

Word length balance in texts: Proportion constancy and word-chain-lengths in Proust's longest sentence

Simone Andersen¹, Düsseldorf

Abstract. Constancy phenomena in word length distributions of texts are demonstrated. The regularity of proportions is shown by intercorrelation of parts under differing kinds of partitioning. Length homogeneity r_A as a measure for the stability of the values of the distribution is developed. Balance number B refers to word-chains in line: Every B words the total number of syllables tends to be equal, indicated by decreased variance.

Keywords: word length, homogeneity, intercorrelation, length balance, word-chain-length, constancy

1. The problem of length proportions

The overall shape of the distribution of word lengths in a given text is well predictable by the word length laws (Zipf 1949; Altmann 1988; Wimmer, Köhler, Grotjahn & Altmann 1994; Wimmer & Altmann 1996, Altmann & Best 1996; Best 2001; Grzybek 2005; <http://www.gwdg.de/~kbest/litlist.htm>).

How are these lengths scattered over the text? Obviously there is no fixed order of short and longer words, as long as we refer to prose. But if their patterns were completely arbitrary, it would be conceivable to find a possibly uneven scattering with heterogeneous components, e.g. a text where all the short words occur at the beginning, so that the longer ones have to crowd together at the remainder to compensate for it at the end. There are concepts in linguistics proposing that only the entire text reveals the true frequency proportions. Orlov presumes (Orlov 1982; Best 2003) that the author of a text organizes the frequency structure of the elements only for the text as a whole. So the proportions do not hold for parts of the text.

Opposed to this, we believe that length balance in speaking and writing becomes visible from the beginning. Whatever the reason for the individual distribution, we suppose that its effects on word length proportions of the text will work in a homogeneous way and presumably should be rather independent of the text producer's talent for organization. Consequently the distribution should be found in the parts as in the whole. We tried to find evidence for this by investigating a text and detecting the degree of homogeneity of the word length proportions characterizing it. Additionally we were looking for hints indicating length balance in very small text segments as well. Tendency towards balance could also be revealed by another constancy

¹ Address correspondence to: AndersenSC@aol.com

phenomenon: We were looking for units in spoken or written text that could be considered as recurring patterns within which the lengths are balanced out so that the pattern units tend to be equal.

2. Method

We partitioned one text in varying ways and observed the properties of the resulting parts. Word length was measured in number of syllables. In order to find indications of length balance we used two kinds of investigation.

The first step was comparing distributions and lengths under differing kinds of text partitioning. In the second step we partitioned the text into a finer grid and examined the length (number of syllables) of word chains - sequences of words occurring after one another in line in the text - regardless of their grammatical or semantic relations, i.e. without taking into account grammatical constituents, clauses or phrases. We studied the variability of the chain lengths depending on the number of words in the chain.

In order to eliminate influences by sentence limits, we looked for a sentence as long as possible. We chose Marcel Proust's longest sentence, detected by the writer Alain de Botton (2001), from Proust's work *A la recherche du temps perdu*.

Using the German translation of the original text makes sure that the word lengths cannot result from the poet's intention or individual taste (sense of rhythm etc.) but are to a greater extent determined by constraints coming from the language system and its properties.

2. Results

The total number of words in this longest sentence is $n = 519$.

The shape of their length distribution is as to be expected (see Table 1), with a slight over-representation of two-syllable words – when compared to one of the Altmann-Fitter distributions (1994; 1997) – probably due to Proust's very detailed descriptions which typically request the use of very specific words.

Table 1
Lengths of single words in the entire sentence

Length x (number of syllables)	Frequency f_x (number of tokens with length x)	Proportion
1	203	0.391
2	194	0.374
3	63	0.120
4	40	0.077
5	16	0.031
6	1	0.002
7	2	0.004
	$n = 519, \bar{x} = 1.99 \quad s^2 = 1.2256$	

In the first step we want to know what happens to the shape of the distribution under varied kinds of text partitioning:

Disregarding the last 19 words, we divide the remaining 500 words of the text into two parts of 250 words and get two distributions of lengths:

Table 2
Split half: Word lengths in the first and second half of text
(without the last 19 words)

Text part	1-syll	2-syll	3-syll	4-syll	5-syll	6-syll	7-syll
(I) 1-250	98	96	27	19	8	1	1
(II) 251-500	99	90	34	19	8	0	0

In the next step we partitioned the whole text into five parts containing 100 words each:

Table 3
Word lengths under text partitioning into 5 parts

Parts	1-s	2-s	3-s	4-s	5-s	6-s	7-s
first: the first 100 words	34	37	15	6	6	1	1
second: words 101-200	45	37	9	8	1	0	0
third: words 201-300	40	38	8	11	3	0	0
fourth: words 301-400	41	38	15	5	1	0	0
fifth: words 401-500	37	36	14	8	5	0	0
last 19 words	6	8	2	2	0	0	1
n = 519	203	194	63	40	16	1	2
Proportion	0.391	0.374	0.120	0.077	0.031	0.002	0.004

We observe a remarkable constancy of proportions as illustrated in Fig. 1.

