Computational Linguistics and Chinese Language Processing
Vol. 8, No. 1, February 2003 , pp. 77-102
© The Association for Computational Linguistics and Chinese Language Processing

Extension of Zipf's Law to Word and Character *N*-grams for English and Chinese

Le Quan Ha*, E. I. Sicilia-Garcia*, Ji Ming* and F. J. Smith*

Abstract

It is shown that for a large corpus, Zipf 's law for both words in English and characters in Chinese does not hold for all ranks. The frequency falls below the frequency predicted by Zipf's law for English words for rank greater than about 5,000 and for Chinese characters for rank greater than about 1,000. However, when single words or characters are combined together with *n*-gram words or characters in one list and put in order of frequency, the frequency of tokens in the combined list follows Zipf's law approximately with the slope close to -1 on a log-log plot for all *n*-grams, down to the lowest frequencies in both languages. This behaviour is also found for English 2-byte and 3-byte word fragments. It only happens when all *n*-grams are used, including semantically incomplete *n*-grams. Previous theories do not predict this behaviour, possibly because conditional probabilities of tokens have not been properly represented.

Keywords: Zipf 's law, Chinese character, Chinese compound word, *n*-grams, phrases.

1. Introduction

1.1 Zipf's law

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

* Computer Science School, Queen's University Belfast, Belfast BT7 1NN, Northern Ireland, UK.
Email: {q.le, e.sicilia, j.ming, fj.smith}@qub.ac.uk

called a Zipf curve), a straight line is obtained with a slope of –1. An example with a small corpus of 250,000 tokens made up of paragraphs chosen at random from the Brown corpus of American English [Francis and Kucera 1964] is given in Figure 1; in this the tokens do not include punctuation marks and numbers. Typographical errors, if any, will appear in the hapax legomenon.

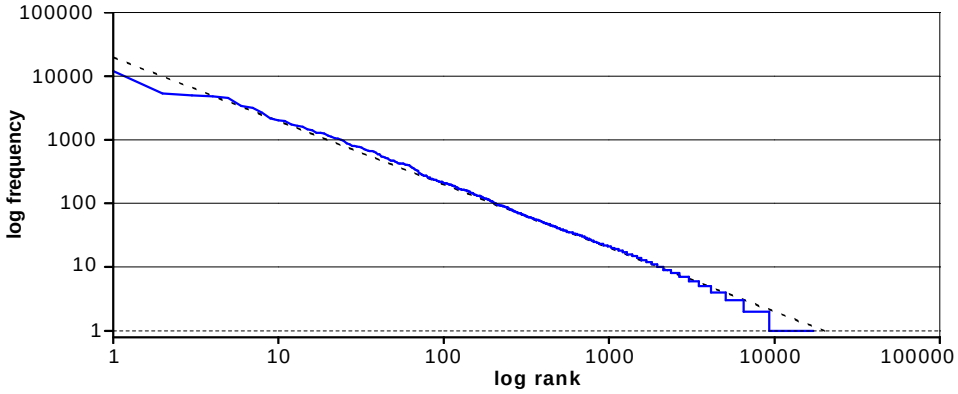

Figure 1 Zipf curve for the unigrams extracted from a 250,000-word tokens corpus.

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], [Baayen 1991], [Chitashvili 1983, 1989] and [Orlov 1983]. It continues to stimulate interest today [Samuelson 1996]; [Baayen 2001]; [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].

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.

1.2 Theoretical developments:

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 that 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 from the straight line, which

suggested several modifications of the law, and in particular one derived theoretically by [Mandelbrot 1953] with the form:

$$f = \frac{k}{(r+a)^b} \qquad (2)$$

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. Exceptions include legal texts which have smaller slopes (≈0.9) showing that lawyers use more word types than other people! [Smith and Devine 1985].

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]. 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.

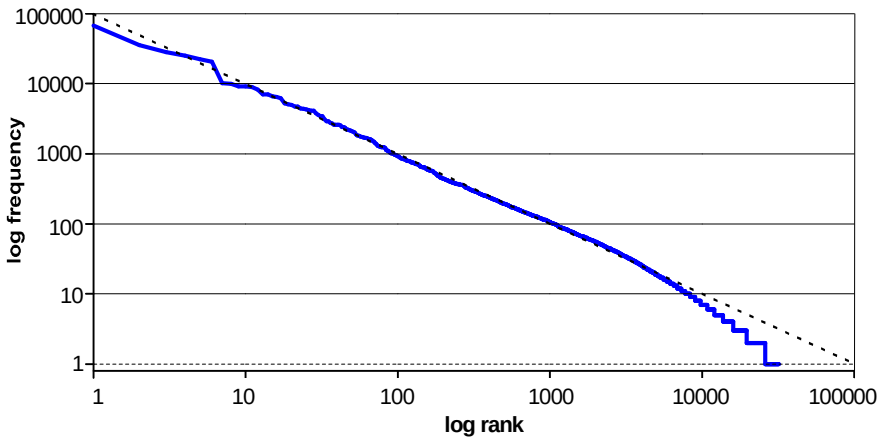

Figure 2 Zipf curve for the unigrams extracted from the 1 million words of the Brown corpus showing that the Zipf curve falls below the line with slope -1 for rank > 5,000.

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, for rank greater than about 5,000. This is illustrated in Figure 2, which shows the Zipf curve for the whole of the Brown corpus (1 million words), again excluding punctuations and numbers.

This deviation from Zipf 's law was interpreted for single-author texts to represent the limited numbers of words in each author's diction. But we see in Figure 2 that a deviation also occurs for a multi-author corpus covering a wide range of domains such as the Brown corpus; so the drop in the curve is not likely to be only due to the limited number of words.

2. Zipf curves for large English corpora

We are going to explore the above deviation from Zipf's law for large corpora in two languages: Chinese and English. We begin with English.

2.1 Single words

The English corpora used in our experiments are the full text of articles appearing in the Wall Street Journal [Paul and 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.

For pre-execution of this corpus, numbers were written as words, e.g. 23 became "twenty three" and punctuation marks were excluded. The characters "=", "#", "~", "<", ">", "|", "+", "-", "^", "*", "@", "/" and "\", etc. were also ignored.

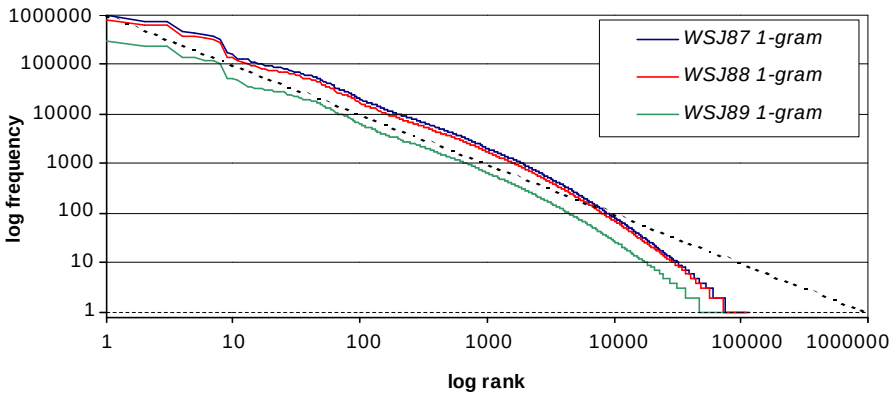

Figure 3 Zipf curves for the unigrams extracted from the 3 training corpora of WSJ

The Zipf curves for the three corpora are parallel, showing similar structures with all 3 curves deviating from Zipf's law for larger r in exactly the same way as the curve for the Brown corpus. Their separation is due to their different sizes.

