WORD LENGTH, SENTENCE LENGTH AND FREQUENCY – ZIPF REVISITED

Bengt Sigurd, Mats Eeg-Olofsson & Joost van de Weijer

Abstract. This paper examines data from English, Swedish and German in order to find a theoretical distribution that describes the observed relation between word length and frequency. In Swedish and English, most word tokens consist of three letters only, while shorter or longer words occur less frequently. We found that the equation with the general form $f_{exp} = a * L^b * c^L$ (a variant of the so-called gamma distribution) approximates the observed frequencies reasonably well. This formula incorporates both the fact that the number of possible words increases with word length, and the fact that longer words tend to be avoided, presumably because they are uneconomic. To our knowledge this formula has not been proposed to describe word frequency data. We examined frequency distributions of word length in Swedish and English, and explored different variants of the equation by systematically varying the *a, b* and *c* parameters. Subsequently, we also applied the formula to the frequency distribution of sentence length in English, and found an almost perfect fit for a corpus consisting of different text genres. Moreover, the data showed that the formula can be used to distinguish between different kinds of text genres.

1. Introduction

George K. Zipf is famous for his law of abbreviation. In *The Psycho-Biology of Language* (1935:38) he says: "In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency. Footnote: Not necessarily proportionate; possibly some non-linear mathematical function."

The law seems to fit data as short words are generally more frequent than long words. However, words shorter than three letters are not taken into account. In the present study we propose a formula which can account for these short words as well. The formula has a natural explanation, and it can, as we will demonstrate, also be used for the approximation of the frequencies of sentences of different length.

2. English and Swedish word frequency data

To illustrate our computation, we used an English and a Swedish language corpus. The English corpus was taken from Kučera and Francis' classical corpus linguistic study *Computational Analysis of Present-Day American English* (1967). It is based on the *Standard Corpus of Present-day Edited American English,* a corpus of language texts assembled at Brown University during 1963–64. The corpus includes about one million running words divided into 500 samples from different genres. Pages 363–367 of the book are devoted to word length and

Studia Linguistica 58(1) 2004, pp. 37–52. © The Editorial Board of Studia Linguistica 2004.
Published by Blackwell Publishing Ltd., 9600 Garsington Road, Oxford OX4 2DQ, UK, and
350 Main Street, Malden, MA 02148, USA

frequency. The left-hand side of the left column (labeled 'observed frequencies') of Table 1 below is a reproduction of the table on page 366 showing the frequencies of word tokens of different lengths. There is also a graph in the book showing the shape of the corresponding curve with a top at the frequency at three letters, but no attempt to find an approximation.

The Swedish corpus is Sture Allén's (1970) *Nusvensk Frekvensordbok 1 (Frequency Dictionary of Present-Day Swedish 1)*. It includes approximately one million words based on samples from different genres in newspapers from 1965, and is a representative corpus for Swedish. It gives a table showing the frequency of words of different lengths in the corpus. This table is reproduced on the left-hand side of the right column of Table 1.

As shown in the table, in both languages three-letter words are most frequent. In English three-letter words have a relative frequency of 21.2%, followed by two-letter words (17.0%) and four-letter words (15.7%). Swedish three-letter words have a relative frequency of 24.6%, followed by words of two letters (12.0%) and words of five letters (11.0%). In both languages, words of one or two letters have lower frequencies than words of three letters, contradicting Zipf's law.

3. Approximation of frequencies of words of three or more letters by a simple exponential function (a geometric distribution)

Zipf did not suggest any function deriving frequency from word length. This problem has interested several researchers (e.g. Wimmer et al. 1994, Best 1996, Wilson & McEnery 1998, Strömqvist et al. 2002).

It is reasonable to approximate the slope starting at three letters by a simple exponential function both for English and Swedish data (suggested by Lars Gårding, Lund). The observed relative frequencies are predicted fairly well by the function $f_{exp} = 21.2 * 0.73^{(L-3)}$ for English, and $f_{exp} = 24.6 * 0.71^{(L-3)}$ for Swedish. The predicted percentages are given on the right-hand sides of the columns in Table 1 below. The correlation coefficients between the observed and the predicted values are almost perfect ($r = 0.996$ for English; $r = 0.976$ for Swedish), suggesting that the theoretical models fit the observed data well.

