Studies
in Quantitative Linguistics
10

Ioan-Iovitz Popescu
Radek Čech
Gabriel Altmann

The Lambda-structure of Texts

RAM - Verlag

The Lambda-structure of Texts

by

Ioan-Iovitz Popescu
Radek Čech
Gabriel Altmann

2011
RAM-Verlag

Studies in quantitative linguistics

1. U. Strauss, F. Fan, G. Altmann, *Problems in quantitative linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in quantitative linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G.Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp

RAM-Verlag
Stüttinghauser Ringstr. 44
D-58515 Lüdenscheid
RAM-Verlag@t-online.de
http://ram-verlag.de

Preface

The problem of frequency structuring of a text is not only very old but it has at present a great number of different aspects. The most popular ones are the studies of vocabulary richness, type-token ratio, rank-frequency distributions and the frequency spectrum, to mention only some of them. Vocabulary richness is a central property of the text useful in the study of language learning, forensic linguistics, style studies, the literary development of a writer etc. Many researchers tried to find a relationship between the number of types and that of tokens (text length), but even if sometimes they succeeded to determine the relation, in the formula a variable remained whose sampling distribution was not known. What is the expected value of V (vocabulary size) and the text length (N)? And even if text length can sometimes be determined in advance (e.g. for a press article), the vocabulary size cannot. How can the standard deviation of the vocabulary size be derived? The answer has never been given and nobody has tried to solve the problem, not even empirically.

The present study shows that if we descend a level deeper, viz. from the vocabulary as a whole to its components, i.e. words and their frequencies, a stable indicator (called lambda) of frequency structure (*cum grano salis* the basis of vocabulary richness) can be set up which does not depend on *N* and whose variance can be asymptotically derived. This fact enables us to set up tests for comparing individual texts, individual authors, genres, and languages, to follow the deployment of a text and the evolution of a writer through years. It allows us to study the jumps in the individual chapters/parts of a text and to express quantitatively different aspects of text dynamics.

Needless to say, even if we exemplify the study using 1185 texts in 35 languages, the research is not finished. Many more texts must be analyzed, new aspect should be discovered and for every aspect test procedures must be devised. Further, frequency structure is not an isolated property. It is associated with other different properties, but first such a connection must be hypothesized and the other properties must be quantified, too, before we begin to set up hypotheses. It can be conjectured that frequency structure is also an element of Köhler´s control cycle but the way to show it will be very long.

Acknowledgements

We are very much obliged to Fazli Can, Fengxiang Fan, Emmerich Kelih, Viktor Krupa, Ján Mačutek, Haruko Sanada, Claudiu Vasilescu and Eric Wheeler who in different ways helped us to prepare this volume. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157.

Contents

1. Introduction

The history of the study of vocabulary richness is an epos describing the battle between vocabulary size and text length (cf. Baayen 1989; Bernett 1988; Brunet 1978; Carroll 1964; Cossette 1994; Covington, McFall 2010; Ejiri, Smith 1993; Guiraud 1954, 1959; Herdan 1960, 1966; Hess, Sefton, Landry 1986, 1989; Holmes 1991, 1992; Holmes, Forsyth 1995; Honore 1979; Köhler, Galle 1993; Kuraszkiewicz 1963; Malvern, Richards 2002; Martel 1986; Martynenko 2010; Menard 1983; Muller 1968, 1971; Müller D. 2002; Müller W. 1971; Orlov 1983; Panas 2001; Ratkowsky, Halstead, Hantrais 1980; Richards 1987; Schach 1987; Serant, Thoiron 1988; Sichel 1975, 1986; Thoiron 1986, 1988; Thoiron, Labbé, Serant 1988; Suprun 1979; Nešitoj 1975; Tešitelová 1972; Tuldava 1977, 1995; Tweedie, Baayen 1998; Wachal, Spreen 1973; Weitzman 1971; Woronczak 1965; Yule 1944 – to mention only some of them). In a previous research, Popescu, Mačutek and Altmann (2010) stated that vocabulary richness can be estimated only if one eliminates the detrimental factor of text length by some transformation. This has been tried in many cases but the solutions did not seem to be satisfactory. The same is the fate of all the indicators of other text properties depending on text length. A typical case is the restriction to the frequently used relativized hapax legomena (V_1/N) as the indicator of richness. Howvere, it has several flaws: (i) hapaxes are not the exclusive indicators of richness; the same holds for dislegomena, etc. (ii) Very short texts can get maximal richness 1, while very long texts in which all words are repeated would have richness 0; texts of intermediate length, say from N = 100 to N = 100000 would still depend on N. Hence taking a special class of words is not sufficient; indicators like V_1/N do not work correctly for any text length even if it is possible to set up a test for comparison for texts of moderate length, i.e. texts that are not too short and not too long. But what is too short and too long? Trying to overcome this problem Popescu, Altmann, Grzybek et al. (2009: 31-34), Tuzzi, Popescu, Altmann (2010) and Popescu, Mačutek, Altmann (2010) used a formula based on Popescu´s (2007) *h*-point. For other different uses of hapax legomena see Popescu, Mačutek, Altmann (2009) where the dependence of hapax legomena on the arc length is shown.

Since hapax legomena are linearly associated with the arc length L joining the highest frequency f_1 with the smallest frequency f_V of the rank-frequency distribution of words and thus capturing the complete vocabulary of the text, Popescu, Mačutek and Altmann (2010) studied the behaviour of the arc length. It can be computed as

$$(1.1) \quad L = \sum_{i=1}^{V-1} [(f_i - f_{i+1})^2 + 1]^{1/2}$$

where f_i are the ordered absolute frequencies (i = 1,2,…V-1) and V is the highest rank (= vocabulary size), i.e. L consists of the sum of Euclidean distances between ranked frequencies. It is not a smooth line; it is a discrete arc.

It has been stated (Popescu, Mačutek, Altmann 2009) that the arc length L formed by the distances of the ranked-frequencies of words in a text also depends on the text length N. The arc length develops as an increasing power function of N; and relative to N, i.e. L/N, it has a form of a decreasing power curve. A linearization e.g. by taking logarithms, may yield a straight line but in this case N remains the independent variable because the straight line is not horizontal. Our aim is, however, to get rid of N in such a way that a transformation of L yields a horizontal straight line. To this end we propose the indicator lambda

$$(1.2) \quad \Lambda = \frac{L(Log_{10}N)}{N}$$

For the data analyzed in the above mentioned article the straight-line slope of this indicator was only –0.0001 but there was only a small number of texts longer than N = 10000. The aim of the present book is to add such texts to our previous computations and use the new indicator for comparing texts, genres and languages. Evidently, the comparison of texts proceeds on a very abstract level (distances between frequencies of the ranks of individual words) but we may give some interpretations based on the results.

In the following chapters we shall analyze 1185 texts (see Appendix) and study the control of frequency structure in one text, one genre, and one language. Furthermore, we shall compare individual texts consisting either of one or of several parts and propose some tests for future investigation.

Writing a text the writer cares for his style but cannot consciously control the rank frequency distribution of words. If in spite of this lack of control one finds some commonalities in all texts pointing in the same direction, one is on the trail of a background mechanism for which a law could be established. Similar phenomena are e.g. the Skinner effect of repetition of some conspicuous elements (Skinner 1939, 1957) or non-conscious word length manipulation in sentence or clause positions. One can imagine a number of factors giving rise to this mechanism: (i) in individual texts, the ability of the author in using his language, as well as the author´s talent, and experience; (ii) in certain genres, the object of description or aesthetical grounds; (iii) in language, the type of grammar “in use”. In case (i) the author´s ability (or talent) is, of course, a relative notion; it can be evaluated only in relation to other authors and, moreover, the judgement makes sense only if the texts of the same theme or “genre” (e.g., a scientific article, poetry, short story) are taken into account. Further, in the case of children’s texts, the age of the intended writer/narrator should be a factor influencing this ability. However, it must be emphasized that lambda is not a direct indicator of the author’s ability or talent. Lambda expresses the structure which emerges as a result of language usage. And an author who uses for example a more synthetic form of the given language – which is indicated by a higher lambda – is not a better language user than the author who uses a more analytic form and vice versa. To be choleric is not "better" than to be stoic (or melancholic or sanguine) just as to have a higher lambda value does not mean to be a "better" author. Lambda expresses a property which can be in relationship to some qualitative characteristics of text but in itself it is no "quality indicator".

In case (iii) a language can be highly synthetic but if the "grammar in use" does not prefer synthetic expressions, our judgement can be false. For example, one could expect that the existence of the past participle and its forms in the Czech grammar should influence Czech to be more synthetic. However, the use of the past participle is so rare that it has no real influence on the typological judgement of Czech, based on the "grammar in use" approach. Hence, we may conjecture the existence of a mechanism but to find all its generating factors is a work for centuries.

Of course, all this must be in the future incorporated in a systemic view of text, but presently we treat *membra disiecta*, and restrict ourselves to a special aspect of the frequency structure of the text.

Lambda cannot be considered directly as indicating vocabulary richness; it is something more. It takes into account not only the extent of the vocabulary – necessary for computing the arc length *L* – but at the same time also the relationship of individual frequencies to their next neighbours in the rank-frequency sequence. If we compare lambda with some indicators like Repeat rate or Entropy, we see the difference: Repeat rate is a measure of concentration, taking into account the frequencies and *V* which yields also its maximal value 1/*V*; Entropy is a measure of uncertainty or dispersion, taking into account the frequencies and *V* which in its logarithmic form *ld V* is its maximal value. Both are global measures. However, lambda takes into account also the differences between neighbouring (ranked) frequencies, that is, not only the use of words but also their respective surplus over the next lower frequency. Hence Repeat rate and Entropy can (*cum grano salis*) be interpreted as vocabulary richness, viz. the smaller the relativized Repeat rate, the greater the richness, and the greater the relativized Entropy, the greater the vocabulary richness. Lambda can be used in the same way, but the problem is a little bit more complicated.

Consider first some relations of lambda to some other text indicators. In a previous publication (Popescu, Kelih et al. 2010) we defined the vector

$$(1.3) \quad P = \left(\frac{f(1)}{h}, \frac{V}{h} \right),$$

containing only the greatest frequency *f*(1), the vocabulary size *V* and the *h*-point defined as

$$(1.4)\quad h = \begin{cases} r, & if\ \ there\ is\, an\ \ r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & if\ \ there\ is\ no\ r = f(r) \end{cases} .$$

i.e. the h-point as that point at which the rank is equal to the frequency, $r = f(r)$. If there is no such point, one takes, if possible, two neighbouring $f(i)$ and $f(j)$ such that $f(i) > r_i$ and $f(j) < r_j$. Mostly $r_i + 1 = r_j$.

These three quantities characterize the rank-frequency distribution by three points. The modulus of P is defined as

$$(1.5)\quad M = \left(\left(\frac{f(1)}{h}\right)^2 + \left(\frac{V}{h}\right)^2\right)^{1/2} = \frac{1}{h}\left(f(1)^2 + V^2\right)^{1/2}.$$

The text size N is not contained in this expression but it can be added for the sake of normalization in the form

$$(1.6)\quad A = M/\mathrm{Log}_{10}N.$$

Evidently, M is a "simplification" of (1.1) taking into account only the two extreme points of the arc in relation to the h-point. Hence there must be some relation between lambda and A (or M). Taking 498 texts in 28 languages we obtain the simple relation

$$\Lambda = 1.2477\ \mathrm{Log}_{10}(1.0208A)$$

yielding $R^2 = 0.63$. The relation would surely change slightly by adding further texts but the dispersion would increase, too. The given relationship could be obtained from a homogeneous differential equation of second order but until further evidence and the knowledge of boundary conditions are given we content ourselves with this preliminary statement. The result is presented in Figure 1.1.

There are other possibilities of eliminating the influence of N and obtaining a horizontal straight line independent of N. For example the product of Λ and A yields

$$(1.7)\quad A\Lambda = LM/N$$

and yields an indicator which could be used for different comparisons but has a very complex sampling distribution. In Figure 1.2 we merely show the result of combining A and Λ.

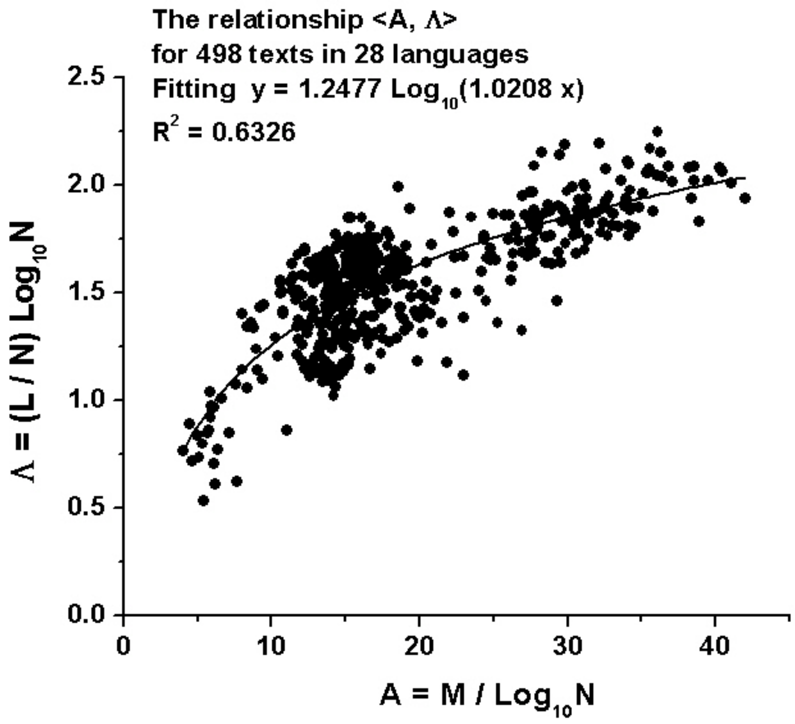


Figure 1.1. The relation of *Λ* to the Adjusted Modulus *A*

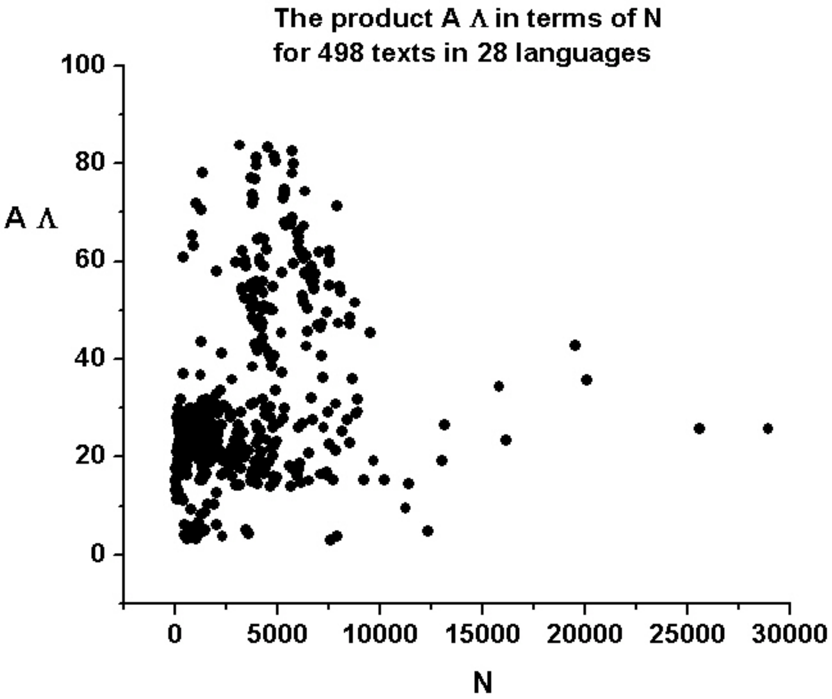


Figure 1.2. The product *AΛ*

Evidently, lambda has some relation to every indicator used for characterising either vocabulary richness or the frequency structure of a text or typological properties of a language. We refer to Chapter 6 in Popescu, Mačutek, Altmann (2009) where different indicators can be found.