2.2 *n*-Grams

Language is not made of individual words, each with its own separate piece of information, but consists of sequences of words, made up of individual words and of phrases of 2, 3 or more words together called *n*-grams. So it is interesting to measure the frequencies of *n*-grams and draw the corresponding Zipf curves.

To do this we allowed *n*-grams to overlap. For example, for the sentence: *"The cat sat on the mat"*, there are four trigrams: (1) *"the cat sat"*, (2) *"cat sat on"*, (3) *"sat on the"* and (4) *"on the mat"*. So semantically incomplete *n*-grams such as *"cat sat on"* are included in our study. No *n*-gram crossed over a punctuation mark. So a fullstop, comma, colon, etc. always ends an *n*-gram and a new *n*-gram starts after the punctuation. Thus the sentence *"Three blind mice, see how they run"* has only three trigrams *"three blind mice"*, *"see how they"* and *"how they run"*.

For each value of *n* between 2 and 5, we thus computed the frequencies of all *n*-grams 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-, 3-, 4- and 5-grams which are shown along with the single word curves in Figure 4, Figure 5 and Figure 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.

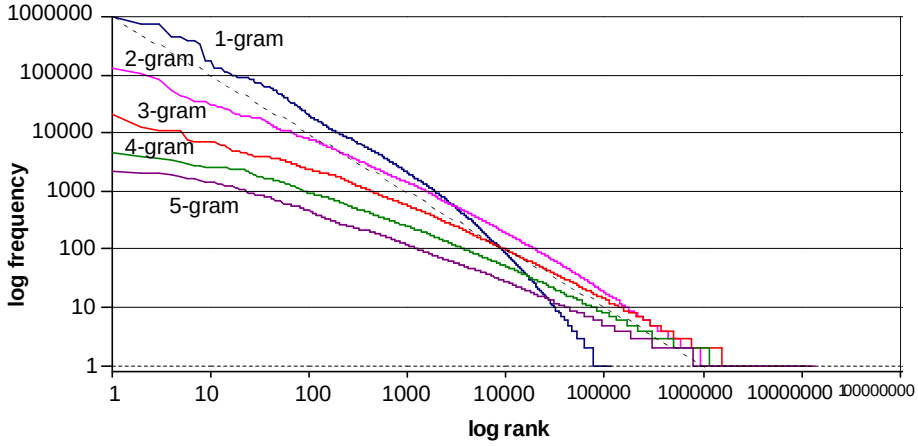

Figure 4 Zipf curves for the WSJ87 corpus

The *n*-gram Zipf curves do not follow straight lines but curve gently downwards. The average slope decreases from about 0.66 for the bigrams to about 0.59 for the 5-grams.

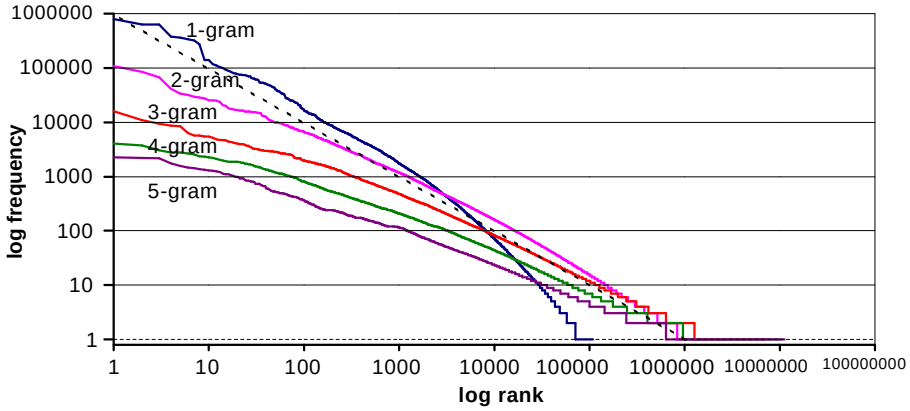

Figure 5 Zipf curves for the WSJ88 corpus

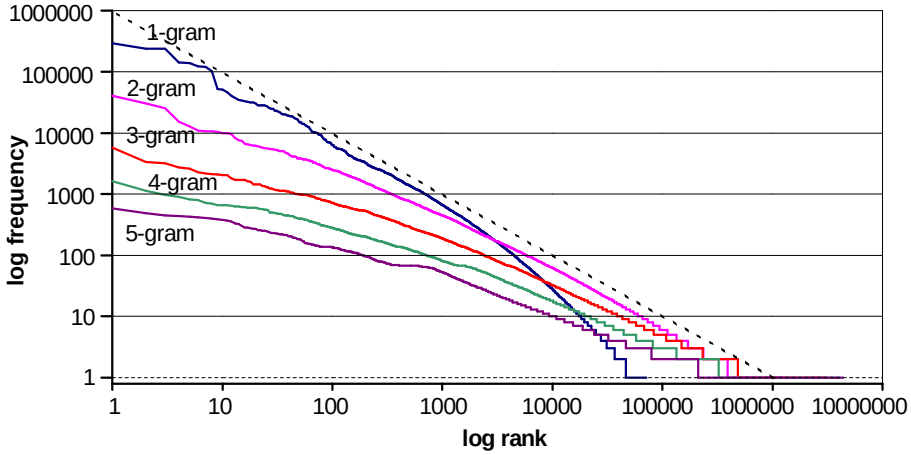

Figure 6 Zipf curves for the WSJ89 corpus

First for WSJ87, the crossing point between the unigram and bigram curves is at rank 2,943 and for the unigram and trigram curves, it is at rank 8,497. For WSJ88, these crossing points are similar, at rank 2,913 and at rank 8,404, and for WSJ89, they are at rank 2,908 and 7,960. So the unigram curves cross the bigram curves when the rank ≅ 3,000 in all 3 cases, and for the unigram and trigram curves, they cross at rank ≅ 8,000.

The ten most common words, bigrams and trigrams in the combined WSJ corpus of 40 million words are listed in Table 1.

