Peter Grzybek (ed.): *Contributions to the Science of Text and Language.*
Dordrecht: Springer, 2007, pp. 277–294

WORD LENGTH AND WORD FREQUENCY

Udo Strauss, Peter Grzybek, Gabriel Altmann

1. Stating the Problem

Since the appearance of Zipf's works, (esp. Zipf 1932, 1935), his hypothesis "that the magnitude of words tends, on the whole, to stand in an inverse (not necessarily proportionate) relationship to the number of occurrences" (1935: 25) has been generally accepted. Zipf illustrated the relation between word length and frequency of word occurrence using German data, namely the frequency dictionary of Kaeding (1897–98).

In the past century, Zipf's idea has been repeatedly taken up and examined with regard to specific problems. Surveying the pertinent work associated with this hypothesis, one cannot avoid the impression that there are quite a number of problems which have not been solved to date. Mainly, this seems to be due to the fact that the fundamentals of the different approaches involved have not been systematically scrutinized. Some of these unsolved problems can be captured in the following points:

i. *The direction of dependence.* Zipf himself discussed the relation between length and frequency of a word or word form – which in itself represents an insufficiently clarified problem – only in one direction, namely as the dependence of frequency on length. However, the question is whether frequency depends on length or vice versa. While scholars such as Miller, Newman, & Friedman (1958) favored the first direction, others, as for example, Köhler (1986), Arapov (1988) or Hammerl (1990), preferred the latter. As to a solution of this question, it seems reasonable to assume that it depends on the manner of embedding these variables in Köhler's control cycle.

ii. *Unit of measurement.* While some researchers – as, e.g., Hammerl (1990) – measured word length in terms of syllable numbers, others – as for example Baker (1951) or Miller, Newman & Friedman (1958) – used letters as the basic units to measure word length. Irrespective of the fact that a high correlation between these two units should seem likely be found, a systematic study of this basic pre-condition would be important with regard to different languages and writing systems.

iii. *Rank or frequency.* Again, while some researchers, as e.g., Köhler (1986), based his analysis on the absolute occurrence of words, others, such as Guiraud (1959), Belonogov (1962), Arapov (1988), or Hammerl (1990) who, in fact, examined both alternatives, considered the frequency rank of word forms. In principle, it might turn out to be irrelevant whether one examines the frequency or the rank, as long as the basic dependence remains the same, and one obtains the same function type with different parameters; still, relevant systematic examinations are missing.

iv. *The linguistic data.* A further decisive point is the fact that Zipf and his followers did not concentrate on individual texts, but on corpus data or frequency dictionaries. The general idea behind this approach has been the assumption, that assembling a broader text basis should result in more representative results, reflecting an alleged norm to be discovered by adequate analyses. However, this assumption raises a crucial question, as far as the quality of the data is concerned. Specifically, it is the problem of data homogeneity, which comes into play (cf. Altmann 1992), and it seems most obvious that any corpus, by principle, is inherently inhomogeneous. Moreover, it should be reasonable to assume that oscillations as observed by Köhler (1986), are the outcome of mixing heterogeneous texts: examining the German LIMAS corpus, Köhler (1986) and Zörnig et al. (1990) found not a monotonously decreasing relationship, but an oscillating course. The reason for this has not been found until today; additionally, no oscillation has been discovered in the corpus data examined by Hammerl (1990).

v. *Hypotheses and immanent aspects.* Finally, it should be noted that Zipf's original hypothesis implies four different aspects; these aspects should, theoretically speaking, be held apart, but, in practice, they tend to be intermingled:

(a) *The textual aspect.* Within a given text, longer words tend to be used more rarely, short words more frequently. If word frequency is not taken into account, one obtains the well-known word length distribution. If, however, word frequency is additionally taken into account, then one can either study the dependence of length from frequency, or the two-dimensional length-frequency distribution. Ultimately, the length distribution is a marginal distribution of the two-dimensional one. In general, one accepts the dependence $L = f(F)$ or $L = f(R)$ [L = length, F = frequency, R = rank].

(b) *The lexematic aspect.* The construction of words, i.e. their length in a given lexicon, depends both on the lexicon size in question and on the phoneme inventory, as well as on the frequential load of other polysemic words. Frequency here is a secondary factor, since it does not play any role in the generation of new words, but will only later result from the load of other words. This aspect cannot easily be embedded in the modeling process because the size of the lexicon is merely an empirical constant whose estimation is associated with great difficulties. It can at best play the role of *ceteris paribus*.

(c) Shortening through usage. This aspect, which concerns the shortening of frequently used words or phrases, has nothing to do with word construction or with the usage of words in texts; rather, the process of shortening, or shortening substitution, is concerned (e.g., references → refs).