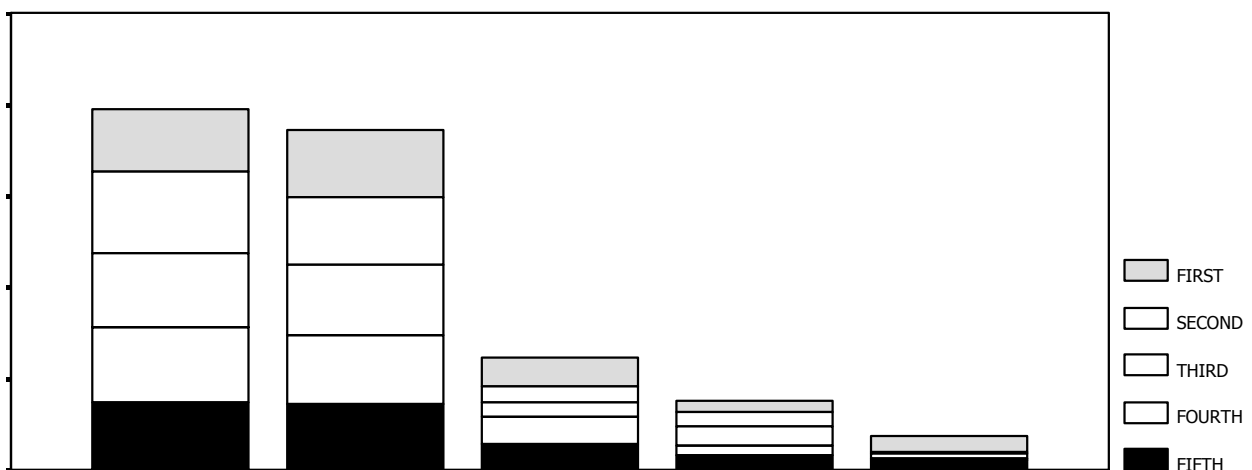


Fig. 1. Proportions of word lengths in the five parts of the text
(Length 5 = 5, 6 or 7 syllables)

The distributions of word lengths in the different parts and in the entire text are shown in Fig.

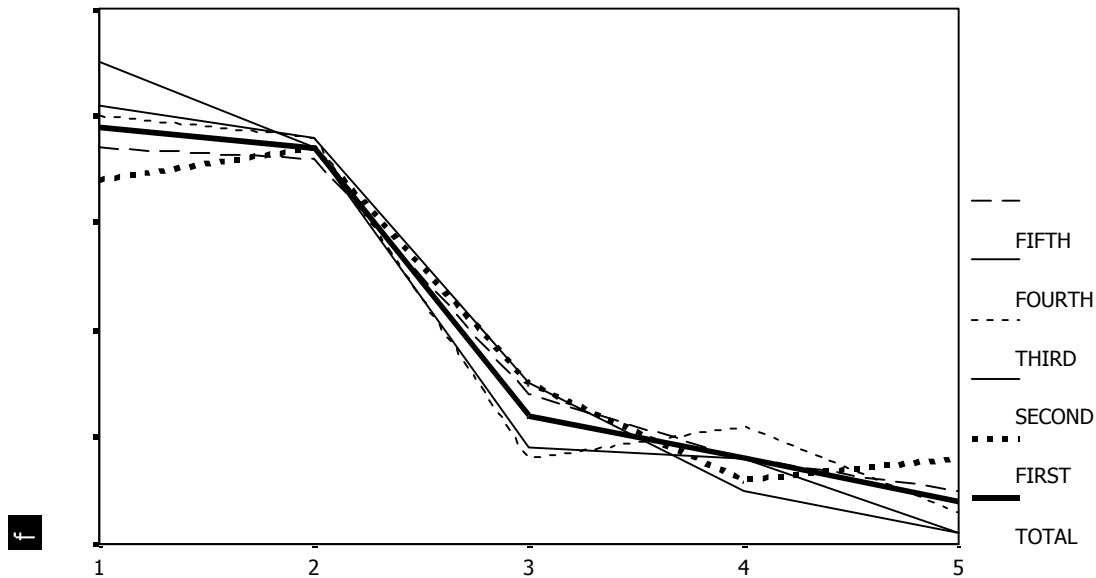


Fig. 2. Length distributions in the five parts and in the entire text (total)

Table 4 shows the proportions for word lengths in the entire text (column 1) in the first and second half (column 2 resp. 3) and in the five parts of 100 words (columns 4 – 8):

Table 4
Proportions of lengths for different parts and entire text

word length (syllables)	total	1.half	2.half	1-100	101-200	201-300	301-400	401-500
1	0.39	0.39	0.40	0.34	0.45	0.40	0.41	0.37
2	0.37	0.38	0.36	0.37	0.37	0.38	0.38	0.36
3	0.12	0.10	0.14	0.15	0.09	0.08	0.15	0.14
4	0.08	0.08	0.08	0.06	0.08	0.11	0.05	0.08
5	0.03	0.03	0.03	0.06	0.01	0.03	0.01	0.05
6	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
	519	250	250	100	100	100	100	100

Something which is very striking is the nearly constant proportion of the two syllable words (see in the second row above): It is nearly constantly $p = 0.37$ which is the overall proportion for the entire text and it can be found in every part, in the first and in the second half as in every 100 words of the text. We will go back to this later.

4. Length homogeneity of texts

Now we are able to calculate the length correlations between the different parts.

The following table (Table 5) shows the correlations r_{pp} between the parts consisting of 100 words and their intercorrelations r_{pt} as the mean of each row (without the diagonal) which is the mean of the correlations between one part and the remaining text (= the four remaining parts, without the last 19 words as explained above). The word lengths of 6 or 7 syllables could be of confounding influence: Because of their extremely low proportions (rounded values of 0.00 almost everywhere) they increase the correlations improperly. So we grouped them together with the 5-syllable-words and counted them as length class no. five (already visible in Fig.1 and Fig.2).

