

# Zipf and Type-Token rules for the English and Irish languages

*Le Quan Ha, Francis J Smith*

Computer Science School, Queen's University of Belfast  
Belfast BT7 1NN, Northern Ireland, United Kingdom  
Tel : +44 (0)28 90 27 47 32 - Fax : +44 (0)28 90 97 56 66  
Email : q.le@qub.ac.uk

## ABSTRACT

The Zipf curve of log of frequency against log of rank for a large English corpus of 500 million word tokens and 689,000 word types is shown to have the usual slope close to  $-1$  for rank less than 5,000, but then for a higher rank it turns to give a slope close to  $-2$ . This is apparently mainly due to foreign words and place names. The Zipf curve for a highly-inflected language (the Indo-European Celtic language, Irish) is also given. Because of the larger number of word types per lemma, it remains flatter than the English curve maintaining a slope of  $-1$  until a turning point of about rank 30,000. 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 English and 30,000 for Irish.

## 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

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

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

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

where  $\alpha$  and  $\beta$  are constants for the corpus being analysed. However, generally the constants  $\alpha$  and  $\beta$  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 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.

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).

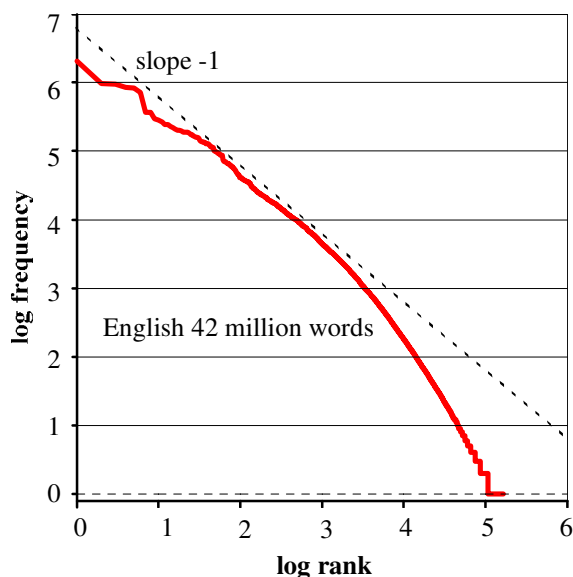


FIG. 1 – Zipf curves for Wall Street Journal.

## 2. ZIPF'S LAW FOR ENGLISH AND IRISH CORPORA

The English corpus used in our experiments is the North American News Text corpus from the Linguistic Data Consortium<sup>1</sup>, 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.

The Irish language is a highly-inflected Indo-European Celtic language. 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 17<sup>th</sup> and 18<sup>th</sup> century Irish from the Royal Irish Academy<sup>2</sup> with sizes 7,122,537 tokens with 449,968 types.

For pre-execution of the corpora in both languages, all numbers were replaced by the symbol #NO and punctuation marks were excluded. The characters "=", "#", "~", "<", ">", "!", "+", "-", "^", "\*", "@", "/" and "\", etc. were also ignored. Typographical errors, if any, appear in the hapax-legomenon (types which occur once only).

<sup>1</sup> <http://www ldc.upenn.edu/>

<sup>2</sup> <http://www ria.ie>

The curve for the large English corpus 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 (although there is an indication that it is just beginning to fall-off with a slightly bigger slope of about 2.2). A similar two-slope behaviour has previously been observed by Ferrer and Solé (2002).