Table 1. The most common unigrams, bigrams and trigrams in the combined WSJ

Unigrams		Bigrams		Trigrams	
Frequency	Token	Frequency	Token	Frequency	Token
2,057,968	the	217,427	of the	42,030	the U. S.
973,650	of	173,797	in the	27,260	in nineteen eighty
940,525	to	110,291	million dollars	24,165	cents a share
853,342	a	89,184	U. S.	18,233	nineteen eighty six
825,489	and	83,799	nineteen eighty	16,786	nineteen eighty seven
711,462	in	76,187	for the	15,316	five million dollars
368,012	that	72,312	to the	14,943	million dollars or
362,771	for	65,565	on the	14,517	million dollars in
298,646	one	63,838	one hundred	12,327	in New York
281,190	is	55,014	that the	11,981	a year earlier

2.3 Hapax legomena and dis legomena

The size of the hapax legomena (tokens with frequency 1) for the *n*-grams rises rapidly with *n* as shown in Table 2a, but it can not rise above the number of tokens; so the rate of increase has slowed when $n = 5$ since almost all tokens are in the hapax legomena. The hapax dis legomena (tokens with frequency 2) is much smaller and reaches a maximum for trigrams from all 3 corpora (see Table 2b) because almost all of the tokens have frequency 1, leaving a smaller number with frequency 2 when $n = 4$ and 5.

Table 2a) Number of hapax legomena for the English corpora.

Corpus		WSJ87	WSJ88	WSJ89
No of Tokens		18,790,794	15,757,051	5,946,585
No of Types		114,581	108,522	71,837
Hapax legomena	Unigram	38,853	36,945	25,162
	Bigram	1,786,290	1,620,385	851,542
	Trigram	6,601,243	5,799,257	2,598,509
	4-gram	10,635,310	9,137,402	3,736,880
	5-gram	12,493,656	10,612,036	4,376,741

Table 2b) Number of hapax dis legomena for the English corpora.

Corpus		WSJ87	WSJ88	WSJ89
No of Tokens		18,790,794	15,757,051	5,946,585
No of Types		114,581	108,522	71,837
Hapax dis- -legomena	Unigram	14,855	14,431	9,861
	Bigram	349,205	314,496	155,068
	Trigram	742,771	632,372	251,435
	4-gram	670,106	546,951	190,947
	5-gram	485,487	389,113	130,544

2.4 The nature of *n*-grams

It can be argued that most of the *n*-grams in the hapax legomena or hapax dis legomena are not meaningful, since they are semantically incomplete. Certainly that meaning may be incomplete and they need the words on either side of them to realise their full meaning. But then it can be argued that this is true of every *n*-gram (and indeed for every word). So we take the view that every *n*-gram taken from a natural language text produced by humans has meaning, though often incomplete.

However, Miller's monkey typing on a word typewriter would produce mainly meaningless *n*-grams, e.g. *"the the the"*, as well as those others which have meaning by accident. The number of possible *n*-grams which the monkey can type is huge. For example, for the WSJ87 corpus there are more than 10^{15} possible trigrams of which less than 7 million produced by humans appear in the Hapax legomenon for the corpus.

Whatever one's views on the meaning of some of these incomplete *n*-grams, we report in this paper on the Zipf curves for all *n*-grams in a corpus. A later paper will include discussion on the equivalent curves for semantically complete phrases.

One of our reasons for including all *n*-grams is that statistical language modellers have been using *n*-grams, similar to the ones we have defined, which include semantically incomplete *n*-grams, with great success in modelling language over the last 20 years [Jelinek and Mercer 1985]; [O'Boyle, Owens and Smith 1994]; [Ney 1999].

3. Zipf Curves for Chinese Corpora

In Chinese, compound words can be created, made up of two or more characters. However, it is not always easy to automatically segment a written sentence in Chinese into compound

words as these are not separated by spaces as in English. Nevertheless, the extraction of a word sequence from a Chinese document has been the subject of study by many authors [Zhu 1981]; [Chen and Shi 1992]; [Bates, Chen, Li, Opie and Tzeng 1993]; [Packard 2000]; [Sproat 2002]; [Tsai and Hsu 2002]; who reference other papers.

Unfortunately, there is still ambiguity in the process of compound word extraction. For example, the following string of characters can be broken into the words: 北京 (Beijing) 城 (city) 里 (in) 交通 (traffic) 繁忙 (busy) (*The traffic in Beijing is very busy*) or into the words 北 (North) 京城 (capital city) 里 (in) 交通 (traffic) 繁忙 (busy) (*The traffic in the north of the capital city is very busy*). Only a human can distinguish which is correct, another example is: 上海 (Shanghai) 边 建设 (build) 边 发展 (develop) (*Shanghai is developing while it is building (up)*) which can also be interpreted as 上 (go to) 海边 (seaside) 建设 边 发展 (*Go to seaside to develop and build*). Once again a human is needed to decide the meaning.

Therefore, it is difficult to write a computer program to extract the correct word sequence, and for a corpus of 250 million syllables, it is impossible to do by hand. So we proceeded as follows: first of all, we used a 50,000 word-syllable dictionary (which can be found at http://www.euroasiasoftware.com/), but the extraction of the words from the text is still partly ambiguous. When a sequence of syllables was found that matched a word in the dictionary, it was usually accepted as a word. When an ambiguity occurs, e.g. 暴风骤雨 which can be one word *hurricane*, or two bi-syllable words: 暴风 骤雨 *storm shower*, then the longer word was accepted 暴风骤雨 *hurricane*. Similarly, 百万富翁 *millionaire* is accepted as one word instead of the three words 百万 富 翁 *million(s) rich elder*.

Although the whole corpus could not be checked manually, the higher frequency *n*-grams can be checked, for example the following 6-gram has been broken into the pattern: 埃及 总 统 穆 巴 拉 克 rather than the pattern 埃及 总统 穆巴拉克 *Egyptian president Mubarak*. This occurs 1,865 times and could be corrected for all 1,865 occurrences in one step all over the corpus. Another example is the 7-gram: 阿 联 酋 乌 兹 别 克 occurring 7 times which should be the 2 names 阿联酋 乌兹别克 *Alanqiu Wuzibieke* to be correct. Because of the multiple occurrence of *n*-grams all of which can be corrected by one change, this speeded-up the manual process considerably. Checking all of the high frequency *n*-grams took more than 2 months work; after this, a check on a test text of 3,117 tokens was found to have 82 errors (2.6%) by an independent native speaker (other than the authors), which we took as acceptable. (The corpus can be made available on request to q.le@qub.ac.uk or fj.smith@qub.ac.uk).

Two corpora were used in our experiments: the TREC corpus and the Mandarin Daily

News corpus. Both are from the Linguistic Data Consortium[1].

There is a small overlap between the Chinese TREC corpus and the Mandarin News corpus (less than 10% of the smaller TREC corpus). This overlap could have been removed, but it was not, to retain the full size of both corpora in the analysis. The effect of overlap will be small.

3.1 TREC Corpus (compound words)

The TREC Corpus was obtained from the full articles in the People's Daily Newspaper from 01/1991 to 12/1993 and from the Xinhua News Agency from 04/1994 to 09/1995.