4. Approximation of the whole range of data

We think it is an interesting task to find a function which can approximate the frequency of the words that are shorter than three letters as well. It is natural to believe that the distribution of word frequencies is related to the number of letters or sounds in the word. The

Table 1: Observed and predicted word frequencies (%) on the basis of word length in letters (L) in English and Swedish. Formulas for the predicted values are $f_{exp} = 21.2 * 0.73^{(L-3)}$ for English, and $24.6 * 0.71^{(L-3)}$ for Swedish

Word Length	English		Swedish	
	Observed	Predicted	Observed	Predicted
1	3.160		3.371	
2	16.975		11.953	
3	21.192	21.200	24.654	24.600
4	15.678	15.476	10.953	17.466
5	10.852	11.297	11.099	12.401
6	8.524	8.247	9.225	8.805
7	7.724	6.020	6.227	6.251
8	5.623	4.395	5.442	4.438
9	4.032	3.208	4.495	3.151
10	2.766	2.342	3.738	2.237
11	1.582	1.710	2.610	1.589
12	0.917	1.248	1.898	1.128
13	0.483	0.911	1.245	0.801
14	0.262	0.665	0.889	0.569
15	0.099	0.486	0.646	0.404
16	0.050	0.354	0.459	0.287
17	0.027	0.259	0.333	0.203
18	0.022	0.189	0.253	0.144
19	0.011	0.138	0.168	0.103
20	0.006	0.101	0.118	0.073
21	0.005	0.073	0.088	0.052
22	0.002	0.054	0.053	0.037
23	0.001	0.039	0.035	0.026
24	0.001	0.029	0.018	0.019
25	0.001	0.021	0.011	0.013
26	0.001	0.015	0.007	0.009
27	0.001	0.011	0.004	0.007
28	0.000	0.008	0.002	0.005
29	0.000	0.006	0.001	0.003
30	0.000	0.004	0.000	0.002
31			0.001	0.002
32	0.000	0.002		0.001
33	0.000	0.002	0.000	0.001
34	0.000	0.001	0.000	0.001
36	0.000	0.001		0.000
37	0.000	0.000		0.000
38	0.000	0.000		0.000
41	0.000	0.000		0.000
44	0.000	0.000		0.000
59			0.000	0.000
61			0.000	0.000

word frequency must increase with the number of potential words, which increases when more letters are allowed in the word. A language with only five vowels and 20 consonants can normally have five words of one letter; 20 * 5 = 100 CV-words and 5 * 20 = 100 VC-words. With three letters the structures CVC, CCV, CVV, VCC,VCV and VVC are possible in Swedish. The shortest words are typically prepositions (*i* 'in', *av* 'of'), conjunctions (*att* 'that'), articles (*en* 'a'), and pronouns (*den* 'it'; cf. Miller 1951:88).

In the vocabulary the number of vowels and the number of consonants in each position vary with the phonotactic context. In Swedish, for instance, only *s* can occur as the first member in a word-initial cluster of three consonants (cf. Sigurd 1965). Without taking the detailed phonotactic constraints into account the number of potential words (W) may be assumed to increase as reflected in $W = L^b$ in which L is the word length measured in the number of letters, and b is a suitable parameter.

The other fact to be taken into account is that the length of words is a burden from several points of view. Longer words take longer time to write and read, to pronounce and perceive. Economic factors press for shorter words, as pointed out by Zipf. Moreover, experimental research on human memory has shown that word length has a direct impact on the amount of items that subjects are able to retain in working memory (Baddeley 1999). Subjects more easily remember shorter words than longer words, presumably because they keep the information active by means of inner articulation during a retention interval.

The decreasing effect of word length is taken into account in the formula suggested by the exponential factor c^L where $0 < c < 1$. The complete formula with these two counteracting factors is then:

$$f_{exp} = a * L^b * c^L \tag{1}$$

This formula corresponds to the general formulation of a Gamma distribution, which has a probability density function (f) that is generally described as

$$f = K * L^{\alpha-1} * e^{\frac{-L}{\beta}} \tag{2}$$

where the normalizing constant K is determined uniquely by the positive parameters α and β and the requirement that the formula defines a proper probability distribution. α is usually called the shape parameter, determining the existence and height of a peak of the graph of the function. β is called the scale parameter, which determines the "spread" of the distribution. Obviously, $b = \alpha - 1$ and $c = e^{\frac{-1}{\beta}}$. For $b > 0$, the function graph has a maximum for $L = b/\ln(1/c)$, where ln designates the natural logarithm. For $b = 0$, the Gamma distribution coincides with

the well-known exponential distribution. Interestingly, when b is a positive integer, a Gamma distribution with parameter b can be viewed as a sum of $b + 1$ independent exponential distributions used e.g. to predict waiting times in queue systems. To our knowledge, this distribution has not been used before to describe word length frequency, although it has been applied to the length of morphs (Creutz 2003).