Here we show only the relation of lambda to the richness indicator B_6 (ibidem 2009: 106). It is defined as

$$(1.8)\quad B_6 = \frac{c}{(V - HL/2)^a}$$

where V is the vocabulary size, HL is the number of hapax legomena and a and c are the parameters of the the Zipfian power function $f(r) = c/r^a$ fitted to the rank-frequency data. The denominator takes into account the vocabulary and the half of the hapax legomena. B_6 is especially good for typological ordering of languages on the synthetism/analytism scale.

We restricted the sample to 100 texts in 20 languages and captured the relation by means of the special exponential function which has been proposed also as a substitute for Zipf´s law (c.f. Popescu, Altmann, Köhler 2009), here

$$(1.9)\quad \Lambda = A + C\exp(-B_6 / D)$$

where A, C and D are parameters. In our case we obtained Λ = 0.29696 + 2.2745exp($-B_6$/1.86016) as can be seen in Figure 1.3.

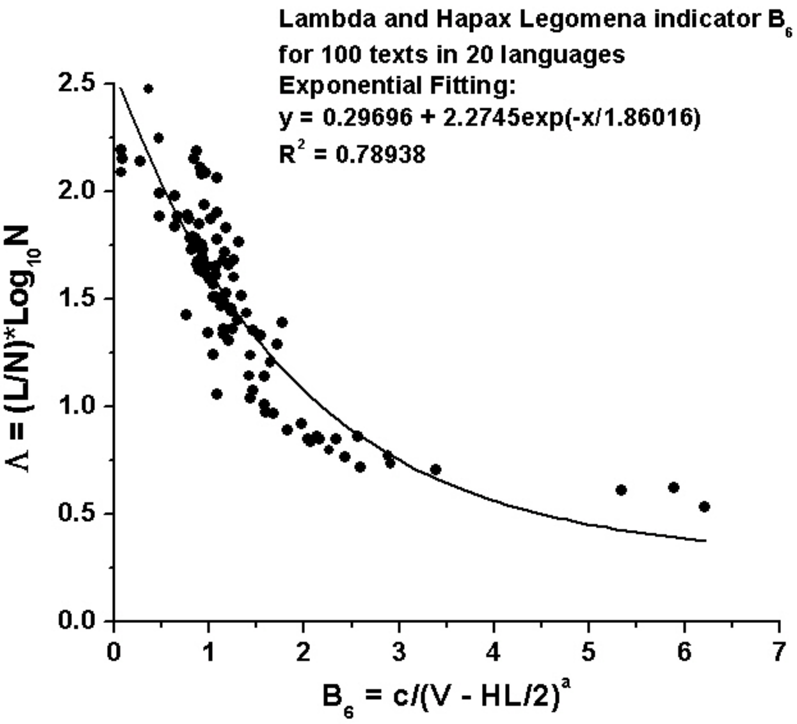


Figure 1.3. The relation of lambda to the richness indicator B_6

It is to be assumed that taking into account more texts and more languages the dispersion around the function could increase, but further exponential components proposed for stratification in the above mentioned publication and the reduction of the texts to one genre could adequately capture this relationship. Lambda seems to express the complete structure of the rank-frequency sequence while the other indicators used for special purposes express individual properties.

In Popescu, Altmann, Grzybek et al. (2009: 29-34) another richness indicator has been shown which does not take into account the hapax legomena at all but uses only the *h*-point and the cumulative relative frequencies, *F*(*h*), up to the *h*-point. Since *F*(*h*) encompasses mostly the proportion of synsemantics, 1 – *F*(*h*) encompasses not only hapax legomena but, as a matter of fact, almost all autosemantics up to the *h*-point, yielding a more realistic picture of vocabulary richness. It has been defined as

(1.10)
$$R_1 = 1 - \left(F(h) - \frac{h^2}{2N} \right).$$

Since $h^2/(2N)$ is a constant for the given text, R_1 yields a simple possibility of comparing texts. Now, it can be shown that lambda has a very direct relation to R_1. In Table 1.1 the values of R_1 and Λ are presented. For R_1 we used the values of Table 3.6 in Popescu, Altmann, Grzybek et al. (2009: 31ff) for 100 texts in 20 languages and for Λ the same texts in the Appendix, and ordered them according to increasing lambda. The identity of individual texts is here irrelevant but can be traced back if necessary.

The relationship is presented graphically in Figure 1.4. It can be shown that using the preliminary function

(1.11) $R_1 = 1 - 0.9455\exp(-1.0140\Lambda)$,

which converges to 1 sufficiently describes the relationship (R^2 = 0.80). The values computed according to (1.11) can be seen in the third column of Table 1.1.

Table 1.1
Lambda and vocabulary richness R_1

Λ	R_1	R_{1th}		Λ	R_1	R_{1th}		Λ	R_1	R_{1th}
0.5336	0.4566	0.4496		1.3592	0.8058	0.7617		1.7498	0.8445	0.8396
0.6071	0.4510	0.4891		1.3605	0.7027	0.7620		1.7505	0.8900	0.8397
0.6215	0.4308	0.4965		1.3869	0.7159	0.7683		1.7640	0.9025	0.8419
0.7031	0.5713	0.5365		1.4009	0.6660	0.7716		1.7767	0.7995	0.8439
0.7161	0.6614	0.5426		1.4250	0.6771	0.7771		1.7805	0.8956	0.8445

0.7327	0.5468	0.5502		1.4337	0.7666	0.7790		1.7831	0.8491	0.8450
0.7615	0.5957	0.5631		1.4416	0.6647	0.7808		1.7850	0.8088	0.8453
0.7725	0.5523	0.5680		1.4533	0.8114	0.7834		1.8314	0.8792	0.8524
0.7964	0.5846	0.5783		1.4631	0.8876	0.7855		1.8342	0.822	0.8528
0.8342	0.5937	0.5942		1.4858	0.8551	0.7904		1.8473	0.8633	0.8547
0.8463	0.6653	0.5991		1.4898	0.8224	0.7913		1.8505	0.8676	0.8552
0.8466	0.612	0.5993		1.5071	0.7545	0.7949		1.8698	0.8043	0.8580
0.8469	0.6169	0.5994		1.5073	0.8619	0.7949		1.8700	0.9026	0.8580
0.8591	0.6083	0.6043		1.5136	0.7385	0.7962		1.8799	0.8499	0.8594
0.8601	0.5343	0.6047		1.526	0.8747	0.7988		1.8826	0.8143	0.8598
0.8863	0.6521	0.6151		1.5649	0.875	0.8066		1.8900	0.8513	0.8609
0.917	0.6231	0.6269		1.5889	0.8485	0.8112		1.9029	0.8866	0.8627
0.9644	0.6744	0.6444		1.5996	0.7789	0.8132		1.9376	0.8936	0.8674
0.9745	0.5940	0.6480		1.6067	0.9227	0.8146		1.9783	0.8635	0.8728
1.0086	0.6272	0.6600		1.6222	0.8371	0.8175		1.9914	0.8402	0.8745
1.0395	0.6805	0.6705		1.6314	0.8777	0.8192		2.0599	0.9186	0.8829
1.054	0.5765	0.6753		1.6505	0.8675	0.8226		2.0783	0.9271	0.8851
1.072	0.6350	0.6811		1.6524	0.8297	0.8230		2.0843	0.8993	0.8858
1.1411	0.6396	0.7027		1.6549	0.7497	0.8234		2.0874	0.7857	0.8861
1.1417	0.6836	0.7029		1.6601	0.7953	0.8244		2.1098	0.9414	0.8887
1.2071	0.6132	0.7220		1.6773	0.8725	0.8274		2.1398	0.8221	0.8920
1.2368	0.8744	0.7302		1.6787	0.8491	0.8276		2.1492	0.8313	0.8930
1.2427	0.6898	0.7318		1.6794	0.8653	0.8278		2.1510	0.9195	0.8932
1.2855	0.8535	0.7432		1.6808	0.8469	0.8280		2.1863	0.9343	0.8970
1.3031	0.6811	0.7477		1.6850	0.8486	0.8287		2.1904	0.8175	0.8974
1.3312	0.695	0.7548		1.7138	0.7945	0.8337		2.2481	0.8718	0.9032
1.3377	0.8667	0.7564		1.7271	0.8300	0.8359		2.4749	0.9087	0.9231
1.3419	0.8037	0.7575		1.7307	0.8697	0.8365				
1.3527	0.7197	0.7601		1.7317	0.7644	0.8367				

We suppose that if further hundreds of texts were evaluated, the two parameters would both attain the value of 1, or at least, the first parameter would attain the inverse value of the second one, but this is merely speculation. For the time being we cannot substantiate this relationship linguistically. It has been chosen because of its simplicity and convergence. That means that, computing lambda, we have to deal not only with the rank-frequency structure of the text but also with its vocabulary richness, which can be computed without recourse to the hapax legomena or some selected frequency classes. In general, the greater the value of lambda, the greater is the vocabulary richness of the given text. The approximate value of the richness can be obtained by transformation (1.11).

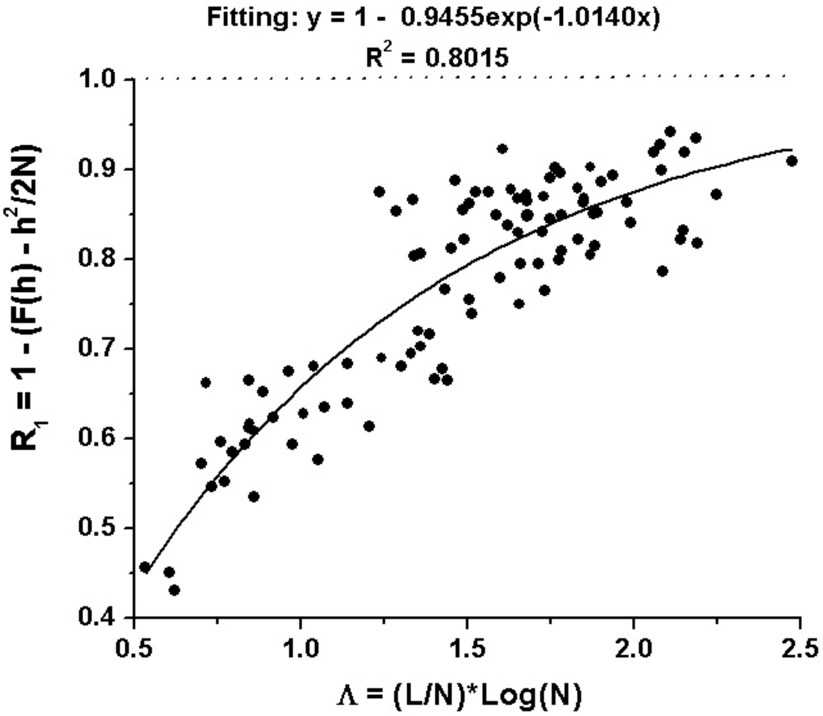


Figure 1.4. The dependence of R_1 on Λ

Since the greatest lambdas were observed in the Latin work by Horatius *Carmina liber III*, namely 2.6565 (not included in the Appendix), and in the Hungarian work *Forgószélben* by Árpás, namely 2.5518, it is perhaps a sign of the strong synthetism of these languages. However, these numbers give us an idea of the empirical maximum value of lambda which can be reached for mid-sized texts rich in hapax-legomena (that is high *L*), having *N* small enough to ensure *L/N* is close to unity, yet high enough to have still a high Log(*N*). At the other end of the scale, with the empirical minimum value of lambda observed so far, we find 0.5336 belonging to the Hawaiian text *Kawelo Mokuna III.* Again the lowest lambda values can be reached by mid-sized texts, this time poor in hapax-legomena (that is with low *L*).

The aim of this book is to show the use of lambda in the study of synchronic states of texts, in the comparison of texts, authors, genres, and languages, and in the study of diachronic processes concerning an author, a child or a whole language. In text or language, only laws represent stable, constant entities. Everything else, like pronunciation, words, rules of grammar, etc. is ephemeral and is kept in balance by the operation of laws. As can be seen in the following chapters, lambda is almost a constant with a small variation interval. The deviations are local phenomena which can be used for the characterisation of texts or their development.

2. Data

We used all the data analyzed in the article mentioned previously (Popescu, Mačutek, Altmann 2010a), i.e. 526 texts in 28 languages and added several other texts and 7 more languages in order to fill the relatively empty space for $N >$ 10000. In this way we obtained the Appendix and the results from 1185 texts as presented in Figure 2.1a,b,c. In order to show different perspectives we present lambda in three variants: (a) up to $N = 5000$ using 702 texts in 28 languages, (b) up to $N = 30000$ using 1116 texts in 35 languages, and (c) up to $N = 150000$ using 1185 texts in 35 languages. Of course, the domain between 100000 and 150000 is, again, not well filled but this is a problem of scaling and has only an illustrative function. More texts could only confirm our assumptions.

