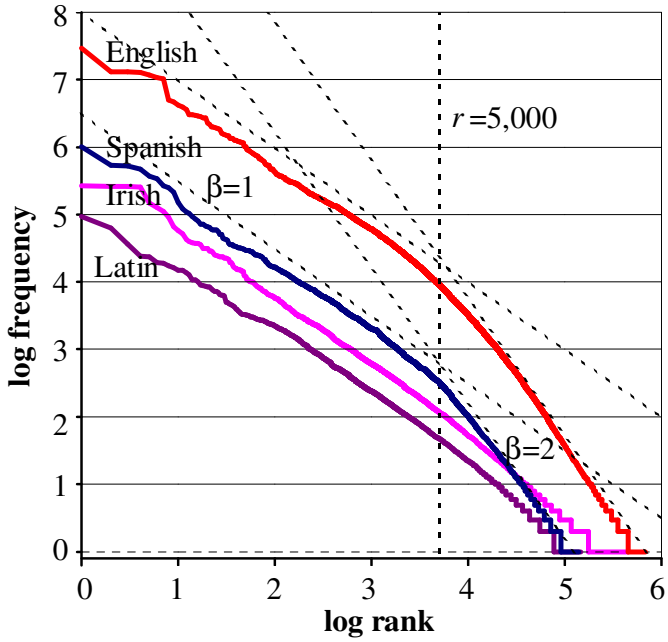

Figure 2. Comparison of Zipf curves for English, Spanish, Irish and Latin.

It is interesting to look at the English word types that occur at different parts of the Zipf curve. Some English samples for rank close to 20, 200, 2,000, 20,000 and 200,000 by Ha and Smith (2004) show that at rank 20, we have mostly the common stop words. At rank 200 are very common words found in newspaper (like WAR) and at rank 2,000 appears less common but everyday words where the slope of the Zipf curve is still –1. At rank 20,000 are some uncommon words and a few proper names; but at 200,000 where the slope is –2, we have mainly foreign words and place names. It is clear that this last set of words is fundamentally different from most of the other words and any theory explaining the slope of –2 needs to take account of their difference.

Because of the identical form of English and Spanish Zipf curves in Figure 2, we re-check by capturing a similar list for Spanish word types at ranks 20, 200, 2,000 and 20,000 in Tables 1*a* and 1*b*.

For Spanish, we also observe that very common stop Spanish words appear at rank 20, common Spanish words in newspaper (like GUINEA, SOCCER) appear at rank 200, less common but everyday Spanish words appear at rank 2,000 and foreign words, names appear after rank 20,000.

Ha and Smith (2004) already showed lists of the most common word tokens for Irish. Because of the numerous word inflections in these languages, their number of word types should be much larger than in English and the Zipf curves should behave differently. This is what is observed in Figure 2. Their curves deviate from a straight-line Zipf curve at larger ranks about 10,000 for Latin and 30,000 for Irish, then appear to have second slopes of about –1.3 (Irish) and –1.35 (Latin). Larger corpora are needed to find if they also have slopes of –2 for very large corpora.

Frequency *f*	Start at *r*=20	Meaning	*f*	Start at *r*=200	Meaning
71,945	LEGAL	LEGAL	9,079	OTRAS	OTHERS
68,611	EDICIÓN	EDITION	9,046	TODO	EVERYTHING
65,903	AVISO	WARNING	9,032	NOTICIA	COMMUNICATION
65,262	NO	NOT	8,935	GUINEA	GUINEA
60,269	TU	YOUR	8,833	FÚTBOL	SOCCER
58,773	AL	TO	8,778	PM	PM
58,465	UNA	ONE	8,762	MÚSICA	MUSIC
58,141	SU	HIS	8,739	COSTA	COAST
55,111	SERVICIOS	SERVICES	8,730	BUSCAR	TO LOOK FOR
50,481	VER	SEE	8,698	ROSA	ROSE

Table 1*a*. The list of 20 words after rank *r*=20, 200 with frequency *f* for the Spanish Zipf curve in Figure 2.