(d) *The paradigmatic aspect.* The best examined aspect is the frequency of forms in a paradigm where the shorter forms are more frequent than the longer ones, or where the frequent forms are shorter. The results of this research can be found under headings such as 'markedness', 'iconism vs. economy', 'naturalness', etc. (cf. Fenk-Oczlon 1986, 1990, Haiman 1983, Manczak 1980). If the paradigmatic aspect is left apart, aspect (d) becomes a special case of aspect (a).

2. The Theoretical Approach

In this domain, quite a number of adequate and theoretically sound formulae have been proposed and empirically confirmed: more often than not, one has adhered to the "Zipfian relationship" also used in synergetic linguistics (cf. Herdan 1966, Guiter 1974, Köhler 1986, Hammerl 1990; Zörnig et al. 1990): consequently, one has started from a differential equation, in which the relative rate of change of mean word length (y) decreases proportionally to the relative rate of change of the frequency (Köhler 1986).

Since in most languages, zero-syllabic words either do not exist, or can be regarded as clitics, the mean length cannot take a value of less than 1. This is the reason why the corresponding function must have the asymptote 1. Finally, the equations get the form (13.1).

$$\frac{dy}{y-1} = -b\frac{dx}{x} \tag{13.1}$$

from which the well-known formula (13.2)

$$y = a \cdot x^{-b} + 1 \tag{13.2}$$

follows, with $a = e^C$ (C being the integration constant). Here, y is the mean length of words occurring x times in a given text. If one also counts words with length zero, the constant 1 must be eliminated, of course, and as a result, at least some of the values (depending on the number of 0-syllabic words) will be lower.

As compared to other approaches, the hypothesis represented by (13.2) has the advantage that the inverse relation yields the same formula, only with different parameters, i.e.

$$x = A \cdot (y-1)^{-B} \tag{13.3}$$

where $A = a^{1/b}$, $B = 1/b$. This means that the dependence of frequency on length can be captured in the same way as can that of length on frequency, only with transformed parameters.

In the present paper, we want to test hypothesis (13.2). We restrict ourselves exclusively to the textual aspect of the problem, assuming that, in a given text, word length is a variable depending on word frequency. Therefore, we concentrate on testing this relationship with regard to individual texts and not – as is usually done – with regard to corpus or (frequency) dictionary material. Though this kind of examination does not, at first sight, seem to yield new theoretical insights with regard to the original hypothesis itself, the focus on the variable text, which, thus far, has not been systematically studied, promises the clarification of at least some of the above-mentioned problems. Particularly, the phenomenon of oscillation as observed by Köhler (1986), might find an adequate solution when this variable is systematically controlled; yet, this particular issue will have to be the special object of a separate follow-up analysis (cf. Grzybek/Altmann 2003).

For the present study, word length has been counted in terms of the numbers of syllables per word, in order to submit the text under study to as few transformations as possible; further, every word form has been considered as a separate type, i.e., the text has not been lemmatized. Since our main objective is to test the validity of Zipf's approach for individual texts, we have chosen exclusively single texts

a) by different authors,

b) in different languages, and

c) of different text types.

Additionally, attention has been paid to the fact that the definition of 'text' itself possibly influences the results. Pragmatically speaking, a 'text' may easily be defined as the result of a unique production and/or reception process. Still, this rather vague definition allows for a relatively broad spectrum of what a concrete text might look like.

Therefore, we have analyzed 'texts' of rather different profiles, in order to gain a more thorough insight into the homogeneity of the textual entity examined:

i. a complete novel, composed of chapters,
ii. one complete book of a novel, consisting of several chapters,
iii. individual chapters, either (a) as part of a book of a novel, or (b) of a whole novel,
iv. dialogical vs. narrative sequences within a text.

It is immediately evident that our study primarily focuses the problem of homogeneity of data, inhomogeneity being the possible result of mixing various texts, different text types, heterogeneous parts of a complex text, etc. Thus, theoretically speaking, there are two possible kinds of data inhomogeneity:

(a) intertextual inhomogeneity,
(b) intratextual inhomogeneity.

Whereas *intertextual inhomogeneity* thus can be understood as the result of combining ("mixing") different texts, *intratextual inhomogeneity* is due to the fact that a given text in itself does not consist of homogeneous elements. This aspect, which is of utmost importance for any kind of quantitative text analysis, has hardly ever been systematically studied.

In addition to the above-mentioned fact that any text corpus is necessarily characterized by data inhomogeneity, one can now state that there is absolutely no reason to a priori assume that a text (in particular, a long text) is characterized by data homogeneity, per se. The crucial question thus is, under which conditions can we speak of a homogeneous 'text', when do we have to speak of mixed texts, and what may empirical studies contribute to a solution of these question?