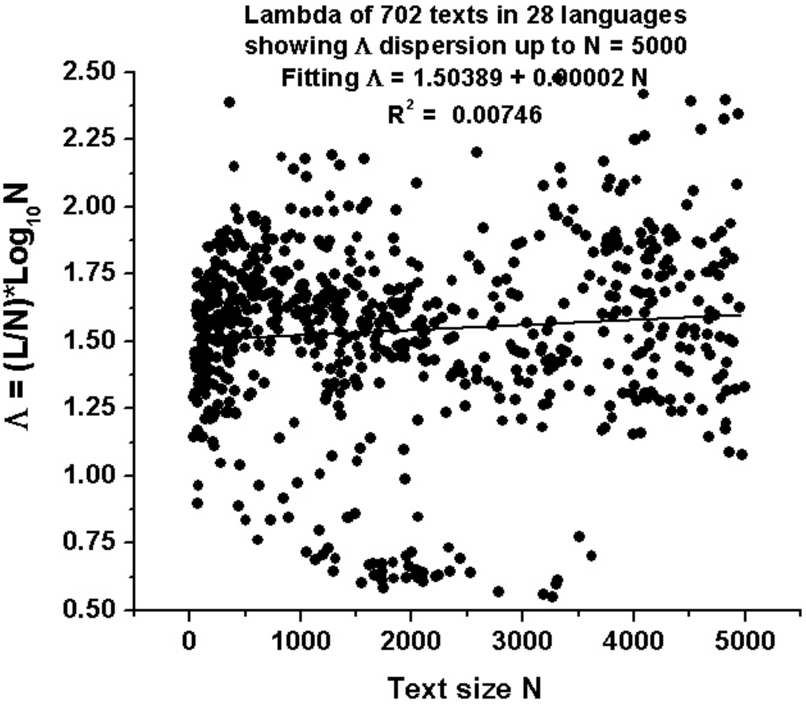


Figure 2.1a. Lambda in 702 texts up to $N = 5000$

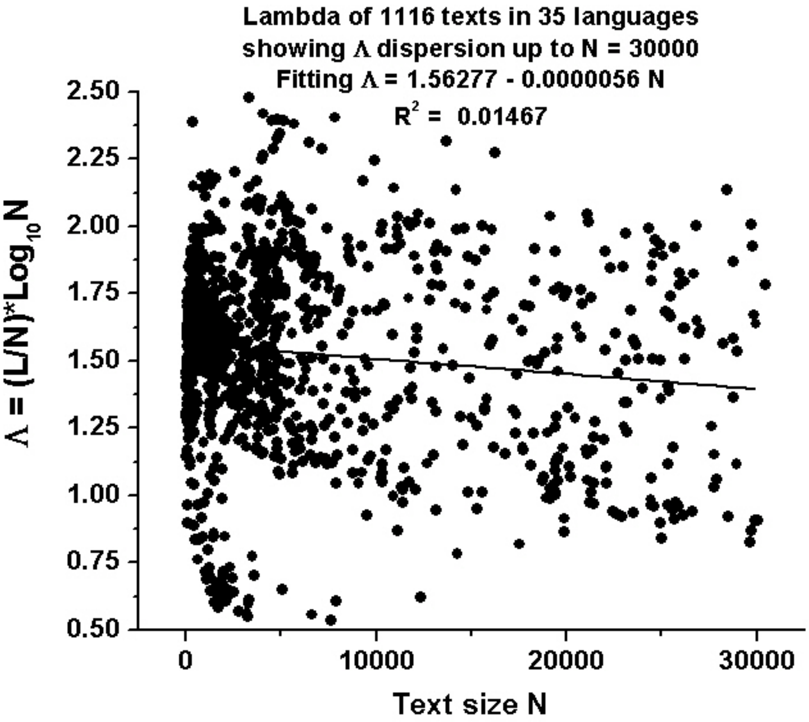


Figure 2.1b. Lambda in 1116 texts up to *N* = 30000

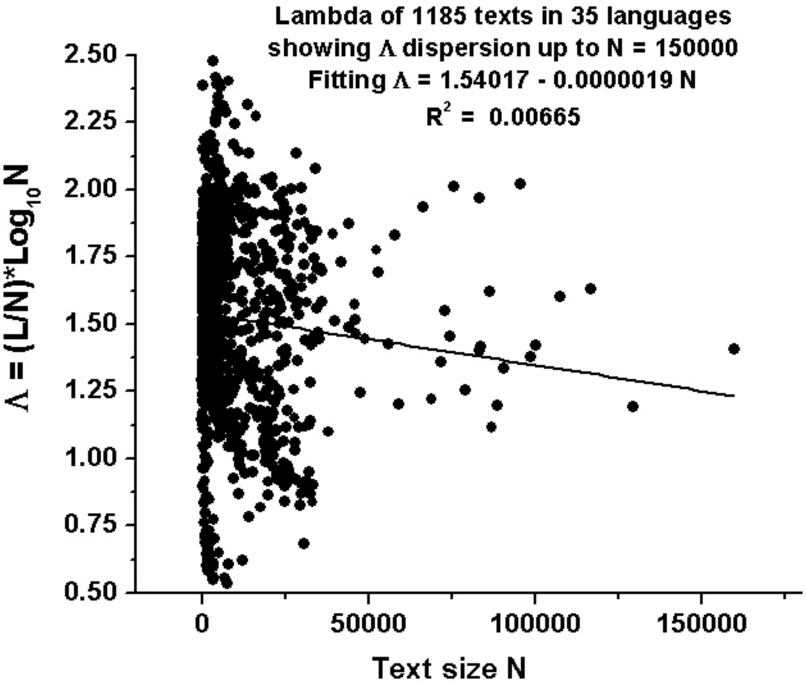


Figure 2.1c. Lambda in 1185 texts up to N = 150000

The data are fitted linearly with $\Lambda = a + bN$, and in Table 2.1 and Figure 2.2 one can see the course of the coefficient b of linear regression in terms of text size N. It converges very rapidly to zero.

Table 2.1
The slope of the linear regression of lambda against text size *N*

N	**slope *b***		***N***	**slope *b***
1000	0.0002300		18000	0.0000020
2000	-0.0000600		20000	-0.0000037
3000	-0.0000700		22000	-0.0000054
4000	-0.0000094		24000	-0.0000060
5000	0.0000200		26000	-0.0000061
6000	0.0000200		28000	-0.0000059
7000	0.0000200		30000	-0.0000057
8000	0.0000200		35000	-0.0000056
9000	0.0000100		40000	-0.0000052
10000	0.0000093		50000	-0.0000046
12000	0.0000060		75000	-0.0000031
14000	0.0000061		100000	-0.0000023
16000	0.0000040		160000	-0.0000019

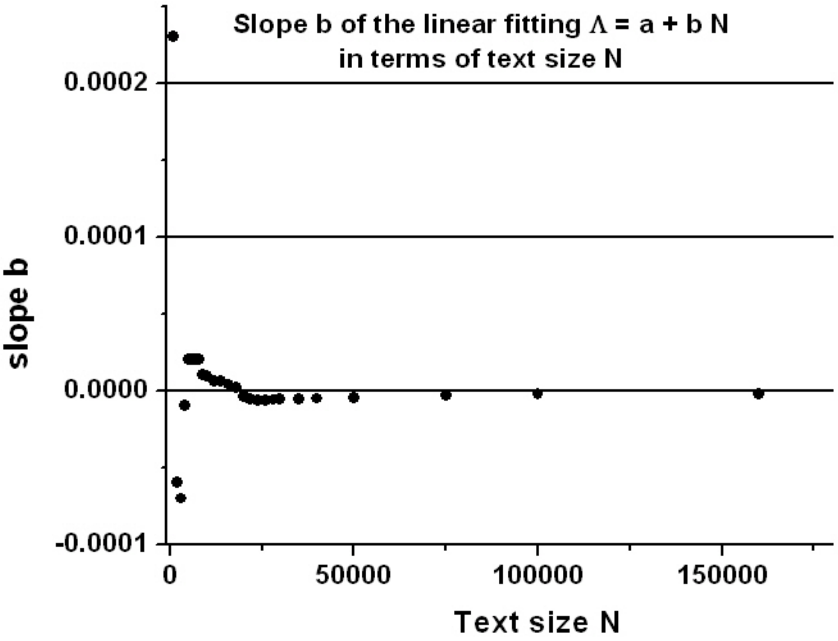


Figure 2.2. The slope of the linear regression of lambda against text size *N*

In all these figures we show that the regression is almost horizontal; the regression coefficient *b* is very near to zero, that is, the influence of text length is ruled out. Thus we obtained an acceptable transformation (normalization) under which the frequency structure of short texts can be compared with those of longer ones. For example, a set of selected poems by the Romanian writer Cărtărescu with length $N = 14700$ has $\Lambda = 1.6971$, while his prosaic text *De ce iubim femeile* with $N = 674$ has $\Lambda = 1.6673$. The lambda values are practically equal.

From the lambda definition, Eq. (1.2), we get simply $\Lambda = LC$, where the proportionality coefficient $C = \Lambda/L = (\log_{10}N)/N$ depends very weakly on N itself. Indeed, as can be seen in Figure 2.3, its maximum is located at $N = 3$, i.e., at a value which is exceptional as text length and can be discarded (Cf. e.g. Chapter 29 of J. Paul´s *Dr. Katzenberger´s Badereise* containing only 5 words, which was the smallest text we met, and skipped, among 1185 ones considered in the Appendix).[1]

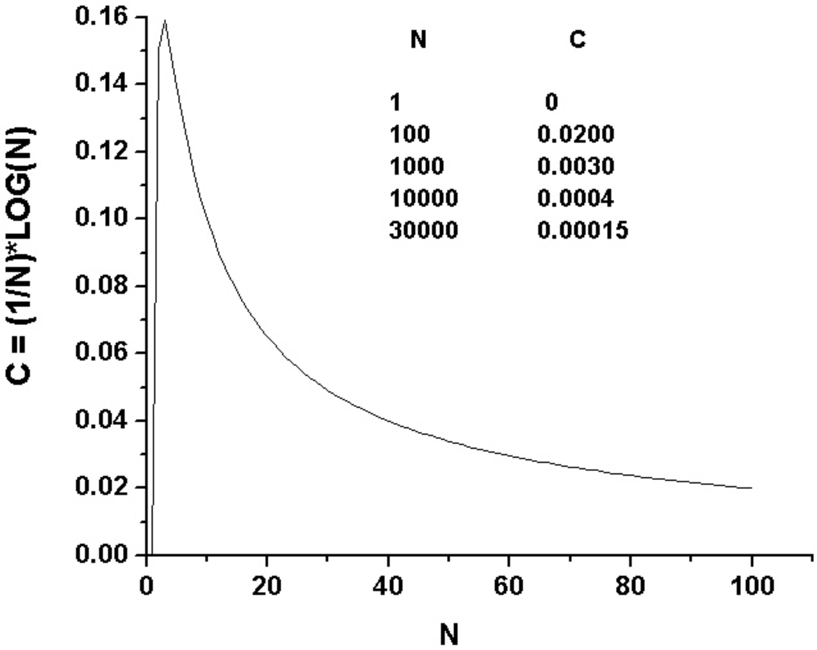


Figure 2.3. Dependence of *C* on *N*

The enormous dispersion in Figures 2.1a,b,c testifies to the fact that in spite of normalization there is a number of factors exerting influence on the rank-frequency distribution, such as personal style, text sort, language type, age of the writer, the year in which the text was written, and a number of other factors which are not even known, and if some of them are known, they have not been

[1] If the integer N is approximated by a continuous variable, the maximum of the function $C = (1/N)\text{Log}(N)$ occurs at $N_{max} = e = 2.71828\ldots$ and has the value $C_{max} = 1/[e\text{Ln}(10)] = 0.15977\ldots$

quantified up to now. As a matter of fact, we have a method which allows different insights in the structure of the text. Since text size becomes irrelevant, we can compare both individual texts using their rank-frequency distributions as well as groups of texts comparing their mean lambdas. Fortunately, for groups we can use the empirical lambda values to compute the empirical variances. Further, the indicator permits us to follow the development of the use of vocabulary size in a text and, with some adaptation, also the historical development of texts in a language. A quite special possibility is the study of language development with children.

In what follows, we shall scrutinize different situations which may arise when one wants to compare the frequency structure of texts, authors, genres or even languages.

As a matter of fact, we take into account not only the number of word-form types (or tokemes, cf. Andersen, Altmann 2006) occurring in a texts but also their frequency arrangement. The distances between individual ranked frequencies can be organized in very different ways but since we use the arc length between them, lambda manifests also something of the way they are ordered. Hence lambda is directly an indicator of frequency structure of a text and indirectly a richness indicator.

In all texts we counted the word-forms and considered each different form a type, or better, a tokeme. This automatically positions texts of strongly synthetic languages at a higher level of “richness” or “organization”, but it is exactly this phenomenon that allows us to order languages according to their synthetism/ analytism degree. The frequency structure of the identical text in a highly synthetic and highly analytic languages yields very different lambdas which in this case signalize the property of language, not a property of text.

The texts were taken randomly – a concept that in our times means “availability” in electronic form, but not everything could be used.

However, it needs an explanatory note regarding the error in word counting produced by apostrophe and hyphen. Some counters respect them and consider a sequence containing an apostrophe or a hyphen as one word, e.g. they differentiate between “its” and “it’s”; other ones decompose “it’s” into “it” and “s”. However, both are “correct”. The apostrophe shows the tendency of two words to coalesce or, in the later case, a morpheme (here “is”) gets a new allomorph (“s”). In some cases, however, the counter yields nonsense, e.g. in “don´t” we may obtain two words: “don” and “t”. The proportion of such “errors” is, however, very small and we did not repair the output manually but rather intervened in the input. Our aim was to show a possible method, not a “correct” segmentation and evaluation of a special language. In some languages the apostrophe and the hyphen can simply be eliminated; in other languages they are very sensitive; in still other ones they could be the object of decades of linguistic discussion as to their function in the formation of a word.

Here a methodological remark must be added. Data are not given; data are constructed by us. Data may not only be wrong, e.g. not corresponding to any grammar, but they may simply be false for testing certain hypothesis. Data are constructed for testing hypotheses – a fact frequently forgotten by corpus linguists – but sometimes seemingly senseless phenomena, e.g. n-grams of letters, may be used for some linguistic purpose, e.g. mechanical language identification. But data obtained or simply "sampled" without a previous hypothesis are not only exposed to the danger of leading to erroneous conclusions, the hypotheses themselves must constantly be examined and if necessary, modified. Hypotheses should be set up in such a way that they – in case of corroboration – should hold for all languages. General hypotheses do not contain empirical terms, e.g. "English"; for general hypotheses "English" is only a medium in which we them. In typological considerations, English and any other language are only points among 6000 other ones – or 60000 if one considers also the dialects and historical changes. In many cases in which a general hypothesis does not hold, boundary conditions must be taken into account. This, however, is situated at a higher level of research because boundary conditions must be captured either by means of a parameter or by means of a new variable. Both procedures make the formulas more complex. Thus, progress in text analysis means increasing complexity just as in all other sciences (cf. Bunge 1963). Nevertheless, simplification is not only allowed, it is inherent in any research because nothing can be captured in its wholeness. We always examine only selected aspects of phenomena: either those that can easily be separated from other ones and scrutinized in isolation, or those that can easily be captured conceptually and translated into the language of mathematics. At the next higher level we link the phenomena because we know that in language no phenomenon is isolated (cf. Altmann 2006). But linking necessitates mathematics and any linking in language is probabilistic. Only grammatical rules are deterministic in a certain time interval. And since text is the only way to approach non-grammatical structures in language, textology is basically a probabilistic discipline.