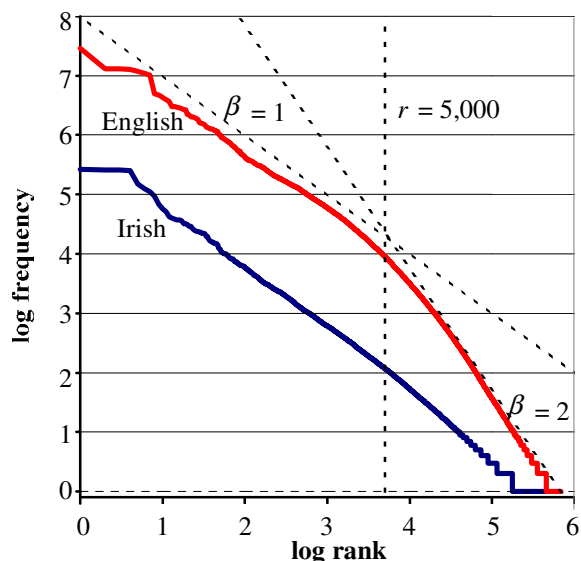


FIG. 2 – Comparison of Zipf curves for English and Irish.

It is interesting to look at the word types that occur at different parts of the Zipf curve. Some samples for rank close to 20, 200, 2,000, 20,000 and 200,000 are shown in Table 1. 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 are 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 numerous word inflections in Irish, the number of word types should be much larger than in English and the Zipf curve should behave differently. This is what is observed in Figure 2. The Irish curve deviates from a straight-line Zipf curve at a larger rank, about 30,000 and then appears to have a second slope of about 1.3. A larger corpus is needed to find if it also has a slope of -2 for very large corpora.

TABLE 1 – The list of 20 words after rank  $r$  with frequency  $f$  for the English Zipf curve in Figure 2.

Frequency $f$	Start at $r=20$	$f$	Start at $r=200$	$f$	Start at $r=2,000$	$f$	Start at $r=20,000$	$f$	Start at $r=200,000$
2,688,370	BY	236,870	TOO	28,746	DECISIONS	1,003	DRIZZLE	8	TARTRE
2,674,178	AT	235,169	AGO	28,706	CABINET	1,003	DIAPER	8	TARPENNING
2,259,799	FROM	234,470	WAR	28,685	WEAK	1,003	DELLA	8	TARNHELM
2,158,633	BE	234,116	OWN	28,674	PALESTINIANS	1,003	CONSTRAINED	8	TARNATION
2,001,699	BUT	233,773	WHITE	28,662	VALLEY	1,002	WORKSTATION	8	TARMIZI
1,996,465	HAVE	233,216	COURT	28,646	FILMS	1,002	STEAKS	8	TARKWA
1,971,744	ARE	232,038	WANT	28,602	WORSE	1,002	POSTCARDS	8	TARITA
1,907,215	HIS	231,361	MONTHS	28,592	GOD	1,002	ORCHESTRAS	8	TARASCO'S
1,881,251	HAS	227,897	PARTY	28,591	FOUNDATION	1,002	INTERSCOPE	8	TARANCON
1,818,538	AN	225,756	BIG	28,553	OKLAHOMA	1,002	CHESTNUT	8	TARABINI
1,626,805	NOT	225,097	SHARE	28,472	ASKING	1,001	SWIPE	8	TAPROOTS
1,575,828	THEY	221,770	SERVICE	28,459	SPOKE	1,001	SINCERELY	8	TAPPET
1,561,028	WILL	221,567	LITTLE	28,453	CAROLINA	1,001	REUNITE	8	TAPPA
1,526,859	DOLLARS	221,352	AMONG	28,389	RETURNS	1,001	RELIC	8	TAPLITZ
1,489,997	THIS	220,063	DAYS	28,376	ALLOWING	1,001	PAIRING	8	TANTOCO
1,472,057	WHO	218,961	SAME	28,355	UNCHANGED	1,001	PAINSTAKING	8	TANKLIKE
1,347,855	NEW	217,311	ME	28,305	MACHINE	1,001	OLIN	8	TANGNEY
1,336,479	ITS	216,926	FRIDAY	28,301	CLOSELY	1,001	MASSE	8	TANGA
1,331,121	I	216,589	AROUND	28,295	WATCHING	1,001	MARTYRS	8	TANEYTOWN
1,296,199	HAD	215,811	LIFE	28,272	CONTACT	1,001	INCRIMINATING	8	TANASESCU