We adopted the following, heuristic approach in finding optimal values for the *a, b* and *c* parameters. We preliminarily tried out several combinations of *a, b* and *c* and compared the observed frequency distribution with the predicted distribution obtained with each combination. The criterion for comparing the two distributions was that the sum of the squared differences between observed and predicted values was as small as possible. This initial step resulted in several combinations that gave seemingly good approximations of the observed values, and several combinations that obviously did not. The best results were obtained with *a* varying between 1 and 100, *b* between 1 and 10, and *c* between 0 and 1. In the following step, we used this information and systematically tried out all possible combinations between these limits. For *a*, we used 99 values ranging from 1, 2, 3 … 99. These were combined with 19 *b* values ranging from 1, 1.5, 2, 2.5 … 10; and 99 *c* values ranging from 0.01, 0.02, 0.03 … 0.99. In total, this resulted in 186,219 combinations. The predictive value of the formula (1) with values suitable for Swedish and English is shown in the Table 2.

As can be seen the approximations by this formula fit the English and Swedish data fairly well. The correlation coefficients between the observed and the predicted values are 0.978 for English, and 0.939 for Swedish. Figure 1 shows the curves for observed and predicted values.

5. Frequency and word length defined by number of phonemes

We applied the same principle to a word frequency distribution in which word length was expressed by the number of phonemes rather than the number of letters. There is no one-to-one correspondence between phonemes and letters in Swedish. A considerable number of phonemes (mainly consonants) are orthographically represented by varying numbers of letters. Many long consonants are written with double letters. The plosive /k/ for instance, can be written with one (e.g., *k, c*) or two (e.g., *ck, ch*) letters. Similarly, the Swedish phoneme /ɧ/ can be written with one (e.g., *j*), two (e.g., *sk*) or three (e.g., *stj*) letters. On the other hand, a few letters are pronounced as two phonemes (e.g., *x* which is usually pronounced as /ks/).

We used an online version of a current pronunciation dictionary of Swedish (Hedelin 1997) and a word frequency list consisting of 34,090

Table 2: Observed and predicted word frequencies (%) on the basis of word length in letters (L) in English and Swedish. The formula for the predicted values in both languages is $f_{exp} = 11.74 * L^3 * 0.4^L$

Word Length	English		Swedish	
	Observed	Predicted	Observed	Predicted
1	3.160	4.696	3.371	4.696
2	16.975	15.027	11.953	15.027
3	21.192	20.287	24.654	20.287
4	15.678	19.235	10.953	19.235
5	10.852	15.027	11.099	15.027
6	8.524	10.387	9.225	10.387
7	7.724	6.598	6.227	6.598
8	5.623	3.939	5.442	3.939
9	4.032	2.244	4.495	2.244
10	2.766	1.231	3.738	1.231
11	1.582	0.655	2.610	0.655
12	0.917	0.340	1.898	0.340
13	0.483	0.173	1.245	0.173
14	0.262	0.086	0.889	0.086
15	0.099	0.043	0.646	0.043
16	0.050	0.021	0.459	0.021
17	0.027	0.010	0.333	0.010
18	0.022	0.005	0.253	0.005
19	0.011	0.002	0.168	0.002
20	0.006	0.001	0.118	0.001
21	0.005	0.000	0.088	0.000
22	0.002	0.000	0.053	0.000
23	0.001	0.000	0.035	0.000
24	0.001	0.000	0.018	0.000
25	0.001	0.000	0.011	0.000
26	0.001	0.000	0.007	0.000
27	0.001	0.000	0.004	0.000
28	0.000	0.000	0.002	0.000
29	0.000	0.000	0.001	0.000
30	0.000	0.000	0.000	0.000
31			0.001	0.000
32	0.000	0.000		
33	0.000	0.000	0.000	0.000
34	0.000	0.000	0.000	0.000
36	0.000	0.000		
37	0.000	0.000		
38	0.000	0.000		
41	0.000	0.000		
44	0.000	0.000		
59			0.000	0.000
61			0.000	0.000