3. Comparison of texts

3.1. Individual comparisons

Comparing the lambdas, i.e. a function of the frequency structure, of two texts, we must take into account all rank-frequencies and compute the variance of each of them by a lengthy procedure. The arc length of the rank-frequency distribution is defined as

$$(3.1)\quad L=\sum_{i=1}^{V-1}[(f_i-f_{i+1})^2+1]^{1/2},$$

where f_i are the individual (ordered) frequencies and V is the vocabulary (highest rank) of the text. The asymptotic variance of arc length, as given in Popescu, Mačutek, Altmann (2010) is

$$(3.2)\qquad Var(L)=\frac{N-f_1}{1-\hat{p}_1}\sum_{r=2}^{V}\hat{a}_r^2\hat{p}_r\left(1-\frac{\hat{p}_r}{1-\hat{p}_1}\right)-2\frac{N-f_1}{\left(1-\hat{p}_1\right)^2}\sum_{r=2}^{V-1}\sum_{s=r+1}^{V}\hat{a}_r\hat{a}_s\hat{p}_r\hat{p}_s$$

where

$$\hat{a}_r=-\frac{\left(N-f_1\right)\left(\frac{\hat{p}_{r-1}-\hat{p}_r}{1-\hat{p}_1}\right)}{\sqrt{\left(N-f_1\right)^2\left(\frac{\hat{p}_{r-1}-\hat{p}_r}{1-\hat{p}_1}\right)^2+1}}+\frac{\left(N-f_1\right)\left(\frac{\hat{p}_r-\hat{p}_{r+1}}{1-\hat{p}_1}\right)}{\sqrt{\left(N-f_1\right)^2\left(\frac{\hat{p}_r-\hat{p}_{r+1}}{1-\hat{p}_1}\right)^2+1}}$$

for r = 2,..,V-1, and

$$\hat{a}_V=-\frac{\left(N-f_1\right)\left(\frac{\hat{p}_{V-1}-\hat{p}_V}{1-\hat{p}_1}\right)}{\sqrt{\left(N-f_1\right)^2\left(\frac{\hat{p}_{V-1}-\hat{p}_V}{1-\hat{p}_1}\right)^2+1}}.$$

An asymptotic test for comparing *two texts of the same genre in the same language* can be set up using the normal approximation

$$(3.3)\quad u=\frac{\Lambda_1-\Lambda_2}{\sqrt{Var(\Lambda_1)+Var(\Lambda_2)}},$$

where the variances of the arcs are computed as shown above and the variance of Λ is

$$(3.4)\quad Var(\Lambda) = \left(\frac{\log_{10} N}{N}\right)^2 Var(L)\,.$$

We present the computation procedure using two German poems, namely Goethe´s *Erlkönig* and his *Elegie 15*. The results are presented in Tables 3.1 and 3.2. In the third column is the result of computing the function $\hat{a}_r$ which is necessary for formula (3.2).

Table 3.1
Rank-frequencies in Goethe´s *Erlkönig* and the computation of $\hat{a}_r$

r	***f**_r*	***â**_r*
1	11	
2	9	-0.894427
3	9	0.894427
4	7	-0.187320
5	6	- 0.707106
6	6	0.707106
7	5	- 0.707106
8	5	0.707106
9	4	- 0.707106
10-14	4	0
15	4	0.707106
16	3	- 0.707106
17-20	3	0
21	3	0.707106
22	2	- 0.707106
23-38	2	0
39	2	0.707106
40	1	-0.707106
41-124	1	0

Table 3.2
Rank-frequencies in Goethe´s *Elegie 15* and the computation of $\hat{a}_r$

r	f_r	$\hat{a}_r$		r	f_r	$\hat{a}_r$
1	18			12-18	4	0
2	16	0.086153		19	4	0.707107
3	11	-0.031897		20	3	-0.707107
4	8	-0.948683		21-31	3	0
5	8	0		32	3	0.707107
6	8	0.707107		33	2	-0.707107
7	7	-0.707107		34-63	2	0
8	7	0		64	2	0.707107
9	7	0.707107		65	1	- 0.707107
10	6	0.187320		66-297	1	0
11	4	-0.894427				

Using the values of $\hat{a}_r$ and $p_r = f_r/N$ in the above tables one can compute the variances as given in Table 3.3.

Table 3.3
Lambdas and their variances in Goethe´s *Erlkönig* and *Elegie 15*

Text	N	V	Λ	$Var(\Lambda)$
Erlkönig	225	124	1.3377	0.003839
Elegie 15	468	297	1.7505	0.000952

Now the u-test for the difference between lambdas of these two texts yields

$$u = \frac{|1.3377 - 1.7505|}{\sqrt{0.003839 + 0.000952}} = 5.96 .$$

In spite of relatively small difference between the two lambdas, the difference is highly significant; however, this fact does not depend on the difference of text lengths. In order to corroborate the statement, we compare *Erlkönig* with Goethe´s *Elegie 2*, which has very similar size (N = 251), but a larger vocabulary (V = 169) and a greater lambda, Λ = 1.6773. The variance of the lambda is $Var(\Lambda)$ = 0.0015658. Inserting these values in formula (3.3) we obtain

$$u(Erlkönig, Elegie2) = \frac{|1.3377 - 1.6773|}{\sqrt{0.003839 + 0.0015658}} = 4.61,$$

which is highly significant, too. The difference is always due to the frequency structure of the two texts. A significant difference, u > 1.96 means that one of the texts is more concentrated; the author uses fewer rhetorical devices and simpler structures, while the "richer" text expresses more emotions or more circumstances (the author "plays" with the text). It may be asked whether the difference of lambdas is parallel to the share of e.g. descriptive and active means (adjectives or active verbs), the well known Busemann-ratio (cf. Altmann 1978):

The computation in absolute values of u is more practical and represents the two-sided test.

The lambda indicator is not only an indicator of vocabulary size but it also captures some features of the structure of texts. This may be exemplified using texts of two *Dada* writers (Table 9a in Appendix). Comparing H. Arp´s *Sekundezeiger* displaying N = 78, V = 25 and Λ = 0.9633 and K. Schwitters´ *Die Puppen* displaying N = 75, V = 31, i.e. a larger number of word types at smaller N, one would expect a greater lambda with Schwitters. However, here Λ = 0.8970, i.e. slightly smaller than with H. Arp. The difference is, of course, not significant (u = 0.37).

Nevertheless, even very small differences may turn out to be significant. It depends on the variances of the compared texts: if they are very small, they may cause a small difference to be significant. Consider, for example the works of two Finnish authors J. Aho, *Juha* written in 1911 and T. Pakkala, *Pieni elämäntarina* written in 1902. Here we have Λ(Aho) = 1.422, var(Λ,Aho) = 0.000041 and Λ(Pakkala) = 1.497, var(Λ,Pakkala) = 0.000060. Using these numbers we obtain u = |1.422 – 1.497|/√(0.000041 + 0.00006) = 7.46, a highly significant difference, though the difference between the two lambdas is very small. Thus, Λ alone is not sufficient for indicating significance but it is sufficient for ordering and classification. Small variances are signs of a kind of regularity of the rank-frequency sequence in long texts in which the writer cannot consciously manipulate the order in creating a special style or evoke a special effect; this is possible in short texts, e.g. poems, or in rank-frequency sequences of entities from small inventories, e.g. phonemes or letters.

It is to be noted that formula (3.3) can be used for the comparison of texts in the same language and same genre where one can assume the equality of expectations. But if one compares texts in different languages and/or different genres, the numerator of (3.3) must contain the difference of expectations, too. Hence it must be written as $(\Lambda_1–E(\Lambda_1)) – (\Lambda_2–E(\Lambda_2)$ or $\Lambda_1 – \Lambda_2 – [E(\Lambda_1) – E(\Lambda_2)]$, and the expectations may be estimated by the mean lambda of the given genre in the given language.

3.2. Group comparisons

If we compare a set of individual texts, we can present the results in tables or by means of weighted graphs. Using the tables in the Appendix we present some results of testing the differences between texts in the same language and in the same genre. In all cases we use the given lambdas and their variances. Consider first the simple case of private letters in Bulgarian presented in Table 3.4. The computing of *u*-tests is identical with the above procedure.

Table 3.4
u-differences between Bulgarian private letters

	Ceneva 1	**Ceneva 2**	**Janko 1**	**Janko 3**
Boris 2	2.08	1.62	0.52	0.58
Ceneva 1		0.6	2.65	1.58
Ceneva 2			2.25	1.07
Janko 1				1.14

This picture differs from that given by mere lambdas according to which the private letters can be ordered as *Janko 1 – Boris 2 – Janko 3 – Ceneva 2 – Ceneva 1* (cf. Table 2a in Appendix). The above matrix is a dissimilarity matrix allowing for all kinds of classifications. Since there are about 500 classification methods, and the larger the matrix, the more different classifications there are, we restrict ourselves to presenting the matrix as a graph leaving the significant differences out. For the data in Table 3.4 we obtain Figure 3.1 containing only those entities that are joined by non-significant *u.*

The individual private letters can be ordered in classes according to the degrees of vertices given by the number of edges touching the vertex. From the graph in Figure 3.1 we obtain

Degrees

4	*Janko 3*
3	*Boris 2, Ceneva 2*
2	*Janko 1, Ceneva 1*

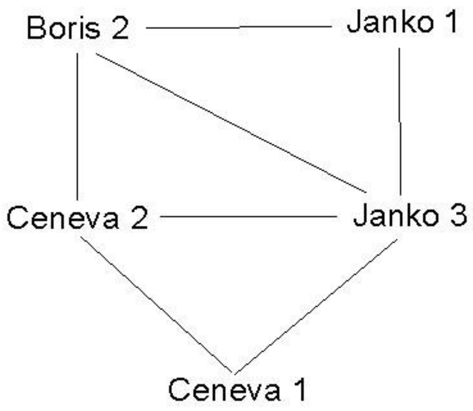


Figure 3.1. The graph of similarities of frequency structures in Bulgarian private letters

Needless to say, the above graph can be made weighted or fuzzy, if one weights each edge by its *u*. The smaller is *u*, the greater is the similarity of the joined vertices.

Consider the Nobel prize lectures of laureates for literature (texts taken from http://nobelprize.org/nobel_prizes/literature/laureates/). We obtain the results in Table 3.5 and the graph in Figure 3.2. The abbreviations of the names are as follows:

Be = Bellow,	Ye = Yeats,	Le = Lewis	Bu = Buck
Ru = Russell	Gl = Golding	Gr = Gordimer	Wa = Walcott
Mo = Morrison	Pi = Pinter	Ls = Lessing	

We insert in the table all *u* values but in the graph only the non-significant ones are considered.

Table 3.5
Nobel prize lectures (for literature)

	Ye	**Le**	**Bu**	**Ru**	**Gl**	**Gr**	**Wa**	**Mo**	**Pi**	**Ls**
Be	6.54	1.38	13.08	5.72	3.54	2.50	5.51	0.31	1.67	12.72
Ye	-	5.37	7.01	1.04	3.33	8.79	12.65	5.31	4.33	6.04
Le		-	12.15	4.48	2.20	3.90	7.26	0.86	0.46	11.71
Bu			-	8.15	10.37	15.00	10.09	11.20	10.59	1.49
Ru					2.36	8.06	12.02	4.55	3.50	7.29
Gl						6.01	9.78	2.67	1.47	9.74
Gr							2.53	2.45	3.91	14.77
Wa								4.89	6.74	19.53
Mo									1.18	10.61
Pi									-	9.96

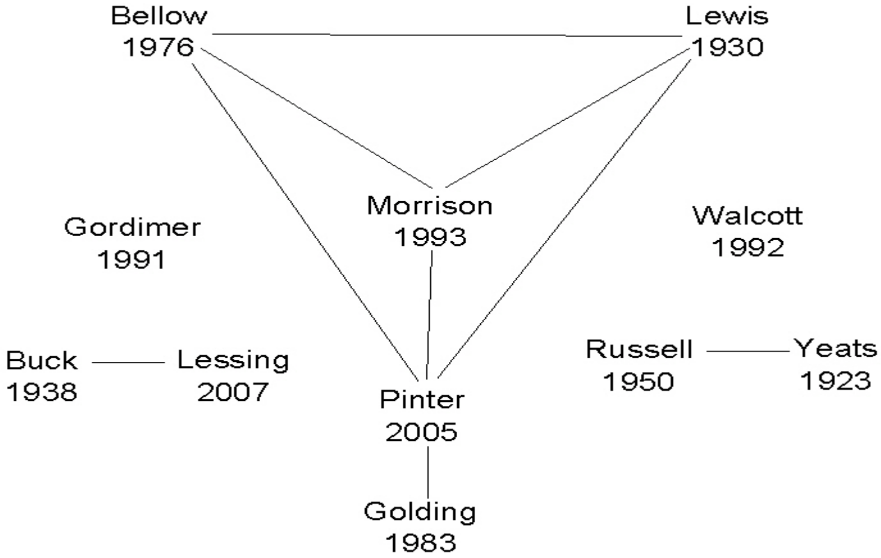


Figure 3.2. The similarities of Nobel prize winners for literature

As can be seen, out of 55 possibilities only 9 display similarity, a fact caused probably by the greater mastery over language by the writers. A comparison of lectures of laureates in other disciplines is, again, an interesting philological problem.

If we abbreviate the French poets as follows:

Ba = Baudelaire, Bo = Boileau, Fo = La Fontaine,
Pr = Prudhomme, Ri = Rimbaud, Ve1 = Verlaine 1, Ve2 = Verlaine 2,

we obtain the matrix presented in Table 3.6.

Table 3.6
French poetry

	Ba	**Bo**	**Fo**	**Pr**	**Ri**	**Ve1**	**Ve2**
Ap	5.66	4.41	2.78	2.20	1.26	4.48	8.50
Ba		0.34	2.06	2.04	6.73	1.38	3.42
Bo			1.49	1.56	5.40	0.77	3.17
Fo				0.23	3.83	0.97	4.81
Pr					3.16	1.09	4.46
Ri						5.61	9.39
Ve1							4.75

The graph of French poetry has three components. Rimbaud and Apollinaire have no similarity with the other ones and Verlaine 2 is quite isolated. Hence we obtain the graph (Figure 3.3).