3. Text Analyses in Different Languages

The results of our analyses are represented according to the scheme in Table 13.1, which contains exemplary data illustrating the procedure: The first column shows the absolute occurrence frequencies (x); the second, the number of words $f(x)$ with the given frequency x; the third, the mean length $L(x)$ of these words in syllables per word. Length classes were pooled, in case of $f(x) < 10$: in the example, classes $x = 8$ and $x = 9$ were pooled because they contain fewer than 10 cases per class. Since the mean values were not weighted, we obtained the new values $x = (8+9)/2 = 8.5$ and $L(x) = (1.5714 + 1.6667)/2 = 1.62$. This kind of smoothing yields more representative classes. In how far other smoothing procedures can lead to diverging results, will have to be analyzed in a separate study.

The following texts have been used for the analyses:

Table 13.1: An Illustrative Example of Data Pooling

x	$f(x)$	$L(x)$	x'	L'
1	2301	27.432	1	27.432
2	354	22.090	2	22.090
3	93	20.645	3	20.645
4	39	19.487	4	19.487
5	29	13.793	5	13.793
6	23	16.087	6	16.087
7	11	11.818	7	11.818
8	7	15.714	8.5	1.62
9	6	16.667		
10	9	12.222	10.5	1.11
11	2	10.000		
...	...	...	...	...

1. **L.N. Tolstoj: Anna Karenina –** This Russian novel appeared first in 1875; in 1877, Tolstoj prepared it for a separate edition, which was published in 1878. The novel consists of eight parts, subdivided into several chapters. Our analysis comprises (a) the first chapter of Part I, and (b) the whole of Part I consisting of 34 chapters.

2. **A.S. Puškin: Evgenij Onegin –** This Russian verse-novel consists of eight chapters. Chapter I first was published in 1825, periodically followed by further individual chapters; the novel as a whole appeared in 1833.

3. **F. Móra: Dióbél királykisasszony [“Nut kernel princess”] –** This short Hungarian children’s story is taken from a children book, published in 1965.

4. **K.Š. Gjalski: Na badnjak [“On Christmas Evening”] –** This Croatian story was first published in 1886, in the volume Pod starimi krovovi. For our purposes, we have analyzed both the complete text, and dialogical and narrative parts separately.

5. **Karel & Josef Čapek: Zářivé hlubiny [“Shining depths”] –** This Czech story is a co-authored work by the two brothers Karel and Josef Čapek. The text appeared in 1913, for the first time, and was then published together with other stories in 1916, in a volume bearing the same title.

6. **Ivan Cankar: Hiša Marije Pomocnice [“The House of Charity”] –** This Slovenian novel was published in 1904. For our purposes, we analyzed the first chapter only.

7. **Janko Král: Zakliata panna vo Váhu a divný Janko [“The Enchanted Virgin in Váh and the Strange Janko”] –** This text is a Slovak poem, which was published in 1844.
8. **Hänsel und Gretel –** This is a famous German fairy tale, which was included in the well-known Kinder- und Hausmärchen by Jacob and Wilhelm Grimm (1812), under the title of “Little brother and little sister”.
9. **Sjarif Amin: Di lembur kuring [“In my Village”] –** This story is written in Sundanese, a minority language of West Java; it was published in 1964. We have analyzed the first chapter of the story.
10. **Pak Ojik: Burung api [“The Fire Bird”] –** This fairy tale from Indonesia (in Bahasa Indonesia), which was published in 1971, is written in the traditional orthography (the preposition *di* being written separately).
11. **Henry James: Portrait of a lady –** This novel, written in 1881, consists of 55 individual chapters. We have analyzed both the whole novel, and the first chapter, only.

Table 13.2 represents the results of the analyses.[1] The first column contains the occurrence frequencies of word forms (x); the next two columns present the observed (y) and the computed (y) mean lengths of word forms having the corresponding frequency in the given individual texts. As described above, words having zero-length – such as for example the Russian preposition *k, s, v*, or the Hungarian *s* (from *és*) – have not been counted as a separate category, and have been considered as proclitics instead. In the last row of Table 13.2 one finds the values for the parameters a and b of (13.2), the text length N, and the determination coefficient R^2.

As can be seen from Table 13.2, hypothesis (13.2) can be accepted in all cases, since the fits yield R^2 values between 0.84 and 0.96, which can be considered very good – independently of language, author, text type, or text length.

[1] The English data for *Portrait of a Lady* have kindly been provided by John Constable; we would like to express our sincere gratitude for his co-operative generosity.– All other texts were analyzed in co-operation with the FWF research project #15485 ≫Word Length Frequency Distributions in Slavic Texts≪ at Graz University (cf. Grzybek/Stadlober 2002).