The proportion of distinct proper name types is higher in the huge English news data than in the smaller Irish corpus. 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 pointed out that 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  $f = Tp(r)$ . Zipf (1938) stated that a word would occur once if

$$1.5 > Tp(r) \geq 0.5 \quad (3)$$

Applying Zipf's law  $f = k/r$ , where  $k$  is a constant we get

$$1.5 > f \geq 0.5 \quad (4)$$

So if  $N$  is the highest rank of any word in the corpus

$$\text{then } \frac{k}{N} = \frac{1}{2}. \text{ So } k = \frac{N}{2}. \quad (5)$$

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

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

Equation (6) was solved as

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

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

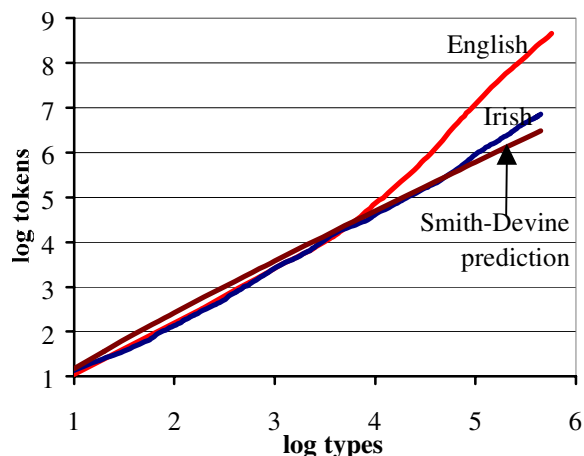


FIG. 3 – Tokens/Types on observed English and Irish words and Smith-Devine law.

The Smith-Devine prediction fits better for the distributions of the highly inflected Irish language 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

$$T = k \left( 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{N} \right) \quad (8)$$

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

$$T = k(\ln N + \gamma) \cong \frac{N}{2} (\ln N + \gamma) \quad (9)$$

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

$$T = \int_{\alpha}^{N+\frac{1}{2}} \frac{k}{r} dr \quad (10)$$

$$\cong k \left[ \ln N + \ln \frac{1}{\alpha} \right] \text{ for large } N, \quad (11)$$

where  $\ln \frac{1}{\alpha} = \gamma$  (12)

In an extended form of the type-token relationship, we break the integral into two parts from  $\alpha$  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

$$T = \begin{cases} \int_{\alpha}^N \frac{k}{r} dr & \text{if } N \leq N_0, \\ \int_{\alpha}^{N_0} \frac{k_1}{r} dr + \int_{N_0}^N \frac{k_2}{r^2} dr & \text{if } N \geq N_0 \end{cases} \quad (13)$$

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

$$\frac{k_1}{N_0} = \frac{k_2}{N_0^2} \Rightarrow k_1 = \frac{k_2}{N_0} = \frac{N^2}{2N_0} \quad (14)$$

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

$$T \cong \begin{cases} \frac{N}{2} (\ln N + \gamma) & \text{if } N \leq N_0, \\ \frac{N^2}{2N_0} (\ln N_0 + \gamma) + \frac{N^2}{2} \left( \frac{1}{N_0} - \frac{1}{N} \right) & \text{if } N \geq N_0 \end{cases} \quad (15)$$

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

Also it would be interesting to withdraw randomly paragraphs from the Wall Street Journal to form an English corpus of 7 million words of similar epoch and content than the Irish corpus, we then put its result in Figure 6.