Frequency *f*	Start at *r*=20	Meaning	*f*	Start at *r*=200	Meaning
913	SABOR	FLAVOR	30	FACTORY	[foreign word]
913	ARMAS	ARMS	30	EXPOMANAGEMENT	[foreign word]
912	SECRETO	SECRET	30	EXPLICATIVA	EXPLANATORY
912	PROPONE	PROPOSES	30	EUSKALTEL	[name]
912	COMPETICIÓN	COMPETITION	30	ESTANCAN	SUSPEND
911	VENECIA	VENICE	30	ESPORT	[foreign word]
911	TREN	TRAIN	30	ESPALDARAZO	ACCOLADE
910	MORENO	BROWN	30	ESCUCHE	LISTEN
908	RECIBE	RECEIVES	30	ESCUCHADO	LISTENED
908	ORGANIZACIÓN	ORGANISATION	30	ESCOLLO	STUMBLING BLOCK

Table 1*b*. The list of 20 words after rank *r*=2,000, 20,000 with frequency *f* for the Spanish Zipf curve in Figure 2.

The proportion of distinct proper name types is higher in the huge English and large Spanish news data than in the smaller Irish and Latin corpora. It is probably not a language-related issue, but related to corpus sources and size.

3 Type-Token relationship

In 1967, Booth investigated the number of types derived from Zipf's law. He stated that a word would occur once if

$$\frac{3}{2} > Tp(r) \geq \frac{1}{2} \qquad (3)$$

Within a training corpus, if $p(r)$ is the probability of the word of rank r and T is the text corpus size, then the number of occurrences of the word of rank r is the frequency $f = Tp(r)$.

Applying Zipf's law (1949) $f = k/r$ into equation (3), where k is a constant we get

$$\frac{3}{2} > f \geq \frac{1}{2} \qquad (4)$$

So if N is the highest rank of words in the corpus then $\frac{k}{N} = f_{\min} \cong \frac{1}{2}$. So $k = \frac{N}{2}$. (5)

Booth's derivation is correct only for small-sized texts not large corpora then needs a further improvement.

In 1985, Smith and Devine used the same logic to investigate the token-type distribution and proposed the integral

$$T = \int_{\frac{1}{2}}^{N} \frac{k}{r} dr \qquad \text{for Zipf's law.} \quad (6)$$

With k given by equation (5), the integral of equation (6) was solved as

$$T \cong \frac{1}{2} N \ln(2N) \qquad (7)$$

This is called Smith-Devine prediction (1985). The comparison of the token-type distribution for English, Irish and Latin languages and the Smith-Devine prediction are shown in Figure 3. Its failure with large corpora is clear.

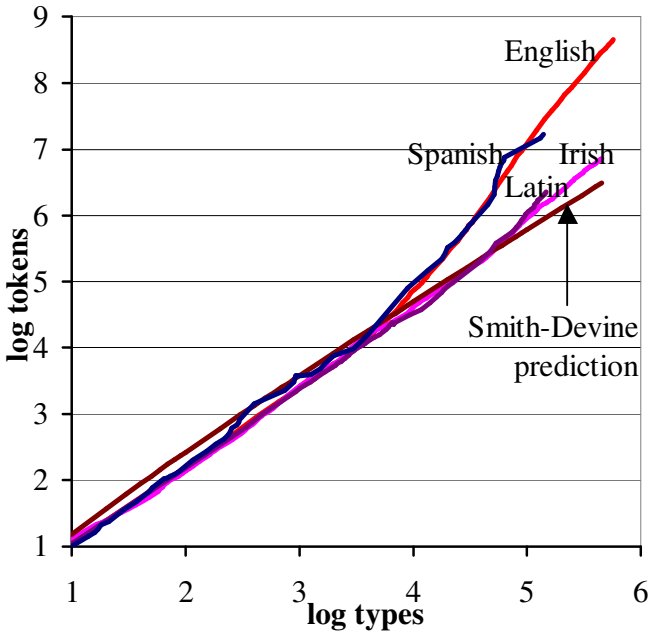

Figure 3. Tokens/Types on observed English, Irish and Latin words and Smith-Devine law.

In order to plot the experimental token-type distributions of all languages in Figure 3, we split the corpus for each language into token sizes 10, 20, 30, …, 90, 100, 200, …, 900, 1,000, 2,000,…, 10^x, 2×10^x, 3×10^x, …, $m\times10^x$ and whole corpus T (for m and x are chosen integers so

that $1 \le m \le 9$, $m \times 10^x \le T < 10^{x+1}$ and $T < (m+1)10^x$) then we count the observed token number and type number of each token size.