Table 13.2: Dependence of Word Form Length on Word Frequency

Russian			Russian			Hungarian		
Anna Karenina (ch. I)			*Evgenij Onegin* (ch. I)			*Dióbél királykisasszony*		
x	y	$\hat{y}$	x	y	$\hat{y}$	x	y	$\hat{y}$
1	2.92	3.03	1	2.66	2.70	1	2.52	2.57
2	2.14	2.04	2	2.13	1.99	2	2.00	1.88
3	2.05	1.70	3	1.78	1.71	3	1.56	1.62
4	1.50	1.53	4	1.42	1.57	4	1.57	1.49
5	1.33	1.43	5.50	1.36	1.45	6	1.33	1.35
6	1.50	1.36	7.50	1.30	1.35	14.66	1	1.17
7	1.67	1.31	11.50	1.35	1.25			
8	1	1.27	39.64	1.09	1.09			
9	1	1.24						
10	1	1.22						
13	1	1.17						
19	1	1.12						
20	1	1.11						
37	1	1.06						
$a = 2.0261, b = 0.9690$			$a = 1.7029, b = 0.7861$			$a = 1.5668, b = 0.8379$		
$R^2 = 0.88, N = 3970$			$R^2 = 0.96, N = 1871$			$R^2 = 0.96, N = 234$		

Croatian			Czech			Slovenian		
Na badnjak			*Zářivé hlubiny*			*Hiša Marije P.* (ch. I)		
x	y	$\hat{y}$	x	y	$\hat{y}$	x	y	$\hat{y}$
1	2.83	2.95	1	2.69	2.76	1	2.71	2.80
2	2.44	2.37	2	2.20	2.17	2	2.35	2.36
3	2.22	2.12	3	2.15	1.92	3	2.23	2.16
4	2.18	1.96	4	1.74	1.77	4	2	2.03
5	1.63	1.86	5	1.74	1.68	5	2	1.94
6	1.76	1.79	6	1.58	1.61	6	2	1.87
7	1.87	1.73	7	1.33	1.55	7	1.86	1.82
8	1.69	1.68	9	1.51	1.48	8	2.14	1.78
9	1.57	1.64	36.23	1.16	1.21	9.5	1.50	1.73
10	1.67	1.61				18.25	1.22	1.56
16.11	1.49	1.48				89.13	1.25	1.30
32.91	1.14	1.33						
127	1	1.17						
$a = 1.9454, b = 0.5064$			$a = 1.7603, b = 0.5921$			$a = 1.7969, b = 0.4023$		
$R^2 = 0.93, N = 2450$			$R^2 = 0.94, N = 1363$			$R^2 = 0.84, N = 1147$		

Table 13.2 (cont.)

Slovak *Zakliata panna*			German *Hänsel & Gretel*			Sundanese *Di lembur kuring*		
x	y	$\hat{y}$	x	y	$\hat{y}$	x	y	$\hat{y}$
1	2.41	2.48	1	2.12	2.17	1	2.79	2.86
2	2.05	1.92	2	1.79	1.82	2	2.38	2.31
3	1.55	1.69	3	1.73	1.67	3	2.05	2.06
4	1.85	1.57	4	1.71	1.58	4.5	1.13	1.86
5	1.50	1.49	5	1.55	1.52	6.5	1.58	1.72
6.50	1.39	1.41	6.5	1.56	1.45	13.29	1.33	1.50
10	1.07	1.30	8.5	1.49	1.40			
24.67	1.11	1.16	10.5	1.08	1.36			
			13.5	1.21	1.31			
			19.67	1.25	1.26			
			50.46	1.15	1.16			
$a = 1.1476, b = 0.675$ $R^2 = 0.88, N = 926$			$a = 1.1688, b = 0.5062$ $R^2 = 0.87, N = 803$			$a = 1.8609, b = 0.5110$ $R^2 = 0.91, N = 431$		

Indonesian *Burung api*			English *Portrait of a Lady* (ch. I)		
x	y	$\hat{y}$	x	y	$\hat{y}$
1	3.34	3.44	1	2.17	2.23
2	3.03	3.04	2	1.78	1.69
3	2.93	2.83	3	1.56	1.49
4	2.78	2.70	4	1.5	1.39
5	2.33	2.61	5	1.37	1.32
6	2.68	2.53	6	1	1.28
7	2.57	2.47	7	1.33	1.24
8	2.53	2.42	8.50	1.28	1.20
9	2.38	2.38	11	1	1.17
10	2.50	2.34	14.50	1.06	1.13
11	2.36	2.31	19.83	1.06	1.10
13	2.17	2.25	27.50	1.13	1.08
17	2	2.17	73.43	1	1.03
$a = 2.4353, b = 0.2587$ $R^2 = 0.92, N = 1393$			$a = 1.2293, b = 0.8314$ $R^2 = 0.89, N = 1104$		

By way of an example, Figure 13.1 illustrates the results of the first chapter of Tolstoj's *Anna Karenina*: the courses both of the observed data and the data computed according to (13.2), can be seen. On the abscissa, the occurrence frequencies from $x = 1$ to $x = 40$ are given, on the ordinate, the mean word lengths, measured in the average number of syllables per word. With a determination coefficient of $R^2 = 0.88$, the fit can be accepted to be satisfactory.