The Zipf curves for the TREC compound words are shown in Figure 7. Note that the unigram curve is different from the curve for English, first with a slope less than 1 then falling rapidly after a rank of about 1,000.

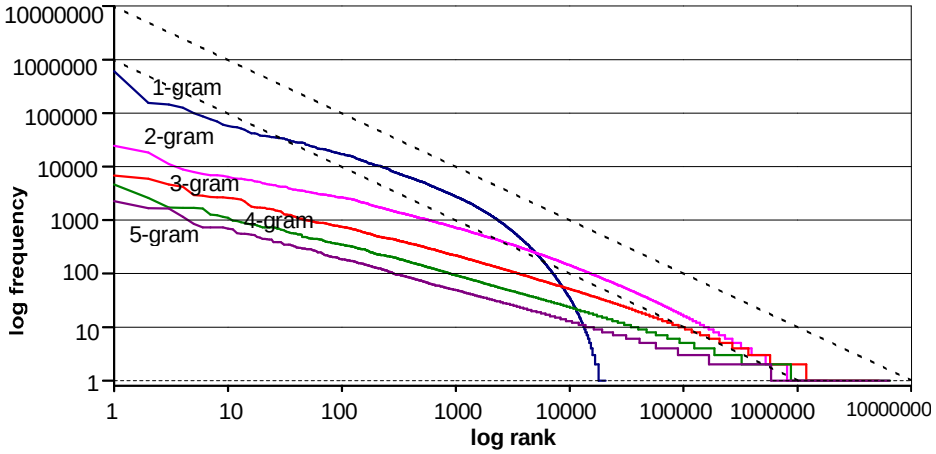

Figure 7 Zipf curves for Mandarin compound words from TREC

The crossing-point between compound word unigrams and bigrams is at rank: 4,999, and between the unigram and trigram curves at rank: 8,589, similar to English.

3.2 Mandarin News corpus (compound words)

The second corpus is the Mandarin News corpus, obtained from the People's Daily Newspaper from 1991 to 1996 (125 million syllables); from the Xinhua News Agency from

[1] http://www.ldc.upenn.edu/

1994 to 1996 (25 million syllables); and from transcripts from China Radio International broadcast from 1994 to 1996 (100 million syllables), altogether over 250 million syllables.

The Zipf curves for the Mandarin News compound words are drawn in Figure 8 and look like those for the TREC corpus. The rapid fall in the curve after rank 10,000 is due to the restricted word dictionary of 50,000 word types used in the experiment. The ten highest frequency Mandarin unigrams, bigrams and trigrams from the Mandarin News are in Table 3 and Table 4.

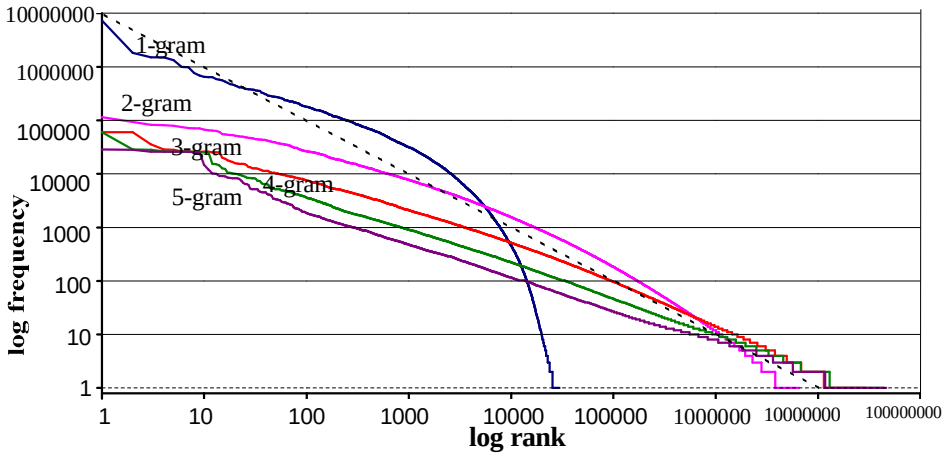

Figure 8 Zipf curves for the Mandarin News corpus (compound words)

The crossing-point between compound word unigrams and bigrams is at rank: 5,544 and between unigrams and trigrams at rank: 9,577 similar to previous values for TREC and English. So these appear to be invariants of language, not just of English.

Table 3 The ten highest frequency unigrams and bigrams from Mandarin News (compound words).

Rank	Unigrams			Bigrams		
	Freq	Token	Meaning	Freq	Token	Meaning
1	7,356,017	的	of	114,910	日 电	daily news
2	1,825,758	在	in / at	92,259	这 一	this one
3	1,515,473	和	and	82,705	这 是	this is
4	1,502,098	了	perfective marker	81,930	中国 的	of China
5	1,331,433	是	yes / right	79,390	的 发展	of development

6	989,235	一	one	75,929	他 说	he says
7	979,211	中国	China	71,922	的 一	of one
8	766,784	中	centre / middle	70,949	新 的	new of
9	686,375	有	have	70,810	日 在	daily at
10	652,004	年	year	67,211	中国 国际	China international

Table 4 The ten highest frequency trigrams from Mandarin News.

Rank	Trigrams		
	Freq	Token	Meaning
1	60,214	国际 广播 电台	international broadcast station
2	60,057	中国 国际 广播	China international broadcast
3	35,584	一 九 九	one nine nine[2]
4	28,589	据 中国 国际	according to China international
5	28,240	广播 电台 报导	broadcast station report
6	26,240	学历 收听 语言	degree listen *(to)* language
7	26,232	年龄 学历 收听	age degree listen *(to)*
8	26,203	收听 语言 备注	listen *(to)* language remarks/notes
9	26,154	职业 年龄 学历	profession age degree
10	26,081	传真 单位 职业	fax department profession

4. Zipf Curves for syllables and character strings

4.1 Chinese syllables

Because of the difficulty in extracting the compound words in Chinese, we decided to draw Zipf curves for the syllables for both Chinese corpora. TREC has 19,546,872 syllable tokens but only 6,300 syllable types, so it is not surprising that the Zipf curve for syllable unigrams

[2] This is how Chinese people read and write the year for example 1993 as "*one nine nine three*"; therefore we elminated numbers but kept the written form.

in Chinese in Figure 9 falls very rapidly after rank about 300. It is similar to previous curves, one for a smaller Chinese corpus of 2 million tokens by [Clark, Lua and McCallum 1986] and one for 10 million tokens by [Sproat 2002]. The Zipf curves for syllable *n*-grams for the TREC corpus are also shown in Figure 9.

