Glottometrics 5, 2002

To Honor G.K. Zipf

RAM - Verlag

Glottometrics 5, 2002, 97-107 *To honor G.K. Zipf*

Oscillation in the frequency-length relationship

Peter Grzybek, Graz[1]
Gabriel Altmann, Lüdenscheid

Abstract. The analysis shows that there is no intrinsic oscillation in the relation between frequency and length of words. The rise of oscillation is caused by using moving averages for smoothing the extremely dispersed data.

Keywords: Frequency-length relation, oscillation

The relationship between the frequency of a word and its length has repeatedly been the object of linguistics studies since Zipf's (1932, 1935) corresponding statements. Subsequent to his hypothesis, stating that the length of a word stands in an inverse relationship to its frequency, many studies have analyzed this problem, based either on texts (or parts of texts), frequency lists, or corpus analyses.

Different models have been suggested to formally describe this particular relationship. In his comprehensive study on the German LIMAS corpus, composed of ca. 500 texts (or parts of texts, respectively), and comprising about one million words, Köhler (1986) tested if the so-called power law, implying the relationship $y = ax^{-b}$, is apt to adequately describe the dependency of word length on word frequency.

As a result of his statistical analysis, Köhler (1986: 137) concluded that his initial hypothesis must not be rejected; in a follow-up study by Zörnig, Köhler, and Brinkmöller (1990: 25), the authors repeat this interpretation, speaking of "a highly positive result". Their interpretation was based on an analysis of variance, yielding $F_{1,158} = 105$ ($p < 0.001$), ($F_{0.01} = 6.8$). Meanwhile, however, it is a well-known fact, that as the sample size increases, the F-test is problematic to reliably interpret results achieved by it. Therefore, it seems to be more reasonable to (at least additionally) evaluate the goodness of regression models by reference to the determination coefficient R^2, although the latter, too, of course, is not free of deficiencies (cf. Grotjahn 1992).

In fact, a re-analysis of Köhler's data shows that his initial fit is far from being good, which is corroborated by the determination coefficient of $R^2 = 0.24$. This poor fit is clearly illustrated in Fig. 1, taken from Köhler (1986: 138).

As a matter of fact, Köhler and his co-authors noticed that, irrespective of the positive F-value, the fit of the exponential function was far from being satisfying. Therefore Köhler (1986: 137) concluded that, given the deviations of the empirical data from the theoretical curve are random, it should be possible to arrive at better results, if one smoothes the data by way of moving averages.

[1] Address correspondence to: Peter Grzybek, Institut für Slawistik, Universität Graz, Merangasse. 70, A-8010 Graz. E-mail: grzybek@uni-graz.at

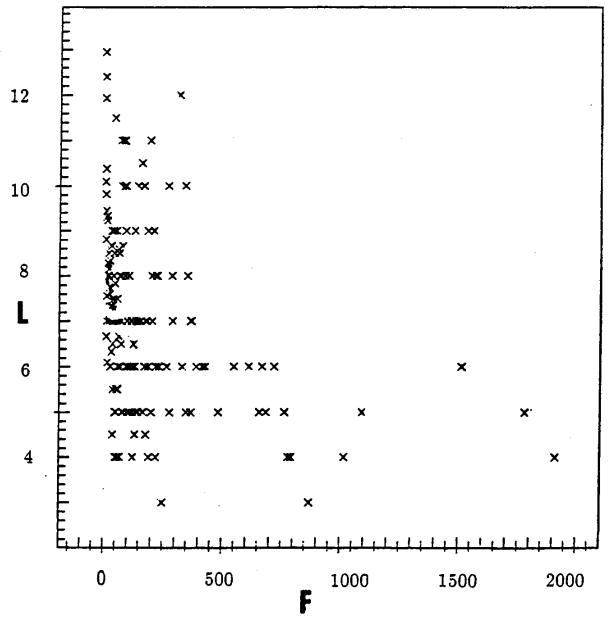

Fig. 1. Corpus data representing the dependence of word length (L) on word frequency (F); cf. Köhler (1986: 138)

Köhler did not systematically pursue this question, but re-analyzing his data, one can indeed show that with an increase of the intervals, the fit of the exponential function becomes stepwise better:

smoothing interval	R^2
none	0.24
20	0.54
50	0.77
100	0.92

Smoothing by way of moving averages, thus seems to be an effective procedure. However, Köhler was not so much interested in the fact of the gradually improving fit, as he was surprised by an oscillating curve around the theoretical hyperbolic function line: This is to say that, after using moving averages with intervals of 20, 50 and 100, a peculiar oscillation appeared which seemed to be very regular (cf. Fig. 2).