Table 5
Homogeneity: Intercorrelations of the parts of the text

Parts	Parts					r_{pt}
	1-100	101-200	201-300	301-400	401-500	
1-100	1.00	0.9527	0.9514	0.9829	0.9882	0.9688
101-200	0.9527	1.00	0.9915	0.9803	0.9855	0.9775
201-300	0.9514	0.9915	1.00	0.9669	0.9802	0.9725
301-400	0.9829	0.9803	0.9669	1.00	0.9971	0.9818
401-500	0.9882	0.9855	0.9802	0.9971	1.00	0.9878
						$r_{int} = 0.9777$

The mean of the last column is

$$r_{int} = (0.9688 + 0.9775 + 0.9725 + 0.9818 + 0.9878)/5 = 0.9777$$

which indicates the degree of intercorrelation of all parts. In analogy to test theoretical scale analysis in psychological diagnostics we could try to interpret the degree of intercorrelation to be a measure of homogeneity related to length proportions.

We will call this length proportion homogeneity or just length homogeneity r_{Λ} with $r_{\Lambda} = r_{int}$ and propose that it may be a useful measure of a given text to characterize the stability of its length distribution. In Proust's sentence length homogeneity r_{Λ} is extremely high (0.9777), but we suppose that any text written or produced by a single author within a narrow time span will yield a considerable length homogeneity. In classical test theory, the concepts of homogeneity or stability converge towards a measure for reliability.

If we look at the proportion of an individual word length in an entire text (for example, the proportion of 0.39 as a score for one-syllable-words), we can put the question of how reliable this value is for each part of the text. Is it the resulting average of very inhomogeneous parts? Or is it a typical proportion value, valid for many text parts?

Thus the length homogeneity is a measure for the precision of assessment in determining the characteristic length distribution in a given text.

An additional step for improving the results could be eliminating those lengths that come to less than 0.01 per cent of the entire text: word lengths of 6 or 7 syllables and more are too rare to be a useful measure. Nearly always producing the same value (here: zero-proportion) means that

they have poor discrimination power: they provide no information, they are levelling the results and will be disregarded – not only in this example but generally. As we observe in Table 5.a, removal of the word lengths of 6 and 7 would not change a lot of the intercorrelation: it would increase by a very small amount.

Table 5a
Homogeneity: Intercorrelations of the parts of the text without lengths of 6 or more

Parts	Parts					
	1-100	101-200	201-300	301-400	401-500	r_{pt}
1-100	1.00	0.9575	0.9562	0.9887	0.9921	0.9736
101-200	0.9575	1.00	0.9915	0.9803	0.9855	0.9787
201-300	0.9562	0.9915	1.00	0.9669	0.9802	0.9737
301-400	0.9887	0.9803	0.9669	1.00	0.9971	0.9834
401-500	0.9821	0.9855	0.9802	0.9971	1.00	0.9887
						$r_{int} = 0.9796$

5. Length homogeneity r_A as a text characteristic

Now we can use r_A as a text characteristic if we observe and determine the decrease of the intercorrelation in dependence of the number of parts t .

Unlike in psychological scale analysis, we have no "natural" units, like the items of a test. Instead, we are able to divide a text into parts of any size, and we can make use of this fact by measuring how far a text can be partitioned into equal parts without losing a considerable amount of homogeneity.

The number N of words within the parts that are to be compared and intercorrelated ranges from the upper limit of $N = n/2$ (with n = the number of words in the entire text) down to $N = 100$, because proportions of less than 10 percent (for words with 4 or more syllables) can be compared meaningfully only if it makes a difference between occurring and non-occurring in the text.

From that it follows that there is the minimum size of $n = 100$ words in a text to determine its word length homogeneity r_A .

N = number of words in the parts; t = number of parts of equal size

$$100 \geq N = n/t$$

Number of parts t	N	r_A
1	519	1.00
2	250	0.9912
3	170	*
4	125	*
5	100	0.9777

(* = has not been calculated here)

In this text, the limit for t is 5, because $100(6) > 519$.

During partitioning up to the individual limit, the intercorrelation does not fall below 0.9. We interpret this as a very high degree of homogeneity of the length distribution in text. Only correlations of less than 0.9 should be considered as loss of homogeneity.

6. Visibility of the frequency distribution and best sample

Additionally we could try to search out those parts that show the highest correlation with the entire text (although this is a rather subtle question, because of the extremely high correlations of all parts). The values are shown in Table 6.

Table 6
Correlations of text parts with the entire text (= total)

	FIRST	SECOND	THIRD	FOURTH	FIFTH
TOTAL	.9814	.9925	.9895	.9927	.9983

We can see that the correlation between part 5 (last part) with the entire text is nearly perfect (0.9983), followed by part 4. Perhaps it could mean that the proportions are being settled best towards the end, so the last part of a text reveals the frequency distribution most apparently and can be considered as the best sample of a given text. Here we approach the point at which we must discuss the question of what causes the length distribution of words in a given text. In view of the fact that the investigated sample is a translated text, and the translator has little choice in respect of determining the lengths of the words even unconsciously, we are minded to propose that the considerable homogeneity provides additional evidence for what we already found and claimed in another context (Andersen 2002): other than in musical composition, the frequencies in texts are to a great extent out of reach for the individual text producer. Probably the even higher correlations of the last parts indicate a small amount of controlled or intentional production in the beginning.

7. Word-chains and their lengths

Let us now look at the text from another perspective. As we could observe already in Table 3 (6th row), the last 19 words in the text do not reveal the typical distribution. Of course, we do not expect that in every text part of any length we will find the proportions above; this would be a very straight pattern. And as we said above, we don't look for proportion constancy in parts smaller than 100 words.

Apart from proportion constancy in larger parts, we are looking for constant or at least similar patterns of length in order to find hints indicating length balance within smaller units of a text. Are we able to find a certain number B (balance) of words where the total number of syllables tends to be equal, regardless of the lengths of the component words?