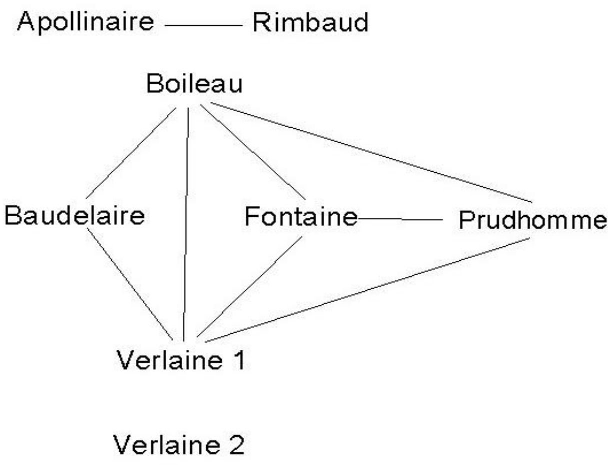


Figure 3.3. Similarities among French poets

Isolated vertices signify some original structuring of frequencies having some correlation with the contents or style, again a problem for philologists.

In this way groups of authors can also be compared, even if they wrote more texts. In that case one compares the means of respective lambdas (see Chapter 4) or, alternately, the corresponding highest lambdas. For example in the latter case the grouping of Romanian poets whose texts were taken mostly from http://www.romanianvoice.com/poezii/ (2010) and are presented in Table 24a of the Appendix yields the following results. The simple ordering by decreasing lambda is presented in Table 3.7.

Table 3.7
Ordering of Romanian poets by decreasing lambda
(text pooling is marked by asterisk)

Author	**Λ decreasing**	***Var(Λ)***
Barbu*	2.1795	0.000699
Dinescu*	2.0413	0.000775
Eminescu	1.9921	0.000412
Arghezi*	1.9920	0.000250
Topârceanu	1.9639	0.001079
Goga	1.9429	0.000576
Labiş	1.9420	0.000560
Doinaş	1.8962	0.000872
Alecsandri*	1.8713	0.000700
Blaga	1.8475	0.000635

Stănescu	1.7666	0.001036
Coşbuc	1.7612	0.001067
Blandiana	1.7124	0.001035
Cărtărescu*	1.6971	0.000035
Paunescu	1.5719	0.003296
Bacovia*	1.5162	0.000239
Sorescu*	1.5063	0.000211

Some texts in this table are pooled. Actually, the ranking of Romanian poets according to single poems having the highest lambda starts with

Eminescu (*Λ* = 1.9921), În căutarea Şeherezadei; – Topârceanu (1.9639), Rapsodii de toamnă; – Barbu (1.9590), Domnişoara Hus; – Goga (1.9429), Clăcaşii; – Labiş (1.9420), Confesiuni; – Arghezi (1.9095), Testament; – Doinaş (1.8962), Orologiul de gheaţă; – Blaga (1.8475), Paşii profetului; – Dinescu (1.8470), Discurs la intrarea unei ţări estice în Europa; and so on.

Comparing, for instance, the authors with data of Table 3.7 we obtain the *u*-tests as given in Table 3.8 and presented in Figure 3.4. The poet with the highest lambda has no similarity with the other ones. From the attached *u*-graph we distinguish the singularity of the mathematician-poet Barbu (highest lambda, no link), followed by a big cluster composed of Eminescu, Dinescu, Arghezi, Topârceanu, Goga, and Labis; and three smaller clusters (Blaga, Alecsandri, Doinaş) (Stănescu, Coşbuc, Blandiana, Cărtărescu) and (Păunescu, Bacovia, Sorescu).

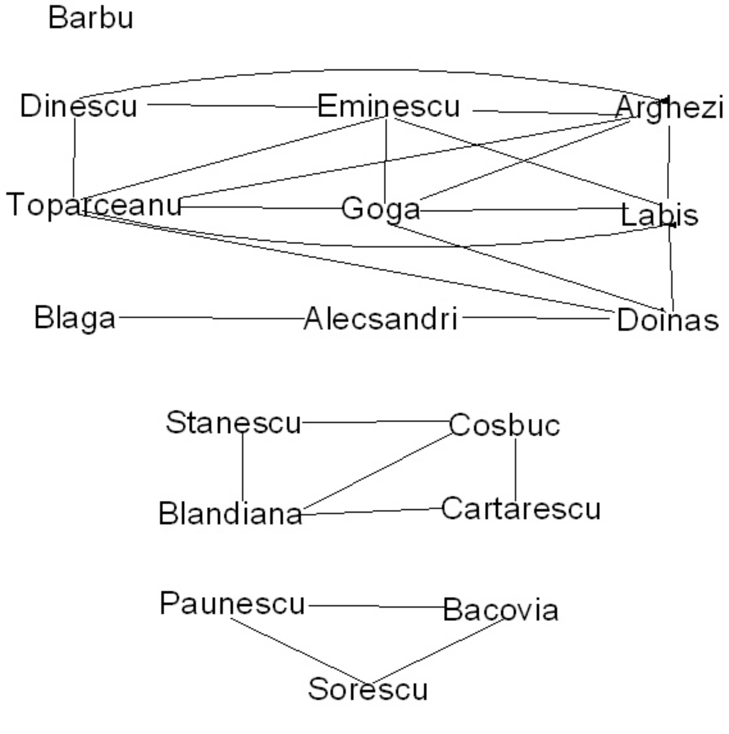


Figure 3.4. The similarities among Romanian poets

Table 3.8
Similarities among Romanian poets

Author	**Alec**	**Argh**	**Baco**	**Barb**	**Blag**	**Blan**	**Cărt**	**Cosb**	**Dine**	**Doin**	**Emin**	**Goga**	**Labi**	**Paun**	**Sore**	Stăn
Arghezi	3.92															
Bacovia	11.59	21.52														
Barbu	8.24	6.09	21.66													
Blaga	0.65	4.86	11.21	9.09												
Blandiana	3.81	7.80	5.50	11.22	3.31											
Cărtărescu	6.43	17.47	10.93	17.81	5.81	0.47										
Coşbuc	2.62	6.36	6.78	9.95	2.09	1.06	1.93									
Dinescu	4.43	1.54	16.49	3.60	5.16	7.73	12.09	6.53								
Doinaş	1.59	2.92	17.14	8.06	2.41	5.69	11.91	4.32	3.60							
Eminescu	3.62	0.00	18.65	5.62	4.47	7.35	13.95	6.00	1.43	2.58						
Goga	2.00	1.71	14.95	6.63	2.74	5.74	9.94	4.48	2.68	0.69	1.57					
Labiş	1.99	1.76	15.06	6.69	2.73	5.75	10.04	4.48	2.72	0.67	1.61	0.03				
Paunescu	4.74	7.05	0.94	9.61	4.40	2.13	2.17	2.87	7.36	5.82	6.90	5.96	5.96			
Sorescu	12.09	22.62	0.47	22.32	11.73	5.84	12.16	7.13	17.04	18.01	19.46	15.56	15.69	1.11		
Stănescu	2.51	6.29	7.01	9.91	1.98	1.19	2.12	0.12	6.46	4.22	5.93	4.39	4.39	2.96	7.37	
Topârceanu	2.20	0.77	12.33	5.11	2.81	5.47	7.99	4.38	1.80	1.12	0.73	0.52	0.54	5.93	12.74	4.29

Alec = Alecsandri

One could continue evaluating the properties of such graphs but we are aware of the fact that this has sense only if the number of texts is greater in the given language and genre. The text having the most similarities (= non-significant differences) with other ones represents the most harmonic frequency structure, i.e. not many irregular jumps in the frequency sequence. So, they are positioned mostly in the midst of all the texts. In significantly deviating structures, philologists would get the possibility of searching for relations between content and form. The graphs could also be used for determining the cohesion of a genre or capturing an aspect of text evolution. All this can be left to philologists.

On the other hand, processing more texts in one genre would lead to insurmountable difficulties. As the number of texts (authors) increases, the tables get larger, but less lucid, and the graphs would look like a skein with which two cats had played for a long time. In order to avoid this difficulty we characterize a writer or works consisting of more parts.

3.3. Characterizing groups

The *u*-test for difference of two lambdas tells us whether we can consider the texts similar or dissimilar with respect to their frequency structures. Studying a novel having several parts, or several texts of an author, we can interpret similarity as inertia or persistence or conservativism: two similar texts/parts have been written under the same frequency regime. Once a text has been written, the experience with its structure remains in the brain of the writer on the subconscious level. If he deviates from this pattern, he strives for originality, he is elastic, innovative, versatile, etc. These properties manifest themselves in dissimilarities with other texts or other parts of his work. In order to capture the behaviour of a writer (comparing all his texts written in the same genre) or the extent of structural changes in one work consisting of several parts, or even the behaviour of a closed group of writers (in the same language and genre), we simply compute the proportion of dissimilarities p in the given set. Then the proportion of similarities is $1-p$.

The simplest possibility of testing the character of similarities/dissimilarities is taking the number of dissimilarities X and treat it as a binomial variable. That is, the number of comparisons of n texts is $n(n-1)/2$, the parameter of the binomial distribution is $p = 0.5$ and we may ask two questions:

(i) if $x > n(n-1)p/2 = n(n-1)/4$ (that is, if the number of dissimilarities is greater than half of the comparisons) we compute the cumulative probability

$$(3.5)\quad P(X \ge x) = \sum_{i=x}^{n(n-1)/2} \binom{n(n-1)/2}{i} 0.5^{n(n-1)/2}.$$

If $P(X \geq x) \leq 0.05$, we can consider the set significantly innovative, versatile, elastic, etc. One can, of course, determine several levels such as 0.05, 0.01 etc.

(ii) If $x < n(n - 1)/4$, i.e. if the number of dissimilarities is smaller than half of the comparisons, we compute the probability

$$(3.6) \quad P(X \leq x) = \sum_{i=0}^{x} \binom{n(n-1)/2}{i} 0.5^{n(n-1)/2}.$$

If now $P(X \leq x) \leq 0.05$ we may consider the set as significantly conservative.

(iii) If $x = n(n - 1)/4$, we obtain both for (1) and (2) the same probability. In this case the number of dissimilarities is neutral; the text is in equilibrium; There are no trials of the writer to distinguish the texts nor to make them uniform.

In this way we obtain exact probabilities, but if n is very large, instead of computing the binomial probabilities we may again use the normal approximation and set up the criterion

$$(3.7) \quad u = \frac{x - n(n-1)/4}{\sqrt{n(n-1)/8}}.$$

If $u > 1.96$, the set is significantly innovative and varied. If $u < -1.96$, the set is significantly conservative and uniform. The approximation is the better, the larger is n. However, here we shall always use the binomial distribution.

Let us illustrate all computations using the set of texts from Chapter 3.2. In Bulgarian private letters (Table 3.4) we have $n = 5$, $n(n - 1)/2 = 10$ and the number of (significant) dissimilarities is $x = 3$. Since $E(x) = n(n - 1)/4 = 5$, we have case (ii), $x < E(x)$, and search for the probability by means of (3.6). We obtain:

$$P_0 + P_1 + P_2 + P_3 =$$
$$= \sum_{i=0}^{3} \binom{10}{i} 0.5^{10} = 0.00098 + 0.00977 + 0.04394 + 0.11719 = 0.17 \ ,$$

which is greater than 0.05, hence Bulgarian private letters do not display any tendency. Since n is very small, the normal test should not be used. But even if we use it, we obtain

$$u = \frac{3-5}{\sqrt{\frac{n(n-1)}{2} 0.5(05)}} = \frac{-2}{\sqrt{10(0.5)0.5}} = -1.26,$$

which is not significant, i.e. the result is the same. The probability of u = -1.26 is 0.10.

In Table 3.5 there are n = 11 Nobel prize laureates in literature, hence we have $n(n - 1)/2$ = 55 comparisons out of which 46 are dissimilarities. The expected number is $E(X) = n(n - 1)/4$ = 27.5, hence the number of dissimilarities is greater than expected.

Computing (1) yields P = 0.000000217. That means, there is a great diversity within the group of Nobel prize laureates in literature.

In Table 3.6 we compare a whole group of French writers, n = 8, $E(X)$ = 14, x = 19 (number of significant dissimilarities), hence the exact result is P = 0.043 which is rather unexpected but conforms with the graph in Figure 3.3. It is to be emphasized that here we do not add the numerical values; we simply count their number.

Analyzing some data contained in the Appendix we obtain the results for individual writers and groups as presented in Table 3.9.

Table 3.9
Variability/neutrality/uniformity in groups of texts or many-part texts
(The table numbers refer to the Appendix)

Text group	***n***	***E(x)***	***x***	***P***
Bulgarian private letters (Table 2b)	5	6	3	0.17
Czech poetry: (Table 4a)	34	280.5	516	$0.9991(10^{-102})$
Czech prose: Hrabal (Table 4b)	15	52.5	66	0.005
Czech scientific texts (Table 4d)	10	22.5	42	$0.4327(10^{-9})$
Dutch prose (Table 5)	5	5	10	0.00098
Nobel lectures of laureates for literature (s. Table 3.4 above)	11	27.5	46	0.0000002
Dos Passos, J., A Pushcart at the Curb. Nights at Bassano. (Table 6a)	6	7.5	8	0,5
Dos Passos, One Man´s Initiation (Table 6b)	11	27.5	45	$0.1029(10^{-5})$
Dos Passos, J., Rosinante to the Road Again (Table 6b)	18	76.5	105	$0.2348(10^{-5})$
Byatt, A., Possession (Table 6b)	10	22.5	38	0.0000016
Ackroyd, P., Hawksmoore (Table 6b)	4	3	4	0.3378
Milton, J., Paradise lost (Table 6b)	4	3	3	0.65
English scientific texts (Table 6d)	10	22.5	42	$0.4327(10^{-9})$
English stories told or written by children*** (Table 6e)	39	370.5	261	$0.3598(10^{-15})$
Finnish prose (Table 7)	5	5	9	0.0107