Köhler (1986) himself and, in the subsequent detail study devoted to this particular problem, Zörnig, Köhler and Brinkmöller (1990), tried to capture the course of the data by adding further power components. They thus first obtained

(1) $L = aF^b + cF^d$

and then the rather complex function

(2) $L = aF^b + cF^d + ke^{m(F-F_0)}\sin(\alpha F)$,

which captured the oscillation in a convincing way, as can be seen in Fig. 3.

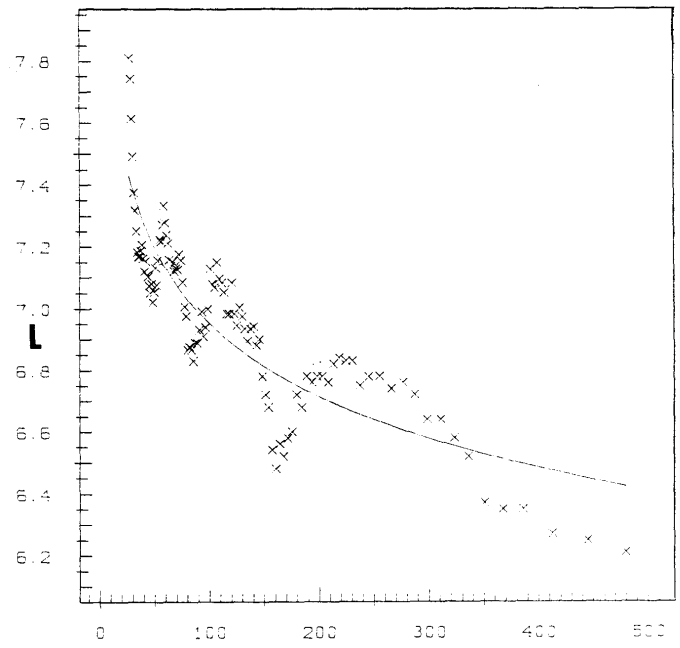

Fig. 2. Smoothing the above data (cf. Fig 1) by moving averages in intervals of 50; cf. Köhler (1986: 141)

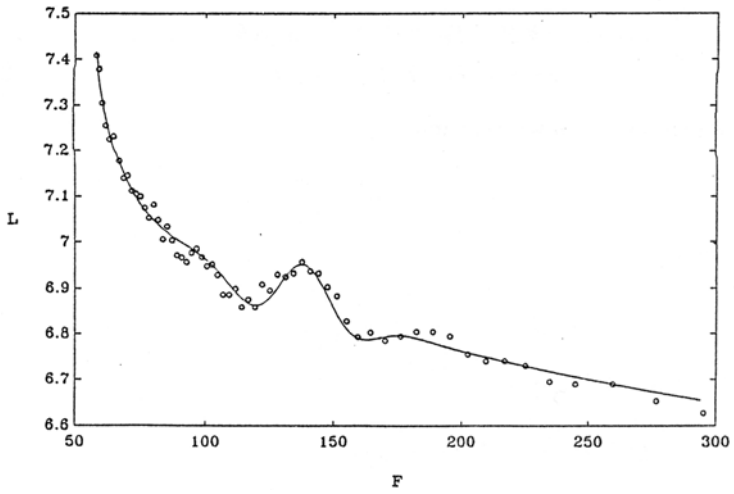

Fig. 3. Observed and computed mean lengths (Zörnig et al. 1990: 37)

What remained open, however, was a linguistic interpretation of this phenomenon, which Zörnig, Köhler, and Brinkmöller (1990: 39) left for "future research". Taking into account the complexity of formula (2), it is not really astonishing that this problem has remained unsolved until today.