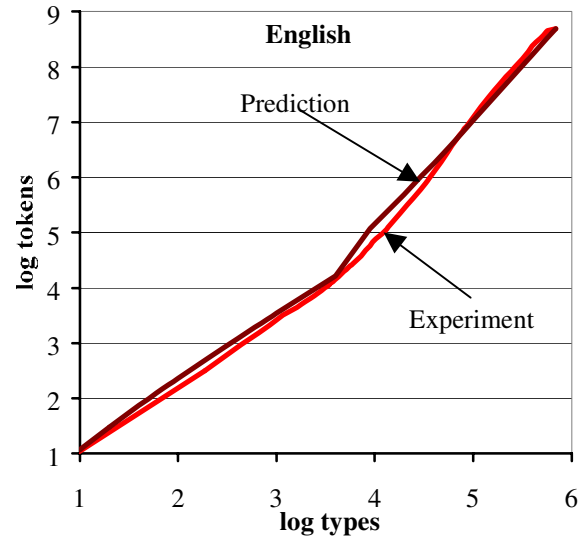


FIG. 4 – Two-slope law for Smith-Devine prediction on English.

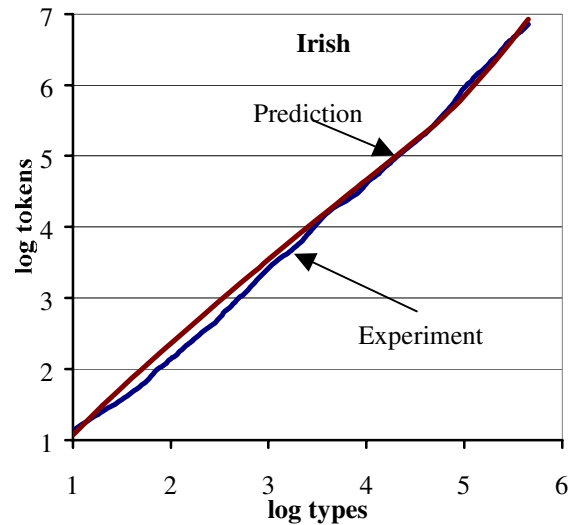
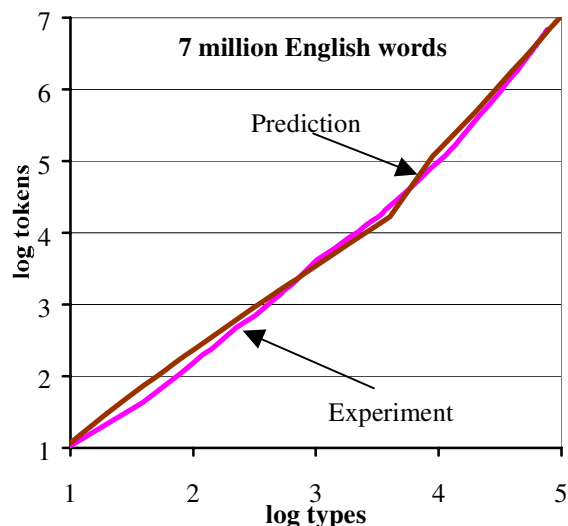


FIG. 5 – Two-slope law for Smith-Devine prediction on Irish.



**FIG. 6** – Two-slope law for Smith-Devine prediction on English corpus of 7 million words.

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.

#### 4. CONCLUSIONS

We have shown that for a very large corpus the Zipf curve for English has two slopes,  $\beta = 1$  for rank less than 5,000 and  $\beta = 2$  for rank above 5,000. The curve for Irish, an inflected language, is flatter with a slope of  $-1$  until a rank of about 30,000. 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 database and to the reviewers who help precious opinions.