French poetry (s. Table 3.5 above)	8	14	19	0.043
French prose (Table 8b)	19	85.50	152	0.2964(10^{26})
Hugo, V., Les misérables (Table 8b)	5	5	7	0.1719
Musset, A.de, Oeuvres completes 7 (Table 8b)	5	5	9	0.0107
Hungarian poetry (Table 11a)	15	52.5	95	0.7927(10^{-18})
Hungarian prose (Table 11b)	30	217.5	387	0.3113(10^{-66})
Hungarian press (Table 11c)	5	5	8	0.0547
Italian prose (Table 13b)	10	22.5	40	0.3739(10^{-7})
Italian poetry (Table 13a)	8	14	27	0.1080(10^{-6})
Italian presidents: End-of-year speeches (Table 13c)	60	885	1229	0.8611(10^{-61})
Indonesian newspapers (Table 12a)	5	5	5	0.6230
Japanese prose (Table 14)	7	10.5	18	0,0007
Droste-Hülshoff, A., 5 poems (Table 9a)	5	5	6	0.3770
Goethe, J.W.v., 7 poems (Table 9a)	7	10.5	7	0.0946
Chamisso, A., Peter Schlemihls wundersame Geschichte (Table 9b)	11	27.5	36	0.0150
Eichendorff, J.F. Aus dem Leben eines Taugenichts (Table 9b)	10	22.5	31	0.008
Kafka, F., Betrachtung (Table 9b)	18	76.5	76	0.5
Lessing, G.E., 10 novels (Table 9b)	10	22.5	22	0.5
Löns, H., Der Werwolf (Table 9b)	13	39	58	0.97515(10^{-5})
Meyer, Der Schuß von der Kanzel*** (Table 9b)	11	27.5	18	0.0072
Novalis, H.O., Heinrich von Ofterdingen (Table 9b)	10	22.5	36	0.00003
Paul, J., Dr. Katzenbergers Badereise*** (Table 9b)	44	473	352	0.1660(10^{-14})
M.Eminescu: 146 poems (Chapter 8.2, Table 8.4)	146	5292.5	6217	0.8421(10^{-72})
Cervantes, M. de, Don Quijote (Table 31)	15	52.5	80	0.3440(10^{-7})
Swedish prose (Table 32)	5	5	10	0.00098
Turkish prose (Table 34)	40	390	741	0.8110(10^{-182})
Hawaiian prose*** (Table 10)	38	351.5	225	0.1587(10^{-12})
Kannada: Social sciences (Table 15a)	9	18	33	0.1136(10^{-6})
Kannada: Commerce (Table 15b)	9	18	30	0.000035

Latin: History & Philosophy (Table 17c)	5	5	9	0.0107
Latin: Rhetorics (Table 17d)	6	7.5	14	0.000488
Maori: Folk narratives (Table 19)	5	5	8	0.0547
Marathi: Poetry (Table 20a)	11	27.5	52	0.770938(10^{-12})
Marathi: Aesthetics (Table 20b)	11	22.5	39	0.270888(10^{-6})
Marathi: Social Sciences (Table 20c)	9	18	25	0.014408
Marathi: Natural Sciences (Table 20d)	10	22.5	39	0.270888(10^{-6})
Marathi: Commerce (Table 20e)	10	22.5	42	0.43749(10^{-9})
Marathi: Offical and Media (Table 20f)	10	22.5	35	0.000124
Marquesan: Folklore texts (Table 21)	7	10.5	16	0.013302
Rarotongan: Prose (Table 23)	5	5	6	0.376953
Romanian: Prose (Table 24b)	15	52.5	94	0.693464(10^{-17})
Russian: Poetry (Table 24a)	5	5	10	0.000976
Russian: Prose (Table 25b	10	22.5	40	0.393919(10^{-7})
Samoan: Prose*** (Table 26)	5	5	2	0.05469
Slovak: Poetry (Table 28a)	9	18	33	0.113607(10^{-6})
Slovak: Prose (Table 28b)	5	5	9	0.05469
Slovenian: Prose (Table 29a)	5	5	10	0.000977
Spanish: Prose (Table 31)	18	76.5	127	0.180969(10^{-16})
Tagalog: Poetry (Table 33a)	11	27.5	30	0.2950
Tagalog: Prose (Table 33b)	16	60	84	0.694841(10^{-5})
Ostrovskij, N., How the steel was tempered: Russian (Table 25b)	10	22.5	40	0.393919(10^{-7})
Belorussian (Table 1)	10	22.5	36	0.00003
Bulgarian (Table 2a)	10	22.5	40	0.3939(10^{-7})
Croatian (Table 3)	10	22.5	38	0.000002
Czech (Table 4b)	10	22.5	38	0.000002
Macedonian (Table 18)	10	22.5	39	0.270888(10^{-6})
Polish (Table 22)	10	22.5	42	0.432749(10^{-9})
Serbian (Table 27)	10	22.5	37	0.768705(10^{-5})
Slovak (Table 28c)	10	22.5	38	0.156065(10^{-5})
Slovenian (Table 29b)	10	22.5	35	0.000124
Sorbian (Table 30)	10	22.5	36	0.000033
Ukrainian (Table 35)	10	22.5	41	0.466744(10^{-8})

As can be seen, there are frequency structures of all kinds:

Uniform text-groups are e.g. Meyer, Paul, English stories told by children, and Hawaiian prose, in which the number of significant differences is smaller than the expectation $E(x)$. They are marked with 3 asterisks.

Neutral structurings are those whose $P > 0.05$, e.g. Bulgarian private letters, Milton, Ackroyd, Dos Passos (*A Pushcard…*), Hugo, Kafka, Lessing, poems by Goethe, poems by Droste-Huelshoff, Samoan prose, Hungarian and Indonesian press texts, Maori folk narratives, Rarotongan prose, Samoan prose, Tagalog poetry.

The rest are *variable* structurings; chapters or texts are written under different frequency regimes. Some of the groups display enormous variability, e.g. Turkish prose, Czech poetry, M. Eminescu et al. The case of Eminescu is especially interesting: all the texts are by the same author but they display enormous variability which would be expected in groups of texts written by different authors e.g. Turkish texts. Cases like this stimulate the philologists to search for other properties associated with the given variability.

As to the translation of Ostrovskij´s *How the steel was tempered* there are but small differences between Slavic languages. All translations have variable structuring; the relatively smallest variability is in the Slovenian translation, and the relatively greatest is in the Polish one.

Though the number of texts analyzed in Table 3.9 is considerable, it is merely the first step in the given direction. Automatically the questions arise: does conservatism in frequency structure correlate with the sequence of events in the texts? Is it correlated with the content? Is it an effect of some brain rhythms or perseveration? Etc. Since the proposed method is simple, it would, perhaps, be appropriate to use it parallel to all other methods of text analysis.

4. Comparison of authors

The comparison of two authors is always somewhat problematic. Even if one has all their works at his disposal, but the works belong to different genres, e.g. one of them wrote prose, poetry, and dramas, the other wrote prose and literary criticism, then one mixes domains whose lambdas may be *eo ipso* different. There is surely something that one could call "frequency structure in Shakespeare" but the number of boundary conditions that have been omitted is so large that the theoretical value of such a structure is very doubtful. Nevertheless, some comparisons are possible. In this case we compare mean lambdas, while the variances about the mean can be estimated empirically. Mean lambdas may be obtained also if a work is divided in several parts (e.g. chapters) or if an author wrote several multi-part works. However, if the genre has a strong influence on the text, then comparison is reasonable only within the given genre.

Let us illustrate the problem in comparing the End-of-year speeches of two Italian presidents, namely Einaudi and Gronchi (for other analyzes of the End-of-year speeches cf. Tuzzi, Popescu Altmann 2010). The lambdas are presented in Table 4.1. For the comparison we no longer need the text sizes because they are normalized. In all subsequent computations *n* means either the number of texts or the number of parts of a text.

Table 4.1
Lambdas of End-of-Year Speeches
of Italian Presidents Einaudi and Gronchi

Year - President	Λ
1949 - Einaudi	1.6928
1950 - Einaudi	1.5781
1951 - Einaudi	1.7686
1952 - Einaudi	1.8488
1953 - Einaudi	1.7489
1954 - Einaudi	1.7303
1955 - Gronchi	1.7038
1956 - Gronchi	1.6672
1957 - Gronchi	1.6194
1958 - Gronchi	1.6236
1959 - Gronchi	1.6718
1960 - Gronchi	1.6703
1961 - Gronchi	1.6677

Einaudi presented n = 6 speeches, Gronchi n = 7. The mean Λ of Einaudi is $\bar{\Lambda}_E$ = 1.7278, the mean Λ of Gronchi is $\bar{\Lambda}_G$ = 1.6605. The variances of means are $s^2(\bar{\Lambda}_E) = s^2(\Lambda)/n$ = 0.008078/6 = 0.001346, and $s^2(\bar{\Lambda}_G) = s^2(\Lambda)/n$ = 0.000874/7 = 0.0001235, both computed directly from the above data. Since Einaudi has a greater mean lambda, we may ask whether it is significantly greater than that of Gronchi. In this case we simply insert the values in the asymptotic formula

$$(4.1) \quad u = \frac{\bar{\Lambda}_1 - \bar{\Lambda}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

representing the normal variable. Since here we have the same language and the same text sort, we automatically assume that the expectations are equal and their difference yields zero. For a two-sided test one simply takes the absolute value of u. In our case we obtain

$$u = \frac{1.7278 - 1.6605}{\sqrt{0.001346 + 0.0001235}} = 1.75,$$

which is not significant in a two-sided test. However, if we compare president Pertini having n = 7, $\bar{\Lambda}_P$ = 1.2578 and $s^2(\bar{\Lambda}_P)$ = 0.003762/7 = 0.000537, with Einaudi, we obtain u = 10.83 which is highly significant. The result depends heavily on the dispersion of values of lambda with each author. Of course, it would be more appropriate to use the t-test but the greater the two n's, the more it approaches the normal distribution. On the other hand, it would be correct to test in advance whether the data can be considered as normally distributed. This can be done by some of the usual tests like Shapiro-Wilk, David, etc. (cf. http://tjkrauss.de/statmeth/normvert.htm). Besides, a number of non-parametric tests can be used in order to state whether the two presidents "are different" but it is not our aim to enlarge the number of different tests which can be found in good textbooks on statistics..

In this way one could compute the normalized differences between all Italian presidents. Though we assume that the speeches of individual presidents are usually written by several persons, the last form and the contents are controlled by the president himself, and hence the given speech is "his text". For the Italian presidents, whose speeches belong to the same text sort, we obtain the results in Table 4.2.

The ordering of Presidents according to the magnitude of mean lambda would yield:

Einaudi – Gronchi – Leone – Segni – Saragat – Napolitano – Ciampi – Cossiga – Scalfaro – Pertini

Table 4.2
Mean lambdas of End-of-Year speeches of Italian Presidents

Years	**President**	***n***	$\bar{\Lambda}$	$s^2(\Lambda)$
1949-1954	Einaudi	6	1.7279	0.008078
1955-1961	Gronchi	7	1.6605	0.000874
1962-1963	Segni	2	1.5854	0.000471
1964-1970	Saragat	7	1.5792	0.002823
1971-1977	Leone	7	1.5943	0.001228
1978-1984	Pertini	7	1.2578	0.003762
1985-1991	Cossiga	7	1.4588	0.008436
1992-1998	Scalfaro	7	1.2991	0,015467
1999-2005	Ciampi	7	1.5299	0.001057
2006-2008	Napolitano	3	1.5749	0.000125

However, the interpretation of the order of mean lambdas of Presidents is not only a linguistic or literary but also a sociological problem. In any case, the mean lambdas are not directly correlated with time.

Performing *u*-tests among all the Presidents we obtain the results in Table 4.3.

Another ordering can be performed according to the position of individual Presidents within the whole field. Here one can proceed in different ways:

(1) One is interested in the overall significance of all differences of a writer and (a) considers $\Sigma u_i/\sqrt{k} \sim N(0,1)$, i.e. a standard normal variable built as the sum of all *u*-values ($i = 1,2,\ldots,k$), or (b) one computes for a single writer the function $\sum(u_i - \bar{u})^2$ for $i = 1,2,\ldots k$) which is distributed like a chi-square with $k–1$ degrees of freedom or (c) one uses the probabilities of the given *u*-tests and forms the function $-2\Sigma\ln P_i$ which is distributed like a chi-square with $2k$ degrees of freedom.

(2) One is not interested in the significance of all combined *u*’s and (a) takes only the sum of all *u*’s of a writer or (b) one restricts oneself only to the number of significant *u*’s of a writer. First we shall use the last procedure (2a) and from Table 4.3 we obtain the following scale

5	Napolitano
6	Segni, Saragat, Leone
7	Cossiga
8	Einaudi, Gronchi, Pertini, Scalfaro, Ciampi

Table 4.3

Comparison of mean lambdas in End-of-Year speeches of Italian Presidents (two-sided *u*-test)

President	**Einaudi**	**Gronchi**	**Segni**	**Saragat**	**Leone**	**Pertini**	**Cossiga**	**Scalfaro**	**Ciampi**	**Napolitano**
mean Λ	1.7279	1.6605	1.5854	1.5792	1.5943	1.2578	1.4588	1.2991	1.5299	1.5749
σ^2/n	0.001346	0.000125	0.000236	0.000403	0.000176	0.000537	0.001205	0.002209	0.0001509	0.0000415
Einaudi										
Gronchi	1.76									
Segni	3.58	3.96								
Saragat	3.56	3.54	0.25							
Leone	3.43	3.82	0.44	0.63						
Pertini	10.83	15.65	11.78	10.48	12.60					
Cossiga	5.33	5.53	3.34	3.00	3.65	4.82				
Scalfaro	7.19	7.48	5.79	5.48	6.04	0.79	2.73			
Ciampi	5.12	7.86	2.82	2.09	3.56	10.37	1.34	4.75		
Napolitano	4.11	6.64	0.63	0.20	1.32	13.18	0.12	5.81	3.24	

This is nothing else but the degrees of the vertices in a graph presenting the network of presidents.

Hence Napolitano has the smallest number of dissimilarities, he stays in the middle of the presidents; but since we have only three of his speeches, he, perhaps, still did not evolve more differentiating properties. If we use method (1a) and from the resulting ones form a new *u* divided by the number of comparisons, √k = √9 = 3, we obtain the following scale:

Saragat	9.75
Cossiga	9.95
Segni	10.86
Napolitano	11.75
Leone	11.83
Ciampi	13.72
Einaudi	14.97
Scalfaro	15.35
Gronchi	18.75
Pertini	30.17

which is more detailed but not very different from the non-weighted scaling. The five presidents with the highest number of significant results (8) have also the highest scores of weighted (normalised) results. Again, this scaling is not correlated with the temporal sequence of presidents.

Consider some mean lambdas in texts from different languages as presented in Table 4.4. Here we present also the standard deviations of lambdas $s(\Lambda)$. The standard deviation of the mean lambda can be computed as $s(\Lambda)/\sqrt{\ n}$, the variance of the mean lambda as $s^2(\Lambda)/n$. Needless to say, different groupings of works would lead to different mean lambdas; hence the present results can be considered only as examples. Thus partitioning Byron´s poetry in 2 parts is nothing but a preliminary convention. Hence, comparing two authors is – as we said at the beginning of the chapter – no search for “truth”. Tables of this kind can be used for monothetic classifications because we compare here only one property: the normalized frequency structure.

It is a task for philologists to find other properties yielding similar classifications and correlate them with the present one. The identity of an author could be determined only if one would find other associated properties and place the frequency structure in a control cycle.