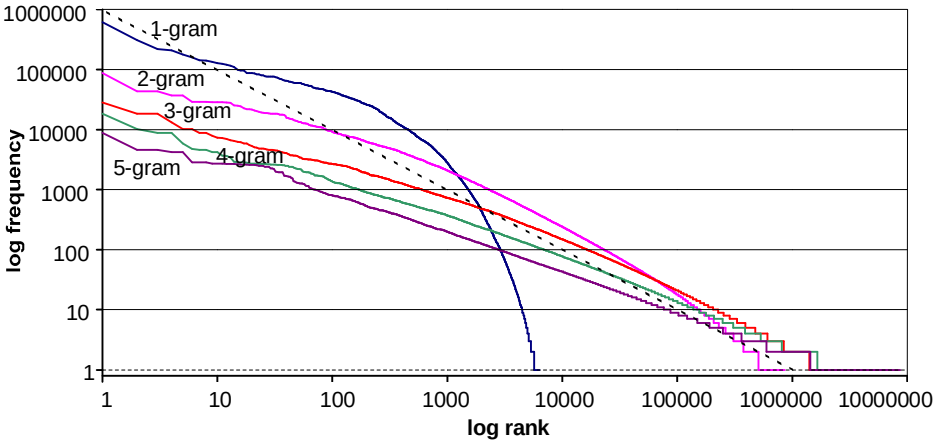

Figure 9 Zipf curves for syllables from the TREC Mandarin corpus

Except for the unigrams, the shapes of the other TREC syllable n-gram Zipf curves are similar to but not quite the same as those for compound words. In particular the syllable bigram curve for Chinese is more curved than the word curve because there are more high-frequency syllable bigrams than word bigrams. The crossing points between the syllable unigram curve and the bigram and trigram curves are at rank: 1,224 and 1,920, respectively, very different from compound words.

The number of syllable-types (i.e. unigrams) in the Mandarin News corpus is 6,800, similar to the TREC corpus. The Zipf curves and crossing points are also similar as shown in Figure 10.

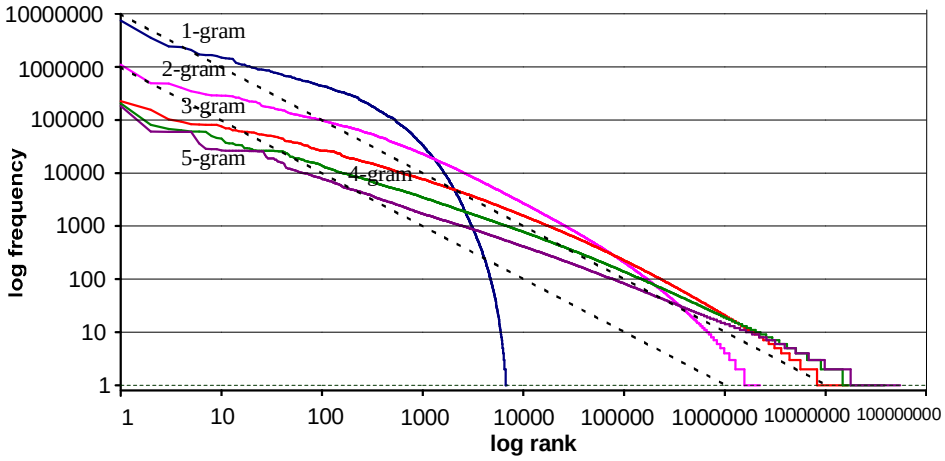

Figure 10 Zipf curves for syllables from the Mandarin News corpus

The hapax legomena and dis legomena for the Chinese corpora Zipf curves are shown in Table 5a and 5b. Their behaviour as *n* increases is similar to the English corpora.

Table 5a) Number of hapax legomena for the Chinese corpora.

Corpus		TREC syllables	TREC compound words	Mandarin News syllables	Mandarin News compound words
No of Tokens		19,720,320	13,467,443	223,222,788	153,942,010
No of Types		6,356	20,587	6,891	29,688
Hapax legomena	Unigram	676	2,642	259	4,192
	Bigram	351,691	1,013,276	667,966	2,671,406
	Trigram	2,447,451	4,009,020	8,462,775	17,794,466
	4-gram	5,309,654	5,661,530	23,812,934	30,885,192
	5-gram	7,279,824	5,875,696	37,348,300	34,617,579

Table 5b) Number of hapax dis legomena for the Chinese corpora.

Corpus		TREC syllables	TREC compound words	Mandarin News syllables	Mandarin News compound words
No of Tokens		19,720,320	13,467,443	223,222,788	153,942,010
No of Types		6,356	20,587	6,891	29,688
Hapax	Unigram	347	1,260	155	1,701

dis-	Bigram	131,619	287,225	294,634	1,001,809
legomena	Trigram	565,069	624,895	2,591,830	4,549,748
	4-gram	834,227	553,965	5,806,633	6,168,487
	5-gram	840,276	417,656	7,874,536	6,057,974

4.2 English byte substring

Following a suggestion by a reviewer of this paper, we built the Zipf curves on English 2-byte and 3-byte substrings to compare them with the Chinese syllable results.

From the WSJ88 corpus, we built a corpus of the first 2 million tokens. Then we took 2-byte and 3-byte moving windows on this corpus ignoring spaces and stopping the 2-bytes or 3-bytes at punctuation marks. As predicted by the reviewer, the results in Figure 11 and Figure 12 show that the Zipf curve for 3-byte substrings looks particularly similar to the Chinese syllable curves.