The Smith-Devine prediction fits better for the distributions of the highly inflected Irish and Latin languages which have more word types because of inflections.

Enhancement of Smith-Devine law

To be more accurate than equation (6), we can make the sum directly from Zipf's law

$$T = k\left(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{N}\right)$$

The sum of this series is well-known and is given by

$$T = k(\ln N + \gamma) \cong \frac{N}{2}(\ln N + \gamma) \qquad (8)$$

where γ = Euler's constant = 0.577. Equation (7) is wrong by ~ $0.11k$. Now it is convenient to write

$$T = \int_{\alpha}^{N+\frac{1}{2}} \frac{k}{r} dr \qquad (9)$$

$$\cong k\left[\ln N + \ln \frac{1}{\alpha}\right] \quad \text{for large } N, \qquad (10)$$

where $$\ln \frac{1}{\alpha} = \gamma \qquad (11)$$

In an extended form of the type-token relationship, we break the integral into two parts from α to N_0, the type number where the curve begins to turn down and from N_0 to N where it is assumed the slope is –2. So

$$T = \begin{cases} \int_{\alpha}^{N} \frac{k}{r} dr & \text{if } N \le N_0, \\ \int_{\alpha}^{N_0} \frac{k_1}{r} dr + \int_{N_0}^{N} \frac{k_2}{r^2} dr & \text{if } N \ge N_0 \end{cases} \qquad (12)$$

where k_1 and k_2 are constants. Now noting that for the last rank N, $f = \frac{1}{2}$; so, $\frac{k_2}{N^2} = \frac{1}{2}$ or $k_2 = \frac{N^2}{2}$.

At the rank $r = N_0$, the two curves join. So

$$\frac{k_1}{N_o} = \frac{k_2}{N_0{}^2} \qquad \Rightarrow \qquad k_1 = \frac{k_2}{N_0} = \frac{N^2}{2N_0} \tag{13}$$

Integrating and substituting for k, k_1 and k_2, we find

$$T \cong \begin{cases} \frac{N}{2}(\ln N + \gamma) & if\ N \le N_0, \\ \frac{N^2}{2N_0}(\ln N_0 + \gamma) + \frac{N^2}{2}\left(\frac{1}{N_0} - \frac{1}{N}\right) & if\ N \ge N_0 \end{cases} \tag{14}$$

This is the extended law. To test this we compare it with the results of experiments on the English, Spanish, Irish and Latin corpora in Figure 4, Figure 5 and Figure 6 with N_0 = 5,000 for English and Spanish, N_0 = 30,000 for Irish and N_0 = 10,000 for Latin.

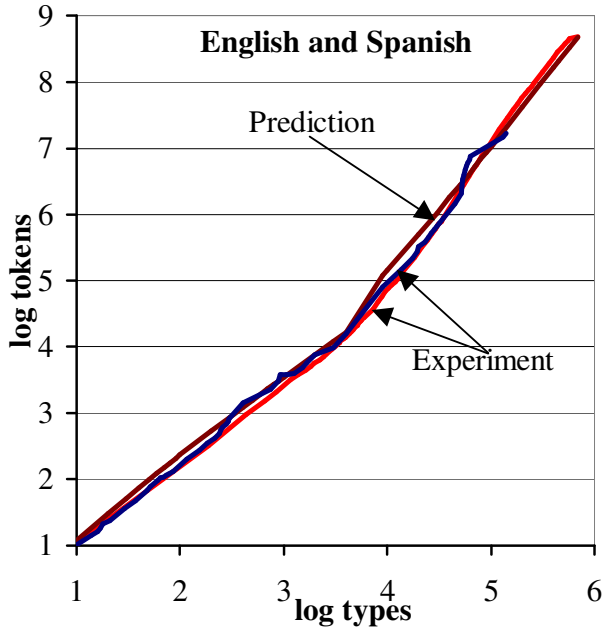

Figure 4. Two-slope law for Smith-Devine prediction on English and Spanish.