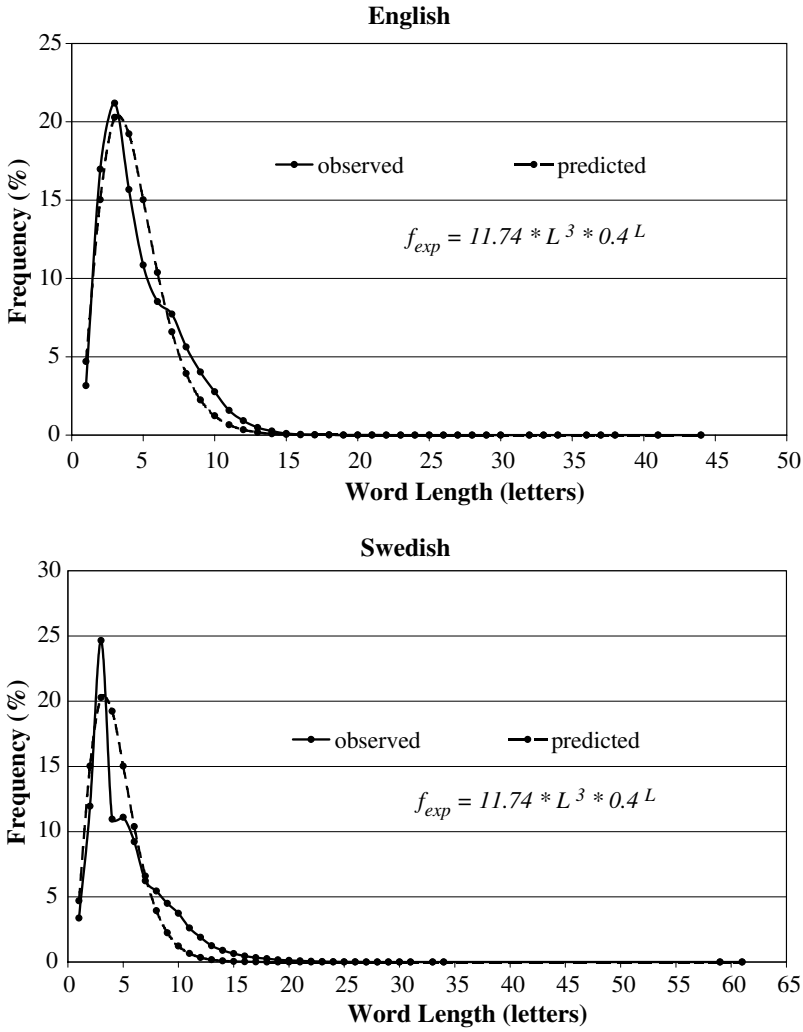

Figure 1: Observed and predicted word frequencies (%) on the basis of word length in letters (*L*) in English (top) and Swedish (bottom).

word types representing nearly 87 million word tokens. The frequency list was based upon a corpus of newspaper texts provided by the Department of Linguistics of Gothenburg University. Using formula (1) on these data yields the results in Table 3 and the curves shown in Figure 2.

The top frequency (37) is greater for the phonetic forms of words and the values in the formula for approximation have to be changed somewhat. Figure 2 below shows the similarities of the curves. The

Table 3: Observed and predicted word frequencies (%) on the basis of word length in phonemes (L) in Swedish. The formula for the predicted values is $f_{exp} = 21.32 * L^4 * 0.25^L$

Word Length	Observed frequency	Predicted frequency
1	8.098	5.330
2	25.660	21.320
3	37.021	26.983
4	11.401	21.320
5	8.222	13.013
6	3.436	6.746
7	2.316	3.124
8	1.529	1.333
9	0.937	0.534
10	0.622	0.203
11	0.330	0.074
12	0.152	0.026
13	0.103	0.009
14	0.069	0.003
15	0.050	0.001
16	0.023	0.000
17	0.016	0.000
18	0.008	0.000
19	0.003	0.000
20	0.002	0.000
21	0.001	0.000
22	0.001	0.000
23	0.000	0.000

correlation coefficient between the observed and the predicted values is equal to 0.929.