So our investigation objects will be *word chains*. Word chains are sequences of words occurring one after another in line in the text, regardless of their grammatical or semantic relations.

The idea of length balance arising when investigating a number of units in line follows from the Menzerath-Altmann-law (Altmann 1980; Altmann & Schwibbe 1989; Hřebíček 2000): The components of the shorter units are longer (per average) than those of the longer units.

We divided the text into single words ("one-word-chains"), two-word-chains (2-w), three-word-chains (3-w) etc. up to 12-word-chains. Instead of considering the distributions of lengths within them we recorded their length, measured by number of syllables in the chain. We are interested in the variability of the chain lengths depending on the number of words in the chain.

To keep the number of tokens constant, in the beginning we considered the first 90 words (tokens) of the text. Values are shown in Table 7.

Table 7
Lengths (number of syllables) of single words, two-word-chains, three-word-chains, 10-word-chains, 11-word-chains with their frequencies *f* (the first 90 tokens)

Length (= number of syllables)	f(1-w)	f(2-w)	f(3-w)	f(10-w)	f(11-w)
1	30	-	-		
2	33	6	-		
3	13	11	2		
4	7	8	4		
5	5	7	4		
6	1	8	5		
7	1	1	4		
8	-	1	4		
9	-	1	4		
10	-	1	2	-	
11	-	-	-	-	-
12	-	-	-	-	-
13	-	-	1	-	-
...				1	-
18				-	-
19				1	1
20				1	2
21				1	-
22				4	1
23				-	-
24				-	2
25				-	2
26				-	-
27				1	-
28				-	1
Number of chains	90	45	30	9	9

To remember the limitations:

Because of combinatorics, the distribution of chain lengths with more than two words per chain cannot show the same shape as the distribution of single words. The shortest chains can hardly be the most frequent, because with an increasing number of words in the chain there are

still fewer possibilities of realization. For example, to get the shortest three-word-chain with a length of 3 syllables, there has to be a coincidence of 1 + 1 + 1 syllables which is only one out of 7^3 possible states - taking the length of 7 syllables as a kind of an upper limit for a word. This is a rare case, regardless of the greater probability of short words, as long as we refer to European languages.

Instead of showing the typical word length shape, the chain lengths should converge towards a favourable length as soon as length balance can be found, or as soon as some typical patterns occur where short and long words are combining in a favoured proportion.

As shown in Tables 8.1 and 8.2, we observe that the variance changes, it is fluctuating. For ten-word-chains (when standardized, see Table 8.2) it is at minimum, and then increases again:

Table 8.1
Number of syllables in word chains (the first 90 tokens)

	chains in the sample	mean	var	s	range
one-word-chains	90	2.233	1.709	1.307	6
two-word-chains	45	4.467	3.618	1.902	8
Three-word-chains	30	6.700	5.528	2.351	10
Nine-word-chains	10	20.10	8.989	2.998	9
Ten-word-chains	9	22.33	7.500	2.739	10
11-word-chains	9**	24.00	8.750	2.958	9
12- word-chains	8*	26.25	14.21	3.770	13

*in this case: 96 words (= 8 x 12)

** 99 words (= 9 x 11)

Table 8.2
Values of Table 8.1, standardized by chain length

	chains in the sample	mean	var	s	range	mean ± s	mean ± s
one-word-chains	90	2.233	1.709	1.307	6	0.926 – 3.540	2.614
two-word-chains	45	2.234	1.809	0.951	4	1.282 – 3.184	1.902
three-word-chains	30	2.233	1.842	0.784	3.33	1.450 – 3.017	1.567
nine-word-chains	10	2.233	0.999	0.333	1	1.900 – 2.566	0.666
Ten-word-chains	9	2.233	0.750	0.274	1	1.959 – 2.507	0.548
11-word-chains	9**	2.182	0.795	0.269	0.82	1.911 – 2.449	0.538
12- word-chains	8*	2.188	1.185	0.314	1.1	1.873 – 2.502	0.629

*in this case: 96 words (= 8 x 12);

** 99 words (= 9 x 11)

The fluctuating variance is striking because it is related to the same 90 words and differs depending on the kind of partitioning. The range (standardized) decreases and increases.

The amount of mean ± s (the "2/3 area" of all values) decreases and increases again with increasing chain length. For ten-word-chains and eleven-word-chains it is at minimum, for 12-word-chains it increases again.

In the first 90 tokens, we considered a sample with constant size but differing number of chains.

In order to eliminate the varying number of cases, we now consider for every chain length the first 10 chains and record their lengths, and their variance (see Tables 9.1 and 9.2):

Table 9.1
The first 10 chains: average number of syllables

	Words in the sample	mean	Variance	St. deviation	Range
two-word-chains	20	4.4	1.822	1.35	3
Three-word-chains	30	6.2	4.844	2.2	7
Nine-word-chains	90	20.1	8.989	2.998	9
Ten-word-chains	100	21.9	8.544	2.923	10
11-word-chains	110	23.6	9.378	3.062	9
12- word-chains	120	25.2	15.96	3.994	13

Table 9.2
Values of Table 9.1, standardized by chain length

	mean	Variance	St. deviation	mean \pm s
two-word-chains	2.200	0.911	0.675	1.520 – 2.875
Three-word-chains	2.066	1.615	0.733	1.330 – 2.799
Nine-word-chains	2.233	0.999	0.333	1.900 – 2.566
Ten-word-chains	2.190	0.854	0.292	1.898 – 2.480
11-word-chains	2.145	0.853	0.278	1.867 – 2.423
12- word-chains	2.100	1.330	0.333	1.767 – 2.433

Again, we find that the standardized variance decreases and increases with increasing chain length.