In a recent study on the dependence of word length on word frequency, Strauss, Grzybek, and Altmann (2003) have examined individual texts, comparing the results to those obtained on the basis of text mixtures. In the tail of the dependence, there were many frequency classes without records, and the recorded ones contained a very small number of cases (mostly 1), thus causing a strong dispersion. Instead of smoothing the data by moving averages, they pooled low-frequency classes in order to obtain more stabile data. By pooling the data in such

a way that each frequency class contained at least 10 records, the authors obtained an unequivocal corroboration of the relationship in all cases (for texts from 10 different languages)

$$L = aF^{-b} + 1 \tag{3}$$

where L = mean length, F = frequency, a and b are coefficients, and the constant 1 is the asymptote of the function (since word length was measured in terms of syllable numbers). Occurring non-syllabic words (such as, e.g., the Russian prepositions *к, с, в*), were considered as proclitics. It was not necessary to take oscillation into account, the fitting quite obviously displayed random residuals.

Irrespective of the satisfying results, it is just the observation of the lack of oscillation which again rises the question of its presence in Köhler's study; retrospectively, the problem pointed out by him remains unanswered till today, and it is not clear whether oscillation arose

(a) due to data mixing, ultimately inherent in any corpus, or
(b) as a result of an increasing sample size, or
(c) whether it was an attribute of the specific data.[2]

In the present study, an attempt shall be undertaken to offer a reason for the rise of oscillation. For our purposes, and by way of a working definition, oscillation can be assumed to be present, if the sequences of neighbouring observed data cross the theoretical curve either too frequently or too rarely. There is an interval within which the number of crossings – or the number of runs above and below the curve – can be considered to be random. We are not concerned with a time series, here, but with a sequence of numbers, capturing the length-frequency relations of words, which are ordered according to their increasing frequency. Since only low frequencies have a sufficient number of records, while higher frequencies have either none or very few records, the results for higher frequencies are insufficiently representative – therefore, great fluctuation is to be expected in this domain.

Fig. 4 illustrates the frequency-length dependence for the complete text of Puškin's verse novel *Evgenij Onegin*. Interpreting the curve, one can say that, generally speaking, high-frequency words tend to be shorter. This tendency holds true only on the average, however: whereas the curve is relatively regular at the beginning, one can observe rather irregular instabilities with the higher frequencies.

The reason for these instabilities is most likely the fact that in this particular part at the curve's tail, individual frequent words tend to have not more but one, two or three syllables. Since these words represent frequency classes in which they may occur alone, great dispersion results are to be expected exactly here. In other words: most probably, the greater dispersion is likely to be due to the insufficient number of records for these data points. Generalizing this observation, it would seem reasonable to assume that the exact part of the curve, *where* the dispersion becomes greater, depends on the distances between the represented classes, or on the frequencies of these classes (or both), which, in turn, might be related to text length (or corpus structure).

The problem at stake is obvious: if one wants to prove the validity of the law postulated by Zipf, one has three options in dealing with the data and their theoretical modelling:

[2] In another follow-up-study, Hammerl (1990) tested Polish corpus data with regard to the previously observed phenomenon of oscillation, but he did not find any.

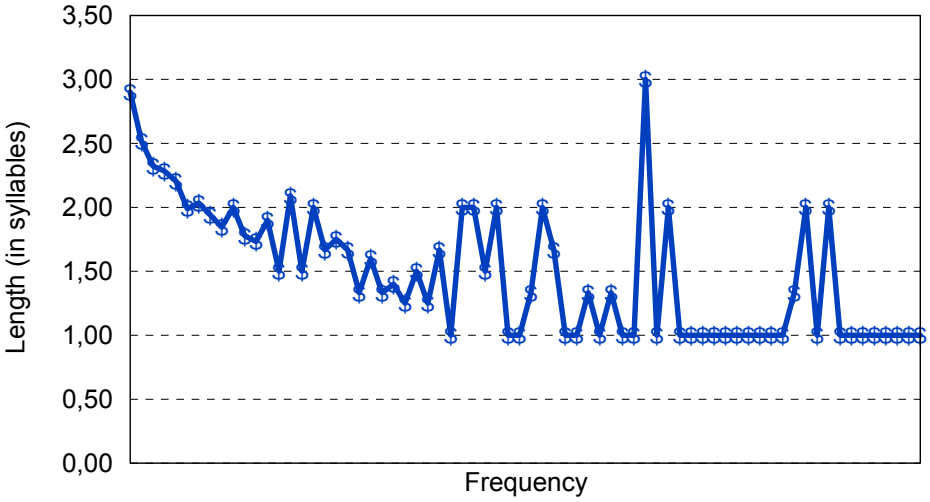

Fig. 4. The frequency-length relationship in *Evgenij Onegin*

(a) one tries to derive a curve capturing this chaotic movement, or
(b) one smoothes the data to obtain a plain course, or
(c) one smoothes the data and tries to capture also such complications (such as oscillation) which may arise during the process of smoothing.