6. Irregularity in the slope

Note that in the slope of the Swedish distribution there is an irregularity in that five-letter words are unexpectedly frequent. The same irregularity is also found in the English distribution, although it is less marked, and shifted towards the right. A potential explanation of this irregularity is that the curve is the combination of two separate distributions, one for the open-class words and one for the closed-class (grammatical) words. Perhaps, the open-class curve has its top further to the right and is less steep than the closed-class curve, resulting in an extra top at five letter words for the overall curve. In order to examine this explanation, we separated the Swedish open-class words from the closed-class words (prepositions, pronouns, conjunctions, articles), and

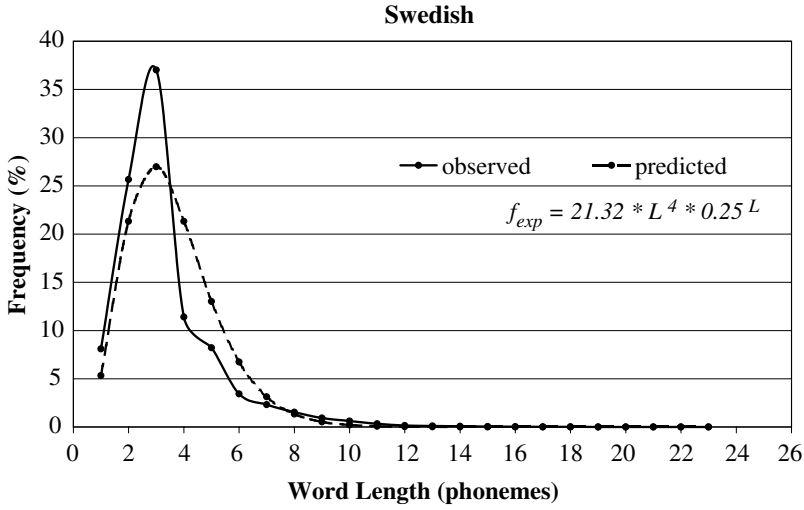

Figure 2: Observed and predicted word frequencies (%) on the basis of word length in phonemes (*L*) in Swedish.

calculated the observed and the expected frequencies for the two word classes separately. The result is shown in Figure 3.

Figure 3 shows that the observed irregularity continues to exist in the frequency distribution of the closed-class words, but has more or less disappeared in the frequency distribution of the open-class words. Our proposed explanation is therefore not entirely correct, and it seems likely that other factors play a role. An alternative explanation is that many more words can be constructed with five phonemes than with four phonemes, so that the type frequency and consequently the token frequency within that group becomes much larger.

7. Zipf's law for German based on word length defined by the number of syllables

Zipf also gives German, Chinese and Latin data for (written) syllables. He mentions Kaeding's investigations of German (Kaeding 1897–98). The percentages given by Kaeding are shown in Table 4. The values are geometrically distributed. For every syllable, the percentage is multiplied by a factor of 0.5. The values can be predicted by the exponential formula

$$f_{exp} = 100 * 0.5^S. \quad (3)$$

in which *S* is the number of syllables. The predicted values are given in Table 4 as well.