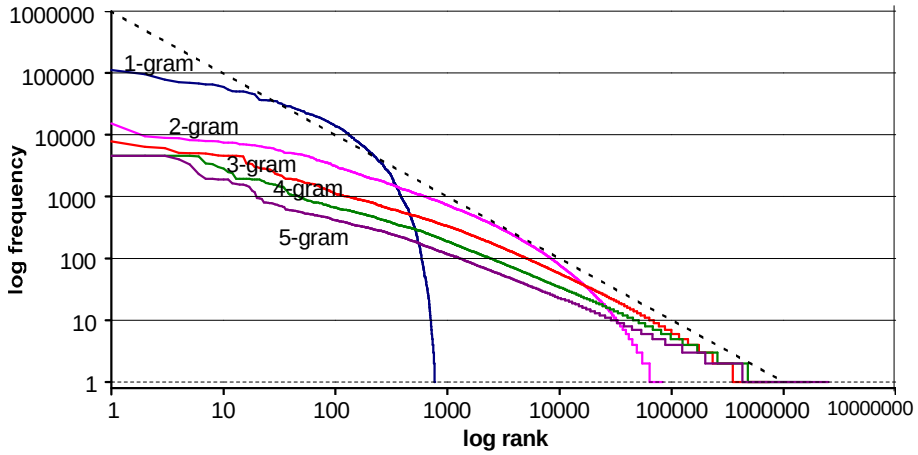

Figure 11 Zipf curves for English 2-byte substrings

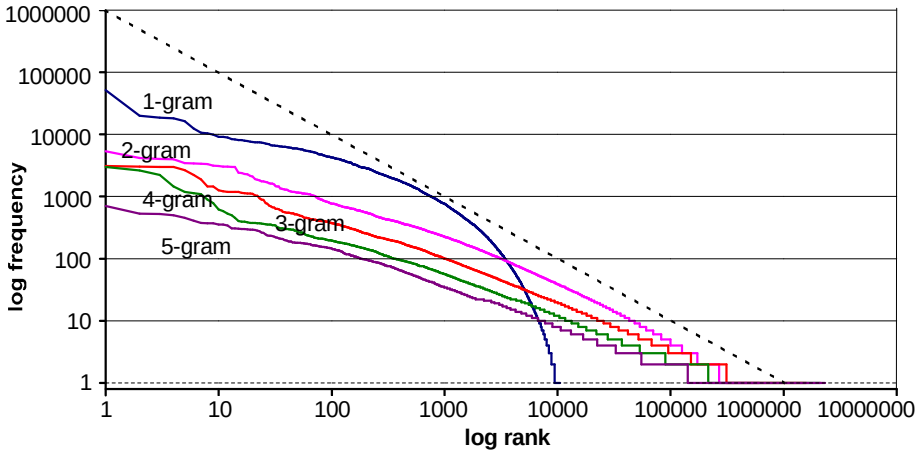

Figure 12 Zipf curves for English 3-byte substrings

Note that the number of 2-byte and 3-byte types in these curves equal 673 and 10,548, compared with the maximum possible numbers $26^2 = 676$ and $26^3 = 17{,}576$.

5. Comparison for all Zipf curves from Chinese and English

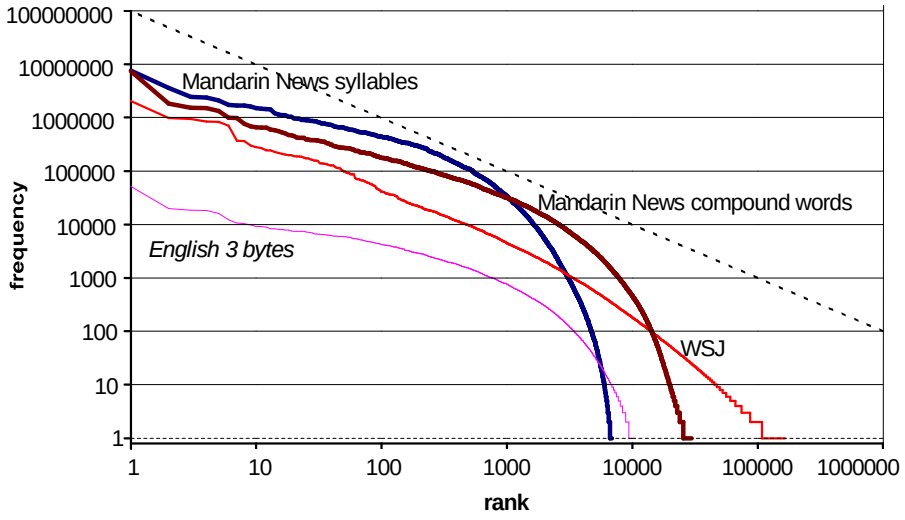

Figure 13 Comparison of Zipf curves for unigrams

The Zipf curves for unigrams for the combined WSJ corpus, the Mandarin News word corpus, the Mandarin News syllable corpus and 3-byte English corpus are compared in Figure 13.

Similarly for 2-grams to 5-grams in Figures 14 to 17.