Table 4.4
Mean lambdas of some texts consisting of several parts (chapters, subdivisions)
Some very long texts were equally partitioned in two or more parts
(Cz = Czech, E = English, F = French, G = German, Hw = Hawaiian, Hu = Hungarian, It = Italian, Lt = Latin, Mr = Marathi, Ro = Romanian, Ru = Russian, Sm = Samoan, Sk = Slovak, Sp = Spanish, Tg = Tagalog)

Lg	**Genre**	**Author**	**Title**	***n***	$\bar{\Lambda}$	***s(Λ)***
Cz	poetry	Blatný	Verše	2	1.6808	0.0864
E	poetry	Browning	Dramatic Romances	2 halves	1.2654	0.0489
E	poetry	Browning	Browning's Shorter Poems	2	1.1763	0.0185
E	poetry	Byron	Poetical Works I.	2 halves	1.0915	0.0850
E	poetry	Dos Passos	A Pushcart at the Curb	6	1.6497	0.0926
E	prose	Ackroyd	Hawksmoor	4 quarters	0.9457	0.0261
E	prose	Barnes	Flaubert's Parrot	3 thirds	1.1271	0.0619
E	prose	Byatt	Possession	10 tenths	1.0768	0.0646
E	prose	Dos Passos	One Man's Initiation	11	1.4811	0.1391
E	prose	Dos Passos	Rosinante to the road again	17	1.5796	0.0814
E	prose	Milton	Paradise Lost	4 quarters	1.0747	0.0374
E	prose	Wells	The Invisible Man	2 halves	0.9742	0.0158
E	prose	Yeats	The Celtic Twilight	2 halves	0.8878	0.0367
F	poetry	Verlaine	Oeuvres Complètes. Vol.1	2 halves	1.2436	0.0406
F	prose	Balzac	Eugenie Grandet	3 thirds	1.0765	0.0821
F	prose	Hugo	Les misérables. I.	5	1.0391	0.1096
F	prose	Maupassant	Boule de Suif	2 halves	1.1694	0.0836
F	prose	Musset	Oeuvres Complètes. 7	5	1.1213	0.0869
F	prose	Prévost	Manon Lescaut	2	0.9380	0.0308
G	poetry	Goethe	Elegien	5	1.7014	0.0475
G	prose	Goethe	Reineke Fuchs	2 halves	1.0493	0.0066
G	prose	Eichendorff	Aus dem Leben eines Taugenichts	10	1.3505	0.0738
G	prose	Chamisso	Peter Schlemihls wundersame Geschichte	11	1.5245	0.0921
G	prose	Hoffmann	Der Sandmann	3	1.4418	0.1127
G	prose	Kafka	Betrachtung	18	1.5444	0.1345

G	prose	Lessing	Fabeln	10	1.5528	0.1027
G	prose	Löns	Der Werwolf	13	1.2246	0.1164
G	prose	Meyer	Der Schuss von der Kanzel	11	1.6844	0.0528
G	prose	Novalis	Heinrich von Ofterdingen	10	1.4897	0.1145
G	prose	Paul	Dr. Katzenbergers Badereise	55	1.6523	0.0758
G	prose	Sealsfield	Das Cajuetenbuch	28	1.3894	0.1391
G	prose	Storm	Der Schimmelreiter	2 halves	1.0071	0.0247
G	prose	Tucholsky	Schloss Gripsholm	5	1.1690	0.0596
G	prose	Wedekind	Mine-Haha	4	1.3010	0.2228
Hw	prose	Legend	Kawelo	4	0.6336	0.1003
Hw	prose	Legend	Laieikawai	35	0.6556	0.0634
Hu	poetry	Arany	Elbeszélő költemények	4	2.0031	0.1043
Hu	poetry	Arany	Versek	3	2.0273	0.0217
Hu	poetry	Petőfi	Poems	3	1.7473	0.0352
Hu	prose	Bársony	Délibáb	2	1.9055	0.0228
Hu	prose	Bródy	Az ezüst kecske	2	1.7714	0.0174
Hu	prose	Gárdonyi	Egri csillagok	5	1.8325	0.0473
Hu	prose	Gárdonyi	Isten Rabjai	3	1.7948	0.0552
Hu	prose	Karinthy	Utazás a koponyám körül	4 quarters	1.9893	0.338
Hu	prose	Karinthy	Utazás	2	1.9940	0.0280
Hu	prose	Tamási	Szülőföldem	4	1.8903	0.0356
Hu	prose	Tornyai	Gyere	2	1.5681	0.0449
Hu	prose	Wesselényi	Balítéletekről	2	1.7791	0.0231
It	poetry	Dante	Divina Commedia	3	1.1116	0.1818
It	poetry	Pellico	Poesie Inedite	4 quarters	1.5299	0.0445
Lt	poetry	Vergilius	Aeneidos	2 halves	1.6948	0.0376
Lt	poetry	Silius Italicus	Punicorum Libri	4	1.9009	0.0692
Lt	history	Tacitus	Historiae	2	2.0245	0.0306
Mr	poetry	Saint Ramdas	Dasbodh	5	1.6326	0.1523
Mr	poetry	Sant	Dnyaneshvari	6	1.9703	0.1006
Mr	commerce	Nene	Bankingshi Manthri	4	1.4929	0.1642
Mr	official	Gadkari	Chaupher	8	1.8449	0.0531
Ro	poetry	Eminescu	Scrisori (Letters)	5	1.8002	0.0552
Ru	prose	Ostrovskij	Kak zakaljalas stal´	10	1.9485	0.1171

Sm	prose	Legends	Tala o le Vavau	5	0.8277	0.0497
Sk	poetry	Rúfus	Dielo	3	1.6777	0.1030
Sp	prose	Cervantes	Don Quijote	15	0.9215	0.0520
Tg	poetry	Tolentino	Dakilang Asal	10	1.4205	0.1139
Tg	prose	De los Reyes	Dalagang Marmol	7	1.4236	0.0514

For the sake of illustration we present the *u*-tests only for German authors mentioned in Table 4.4. All *u* values greater than 1.96 are significant, i.e. they mean significant dissimilarity of frequency structure. In all cases we use the two-sided test, i.e. *u* is always positive. The results are presented in Table 4.5.

Here we compare 15 authors and obtain the following degrees of dissimilarity according to the method (2b):

1 Keller
7 Droste-Huelshoff, Wedekind
8 Raabe
9 Chamisso, Eichendorf, Kafka,. Lessing, Novalis, Schnitzler, Sealsfield
10 Löns
11 Meyer, Paul
12 Tucholsky

Thus, within the set of compared writers, Tucholsky seems to have the most dissimilarities, i.e. to be original in his frequency structures. This has, of course, nothing to do with the literary quality of his writings. But the fact that his “writing behaviour” significantly differs from that of other authors opens the question for an historian of literature or a literary critic. Perhaps, this method, which is in principle focused on a global property of the author´s style, could inspire literary critics, who analyse literary works qualitatively, to rethink both the literary works and the role of a particular author in the given period. On the other end of the scale, Keller has only one dissimilarity, i.e. his work is somewhere in the middle of German literature.

Further, 15 authors taken randomly from hundreds are not sufficient to display the properties of the dissimilarity graph. The degree distribution shown above displays a rather peculiar regularity.

Again, the unweighted ordering can be compared with the weighted one. The weighted ordering, as shown below, yields more detailed information but in general, it agrees with the unweighted one.

The weights are, of course, preliminary and relative: as soon as one adds a next writer or takes into account one more work of the given authors, the ordering can change.

The weighted ordering of 15 German writers is as follows

Keller	13.33
Wedekind	26.56
Droste-Huelshoff	26.68
Schnitzler	30.01
Raabe	34.68
Novalis	42.10
Kafka	43.61
Lessing	44.03
Chamisso	45.05
Sealsfield	48.79
Eichendorff	56.73
Löns	72.27
Paul	86.81
Meyer	92.85
Tucholsky	96.71

Table 4.5
Differences between German prose writers

Author	Chamisso	Droste-Huelshoff	Eichendorff	Kafka	Keller	Lessing	Löns	Meyer	Novalis	Paul	Raabe	Schnitzler	Sealsfield	Tucholsky
Droste-Huelshoff	0.89													
Eichendorff	4.79	2.74												
Kafka	0.24	0.99	4.45											
Keller	0.35	0.82	0.90	0.29										
Lessing	0.61	0.61	4.90	0.81	0.54									
Löns	7.03	3.96	3.16	6.72	1.78	6.99								
Meyer	5.12	0.84	11.99	5.29	1.55	3.69	12.89							
Novalis	0.34	1.03	4.18	0.11	0.25	0.89	6.44	5.19						
Paul	4.29	0.45	11.86	4.49	1.29	2.86	12.61	1.95	4.40					
Raabe	3.24	2.75	0.65	3.08	1.12	3.50	1.21	6.07	2.98	5.59				
Schnitzler	2.99	2.32	0.58	2.77	0.68	3.29	2.96	7.15	2.62	6.57	0.96			
Sealsfield	3.53	2.31	1.11	3.24	0.62	3.79	3.95	9.76	3.02	9.33	1.23	0.24		
Tucholsky	9.23	4.61	5.14	8.85	2.20	8.90	1.34	16.78	8.45	16.96	2.10	4.31	5.90	
Wedekind	2.40	2.36	0.28	2.28	0.94	2.65	1.23	4.58	2.20	4.16	0.20	0.57	0.76	1.94

5. Comparison of genres

Genre is a fuzzy concept encompassing whole families of individual concepts which can be combined and whose boundaries are set up by definition. For example a novel can concern adventure, science fiction, psychology, horror, Wild West, youth, love, history, etc. But the possibility of a combination of all these aspects of our life makes the classification impossible. Genre specialists did not even define the necessary components in the domain of individual genres. The story of Robin Hood is a mixed “historical adventure” novel but *The History of Great Britain* is a scholarly text. Why? What are the entities or units that make a text belong to genre A+B or to C? What is the proportion of science fiction, horror and adventure in the famous film *Alien*? Evidently, the concept of genres is still in a proto-scientific state without any clear definitions, quantifications and hypothesis formation. In spite of this shortage one can always draw a line between some “clearly” defined texts like prose, poetry, scientific writing, e.g. one can say that a non-poetic and non-scientific text in a sample – in which there are no stage plays – is prose.

Our aim is to differentiate genres using an elementary quantification without recourse to the content or style of a text. The study is preliminary and merely sketches one of the possible ways of analysis.

Comparing different genres defined in advance must be done separately for every language. But even in that case the dispersion within the given genre in one language is so great that one cannot expect significant differences between the mean lambdas. However, for differentiating two genres even the difference of their variances can be taken into account.

In the data we collected, there are private letters, prose, poetry, stories told and written by children, Nobel lectures, journalistic texts, and scientific texts. We shall compare them explicitly in each language. The results are presented in Table 5.1. It contains only those languages in which we had at our disposal at least two different genres.

Let us illustrate the testing using Czech data. In Czech we compare *prose* having n = 37 texts whose $\bar{\Lambda}$ = 1.7198, $s^2(\bar{\Lambda})$ = 0.000699 with n = 34 pieces of *poetry* having $\bar{\Lambda}$ = 1.8625 and $s^2(\bar{\Lambda})$ = 0.001520 yielding u = 3.01 which is as significant as could be expected, i.e. in Czech, poetry has a more complex frequency structure or a greater vocabulary richness than prose. If we test the dispersions, we first multiply $s^2(\bar{\Lambda})$ with n in order to obtain $s^2(\Lambda)$, then we use the F-criterion

$$(6.1) \quad F\left(n_1 - 1, n_2 - 1\right) = \frac{s_1^2}{s_2^2} \ ,$$

where we put the greater variance in the numerator and the smaller one in the denominator. Of course, we suppose that the data are normally distributed since we sampled the texts randomly. For Czech we obtain s^2(prose) = 0.000699(37) = 0.025863, and s^2(poetry) = 0.001520(34) = 0.05168. From this we obtain F(33,36) = 0.05168/0.02586 = 2.00 yielding the probability 0.022 which indicates a significant difference.

Table 5.1
Comparison of genres using the *u*-test
(Ae = Aesthetics, Ch = Children stories, Co = commerce, HP = History and philosophy, J = journalistic texts, L = letters, NL = Nobel lectures, NS = Natural sciences, Of = official and media texts Po = Poetry, Pr = Prose, PS = Presidential speeches, Rh = Rhetorics, Sc = Scientific texts, SS = Social sciences, Tr = Prose translation)

Genre	***n***	$\bar{\Lambda}$	$s^2(\bar{\Lambda})$	$\lvert u\rvert$
Bulgarian				
L	5	1.5747	0.000925	(L,Tr) = 0,82
Tr	10	1.6134	0.001316	
Czech				
Pr	37	1.7198	0.000699	(Po,Pr) = 3.07, (Po,Sc) = 9.4, (Po,Tr) = 0.76, (Po,Ch) = 6.98, (Pr,Sc) = 8.45, (Pr,Tr) = 4.27, (Pr,Ch) = 4.92, (Sc,Tr) = 11.28, (Sc,Ch) = 4.51, (Tr,Ch) = 8.55
Po	34	1.8646	0.001520	
Sc	10	1.3264	0.001469	
Tr	10	1.9038	0.001154	
Ch	68	1.5363	0.000693	
English				
Po	18	1.4450	0.003254	(Po,Pr) = 2.17, (Po,NL) = 2.03, (Po,Sc) = 5,70, (Po,Ch) = 2.99, (Pr,NL) = 0.19, (Pr,Sc) = 4,63, (Pr,Ch) = 1.05, (NL,Sc) = 4.82, (NL,Ch) = 1.05, (Sc,Ch) = 4.94
Pr	56	1.2981	0.001320	
NL	21	1.3079	0.001315	
Sc	10	1.0528	0.001486	
Ch	39	1.2651	0.000359	
French				
Po	8	1.2817	0.000210	(Po,Pr) = 4.67
Pr	19	1.1096	0.001147	
German				
Po	25	1.6115	0.002226	(Po,Pr) = 2.88
Pr	238	1.4708	0.000156	
Hungarian				
Po	15	1.9924	0.003189	(Po,Pr) = 2.46, (Po,J) = 1.81, (Pr,J) = 6.70
Pr	30	1.8444	0.000420	
J	5	2.1116	0.001172	
Italian				
Po	8	1.4111	0.010296	(Po,Pr) = 0.37,

Pr	10	1.3668	0.000964	(Po,PS) = 1.01, (Pr,PS) = 3.11
PS	60	1.5154	0.000443	
Kannada				
SS	9	1.9369	0.002649	(SS,Co) = 5.08
Co	9	1.6006	0.001725	
Latin				
Po	18	1.8754	0.003307	(Po,Pr) = 7,60, (Po,HP) = 0.80, (Po,Rh) = 5.59, (Pr,HP) = 3.04, (Pr,Rh) = 20.5, (HP,Rh) = 0.82
Pr	12	2.3335	0.000322	
HP	5	1.6748	0.046591	
Rh	6	1.4949	0.001327	
Marathi				
Po	11	1.8168	0.004128	(Po,Ae) = 1.66, (Po,SS) = 2.31, (Po,NS) = 3.48, (Po,Co) = 2.77, (Po,Of) = 1.10, (Ae,SS) = 6.28, (Ae,NS) = 6.47, (Ae,Co) = 5.13, (Ae,Of) = 0.71, (So,NS) = 2.10, (SS,Co) = 1.25, (SS,Of) = 5.02, (NS,Co) = 0.48, (NS,Of) = 5.63, (Co,Of) = 4.46
Ae	10	1.9379	0.001225	
SS	9	1.6540	0.000818	
NS	10	1.5276	0.002796	
Co	10	1.5672	0.003988	
Of	10	1.9003	0.001593	
Romanian				
Po	42	1.8462	0.000599	(Po,Pr) = 2.41
Pr	15	1.6369	0.006960	
Russian				
Po	5	1.6492	0.008299	(Po,Pr) = 1.23
Pr	20	1.7757	0.002328	
Slovak				
Po	9	1.6506	0.002237	(Po,Pr) = 2.13, (Po,Tr) = 4.13, (Pr,Tr) = 6.61
Pr	5	1.5071	0.002294	
Tr	10	1.8842	0.000960	
Slovenian				
Pr	5	1.6109	0.006209	(Pr,Tr) = 2.19
Tr	10	1.7937	0.000785	
Tagalog				
Po	11	1.4477	0.001800	(Po,Pr) = 2.21
Pr	16	1.3104	0.002046	

Without any test, we can state that in all the languages we studied, except for Latin where we took only one prose writer (Apuleius) and Russian where we took only one poetry writer, there is the same order of genres as expressed by lambda:

poetry > prose.