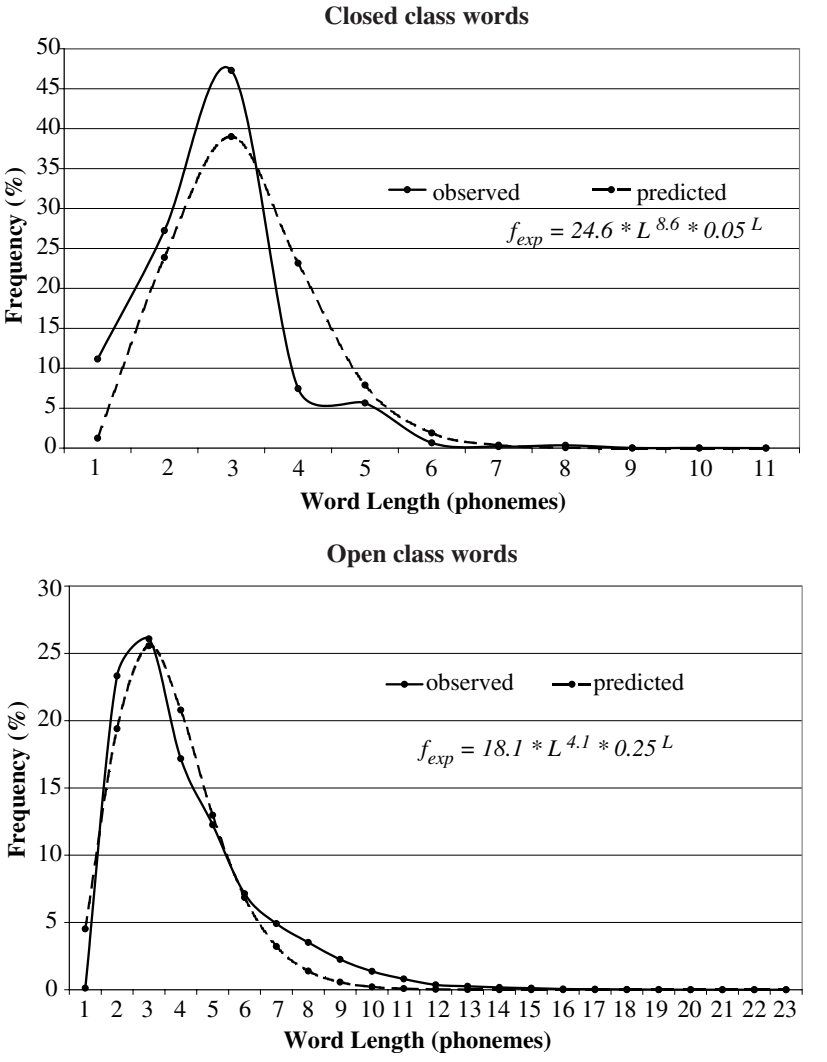

Figure 3: Observed and predicted word frequencies (%) on the basis of word length in letters (*L*) in Swedish for closed-class words (top) and open-class words (bottom).

Clearly Zipf's law holds for the German data if words are defined by the number of syllables. Tentative investigations show that the formula holds well for English and Swedish word length defined by number of syllables too.

Table 4: Observed and predicted word frequencies (%) on the basis of word length in syllables (S) in German. The formula for the predicted values is $f_{exp} = 100 * 0.5^S$

Word Length	Observed frequency	Predicted frequency
1	49.76	50
2	28.94	25
3	12.93	12.5
4	5.93	6.25
5	1.72	3.1
6	0.5	1.5
7-15	0.22	1.6

8. Sentence length and frequency

Sentence length defined by the number of words included in the sentence has attracted much interest in linguistic and literary studies. Usually the average length is mentioned, sometimes with values for standard deviations although the distribution is not generally a standard distribution, see e.g. Kučera & Francis (1967), Loman & Jörgensen (1971).

It is interesting to see how well the formula (1) fits the frequency of sentences of different length. The arguments for using the formula are similar to the arguments when using it for word length. More sentences can be constructed with many words than with few words which naturally increases frequency. On the other hand longer sentences are more inconvenient than short sentences. They take longer time and more effort to pronounce, write, read or interpret. Table 5, reproduced from Kučera & Francis (1967:380, 381), shows the frequency of sentences of different length defined in the entire Brown corpus. The figures to the right are predictions according to the formula: $f_{exp} = 1.1 * L^1 * 0.90^L$. The correlation coefficient between the observed and the predicted values is 0.992. Figure 4 shows the curves for the observed and the predicted frequencies.

Kučera & Francis (1967) include tables and graphs which indicate the variation between different genres. Table 5 is based on their figures for Mystery and Detective stories. Figure 5 shows the corresponding curves. It has the typical skewed profile and can be approximated well by the function: $f_{exp} = 3 * L^1 * 0.85^L$. The parameters ($a = 3$; $b = 1$; $c = 0.85$) are slightly different from the values for the whole text. But the correlation coefficient between the observed and the predicted values is again very high: 0.993. We expect that our approach will be of value in stylistic studies.