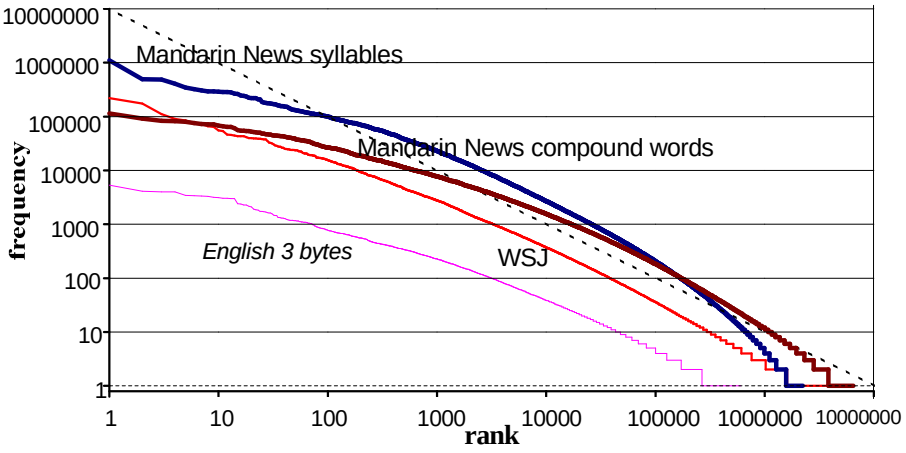

Figure 14 Comparison of Zipf curves for bigrams

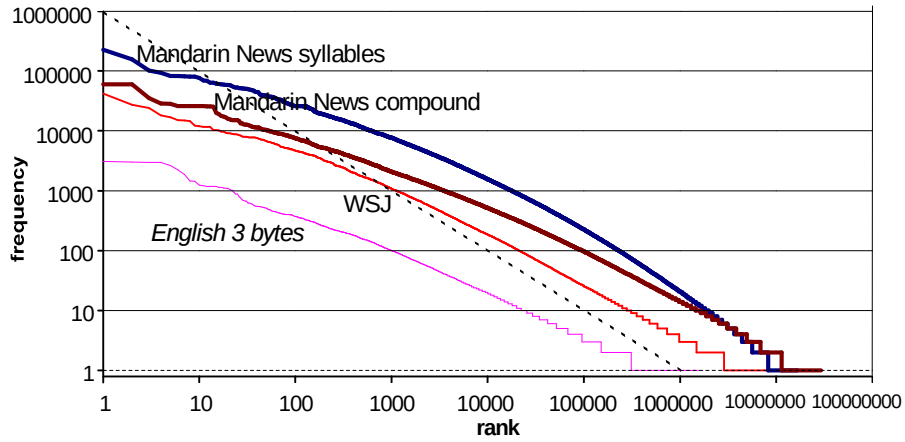

Figure 15 Comparison of Zipf curves for trigrams

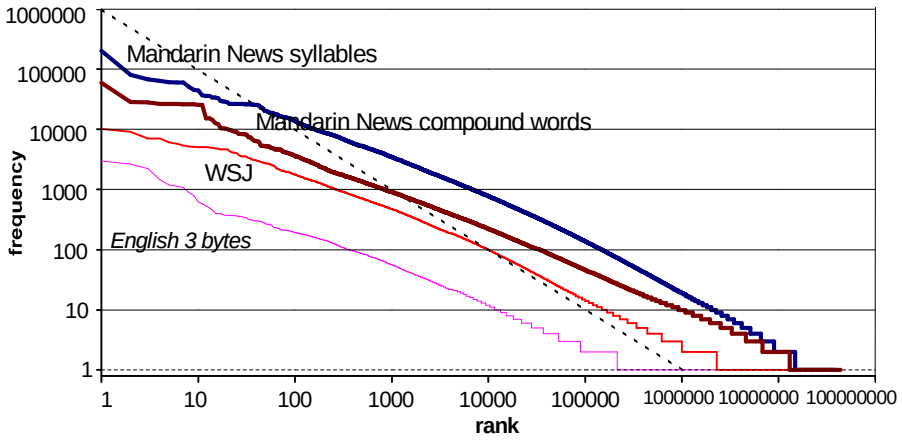

Figure 16 Comparison of Zipf curves for 4-grams

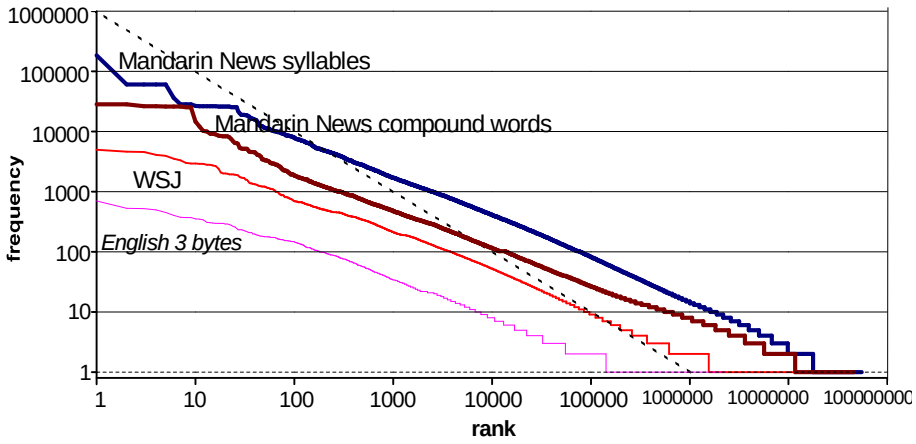

Figure 17 Comparison of Zipf curves for 5-grams

These Figures show two things as *n* increases. First, the curves straighten out for high *n*. Secondly, the number of hapax legomena becomes very large, often larger than one would expect from the last 10 steps of the rank-frequency step function. This is exactly the pattern one gets when Markov models are used to generate data sets [Baayen 1991, 2001].

6. Combined *n*-grams

The theoretical justifications for Zipf's law by Mandelbrot, Miller, Simon and others were based on single word tokens and they worked quite well for small corpora, but none of them could predict the drop in the Zipf curve below Zipf 's law for English and Chinese when the rank is greater than 5,000 word types. In the case of Chinese syllables, Zipf 's law could not hold for rank greater than about 100, but when these syllables are combined into compound words then Zipf 's law is valid for a wider range, up to about rank 1,000. Therefore, by combining Chinese syllables into larger units, Zipf 's law was extended from rank 100 to rank 1,000. This led us to combine all syllable *n*-grams, to see if the law could be extended to even higher rank and to combine word *n*-grams in Chinese and English for the same purpose.

We therefore put all unigrams and *n*-grams together with their frequencies into one large file, sorted on frequency and put in rank order as previously. The resulting combined Zipf curve is shown with the unigram curve for English words for the combined WSJ corpus in Figure 18 and for the Chinese syllables for Mandarin News in Figure 19.

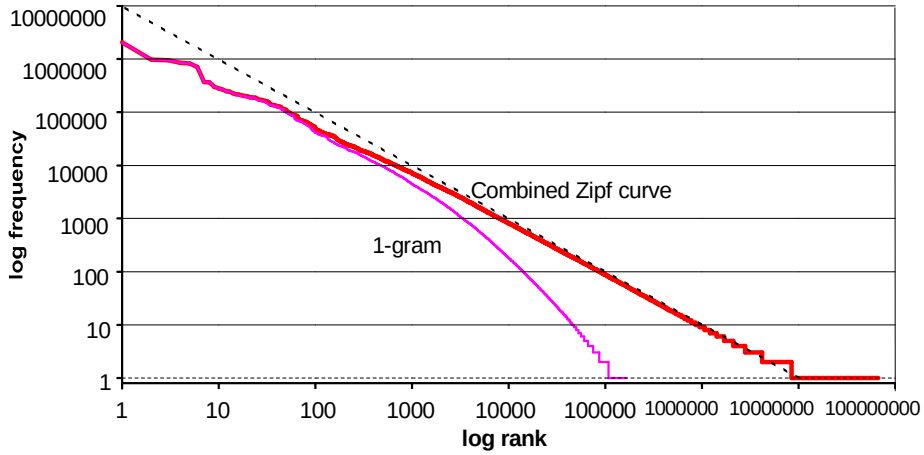

Figure 18 The unigram and combined curves for the combined WSJ corpus

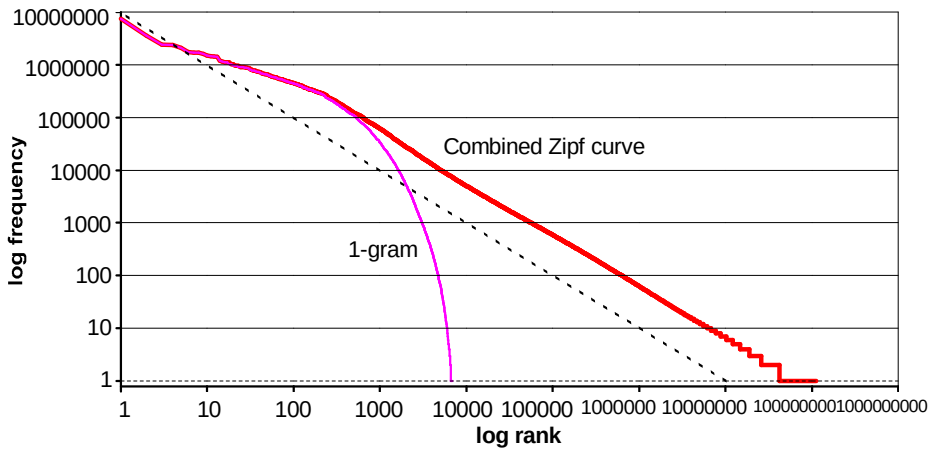

Figure 19 The unigram and combined curves for the Mandarin News syllable corpus

This shows the remarkable result that as the unigram curve drops away from Zipf 's slope of -1, the shortfall is made up almost exactly by the *n*-grams in both cases, even though those shortfalls are very different in the two cases. So when all *n*-grams are combined together, including unigrams, Zipf's law is found to be approximately correct with a slope close to -1 for all ranks. If semantically incomplete *n*-grams had been excluded from this analysis, this result would not have been obtained.

The resulting Zipf curves for the combined *n*-grams from all of the corpora are shown in Figure 20.