8. Comparing the variances

Let us now look at the variability of the data values. We are not allowed to calculate a statistical Analysis of variance, because the assumptions required are not fulfilled (independent groups, normal distribution). But we do not need it either, because we are not really interested in the question of whether the chain means are equal or not. Of course, they are not. Rather, we are interested in finding evidence for greater homogeneity of the first ten 10-word chains compared to the first ten 3-word-chains.

We want to know if the variability of each can be attributed to the variability among the chains, or rather to some characteristics of the individual chains. So we must compare the variance within each chain to the variance between the chains. For this purpose we make use of the starting procedure of an analysis of variance.

10-word-chains:

In Table 9.1 we find the variance between the chains: $s^2(\text{betw})_{10} = 8.544$.

If we divide this by the number of words per chain (here: 10 words), we get the average variance *between* the chains: $s_w^2(\text{betw})_{10} = 0.854$ (see Table 9.2).

To get the variance *within* the chains, we have to sum up the ten single variances (see Table 10, last row): $s^2(\text{in})_{10} = 16.85$ and divide it by the number of chains, so we get the mean variance in a chain as: $s_w^2(\text{in})_{10} = 16.85 : 10 = 1.685$

Table 10
Distributions of word lengths (number of syllables) in 10-word-chains

Syll	Chain no.									
	1	2	3	4	5	6	7	8	9	10
1	3	2	7	4	2	4	4	3	1	4
2	5	4	1	3	4	1	5	4	6	4
3	0	3	0	1	2	4	0	1	2	2
4	2	1	1	0	0	1	0	1	1	0
5	0	0	1	2	1	0	0	1	0	0
6	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	1	0	0	0	0	0
Total	10	10	10	10	10	10	10	10	10	10
Mean	2.1	2.3	1.8	2.3	2.8	2.2	2.0	2.3	2.3	1.8
Var	1.21	0.9	2.18	2.46	3.51	1.29	2.22	1.79	0.68	0.62

This is certainly not the same value as: $s_w^2(\text{betw})_{10} = 0.854$, but as we said before, we do not want to calculate an ANOVA or an F-test.

We only want to compare the variance that is due to the length variation of the words within a chain to the variance due to the length variation between the chains.

variance for 10-word-chains
 between within chains
 0,854 1,685

If we do the same for the 3-word-chains, there is a noticeable difference:

3-word-chains:

Calculating the variance within the chains, we sum up the last row of Table 11 (the single variances of the 3-word-chains) and get $s^2(\text{in})_3 = 12.664$. If we standardize, we get 1.266 as the average variance within the chains.

We compare to the variance between the chains which is $s^2(\text{betw})_3 = 4.844$ (see Table 9.1) then standardized as average variance between: $s_w^2(\text{betw})_3 = 1.615$ (see Table 9.2)

Variance for 3-word-chains
 between within chains
 1.615 1.266

Table 11
Distributions of word lengths (number of syllables) in 3-word-chains:

Syll	Chain no.									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	0	0	1	2	3	1	2
2	1	1	2	2	1	1	1	0	0	1
3	0	0	0	0	2	1	0	0	0	0
4	1	1	0	1	0	0	0	0	1	0
5	0	0	0	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0
Total	3	3	3	3	3	3	3	3	3	3
Mean	2.33	2.33	1.67	2.67	2.67	2	1.33	1	3.33	1.33
Var	2.33	2.33	0.33	1.33	0.33	1	0.33	0	4.33	0.33

We observe that the variance between the 3-word-chains exceeds the variance within them, which means that the 3-word-chains are more "individual": there are typical longer chains and typical shorter ones, and for the length of a word it is more important in which 3-word-chain it occurs than the fact that it occurs in a 3-word-chain. And conversely: 10-word-chains determine the lengths of their words more than 3-word-chains do. The 10-word-chains are more similar to one another than the 3-word-chains are. In 10-word-chains there seems to be a balancing influence that effects the lengths of their words.

Because of that we take a look at the variance of all ten-word-chains in the entire text. We divided the complete text (519 words) into 52 ten-word-chains. Their lengths are given in Table 12.

Table 12
Lengths of all 52 ten-word-chains

Length (number of syllables) x	Frequency f_x	Proportion observed f_x/n
14	1	0.019
15	4	0.077
16	6	0.115
17	2	0.038
18	6	0.115
19	6	0.115
20	4	0.077
21	3	0.058
22	3	0.058
23	9	0.173
24	5	0.096
25	1	0.019
26	1	0.019
27	1	0.019
Sum 1040	52	1.00
Mean = 20.00, Std. dev. = 3.331, Variance = 11.098		

As it became already visible in Table 7, the length of 23 syllables is a preferred length for a 10-word-chain. Another typical length seems to be the number of 16, 18 or 19 syllables. Half of the observed chains are showing one of these lengths.

The observed length of a 10-word-chain ranges between 14 and 27 syllables, as illustrated in Fig. 3.