It is obvious, that no one would ever try to go the first way, because data of this kind are extremely dispersing. Therefore, some kind of smoothing is necessary, since fitting curve (3) to the empirical data; without any smoothing, yields a very poor result (in this case $a = 2.1682$, $b = -0{,}4524$; $R^2 = 0.49$).

Now, as to the concrete manner of smoothing, different options are available: whereas Strauss, Grzybek and Altmann (2003) pooled the data and corroborated (3) in every case, another way was chosen by Köhler (1986) and Zörnig, Köhler, Brinkmöller (1990), who used moving averages and obtained the oscillating curve described above; ultimately succeeded in modelling this oscillation (see above), they had to leave open the question of its rise.

It seems most reasonable that, in one way or another, the concrete manner of smoothing is related to the phenomenon; this is not to say that oscillation necessarily is a consequence of smoothing by moving averages; yet, this might be the case in combination with a particular data structure. In an attempt to test this assumption, we will try to reproduce Köhler's finding for Puškin's *Evgenij Onegin*, and to "artificially" generate the oscillating phenomenon.

Let us start by replicating the smoothing method applied by Strauss, Grzybek, and Altmann (2003). This is to say, we first have to compute the mean length of words occurring exactly x times. For the sake of data homogeneity, we will initially concentrate on the first chapter, only; the values thus obtained are represented in Table 1 (see below). We then have to pool the data as described above, i.e. in such a way that each frequency class is based on at least ten records;[3] the resulting values of this pooling procedure are represented in Table 1.

As can be seen from Table 1, smoothing by way of pooling the classes, as described above, yields a very good result of $R^2 = 0.96$, which is graphically represented in Fig. 5.

[3] If the present results slightly differ from those presented by Strauss, Grzybek, and Altmann (2003), the reason for this is, first, that classes are pooled here "bottom-up", whereas they were pooled "top-down" in the article mentioned; and second, that the means obtained here are weighted means both for frequency and length, whereas unweighted means were calculated in the previous study.

Table 1
Fitting (3) to the data of *Evgenij Onegin*: smoothing by pooling

F	*L*	*L**
1	2.6595	2.7123
2	2.1256	1.9894
3	1.7800	1.7178
4	1.4800	1.5716
5	1.3750	1.4791
6.46	1.5385	1.3911
9.40	1.1333	1.2907
15.10	1.2000	1.1998
45.50	1.1000	1.0835
a = 1.7123, b = -0.7914, R^2 = 0.96		

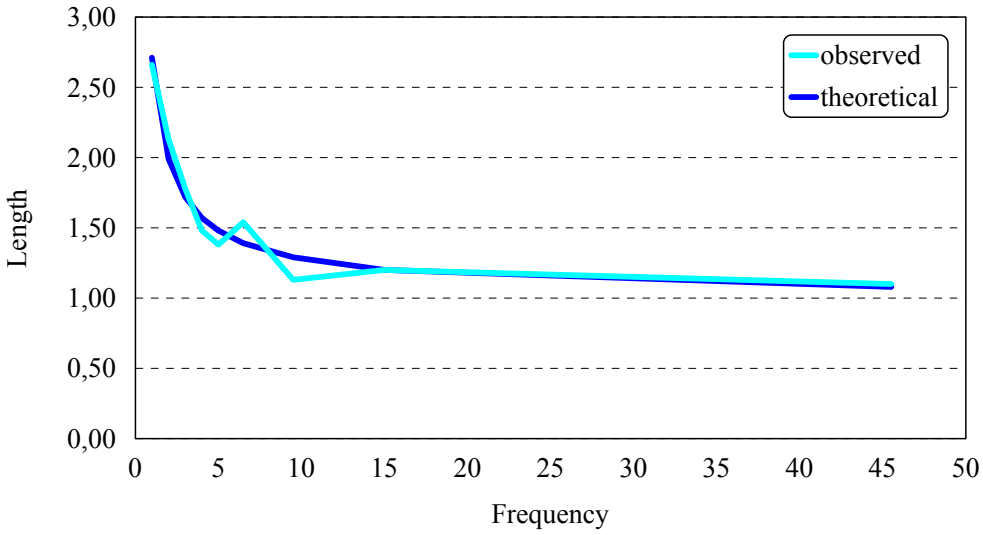

Fig. 5. Observed and computed mean lengths in *Evgenij Onegin* (ch. I)