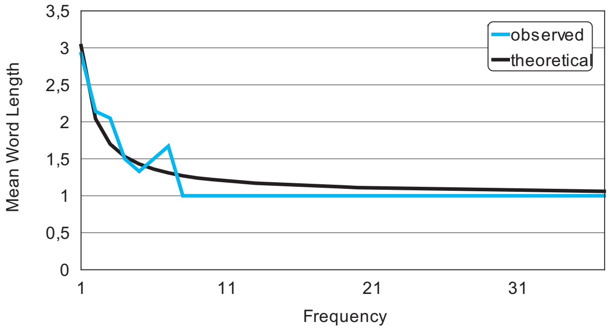


Figure 13.1: Observed and Computed Mean Lengths in *Anna Karenina* (I,1)

4. The Parameters

Since all results can be regarded to be good ($R^2 > .85$), or even very good ($R^2 > .95$), the question of a synthetical interpretation of these results quite naturally arises. First and foremost, a qualitative interpretation of the parameters a and b, as well as a possible relation between them, would be desirable. Figure 13.2 represents the course of all theoretical curves, based on the parameters a and b given in Table 13.3.

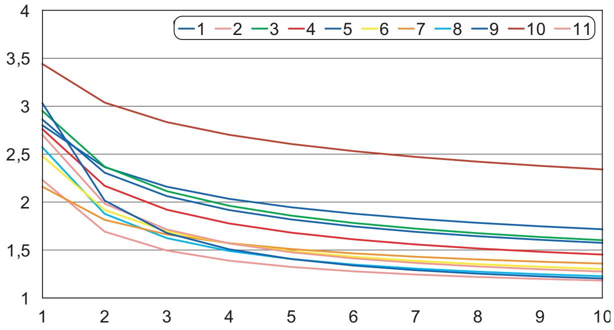


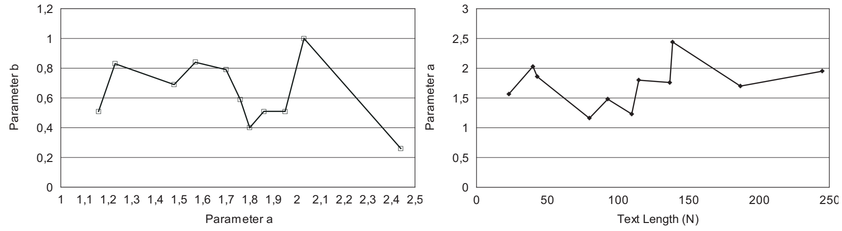
Figure 13.2: Course of Theoretical Curves (Dependence of Word Form Length on Frequency; cf. Table 13.2)

Since the curves representing the individual texts intersect, it is to be assumed that no general judgment, holding true for all texts in all languages, is possible. In Table 13.3 the parameters and the length of the individual texts are summarized.

Table 13.3: Parameters and Text Length of Individual Texts

Text	**Language**	a	b	N
Anna Karenina (I,1)	Russian	2.03	0.97	397
Evgenij Onegin (I)	Russian	1.70	0.79	1871
Na badnjak	Croatian	1.95	0.51	2450
Zářivé hlubiny	Czech	1.76	0.59	1363
Hiša Marije Pomocnice (I)	Slovenian	1.80	0.40	1147
Zakliata panna	Slovak	1.48	0.69	926
Hänsel und Gretel	German	1.16	0.51	803
Fairy Tale by Móra	Hungarian	1.57	0.84	234
Di lembung kuring	Sundanese	1.86	0.51	431
Burung api	Indonesian	2.44	0.26	1393
Portrait of a Lady (I)	English	1.23	0.83	1104

From Figure 13.3(a) it can easily be seen that there is no clear-cut relationship between the two parameters a and b. The next question to be asked, quite logically concerns a possible relation between the parameters a and b, and the text length N; yet, the answer is negative, again. As can be seen in Figure 13.3(b), the relation rather seems to be relatively constant with a great dispersion; consequently, no interpretable curve can capture it.



(a) Parameters a and b (

(b) Text Length N and Parameter a

Figure 13.3: Relationship between Parameters a and b and Text Length N (cf. Table 13.3)

It is evident that the fact of a missing relationship between the parameters a and b, and text length N, respectively, can be accounted for by the obvious data inhomogeneity: since the texts come from different languages and various text types, the *ceteris paribus* condition is strongly violated, and the data in this mixture are not adequate for testing the hypothesis at stake.

5. The Homogeneity of a ‘Text’

In order to avoid the encroachment caused by the different provenience of the texts, we will next examine the problem using texts whose linguistic and textual homogeneity can, at least hypothetically, a priori be taken for granted. However, even here, the problem of homogeneity is not ultimately solved. Let us therefore compare the results for Chapter I of Tolstoj’s *Anna Karenina* with those for the complete Book I, consisting of 34 chapters, as represented in Table 13.4.