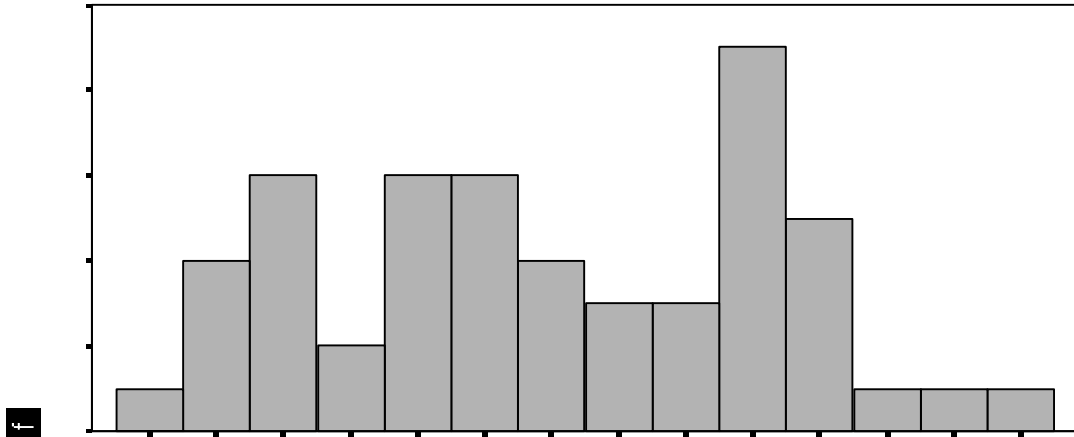


Fig. 3. Syllable numbers of all 52 ten-word-chains

We want to compare the variance between the chains and the variance within them for the entire text: The variance between all 52 chains is given in Table 12: $s^2(\text{betw})_{10} = 11.098$

If we divide it by the number of words per chain (here: 10 words), we get the variance between the chains per word on average: $s_w^2(\text{betw})_{10} = 1.1098$

To get the variance within the chains, we have to sum up the 52 single variances (data see appendix): $s^2(\text{in})_{10} = 63.556$ and divide it by the number of chains, so we get the mean variance within a chain as: $s_w^2(\text{in})_{10} = 63.556 : 52 = 1.222$.

variance for all 52 ten-word-chains	
between	within chains
1.1098	1.222

The variance within the 10-word chains exceeds the variance between the 10-word-chains. In a classical test of analysis of variance, one usually would try to corroborate the hypothesis that the units (here for example: the chains) are differing significantly from one another by showing that the variance between them is significantly greater than the variance within them. Here it is the reverse. Not only are we unable to find the variance between the chains significantly greater, it is in fact even smaller than the variance within the chains. Furthermore, it is also smaller than the total variance of all the words in the text that equals the variance within (see Table 1).

So we can state that the chains are more similar to one another than the words within the chains, and more similar than could be expected by the total variance of word lengths in the text as a whole.

9. Constancy of the proportion of two-syllable-words

In search of possible causes for this constancy we want to take a second look at the phenomenon of Table 4: the constant proportion of two-syllable-words.

We determined the occurrence of them in all of the 52 ten-word-chains and compared their proportion to the values of the binomial distribution with $n = 10$ and $p = 0.374$ (the proportion of two-syllable-words in the text, see Table 1 above). Values are given in Table 13:

Table 13
Occurrence (number x) of two-syllable-words in a 10-word-chain

Occurrence of 2-syllabic words in a 10-word-chain x	Cases (number of chains) f_x	P_{observed} f_x/N	P_{exp}	Expected number (binomial) NP_{exp}
0	0	0.0	0.009	0.48
1	3	0.057	0.055	2.87
2	8	0.154	0.148	7.72
3	8	0.154	0.236	12.30
4	20	0.385	0.247	12.86
5	8	0.154	0.177	9.22
6	4	0.078	0.088	4.59
7	1	0.019	0.030	1.57
8	0	0.0	0.0067	0.35
9	0	0.0	0.0009	0.05
10	0	0.0	≈ 0	0.003
Sum	52	1.00	1.00	52

Our aim was not to fit the distribution or to determine the goodness of fit. But we can see that everything happens as to be expected – with the exception of the case of four 2-syllable words. Again we find (see row no. 4) that this event is far more frequent than expected by chance. Nearly half of the ten-word-chains contain exactly four two-syllable-words. This is deviating clearly from the expectable proportion.

To be sure that this result is not due to the fact of a too small sample:

Taking the theoretical probability from the binomial distribution as $P = 0.247$ and using it as the parameter p we calculate the probability for such a result to be created at random:

In two out of five chains ($n = 5$ and $x = 2$) a probability would result which is $P = 0.260$, so this could be possible.

Even in 5 out of 13 chains (a quarter of the sample) ($n = 13$ and $x = 5$) the probability would yield $P = 0.1222$ which is still conceivable.

But in 20 out of 52 chains we get $P = 0.0103$ which has to be considered as very improbable to be produced by chance.

Ten-word-chains with just 4 two-syllable-words are definitely preferred compared to the random situation ($P(X = 4) = 0.247$).

For every 10 words, a rather constant pattern can be found that cannot be explained by chance.

10. Conclusion

Length balance in a text can be proven and characterized by the measures of r_A and B .

The index of length homogeneity r_A indicates the degree of intercorrelation of the proportions of word lengths in parts of the text. In Proust's sentence, length homogeneity $r_A = 0.9777$ with $t = 5$ (individual limit).

We suppose that for every author (or text, sort of text, time, genre, style etc.) there is a style characteristic rhythm number B , so that in every B words the lengths of the resulting word-chains are balanced out and tend to be constant. In Proust's longest sentence, $B = 10$.

Further research should be done to corroborate the observation of B in any texts and to investigate the exceptional regularity of the two-syllable-words.

Length homogeneity r_A as a measure for the reliability of assessment in recording word length proportions should be determined, at least at split half level ($t = 2$), when investigating frequency distributions of word lengths in texts.