Based on these procedures of the previous work, we may now focus the question of oscillation and extend our ruminations. Theoretically speaking, if a fitting is satisfactory, then not only small sums of squared deviations should be attained, but additionally, the empirical values should display random fluctuations around the theoretical curve. In other words: in this case, there must be neither too many nor too few runs of values on both sides of the curve. If this should still be the case, then the data either contain an intrinsic oscillation (if there are too many runs) or they display a slow wavelike motion (if there are too few runs). The oscillation must be caught by superposed curves since a simple curve cannot capture it adequately. However, if the wavelike motion arises by manipulation of data, it is not real and a simple curve is sufficient to capture it.

We can easily test this by applying the theory of runs (cf. Grotjahn 1979: 143ff., 1980). The basic idea, here, would be to test the number of sequences above and below the theoretical curve; in our case, it would be sufficient to know, if there are two few sequences (runs). In order to test this statistically, we need

n_1 number of data points above the theoretical curve (+)
n_2 number of data points below the theoretical curve (–)
r_1 number of (+)-sequences
r_2 number of (–)-sequences
$n = n_1 + n_2$
$r = r_1 + r_2$

Since we are interested in the question if there are too few runs, we test the one-sided hypothesis. As is well known, the approximation to the normal distribution may be used, for larger n ($n > 30$). Since the number of runs does not exceed 30, however, we have to calculate the exact (cumulated) probabilities which can also be taken from existing tables.

In order to show the rise of the wave, we first test the number of runs based on the data given in Table 2.

Table 2
Fitting (3) to the raw data in *Evgenij Onegin* (Chapter I)

F	*L*	*L**		*F*	*L*	*L**		*F*	*L*	*L**	
1	2.66	2.70	–	10	1	1.28	–	21	2	1.16	+
2	2.13	1.99	+	11	1	1.26	–	24	1	1.14	–
3	1.78	1.72	+	12	3	1.24	+	25	1	1.14	
4	1.42	1.57	–	13	1	1.23	–	32	1	1.11	
5	1.29	1.48	–	14	1	1.21	–	38	1	1.10	
6	1.43	1.42	+	15	1	1.20	–	45	1	1.09	
7	1.43	1.37	+	17	1	1.18	–	49	1	1.08	
8	1.17	1.33	–	19	1	1.17	–	68	1	1.06	
9	1.00	1.30	–	20	1	1.16	–	155	1	1.03	
$a = -0.7846$, $b = 1.6977$, $R^2 = 0.43$											

According to the description above, the positive deviations, i.e., those values which lie above the theoretical curve, are marked by (+), the negative ones, i.e., those that lie below it, by (–). Since in the whole tail, the theoretical curve lies above the empirical data, we cut off the both after the first negative sign of the last run; the number of runs thus remains constant, but the number of elements decreases thus requiring more extreme test results. As can be seen, we have

$n_1 = 6$	(6 times "+")	$r_1 = 4$	(4 runs of "+")
$n_2 = 14$	(14 times "–")	$r_2 = 5$	(5 runs of "–")
$n = n_1 + n_2 = 20$		$r_1 + r_2 = 9$	

Since the number of runs is relatively small ($r < 30$), we have to calculate the exact cumulative probability, and we thus obtain $P(R \leq r) = 0.5204$, which is not significant, of course: this is to say that the number of runs does not differ from the expected one.

Now, let us smooth the data using moving averages with larger intervals. Table 3, represents the results of smoothing with moving averages on the basis of different intervals. In the following tables interval 1 means no smoothing.

Table 3
Building moving averages and testing the runs

Interval of the moving average	n_1	n_2	n	r_1	r_2	r	$P(R < r)$
1	14	6	20	5	4	9	0.5204
2	15	5	20	4	3	7	0.2722
3	14	6	20	3	2	5	0.0173
4	12	8	20	3	2	5	0.0063

It can clearly be seen that, with an increase of the intervals, the probability for oscillation to come into play soon rises. Figure 6 convincingly illustrates this tendency, juxtaposing the results for both manners of smoothing for the sake of comparison.