Table 13.4: The Length-Frequency Curves for Chapter I and the Complete Text of *Anna Karenina*

Chapter 1			Complete text					
x	y	$\hat{y}$	x	y	$\hat{y}$	x	y	$\hat{y}$
1	2.92	3.03	1	3.38	3.60	22	2.20	2.13
2	2.14	2.04	2	3.04	3.16	23	1.94	2.12
3	2.05	1.70	3	2.84	2.94	24.50	2.27	2.10
4	1.50	1.53	4	2.76	2.79	26.50	2.04	2.08
5	1.33	1.43	5	2.84	2.69	28.50	2.27	2.05
6	1.50	1.36	6	2.65	2.61	30.50	2.07	2.04
7	1.67	1.31	7	2.45	2.54	33.50	2.29	2.01
8	1	1.27	8	2.57	2.49	38	1.70	1.98
9	1	1.24	9	2.47	2.44	42.50	2.13	1.95
10	1	1.22	10	2.64	2.40	51.50	1.67	1.90
13	1	1.17	11	2.59	2.36	57.50	1.83	1.87
19	1	1.12	12	2.41	2.33	62	1.64	1.85
20	1	1.11	13	2.50	2.30	73.25	1.88	1.82
37	1	1.06	14	2.20	2.28	91.71	1.43	1.77
			15	2.43	2.25	106.86	1.33	1.74
			16.50	2.11	2.22	137.70	1.70	1.69
			18	2.35	2.19	229.11	1.28	1.60
			19.50	2.32	2.17	458.75	1.38	1.50
			21	2.20	2.14			

In Figure 13.4, the empirical data and the theoretical curve are presented for the sake of a graphical comparison. One can observe two facts:

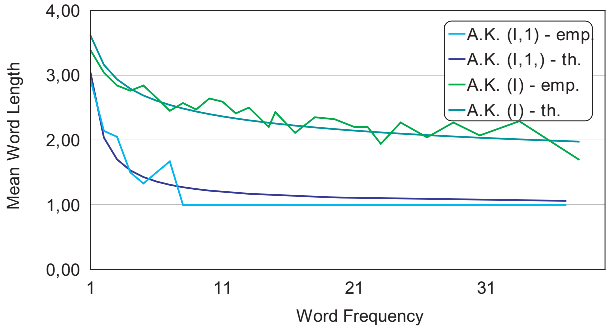


Figure 13.4: Comparison of *Anna Karenina*, Chap. I and Book I

1. The empirical and, consequently, the theoretical values of the larger sample (i.e., Book I, 1-34), are located distinctly higher. For the theoretical curve this results in an increase of a and a decrease of b.
2. The fitting for the greater sample is still acceptable, but clearly worse ($R^2 = 0.86$) as compared to the smaller sample ($R^2 = 0.97$).

	A.K. (I,1)	**A.K. (I)**
Tokens	702	38226
Types	397	8661
a	2.03	2.60
b	0.97	0.27
R^2	0.97	0.86

The important finding, that a more comprehensive text unit leads to a worse result than a particular part of this ‘text’, can be demonstrated even more clearly, comparing the single chapter of a novel with the whole novel. Testing the complete novel *Portrait of a Lady* to this end, one obtains a determination coefficient of merely $R^2 = 0.58$, even after smoothing the data as described above. As compared to the first chapter of this novel taken separately, yielding a determination coefficient of $R^2 = 0.89$, (cf. Table 13.2), this is a dramatic decrease. In fact, an extremely drastic smoothing procedure is necessary in order to obtain an acceptable result (with $a = 1.34, b = 0.47; R^2 = 0.92$), as shown in Table 13.5 and Figure 13.5.

Thus, appropriate smoothing of the data turns out to be an additional problem. On the one hand, some kind of smoothing is necessary because the frequency class size should be “representative” enough, and on the other hand, the particular kind of smoothing is a further factor influencing the results.

Table 13.5: Results of Fitting Equation (13.2) to *Portrait of a Lady*, Using the Given Pooling of Values

Class					**Class**				
lower limit	*upper limit*	x'	y	$\hat{y}$	*lower limit*	*upper limit*	x'	y	$\hat{y}$
1	10	1	2.22	2.34	601	700	70	1	1.18
11	20	2	1.92	1.97	701	800	80	1	1.17
21	30	3	1.84	1.80	801	900	90	1	1.16
31	40	4	1.81	1.70	901	1000	100	1	1.15
41	50	5	1.71	1.63	1001	2000	200	1.17	1.11
51	60	6	1.71	1.57	2001	4000	400	1	1.08
61	70	7	1.66	1.53	4001	5000	500	1	1.07
71	80	8	1.49	1.50	5001	6000	600	1	1.06
81	90	9	1.55	1.47	6001	7000	700	1	1.06
91	100	10	1.30	1.45	7001	8000	800	1	1.06
101	200	20	1.40	1.32	8001	9000	900	1	1.05
201	300	30	1.25	1.27					
301	400	40	1.40	1.23					
401	500	50	1.25	1.21					
501	600	60	1.18	1.19					