#### REFERENCES

- [1] Baayen, H. "A Stochastic Process for Word Frequency Distributions", *In Proceedings of the 29<sup>th</sup> Annual Meeting of the Association for Computational Linguistics (ACL-29)*, pages 271-278, Berkeley, California, USA, 1991.
- [2] Baayen, H. "Word Frequency Distributions", *Kluwer Academic Publishers*, 2001.
- [3] Booth, A. D. "A Law of Occurrences for Words of Low Frequency", *Information and Control*, Vol. 10, No. 4, pages 386-393, April 1967.
- [4] Carroll, J. B. "A Rationale for an Asymptotic Lognormal Form of Word Frequency Distributions", *Research Bulletin -- Educational Testing Service*, Princeton, November 1969.
- [5] Chen, S., and Shi, D-X. "On the feeding relation between syntax and morphology: Evidence from Chinese V-N compounds", *In Proceedings of the Third International Symposium on Chinese*

*Languages and Linguistics*, Taiwan: Chinghua University, 1992.

- [6] Clark, J. L., Lua, K. T. and McCallum, J." Using Zipf's Law to Analyse the Rank Frequency Distribution of Elements in Chinese Text ", *In Proceedings of International Conference on Chinese Computing*, pages 321-324, Singapore, August 1986.
- [7] Evert, S. "A Simple LNRE Model for Random Character Sequences", *Proceedings of the 7<sup>emes</sup> Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 411-422, 2004.
- [8] Fedorowicz, J. "A Zipfian Model of an Automatic Bibliographic System: an Application to MEDLINE ", *Journal of American Society of Information Science*, Vol. 33, pages 223-232, 1982.
- [9] Ferrer i Cancho, R., Solé, R. V. "Two Regimes in the Frequency of Words and the Origin of Complex Lexicon ", *Journal of Quantitative Linguistics*, Vol. 8, No. 3, pages 165 – 173, 2002.
- [10] Francis, W. N. and Kucera, H. "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, 1964.
- [11] Good, I. J. "The population frequencies of species and the estimation of population parameters ", *Biometrika*, 40 (3 and 4): pages 237-264, 1953.
- [12] Guiter, H. and Arapov, M. editors "Studies on Zipf's Law ", Brochmeyer, Bochum, 1982.
- [13] Hatzigeorgiu, N., Mikros, G., and Carayannis, G. "Word Length, Word Frequencies and Zipf's Law in the Greek Language ", *Journal of Quantitative Linguistics*, Vol. 8, No. 3, pages 175 - 185, 2001.
- [14] Jelinek, F., Mercer, R. L. "Probability distribution estimation from sparse data ", *IBM Technical Disclosure Bulletin*, Vol. 28, No. 6, November 1985.
- [15] Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. "Perplexity - a measure of difficulty of speech recognition tasks ", *94<sup>th</sup> Meeting of the Acoustical Society of America*, Miami Beach, FL, 1977.
- [16] Ha, L. Q., Sicilia-Garcia, E. I., Ming, J. and Smith, F. J. "Extension of Zipf's Law to Word and Character  $N$ -Grams for English and Chinese ", *Journal of Computational Linguistics and Chinese Language Processing*, Vol. 8, No. 1, pages 77-102, February 2003.
- [17] Li, W. "Zipf's Law in Importance of Genes for Cancer Classification Using Microarray Data ", Laboratory of Statistical Genetics, Rockefeller University, New York, 2001.
- [18] Mandelbrot, B. "An Information Theory of the Statistical Structure of Language, Communication Theory ", edited by Willis Jackson, New York:

- Academic Press, pages 486-502, 1953.
- [19] Mandelbrot, B. "Simple Games of Strategy Occurring in Communication through Natural Languages", *Transactions of the IRE Professional Group on Information Theory, Vol. 3*, pages 124-137, 1954.
- [20] Mandelbrot, B. "A note on a class of skew distribution function analysis and critique of a paper by H. A. Simon", *Information and Control, Vol. 2*, pages 90-99, 1959.
- [21] Mandelbrot, B. "Final note on a class of skew distribution functions: analysis and critique of a model due to H. A. Simon", *Information and Control, Vol. 4*, pages 198-216, 1961.
- [22] Mandelbrot, B. B. "Post Scriptum to 'final note'", *Information and Control, Vol. 4*, pages 300-304, 1961.
- [23] Miller, G. A. "Communication", *Annual Review of Psychology, 5*, pages 401-420, 1954.
- [24] Miller, G. A. "Some effects of intermittent silence", *The American Journal of Psychology, 52*, pages 311-314, 1957.
- [25] Miller, G. A., Newman, E. B. and Friedman, E. A. "Length-Frequency Statistics for Written English", *Information and control, Vol. 1*, pages 370-389, 1958.
- [26] Montemurro, M. "Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics", *Physica A: Statistical Mechanics and its Applications, Vol. 300, Issues 3-4*, pages 567-578, November 2001.
- [27] Ney, H. "The Use of the Maximum Likelihood Criterion in Language Modelling", *K. Ponting (\*ed.): Computational Models of Speech Pattern Processing*, pages 259-279, Springer, Berlin, Germany, 1999.
- [28] O'Boyle, P., Owens, M. and Smith, F. J. "A weighted average  $n$ -gram model of natural language", *Computer Speech and Language, Vol. 8*, pages 337-349, 1994.
- [29] Orlov, J. K. and Chitashvili, R. Y. "Generalized Z-distribution generating the well-known 'rank-distributions'", *Bulletin of the Academy of Sciences, 110.2*, pages 269-272, Georgia, 1983.
- [30] Packard, J. L. "The Morphology of Chinese A Linguistic and Cognitive Approach", Cambridge University Press, United Kingdom, 2000.
- [31] Paul, D. B. and Baker, J. M. "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, October 1992.
- [32] Samuelson, C. "Relating Turing's Formula and Zipf's Law", *In Proceedings of the 4<sup>th</sup> Workshop on Very Large Corpora*, Copenhagen, Denmark, 1996.
- [33] Sichel, H. S. "On a Distribution Law for Word Frequencies", *Journal of the American Statistical Association, 70*, pages 542-547, 1975.
- [34] Sichel, H. S. "Word Frequency Distributions and Type-Token Characteristics", *Mathematical Scientist, 11*, pages 45-72, 1986.
- [35] Sichel, H. S. "Modelling Species-Abundance Frequencies and Species-Individual Functions with the Generalized Inverse Gaussian-Poisson Distribution", *South African Statistical Journal, 31*, pages 13-37, 1997.
- [36] Silagadze, Z. K. "Citations and the Zipf-Mandelbrot Law", *Complex Systems, Vol. 11, No. 6*, pages 487-499, 1997.
- [37] Simon, H. A. "Some Further Notes on a Class of Skew Distribution Functions", *Information and Control, Vol. 3*, pages 80-88, 1960.
- [38] Simon, H. A. "Reply to Dr. Mandelbrot's post Scriptum", *Information and Control, Vol. 4*, pages 305-308, 1961.
- [39] Simon, H. A. "Reply to 'final note' by Benoit Mandelbrot", *Information and Control, Vol. 4*, pages 217-223, 1961.
- [40] Simon, H. A. "On a Class of Skew Distribution Functions", *Biometrika, Vol. 42*, pages 425-440, 1995.
- [41] Smith, F. J. and Devine, K. "Storing and Retrieving Word Phrases", *Information Processing and Management, Vol. 21, No. 3*, pages 215-224, 1985.
- [42] Sproat, R. "Corpus-Based methods in Chinese Morphology", *Tutorial of the 19<sup>th</sup> International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, August 2002.
- [43] Yonezawa, Y. and Motohosi, H. "Zipf-Scaling Description in the DNA Sequence", *In Proceedings of the 10<sup>th</sup> Workshop on Genome Informatics*, Japan, December 1999.
- [44] Zhu, D. X. "Yufa Jiangyi (Chinese Syntax)", *Shanghai: The Commercial Publisher, China*, 1981.
- [45] Zipf, G. K. "Human Behaviour and the Principle of Least Effort", Reading, MA: Addison- Wesley Publishing Co., 1949.