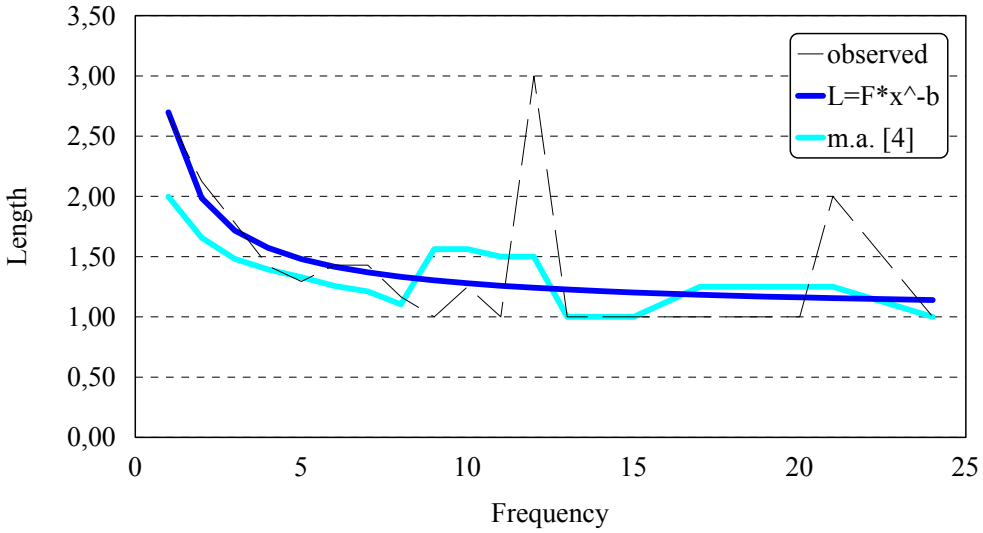

Fig. 6. Observed and computed mean lengths in *Evgenij Onegin* (ch. I)

For the sake of generalisation, let us finally extend this procedure to a broader text basis. Table 4 represents the results for each of the eight chapters of *Evgenij Onegin;* in order to eventually compare the exact results to those obtained by approximation to the normal distribution, both values are presented in parallel. The comparison of the smoothed values with the theoretical curve yields the following results (see Table 4):

Table 4
Test for the number of runs with different smoothing intervals
(*Evgenij Onegin*, chs. I–VIII)

	N	Interval	n_1	n_2	*n*	r_1	r_2	*r*	$P(R < r)$
EO 1	3086	1	15	5	20	5	4	9	0.7417
		2	14	6	20	4	3	7	0.1514
		3	14	6	20	3	2	5	0.0173
EO 2	2235	1	12	7	19	5	4	9	0.4276
		2	12	7	19	5	4	9	0.4276
		3	13	6	19	3	2	5	0.0217
		4	11	8	19	2	1	3	0.0003
EO 3	2702	1	9	7	16	3	2	5	0.0350
		2	10	6	16	3	2	5	0.0470
		3	12	4	16	2	1	3	0.0088
EO 4	2441	1	7	6	13	3	2	5	0.1212
		2	7	6	13	3	2	5	0.1212
		3	8	5	13	3	2	5	0.1515
		4	10	3	13	2	1	3	0.0455
EO 5	2310	1	10	6	16	6	5	11	0.9580
		2	10	6	16	3	2	5	0.0470
		3	9	7	16	3	2	5	0.0350
		4	11	5	16	3	2	5	0.0769
		5	11	5	16	2	1	3	0.0037
EO 6	2471	1	10	5	15	5	4	9	0.8741
		2	9	6	15	4	3	7	0.3427
		3	10	5	15	4	3	7	0.4545
		4	10	5	15	3	2	5	0.0949
		5	9	5	14	2	1	3	0.0070
EO 7	2922	1	14	8	22	6	5	11	0.5573
		2	14	8	22	4	3	7	0.0408
		3	13	9	22	4	3	7	0.0294
		4	15	7	22	3	2	5	0.0055
EO 8	3217	1	14	10	24	5	4	9	0.0857
		2	14	10	24	5	4	9	0.0857
		3	17	7	24	4	3	7	0.0450
		4	19	5	24	3	2	5	0.0209

As can clearly be seen, in most cases, intervals of three or four radically change the situation: It is almost self-evident that values of $P < 0.05$ signalize significantly few runs, i.e., the rise of a slow wave motion (with a one-sided hypothesis). As a matter of fact, by prolonging the intervals, one obtains ever longer waves, what need not be demonstrated in detail, here.