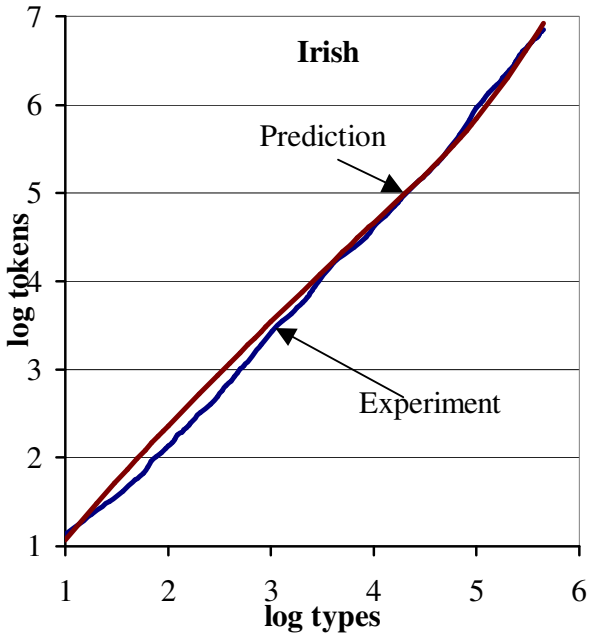

Figure 5. Two-slope law for Smith-Devine prediction on Irish.

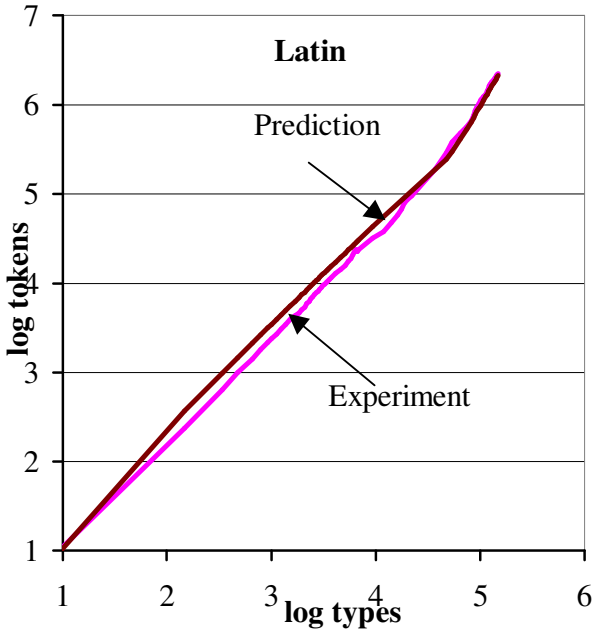

Figure 6. Two-slope law for Smith-Devine prediction on Latin.

A comparison between the number of tokens calculated by the extended law and the correct number for different types is given in Figure 4, Figure 5 and Figure 6. Reasonable agreement is

obtained here; we expect to investigate this agreement further with larger corpora in Irish and Latin in the future.

4 Conclusions

We have shown that for a very large corpus the Zipf curves for English and Spanish have two slopes, $\beta = 1$ for rank less than 5,000 and $\beta = 2$ for rank above 5,000. The curves for Irish and Latin, inflected languages, are flatter with a slope of -1 until a rank of about 30,000 and 10,000. An extended law for the type-token relationship is derived and tested.

5 Acknowledgements

Our thanks go to the Royal Irish Academy for their Irish and Latin databases.

References

Baayen, H. 1991. A Stochastic Process for Word Frequency Distributions. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL-29)*, pages 271-278, Berkeley, California, USA.

Baayen, H. 2001. Word Frequency Distributions. Kluwer Academic Publishers.

Booth, A. D. 1967. A Law of Occurrences for Words of Low Frequency. *Information and Control*, 10(4):386-393.

Carroll, J. B. 1969. A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions. *Research Bulletin -- Educational Testing Service*, Princeton.

Devine, K. and Smith, F. J. 1985. Storing and Retrieving Word Phrases. *Information Processing and Management*, 21(3):215-224.

Evert, S. 2004. A Simple LNRE Model for Random Character Sequences. In Proceedings of the 7èmes Journées Internationales d'Analyse Statistique des Données Textuelles, pages 411-422.

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

Ferrer i Cancho, R. and Solé, R. V. 2002. Two Regimes in the Frequency of Words and the Origin of Complex Lexicon. *Journal of Quantitative Linguistics*, 8(3):165 – 173,.

Francis, N. W. and 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.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3&4):237-264.

Guiter, H. and Arapov, M. V. editors. 1982. Studies on Zipf's Law. Brochmeyer, Bochum.