Table 5: Observed and predicted sentence frequencies (%) on the basis of sentence length in words (L) in English. The formula for the predicted values is $f_{exp} = 1.1 * L^1 * 0.90^L$

Sentence length	Observed frequency	Predicted frequency
1	0.806	0.990
2	1.370	1.782
3	1.862	2.406
4	2.547	2.887
5	3.043	3.248
6	3.189	3.508
7	3.516	3.683
8	3.545	3.788
9	3.286	3.835
10	3.533	3.835
11	3.562	3.797
12	3.788	3.728
13	3.669	3.635
14	3.751	3.523
15	3.518	3.397
16	3.541	3.261
17	3.434	3.119
18	3.305	2.972
19	3.229	2.823
20	3.103	2.675
21	2.867	2.528
22	2.724	2.383
23	2.647	2.242
24	2.526	2.106
25	2.086	1.974
26	2.178	1.848
27	2.128	1.727
28	1.801	1.612
29	1.690	1.503
30	1.556	1.399
31	1.512	1.301
32	1.326	1.209
33	1.277	1.122
34	1.062	1.040
35	1.051	0.964
36	0.901	0.892
37	0.838	0.825
38	0.764	0.763
39	0.683	0.705
40	0.589	0.650
41	0.624	0.600

Table 5: Continued

Sentence length	Observed frequency	Predicted frequency
42	0.488	0.553
43	0.477	0.510
44	0.406	0.469
45	0.390	0.432
46	0.350	0.397
47	0.318	0.366
48	0.241	0.336
49	0.224	0.309
50	0.220	0.283
51	0.262	0.260
52	0.207	0.239
53	0.174	0.219
54	0.174	0.201
55	0.128	0.184
56	0.121	0.169
57	0.103	0.155
58	0.117	0.142
59	0.124	0.130
60	0.082	0.119
61	0.088	0.109
62	0.061	0.099
63	0.061	0.091
64	0.075	0.083
65	0.063	0.076
66	0.056	0.069
67	0.052	0.063
68	0.057	0.058
69	0.031	0.053
70	0.029	0.048
71	0.021	0.044
72	0.017	0.040
73	0.021	0.037
74	0.034	0.033
75	0.031	0.031
76	0.011	0.028
77	0.011	0.025
78	0.008	0.023
79	0.006	0.021
80-240	0.232	0.213

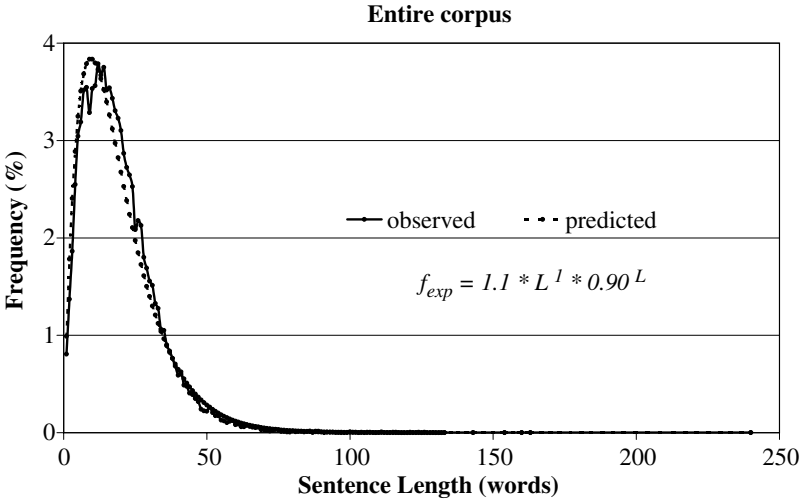

Figure 4: Observed and predicted sentence frequencies on the basis of sentence length in words (L) in the entire Brown corpus.

9. Conclusions

Zipf's ideas about the economy of language are very suggestive and they are confirmed by our data to a certain extent. For word length defined by number of letters or phonemes only the words with more than 3 letters follow Zipf's simplest principle in English and Swedish. A more complicated formula (similar to the Gamma distribution) is needed for the whole distribution as the shortest words of one or two letters do not have the high frequency expected by Zipf's law of abbreviation (as for the relation between abbreviation and frequency, cf. Sigurd 1978).

The formula suggested in this paper takes the potential of many letters/sounds in longer words into account, but also the increasing effort needed in longer words. It has a certain explanatory value. The same formula can probably be used to approximate the length of words defined by the number of morphemes included. But morphemic analysis is not a straightforward matter, so we refrain from that project here.