References

- Altmann, G.** (1980). Prolegomena to Menzerath's Law. In: Grotjahn, R. (ed.), *Glottometrika 2, 1-10*. Bochum: Brockmeyer.
- Altmann, G.** (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.
- Altmann, G. & Best, K.-H.** (1996). Zur Länge der Wörter in deutschen Texten. In: Schmidt, P. (ed.) *Glottometrika 15, 166-180*. Trier: Wissenschaftlicher Verlag Trier.
- Altmann, G. & Schwibbe, M.H.** (1989). *Das Menzerathsche Gesetz in informationsverarbeitenden Systemen*. Hildesheim: Olms.
- Altmann-Fitter** (1994). Lüdenscheid: RAM-Verlag.
- Altmann-Fitter** (1997). *Iterative Fitting of Probability Distributions*. Lüdenscheid: RAM-Verlag.
- Andersen, S.** (2002). Freedom of choice and the psychological interpretation of word frequencies in texts. *Glottometrics 2, 45-52*.
- Best, K.-H.** (ed.) (1997). *The Distribution of Word and Sentence Length (= Glottometrika 16)*. Trier: Wissenschaftlicher Verlag Trier.
- Best, K.-H.** (ed.) (2001). *Häufigkeitsverteilungen in Texten*. Göttingen: Peust & Gutschmidt.
- Best, K.-H.** (2003). *Quantitative Linguistik. Eine Annäherung*. Göttingen: Peust & Gutschmidt.
- Botton, A. de** (1997). *How Proust Can Change Your Life*. London: Picador Macmillan. (1998). *Wie Proust Ihr Leben verändern kann*. Frankfurt: S. Fischer.
- Grzybek, P.** (ed.) (2005). *Word length studies and related issues*. Boston/Dordrecht: Kluwer.
- Hřebíček, L.** (2000). *Variation in sequences*. Prague: Oriental Institute.
- Orlov, Ju. K.** (1982). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Orlov, Ju.K., Boroda, M. G., & Nadarejşvili, I.Ş., *Sprache, Text, Kunst. Quantitative Analysen: 118-192*. Bochum: Brockmeyer.
- Wimmer, G., Köhler, R., Grotjahn, R., & Altmann, G.** (1994). Towards a theory of word length distribution. *Journal of Quantitative Linguistics 1, 98-106*.

- Wimmer, G. & Altmann, G.** (1996). The theory of word length distribution: some results and generalizations. In: Schmidt, P. (ed.), *Glottometrika 15*, 112-133. Trier: Wissenschaftlicher Verlag Trier.
- Zipf, G. K.** (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley.

1. Appendix

I.

Proust's longest sentence from his work "A la recherche du temps perdu" in German translation

(detected by Alain de Botton 1997; 1998)

Diejenigen der alten Verdurinschen Möbel, die hier, manchmal sogar unter Beibehaltung einer bestimmten Anordnung, erneut Platz gefunden hatten und denen ich selbst in La Raspelière wiederbegegnet war, fügten in den gegenwärtigen Salon Teile des alten ein, die augenblicksweise mit nahezu halluzinatorischer Deutlichkeit jenen früheren noch einmal heraufbeschworen, gleich darauf aber fast unwirklich schienen, weil sie inmitten der umgebenden Wirklichkeit Bruchstücke einer untergegangenen Welt, die man an einem andern Orte wähte, wiedererstehen ließen: ein aus Träumen entstiegene Kanapee zwischen neuen, sehr wirklichen Sesseln, kleine, mit rosa Seide bezogene Stühle, eine durchwirkte Tischdecke auf dem Spieltisch, die zur Würde einer Person erhoben schien, denn wie eine Person besaß sie eine Vergangenheit, ein Gedächtnis, behielt sie doch im kalten Dunkel des Salons am Quai Conti jene Bräunung bei, welche die durch die Fenster der Rue Montalivet einfallende Sonnenstrahlung (deren genaue Stunde die Decke ebenso gut kannte wie Madame Verdurin selbst) bewirkt hatte, sowie die, die durch die Glasfenster der Gegend bei Deauville sich ergoß – wohin man jenes Requisit mitgenommen und wo es den ganzen Tag über den Blumengarten hinweg das tiefe Tal überschaut hatte in Erwartung der Stunden, da Cottard und der Geiger ihre Kartenspiele absolvieren würden – oder auch ein Strauß aus Veilchen und Stiefmütterchen in Pastell, Geschenk eines befreundeten großen Künstlers, der seither verstorben war, einziges hinterbliebenes Fragment eines Lebens, das sonst keine Spuren hinterlassen hatte; jetzt sprach nur dieses Bild noch – in ganz summarischen Zügen – von einem großen Talent und von einer langen Freundschaft, als einziges Überbleibsel erinnerte es noch an Elstirs sanften Blick, an die schöne, füllige und traurige Hand, mit der er immer gemalt hatte; ein gefälliges Durcheinander, eine Wirrnis aus Geschenken der Getreuen, die der Hausherrin überallhin gefolgt waren und schließlich die feste Prägung eines Charakterzuges, einer Schicksalslinie angenommen hatten, eine Fülle von Blumensträußen und Pralinenschachteln, die hier wie dort in einer ganz gleichen Art von üppigem Wachstum wuchernd sich entfalteten; eine merkwürdige Einsprengung aus sonderbaren und überflüssigen Objekten, jenen Dingen, die noch aussehen, als kommen sie eben erst aus der Verpackung hervor, in der sie als Geschenk überreicht worden sind, und die das ganze Leben hindurch bleiben, was sie zunächst gewesen sind, nämlich Geschenke zum 1. Januar, alle jene Gegenstände endlich, die man von den anderen nicht hätte trennen können, die aber für Brichot, den alten Besucher der Verdurinschen Feste, eine Patina und Weichheit bekommen hatten, wie sie Dingen eigen sind, denen ein geistiges Abbild ihrer selbst in unserem Innern eine Art von Tiefe hinzuzufügen scheint – alles dies ließ perlend in ihm jeweils Töne erwachen, welche in seinem Herzen geliebte Anklänge weckten: verworrene Erinnerungen,