Ha, L. Q. and Smith, F. J. 2004. Zipf and Type-Token rules for the English and Irish languages. *MIDL workshop*, Paris.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J. and Smith, F. J. 2002. Extension of Zipf's Law to Words and Phrases. In *Proceedings of the 19th International Conference on Computational Linguistics*, 1:315-320.

Ha, L. Q., Sicilia-Garcia, E. I., Ming, J. and Smith, F. J. 2003. Extension of Zipf's Law to Word and Character N-Grams for English and Chinese. *Journal of Computational Linguistics and Chinese Language Processing*, 8(1):77-102.

Harvey, A., Devine, K. and Smith, F. J. 1994. Archive of Celtic-Latin Literature ACLL-1Royal Irish Academy, Dictionary of Medieval Latin from Celtic sources. Brespols.

Hatzigeorgiu, N., Mikros, G. and Carayannis, G. 2001. Word Length, Word Frequencies and Zipf's Law in the Greek Language. *Journal of Quantitative Linguistics,* 8(3):175 – 185.

Jelinek, F. and Mercer, R. L. 1985. Probability distribution estimation from sparse data. *IBM Technical Disclosure Bulletin*, 28(6).

Jelinek, F., Mercer, R. L., Bahl, L. R., and K. B. J. 1977. Perplexity --- a measure of difficulty of speech recognition tasks. *94th Meeting of the Acoustical Society of America*, Miami Beach, FL.

Li, W. 2001. Zipf's Law in Importance of Genes for Cancer Classification Using Microarray Data. Laboratory of Statistical Genetics, Rockefeller University, New York.

Mandelbrot, B. 1953. An Information Theory of the Statistical Structure of Language. Communication Theory, edited by Willis Jackson, New York: Academic Press, pages 486-502.

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. 1959. A note on a class of skew distribution function analysis and critique of a paper by H. A. Simon. *Information and Control*, 2:90-99.

Mandelbrot, B. 961. Final note on a class of skew distribution functions: analysis and critique of a model due to H. A. Simon. *Information and Control*, 4:198-216.

Mandelbrot, B. B. 1961. Post Scriptum to 'final note'. *Information and Control*, 4:300-304.

Miller, G. A. 1954. Communication. *Annual Review of Psychology*, 5:401-420.

Miller, G. A. 1957. Some effects of intermittent silence. *The American Journal of Psychology*, 52:311-314.

Miller, G. A., Newman, E. B. and Friedman, E. A. 1958. Length-Frequency Statistics for Written English. *Information and control*, 1:370-389.

Montemurro, M. A. 2001. Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics. *Physica A*: *Statistical Mechanics and its Applications*, 300(3&4):567-578.

Orlov, J. K. and Chitashvili, R. Y. 1983. Generalized Z-distribution generating the well-known 'rank-distributions'. *Bulletin of the Academy of Sciences*, 110.2:269-272, Georgia.

Paul, D. B. and Baker, J. M. 1992. The Design for the Wall Street Journal-based CSR Corpus. *In Proceedings of International Conference on Spoken Language Processing* (ICLSP), pages 899-902, Banff, Alberta, Canada.

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

Sichel, H. S. 1975. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, 70:542-547.

Sichel, H. S. 1986. Word Frequency Distributions and Type-Token Characteristics. *Mathematical Scientist*, 11:45-72.

Sichel, H. S. 1997. Modelling Species-Abundance Frequencies and Species-Individual Functions with the Generalized Inverse Gaussion-Poisson Distribution. *South African Statistical Journal*, 31:13-37.

Silagadze, Z. K. 1997. Citations and the Zipf-Mandelbrot Law. *Complex Systems*, 11(6):487-499.

Simon, H. A. 1960. Some Further Notes on a Class of Skew Distribution Functions. *Information and Control*. 3:80-88.

Simon, H. A. 1961. Reply to Dr. Mandelbrot's post Scriptum. *Information and Control*, 4:305-308.

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

Simon, H. A. 1995. On a Class of Skew Distribution Functions. *Biometrika*. 42:425-440.

Yonezawa, Y. and Motohasi, H. 1999. Zipf-Scaling Description in the DNA Sequence. *In Proceedings of the 10th Workshop on Genome Informatics*, Japan.

Zipf, G. K. 1949. Human Behaviour and the Principle of Least Effort. Reading, MA: Addison-Wesley Publishing Co.