However, there is a clear tendency according to which the individual chapters of a novel abide by their own individual regimes organizing the length-frequency relation. This boils down to the assessment that even a seemingly homogeneous novel-text is an inhomogeneous text mixture, composed of diverging superpositions. As to an interpretation of this phenomenon, it seems most likely that after the end of a given chapter, a given ruling order ends, and a new order (of the

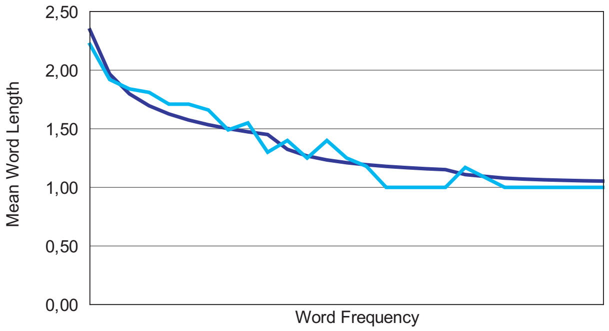


Figure 13.5: Fitting Equation (13.2) to *Portrait of a Lady* (cf. Table 13.5)

same organization principle) begins. The new order superposes the preceding, balances or destroys it. Theoretically speaking, one should start with as many components of $y = a_1x^{b_1} + a_2x^{b_2} + a_3x^{b_3} + \ldots$, as there are chapters in the text. Whether this is, in fact, a reasonable procedure, will have to be examined separately).

As a further consequence, one must even ask if one individual chapter of a novel, or a short story, etc. is a homogeneous text, or if we are concerned with text mixtures due to the combination of heterogeneous components. In order to at least draw attention to this problem, we have separately analyzed the narrative and the dialogical sequences in the Croatian story *Na badnjak*. As a result, it turned out that the outcome is relatively similar under all circumstances: for the dialogue sequences we obtain the values $a = 1.61, b = 0.84, R^2 = 0.96$, for the narrative sequences $a = 1.93, b = 0.54$, and $R^2 = 0.91$ (as compared to $a = 1.95, b = 0.51, R^2 = 0.93$ for the story as a whole). It goes without saying, that more systematic examination is necessary to attain more reliable results.

While on the one hand, it turns out that a longer text does not necessarily yield better results, on the other hand, increasing text length need not necessarily yield worse results. By way of an example, this can be shown on the basis of cumulative processing of *Evgenij Onegin* and its eight chapters (i.e., chapter 1, then chapter 1+2, 1+2+3, etc.). In this way, one obtains the results shown in Table 13.6; the curves corresponding to the particular parts are displayed in Figure 13.6.

Table 13.6: Parameters of the Frequency-Length Relation in *Evgenij Onegin*

	Parameters		*Types*	*Tokens*	*Fit*
Chapter	a	b	N	M	R^2
1	1.703	0.786	1871	3209	0.96
1-2	1.838	0.691	2918	5546	0.88
1-3	1.921	0.574	3951	8359	0.88
1-4	1.967	0.525	4851	10936	0.92
1-5	1.954	0.476	5737	13376	0.94
1-6	1.968	0.52	6509	15978	0.94
1-7	2.031	0.425	7476	19061	0.86
1-8	2.049	0.399	8329	22482	0.88

As can be seen, the curves do not intersect under these circumstances. The displacement of the curve position with increasing text size can be explained

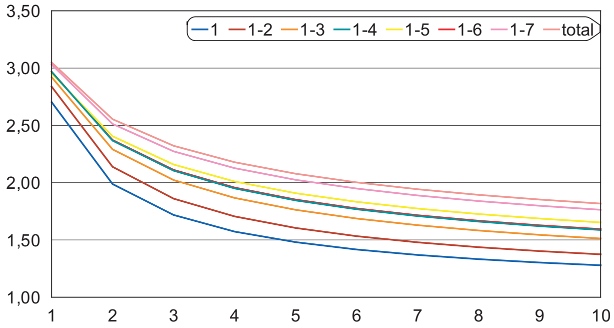
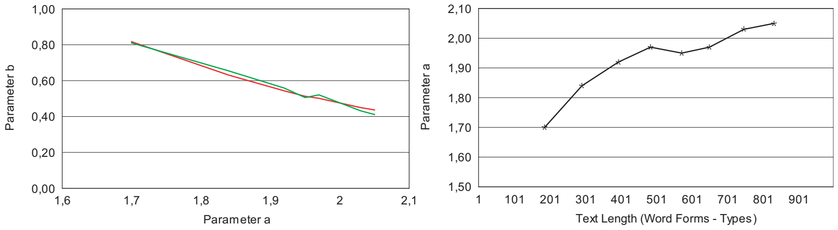


Figure 13.6: Fitting (13.2) to the Text Cumulation of *Evgenij Onegin*

by the fact that words from classes with low frequency wander to higher classes and are substituted by ever longer words. In Figure 13.7(a) the dependency between the parameters a and b is shown for the cumulative processing (b being represented by its absolute value).