We also suggest the formula for approximation of the frequencies of sentences of different length. The reason for applying it is that more sentences can be constructed with many than with few words which naturally increases frequency, but on the other hand long sentences are more inconvenient and less economic. The paper shows that the frequencies of language units and their structures seem to be a compromise between the desire to have many alternatives of expression

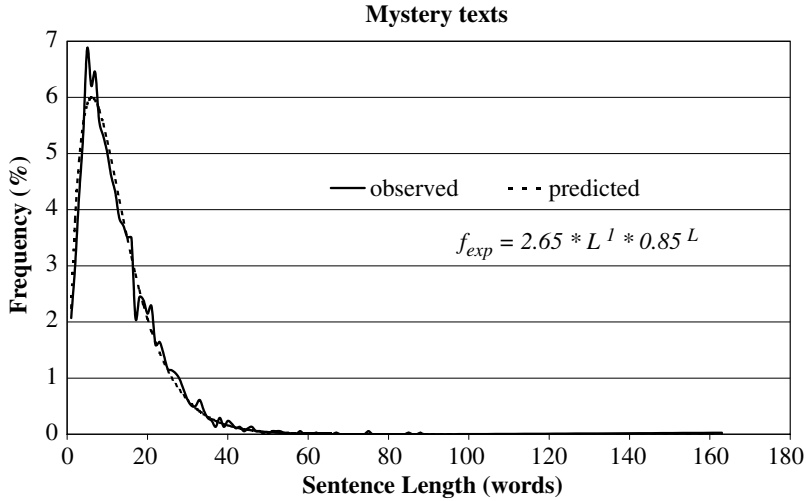

Figure 5: Observed and predicted sentence frequencies on the basis of sentence length in words (L) in the Mystery stories sample in the Brown corpus.

and the desire to minimize the effort of producing and interpreting them. This should be in line with Zipf's ideas.

References

Allén, S. 1970. *Nusvensk frekvensordbok*. Stockholm: Almqvist & Wiksell.

Baddeley, A. 1999. *Essentials of human memory*. Sussex: Psychology Press.

Best, K.-H. 1996. Word length in Old Icelandic songs and prose texts. *Journal of Quantitative Linguistics* 3, 97–105.

Creutz, M. 2003. Unsupervised segmentation of words using prior distributions of morph length and frequency. *Proceedings of ACL-03, The 41st Annual Meeting of the Association of Computational Linguistics*, 280–287, Sapporo, Japan, 7–12 July.

Hedelin, P. 1997. *Norstedts svenska uttalslexikon [Norstedt's Swedish pronunciation dictionary]*. Norstedts Ordbok.

Kaeding, F. W. 1897–98. *Häufigkeitswörterbuch der deutschen Sprache*. Steglitz bei Berlin: Selbstverlag des Herausgebers.

Kučera H. & Francis, W. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Loman, B. & Jörgensen, N. 1971. *Manual för beskrivning av makrosyntagmer*. Lund: Studentlitteratur.

Miller, G. A. 1951. *Language and communication*. New York: McGraw-Hill.

Sigurd, B. 1978. Arbitrariness, frequency and abbreviation. *Studia Linguistica* 32, 169–173.

Sigurd, B. 1965. *Phonotactic structures in Swedish*. Lund: Uniskol.

Strömqvist, S., Johansson, V., Kriz, S., Ragnarsdóttir, H., Aisenman, R. & Ravid, D. 2002. Towards a crosslinguistic comparison of lexical quanta in speech and writing. *Written Language and Literacy* 5, 45–68.

Wilson, A. & McEnery, T. 1998. Word length distribution in biblical and medieval Latin. *The Prague Bulletin of Mathematical Linguistics 70*. Prague: Charles University.

Wimmer, G., Köhler, K., Grotjahn, R. & Altmann, G. 1994. Towards a theory of word length distribution. *Journal of Quantitative Linguistics* 1, 98–106.

Zipf, G. K. 1935. (reprinted 1965). *The psycho-biology of language*. Cambridge MA: MIT Press.

Received June 10, 2003
Accepted November 17, 2003

Bengt Sigurd, Mats Eeg-Olofsson
& Joost van de Weijer
Department of Linguistics and Phonetics
Lund University
Helgonabacken 12
S-22362 Lund
Sweden
Bengt.Sigurd@ling.lu.se
Mats.Eeg-Olofsson@ling.lu.se
vdweijer@ling.lu.se