die gerade hier in diesem ganz und gar die Gegenwart verkörpernden Salon, indem sie vereinzelt Lichtflecke schufen – so wie an einem schönen Tage die Sonne im Viereck geradezu in die Atmosphäre eines Raumes hineingezeichnet – die Möbel und Teppiche gleichsam ausschnitten und mit einer Rahmenlinie umzogen, wobei sie von einem Kissen zu einer Blumenvase, einem Hocker zu einem noch lose anhaltenden Duft, einer Beleuchtungsart zu einem Vorherrschen bestimmter Farben hinübereilten und in plastischer und gleichzeitig beseelter Gestalt eine Form vor Augen rückten, welche gleichsam die ideale, allen aufeinanderfolgenden Heimen anhaftende Urgestalt des Salons der Verdurins war.

Counting modalities:

French proper names were counted:

- syllable number by sound (Deauville = 2 syllables, Madame = 2 syllables)

- as entire word, if German translation would yield one (Rue Montalivet as "Montalivetstrasse" = 1 word, 5 syllables; Quai Conti as "Contiquai" oder "Contiufer" = 1 word, 3 syllables; La Raspeliere = 1 word, 5 syllables; but Madame Verdurin = 2 words, 2 and 3 syllables: "Frau Verdurin")

II.

52 ten-word-chains c with lengths of the words (number of syllables)

c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
4	4	1	2	3	3	6	2	2	1
1	2	1	2	2	2	1	1	2	1
2	3	1	1	3	1	1	1	1	2
4	3	5	2	1	1	1	2	2	1
2	2	5	1	2	3	1	4	2	1
1	1	1	1	5	1	2	3	4	2
1	3	2	5	1	4	2	2	2	2
2	2	1	1	2	3	2	2	2	2
2	1	1	3	2	3	2	1	3	3
2	2	5	7	1	3	5	3	3	1
c11	c12	c13	c14	c15	c16	c17	c18	c19	c20
1	2	2	4	1	1	2	1	1	2
1	1	2	4	2	3	3	4	2	2
2	1	1	2	3	1	4	2	1	1
2	1	2	3	1	2	1	1	2	1
2	2	1	2	2	1	1	2	1	1
1	2	1	1	2	2	1	1	1	1
2	1	1	2	2	1	1	3	2	2
4	2	2	3	1	2	2	2	2	1
1	1	1	1	1	2	1	1	4	4
3	3	5	2	1	1	2	3	4	1

c21	c22	c23	c24	c25	c26	c27	c28	c29	c30
2	3	2	2	1	1	2	1	1	2
2	5	1	1	3	1	2	3	2	1
2	2	1	2	4	2	2	1	2	4
4	2	1	2	4	3	1	1	2	1
2	2	2	2	1	1	4	3	5	5
2	1	1	1	1	3	4	4	2	1
1	1	1	1	1	1	2	2	4	1
2	2	1	2	2	1	2	2	4	1
3	2	1	2	2	1	1	1	2	1
1	4	4	2	1	1	3	2	2	1

c31	c32	c33	c34	c35	c36	c37	c38	c39	c40
2	2	1	3	1	3	2	1	2	1
1	4	1	2	1	1	1	2	3	2
2	3	3	1	1	2	1	1	1	1
1	1	1	1	2	3	1	2	2	3
1	4	2	1	2	1	1	1	3	2
3	1	1	1	2	2	3	2	2	2
2	5	2	2	2	3	1	3	1	1
2	3	1	3	1	2	2	1	1	1
1	2	1	2	1	2	2	4	2	3
4	2	1	1	2	4	2	2	2	2

c41	c42	c43	c44	c45	c46	c47	c48	c49	c50	c51	c52
2	1	3	1	2	2	1	1	2	3	2	7
1	1	2	1	1	4	3	1	1	2	1	2
1	2	4	1	1	1	2	2	2	5	1	4
2	2	5	3	1	1	3	2	4	1	2	3
5	3	1	4	2	4	1	1	1	1	2	1
1	2	3	2	2	2	1	2	2	3	2	2
2	1	1	2	2	2	2	4	4	1	2	1
1	2	1	1	1	5	4	2	1	3	1	3
1	2	2	4	2	1	3	2	2	3	3	1
2	3	1	3	1	2	2	1	3	2	2	1

III.

Variances of 52 ten-word-chains:

C1	1.211	C3	3.567	C5	1.511
C2	0.900	C4	4.056	C6	1.156

C7 3.122
C8 0.989
C9 0.678
C10 0.489
C11 0.989
C12 0.489
C13 1.511
C14 1.156
C15 0.489
C16 0.489
C17 1.067
C18 1.111
C19 1.333
C20 0.933
C21 0.767
C22 1.600

C23 0.944
C24 0.233
C25 1.556
C26 0.722
C27 1.122
C28 1.111
C29 1.600
C30 2.178
C31 0.989
C32 1.789
C33 0.489
C34 0.678
C35 0.278
C36 0.900
C37 0.489
C38 0.989

C39 0.544
C40 0.622
C41 1.511
C42 0.544
C43 2.011
C44 1.511
C45 0.278
C46 2.044
C47 1.067
C48 0.844
C49 1.289
C50 1.600
C51 0.400
C52 3.611

$\Sigma = 63.556$

$63.556 / 52 = 1.222$