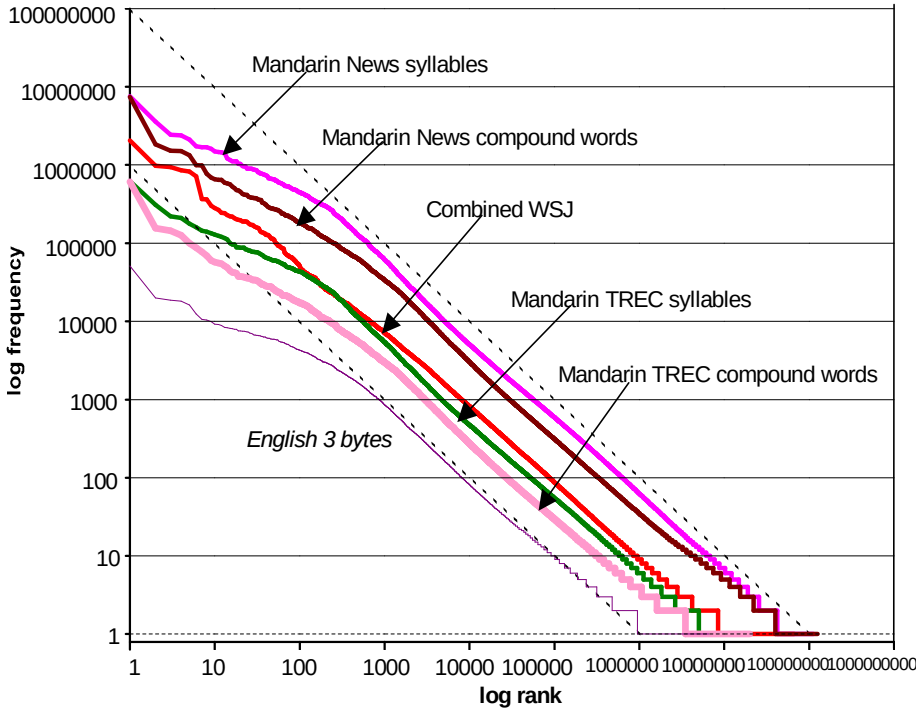

Figure 20 Combined Zipf curves for both of the languages

This shows that the 6 combined Zipf curves are all approximately straight lines with slopes close to –1 for all ranks > 1,000. For ranks < 1,000, the unigram curves dominate and are not so straight. As in Figure 18 and 19, the *n*-grams ($n \geq 2$) almost exactly make up for the deviation of the unigram Zipf curve from Zipf 's law for the six very different unigram curves. So the results in Figure 20 are a new confirmation of Zipf's original law in an extended form.

7. Summary and Conclusions

This paper reports on the results of some experiments conducted on Zipf curves for English and Chinese corpora. It was confirmed that Zipf curves on a log-log graph for single word unigram distributions for both languages fall below the straight line with slope -1 as predicted by Zipf 's law. The deviation from Zipf 's law occurs at a rank close to rank = 5,000, for the 3 corpora in English and 2 corpora in Chinese. This rank (5,000) is also the rank near which the unigram and bigram Zipf curves cross for all 5 corpora.

The more significant result was the discovery that when the frequency distribution of

words is combined with the distributions of all 2-, 3-, 4- and 5-grams, the combined Zipf curve approximately obeys Zipf 's law for all ranks and frequencies for both languages. This effectively extends Zipf 's law, with the higher *n*-grams almost exactly making up for the fall-off in the Zipf curve for words. Furthermore, this extended form of Zipf 's law also holds for the syllables of Chinese (as well as for 2-byte and 3-byte word fragments in English), even though the distribution of syllable unigrams is very different from the distribution for words.

This paper does not explain why Zipf 's law in an extended form is valid for large corpora or what this result means. This must be left for further experiments and other researchers. However, preliminary results, not yet complete, for other languages suggest that these results are universal for all languages. We also know that they do not hold for all artificial distributions of words, because some experiments with computer generated artificial distributions did not yield an extended Zipf curve, (with a random distribution, and with Zipf distributions for words with slopes $\beta = 2$ and $\beta = 0.5$).

The earlier derivations of Zipf's law due to Mandelbrot, Miller, Simon and others fail to predict the fall-off in the Zipf curve from about rank 5,000 and to predict the extended form of Zipf's law for the combined *n*-gram curves. We believe that this is because these derivations do not properly take account of the fact that each token is part of a sequence and its information is dependent on a conditional probability, conditional on the words or characters around it; this can be approximated in terms of the frequency of *n*-grams [O 'Boyle, Owens and Smith 1994].

Acknowledgement

The authors would like to express their appreciation to reviewers of this paper whose comments and suggestions made a great improvement to the paper and to Dr Xiaoyu Qiao for her contribution of testing and standardising the Chinese morphology.

References

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

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

Bates, E., Chen, S., Li, P., Opie, M. and Tzeng, O. "Where is the boundary between compounds and phrases in Chinese ? A reply to Zhou et al.", *Brain and Language*, 45, 1993, pp. 94-107.

Booth, A. D. “A Law of Occurrences for Words of Low Frequency”, *Information and Control*, Vol. 10, No. 4, April 1967, pp. 386-393.

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

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

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: Chinghwa University, 1992.

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

Ferrer i Cancho, R., Solé, R. V., "Two Regimes in the Frequency of Words and the Origin of Complex Lexicons", *Journal of Quantitative Linguistics*, Vol. 8, No. 3, 2002, pp. 165 - 173.

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.

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

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, 2001, pp. 175 - 185.

Jelinek, F., Mercer, R. L. "Probability distribution estimation from sparse data", *IBM Technical Disclosure Bulletin*, Vol. 28, No. 6, November 1985.

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

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

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

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, 1959, pp. 90-99.

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, 1961, pp. 198-216.

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

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

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

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

Montemurro, M. "Beyond the Zipf-Mandelbrot Law in Quantitative Linguistics", *Physica A*: *Statistical Mechanics and its Applications*, Vol. 300, Issues 3-4, November 2001, pp. 567-578.

Ney, H. "The Use of the Maximum Likelihood Criterion in Language Modelling", In *K. Ponting (*ed.): Computational Models of Speech Pattern Processing*, Springer, Berlin, Germany, 1999, pp. 259-279.

O 'Boyle, P., Owens, M. and Smith, F. J. "A weighted average *n*-gram model of natural language", *Computer Speech and Language*, Vol. 8, 1994, pp. 337-349.

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

Packard, J. L., "The Morphology of Chinese A Linguistic and Cognitive Approach", Cambridge University Press, 2000, UK.

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 (ICSLP)*, Banff, Alberta, Canada, October 1992, pp. 899-902.

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

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

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

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

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

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

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

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

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

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

Sproat, R., "Corpus-Based methods in Chinese Morphology", *Tutorial of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, August 2002.

Tsai, J-L., Hsu, W-L. "Applying an NVEF Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem", In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, August 2002, pp. 1016-1022.

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

Zhu, D. X. "Yufa Jiangyi (Chinese Syntax)", Shanghai: The Commercial Publisher, China, 1981.

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