Additionally, in order to at least raise the questions of text length or data mixture, Table 5 represents the results for a successive cumulation of chapters I-VIII of *Evgenij Onegin*.

Table 5
Test for the number of runs with different smoothing intervals
(*Evgenij Onegin*, cumulated chs. I–VIII)

	N	Interval	n_1	n_2	n	r_1	r_2	r	$P(R < r)$
EO 1-2	5321	1	17	13	30	10	9	19	0.9238
		2	17	13	30	7	6	13	0.1980
		3	17	13	30	7	6	13	0.1980
		4	18	12	30	5	4	9	0.0106
EO 1-3	8023	1	21	15	36	11	10	21	0.8521
		2	21	15	36	7	6	13	0.0404
		3	25	11	36	6	5	11	0.0306
		4	27	9	36	4	3	7	0.0012
EO 1-4	10464	1	20	22	42	10	9	19	0.2204
		2	22	20	42	7	6	13	0.0036
EO 1-5	12774	1	27	21	48	14	13	27	0.8026
		2	29	19	48	10	9	19	0.0872
		3	31	17	48	7	6	13	0.0013
EO 1-6	15245	1	30	25	55	14	13	27	0.4153
		2	27	28	55	10	9	19	0.0067
EO 1-7	18167	1	35	28	63	10	9	19	0.0005
EO-tot	21401	1	33	30	63	13	12	25	0.0383
		2	32	31	63	6	5	11	0.0000

Again, one clearly sees the impact of smoothing by moving averages on the rise of oscillation; additionally, it can easily be observed that oscillation is more likely to arise for the larger, cumulated samples. If this is due to sample size, or data mixture, or to a combination of both factors, will have to be the topic of a detail study particularly devoted to this problem.

Summarizing we can state that the relation between frequency and length of words does not contain an intrinsic oscillation based on a linguistic cause, as has previously been suspected. It is simply the consequence of a special kind of smoothing.

References

Grotjahn, R. (1979). *Linguistische und statistische Methoden in Metrik und Textwissenschaft*. Bochum: Brockmeyer.

Grotjahn, R. (1980). The theory of runs as an instrument for research in quantitative linguistics. *Glottometrika 2*, *14-43.*

Grotjahn, R. (1992). Evaluating the adequacy of regression models: Some potential pitfalls. *Glottometrika 13*, *121-172.*

Hammerl, R. (1990). Länge – Frequenz, Länge – Rangnummer: Überprüfung von zwei lexikalischen Modellen. *Glottometrika 12, 1-24.*

Köhler, R. (1986). *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum: Brockmeyer.

Strauss, U., Grzybek, P., Altmann, G. (2003). The more the better? Word length and word frequency. [In print]

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, Mass.: Harvard University Press.

Zipf, G.K. (1935). *The psycho-biology of language: An introduction to dynamic philology*. Boston: Houghton Mifflin.

Zörnig, P., Köhler, R., Brinkmöller, R. (1990). Differential equation models for the oscillation of the word length as a function of the frequency. *Glottometrika 12*, *25-40.*

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitschrift für die quantitative Erforschung von Sprache und Text
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden
Glottometrics kann aus dem **Internet** heruntergeladen, auf **CD-ROM** (in PDF Format) oder in **Buchform** bestellt werden

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors
Glottometrics can be downloaded from the **Internet**, obtained on **CD-ROM** (in PDF) or in form of **printed copies**

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an
Orders for CD-ROM´s or printed copies to

RAM-Verlag RAM-Verlag@t-online.de

Herunterladen / Downloading: http://www.ram-verlag.de

Die Deutsche Bibliothek – CIP-Einheitsaufnahme

Glottometrics. –5 (2002) –. – Lüdenscheid: RAM-Verl., 2002
Erscheint unregelmäßig. – Auch im Internet als elektronische Ressource unter der Adresse http://www.ram-verlag.de verfügbar.-
Bibliographische Deskription nach 5 (2002)
ISSN 1617-8351

Contents