(a) Parameters a and b

(b) Text Length N and Parameter a

Figure 13.7: Relationship between Parameters a and b and Text Length N in *Evgenij Onegin* (cf. Table 13.6)

Evidently, b depends on a. Although there seems to be a linear decline, the relation between a and b cannot, however, be linear, since b must remain greater than 0. The power curve $b = 4.9615a^{-3.3885}$ yields a good fit with $R^2 = 0.92$. In the same way, b depends on text length N. The same relationship yields $b = 21.4405N^{-0.4360}$ with $R^2 = 0.96$. The dependence of a on N can be computed by mere substitution in the second formula, yielding $a = 0.6493N^{0.1286}$ whose values are almost identical with the observed ones. It is irrelevant whether one considers types or tokens since they are strongly correlated ($r = 0.997$). Fig. 13.7(b) shows the relationship between text length N and parameter a.

It can thus be concluded that, in a homogeneous text, i.e., in a text in which one can reasonably assume the *ceteris paribus* condition to be fulfilled, the relationship between frequency and length remains intact: with an increasing text length, the curve is shifted upwards and becomes flatter. The parameters are joined in form of $a = f(N), b = g(a)$ or $b = h(N)$, respectively, f, g, h being functions of the same type.

6. Conclusion

Let us summarize the basic results of the present study. With regard to the leading question as to the relationship between frequency and length of words in texts, we have come to the following conclusions:

I. The above hypothesis (2) is corroborated in the given form by our data;

II. A homogeneous text does not interfere with linguistic laws, an inhomogeneous one can distort the textual reality;

III. Text mixtures can evoke phenomena which do not exist as such in individual texts: In text mixtures, the *ceteris paribus* condition does not hold; short texts have the disadvantage of not allowing a property to take appropriate shape; without smoothing, the dispersion can be too strong. Long texts contain mixed generating regimes superposing different layers. In text corpora, this may lead to “artificial” phenomena as, probably, oscillation. Since these phenomena do not occur in all corpora, it seems reasonable to consider them as a result of mixing.

IV. With increasing text size, the resulting curve of frequency-length relation is shifted upwards; this is caused by the fact that the number of words occurring only once increases up to a certain text length. If this assumption is correct, then b converges to zero, yielding the limit $y = a$.

References

Altmann, G.
1992 "Das Problem der Datenhomogenität." In: *Glottometrika 13*. Bochum. (287–298).

Arapov, M.V.
1988 *Kvantitativnaja lingvistika.* Moskva.

Baker, S.J.
1951 "A linguistic law of constancy: II", in: *The Journal of General Psychology, 44*; 113–120.

Belonogov, G.G.
1962 "O nekotorych statističeskich zakonomernostjach v russkoj pis'mennoj reči", in: *Voprosy jazykoznanija, 11(1)*; 100–101.

Fenk-Oczlon, G.
1990 "Ikonismus versus Ökonomieprinzip. Am Beispiel russischer Aspekt- und Kasusbildungen", in: *Papiere zur Linguistik, 42*; 49–68.

Fenk-Oczlon, G.
1986 "Morphologische Natürlichkeit und Frequenz." Paper presented at the 19th Annual Meeting of Societas Linguistica Europae, Ohrid.

Grzybek, P.; Altmann, G.
2003 "Oscillation in the frequency-length relationship", in: *Glottometrics, 5*; 97–107.

Grzybek, P.; Stadlober, E.
2003 "The Graz Project on Word Length Frequency (Distributions)", in: *Journal of Quantitative Linguistics, 9(2)*; 187–192.

Guiraud, P.
1954 *Les caractères statistiques du vocabulaire. Essai de méthodologie.* Paris.

Guiter, H.
1977 "Les relations /fréquence – longueur – sens/ des mots (langues romanes et anglais)." In: *XIV Congresso Internazionale di linguistica e filologia romanza, Napoli, 15-20 aprile 1974.* Napoli/Amsterdam. (373–381).

Haiman, J.
1983 "Iconic and economic motivation", in: *Language, 59*; 781–819.

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

Herdan, G.
1966 *The advanced theory of language as choice and chance.* Berlin.

Kaeding, F.W.
1897–98 *Häufigkeitswörterbuch der deutschen Sprache.* Steglitz.

Köhler, R.
1986 *Zur linguistischen Synergetik. Struktur und Dynamik der Lexik.* Bochum.

Manczak, W.
1980 "Frequenz und Sprachwandel." In: Lüdtke, H. (ed.), *Kommunikationstheoretische Grundlagen des Sprachwandels.* Berlin/New York. (37–79).

Miller, G.A.; Newman, E.B.; Friedman, E.A.
1958 "Length-frequency statistics for written English", in: *Information and Control, 1*; 370–389.

Zipf, G.K.
1932 *Selected studies of the principle of relative frequency in language.* Cambridge, Mass.

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

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