However, even Latin is no exception to this general rule if we consider Latin poetry represented by the best poems of Horatius and Vergilius, as given, for instance, in Table 5.2.

Table 5.2
Some Latin poems

Author	**Text**	***N***	***V***	***L***	***Λ***
Horatius, Flaccus	Carmina. Liber I	3969	2679	2782.1327	2.5226
Horatius, Flaccus	Carmina. Liber II	2412	1813	1886.0513	2.6448
Horatius, Flaccus	Carmina. Liber III	4359	3020	3181.7865	2.6565
Horatius, Flaccus	Carmina. Liber IV	2882	2056	2141.2471	2.5705
Horatius, Flaccus	Ars Poetica	3089	2046	2177.3291	2.4599
Horatius, Flaccus	Epodes	3023	2048	2108.5300	2.4276
Horatius, Flaccus	Carmen Saeculare	311	278	288.6956	2.3140
Vergilius, Publius	Georgicon Liber I	3306	2207	2324.4387	2.4744
Vergilius, Publius	Georgicon Liber II	3518	2263	2393.3385	2.4126
Vergilius, Publius	Georgicon Liber III	3698	2349	2479.5678	2.3924
Vergilius, Publius	Georgicon Liber IV	3658	2365	2488.7355	2.4243

Indeed, the mean lambda = 2.4818 of the above 11 poetry positions of Horatius plus Vergilius is significantly greater than the mean lambda = 2.3335 of the 12 prose positions of Apuleius considered in the Appendix. We note that lambda = 2.6565 of the poem *Carmina Liber III* by Horatius is the greatest lambda value we have observed so far in general. This is closely followed by lambda = 2.5518 of the Hungarian poem *Forgószélben* by Árpás.

However, in the languages where public texts were analyzed, these types of text were always in first place. In this research, public texts are journalistic texts, speeches of Italian presidents, i.e. texts for direct public use. They appear either on TV or in newspapers. Nobel lectures are public texts but written under different style regimes. Evidently, this ordering is correlated with the thematic/ lexical diversity of texts. In public texts one finds a great thematic diversity and/or very detailed, quick information. The poet usually expresses his idea in a many-coloured language in order to endow his text with plasticity. Prose is monothematic and describes something in a long-winded manner. Scientific texts are monothematic over long passages; they do not have a rich vocabulary and stand, wherever we analyzed them, below prose. Children have a restricted vocabulary and a restricted theme and stay in English in the last place. Thus different genres convey information in different ways and in different quantities, and use different vocabulary leading to different frequency structuring. In general, we have the following order

public texts > poetry > prose > children texts > scientific texts

This order can occasionally be disturbed by an outlier striving for originality, but by and large the above ordering is a background for studying stylistic and lexical deviations of individual texts. Idiosyncrasy can be stated only against a well-known background. Unfortunately, one cannot set up lambda-intervals for individual genres because they differ from language to language. For example, a 95% confidence interval for Italian poetry is <1.2122, 1.6100> while for Latin poetry we obtain <1.7627, 1.9881>, and for French poetry <1.2533, 1.3101> which lies within the Italian interval, and both French and Italian are much lower than Latin – a fact holding also for the prose. One can, of course, ask whether this is associated with the gradual loss of Latin synthetism and at the same time with the growth of dictionary in modern languages, but in order to arrive at reasonable statements a thorough historical study would be necessary. In any case, lambda can be used as one of the several indicators of the change in frequency structuring which is linked with the overall language type and its history.

It must be emphasized that the above intervals are set up for means, not for individual texts. Our data are not sufficient for setting up intervals for individual texts in every genre and every language. Since we treat the data as entities abiding by normality, only the use of corpora could be helpful for solving at least some aspects of the problem. But even with corpora the problem remains complicated because one must distinguish not only subcategories of genres but also the time of origin of every work. If a certain genre developed 250 years, then putting the oldest text in the same class as the newest one of the same category would mean ignoring an important factor, viz. time. But any segmentation of time in historical or political or cultural epochs is artificial. Possibly the lambda technique could be of some help here.

Special attention should be paid to translations. Here we cannot compare the translation of N. Ostrovskij´s *How the steel was tempered* in all Slavic languages because we do not have other original prose texts in these languages at our disposal. However, in those we had and analyzed here (Bulgarian, Czech, Slovak, Slovenian), the translation has a higher mean lambda than the original texts. This is caused by the fact that the translators were bound to a foreign frequency structure and a foreign vocabulary which must be translated exactly and sometimes with paraphrasing.

The study of lambda can be used for intra-genre studies, too. In this case, it seems reasonable to observe the lambda values of individual complete works, and not mixtures of texts like collected works, anthologies, or arbitrarily segmented texts. This kind of study can reveal the structure of genre based on the frequency structure of particular texts. Obviously, for an appropriate analysis a very large database is necessary. The potential fruitfulness of the approach can be illustrated in our database of Czech prose which contains only complete works.

If the texts from Table 4c in the Appendix are ranked according to decreasing lambda values, as is shown in Table 5.3, a great range of lambda values appears, namely from Λ = 1.3179 to Λ = 2.0143.

Table 5.3
Czech prose ranked according to decreasing lambda

Author	**Text**	***N***	***Λ***
Hájek, P.	Kráska a netvor	11695	2.0143
Hrabal, B.	Expozé panu ministru (Jarmilka, 44–47)	1044	1.9783
Jedlička, J.	Kde život náš je v půli se svou poutí	24666	1.9489
Škvorecký, J.	Eva byla nahá	13106	1.9076
Weil, J.	Žalozpěv za 77 297 obětí	3565	1.8921
Hrabal, B.	Pogrom	1839	1.8835
Hrabal, B.	Protokol (Jarmilka, 129–131)	999	1.8826
Hrabal, B.	Veselé vánoce	1264	1.8748
Flos, F.	Lovci kožešin	4568	1.8707
Hrabal, B.	Květnové idy	1196	1.8580
Hrabal, B.	Pohádka o zlaté Praze	1448	1.8358
Hrabal, B.	Modrý pondělí	2520	1.8141
Hrabal, B.	Praha, město utajených infarktů	2004	1.7943
Hrabal, B.	Lednová povídka (Jarmilka, 58–61)	984	1.7831
Hrabal, B.	Blitzkrieg (Jarmilka, 86–87)	522	1.7805
Hrabal, B.	Modrý pokoj	1668	1.7598
Páral, V.	Veletrh splněných přání	21275	1.7360
Hrabal, B.	Únorová povídka (Jarmilka, 62–69)	2858	1.7317
Hrabal, B.	Zavražděný kohout	1435	1.7313
Hrabal, B.	Kůň truhláře Bárty	1500	1.7045
Hrabal, B.	Česká rapsodie	1989	1.6966
Fejt, V.	Věž	15818	1.6900
Hovorka, J.	Okarína do re mi	23402	1.6865

Klíma, L.	Sus triumfans	10073	1.6854
Viewegh, M.	Názory na vraždu	25476	1.6616
Krupička, J.	Stará pevnost	29947	1.6371
Karel, M.	Gypsová dáma	17654	1.6123
Klíma, L.	Melia	8264	1.5979
Pecka, K.	Pasáž	28758	1.5830
Vávra, V.	Muž v jiných končinách světa	23528	1.5782
Vodňanský, J.	Velký dračí propadák aneb Král v kukani	16145	1.5766
Bondy, E.	Leden na vsi	23643	1.5600
Správcová, B.	Spravedlnost	22041	1.5402
Fučík, J.	Reportáž psaná na oprátce	23815	1.5047
Körner, V.	Adelheid	24943	1.5019
Uhde, M.	Modrý anděl	5459	1.4198
Fischl, V.	Strýček Bosko	15738	1.3179

Here, it is striking that the lambdas of Hrabal’s short stories, except for *Expozé panu ministru* (*Jarmilka*, 44–47), lie within a “narrow” interval <1.6966, 1.8835>. Unfortunately, the database does not comprise Hrabal’s novels and, consequently, it is not possible to explore whether the clustering of Hrabal’s works is caused only by the authorship or whether the type of prose (short story vs. novel) also influences the frequency structure of Hrabal’s texts.

Moreover, one can expect that there are some other text properties which may have an impact on lambda, for example the thematic concentration of text as defined by Popescu et al. (2008) or the vocabulary richness (see Chapter 1). A multifactor intra-genre analysis based on quantitative characteristics of texts is in principle possible and it could reveal the potential structure of the genre. Perhaps, this kind of observation will shed more light on the concept of genre itself.

6. Comparison of languages

6.1. The same text

If we treat different texts in a language, then the mean lambda may vary considerably, according to the characteristics of the analyzed texts. Even if we use a corpus, the variation of lambdas is so large that a 95% confidence interval of one language may encompass a typologically different language. Nevertheless, an ordering according to lambda is always possible. We suppose that by increasing the number of texts in one language the mean lambda will – perhaps – approach a stable value but the variance may increase, too.

Consider, for example the lambda in Slavic languages. If we take only Chapter 4 with peak lambda (cf. Figure 7.1) of the translation of Ostrovskij´s *How the steel was tempered*. into other languages, we have the lambdas as presented in Figure 6.1 (dashed line). However, if we consider all the ten chapters, we obtain lower values, as can be seen in Figure 6.1 (solid line).

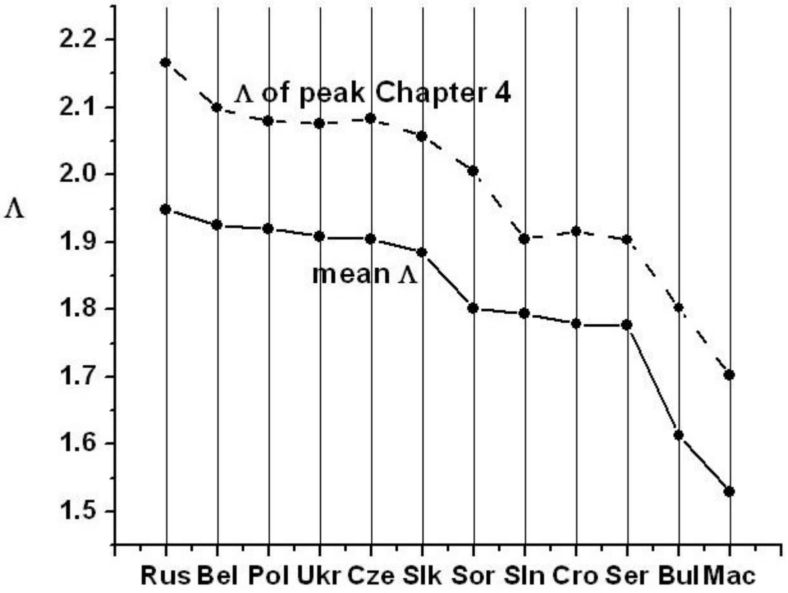


Figure 6.1. Mean lambda ranking in Slavic languages

The consequence of this fact forces us to compare texts of the same genre in different languages and define some kind of difference as a function of all genre differences. Needless to say, comparisons of different genres even in different languages are possible without much additional effort, provided we have relatively stable estimations of the means of the given genres in the given languages. They must be used as expectations, as has been shown at the end of 3.1.

We begin the analysis with comparing the mean lambdas of individual translations of the novel *How the steel was tempered* by N. Ostrovskij written in Russian in 1932–34 but revised and ideologically modified several times by Russian publishers. It has been translated into 11 Slavic languages in different years: Slovenian and Croatian 1945, Czech 1948, Serbian 1949, Belorussian 1950, Sorbian 1960, Slovak 1966, Polish and Ukrainian 1974, Bulgarian 1976, and Macedonian 1988. Kelih (cf. 2009a,b) prepared a corpus of 10 chapters in each language which is available on the Internet. The analysis yielded the results presented in Table 6.1. The lambdas of individual chapters can be found in the Appendix. Here we present only the means and the variances of mean lambdas. Since in each language there are 10 chapters, the variance of the ten chapters can be obtained as $s^2(\Lambda) = 10s^2(\bar{\Lambda})$.

Table 6.1
Mean lambdas of 10 chapters of N. Ostrovskij´s novel *How the steel was tempered* translated from Russian in 11 Slavic languages

Language	$\bar{\Lambda}$	$s^2(\bar{\Lambda})$
Belorussian	1.9247	0.001177
Bulgarian	1.6134	0.001316
Croatian	1.7786	0.001065
Czech	1.9038	0.001154
Macedonian	1.5290	0.001453
Polish	1.9195	0.001016
Russian	1.9485	0.001371
Serbian	1.7761	0.000996
Slovak	1.8842	0.000962
Slovenian	1.7937	0.000784
Sorbian	1.8024	0.001933
Ukrainian	1.9089	0.001425

We performed the *u*-test for comparing the means. The results are given in Table 6.2. Using only the number of *significant dissimilarities* for each language we obtain the classification

5	Czech, Slovak, Sorbian, Ukrainian
6	Belorussian, Polish, Russian
8	Croatian, Serbian, Slovenian
10	Bulgarian, Macedonian

Table 6.2
The u-test for mean lambdas in 12 Slavic languages

	Bel	**Bul**	**Cr**	**Cz**	**Mac**	**Pol**	**Rus**	**Ser**	**Slk**	**Slov**	**Sor**
Bul	6.23										
Cr	3.09	3.39									
Cz	0.43	5.84	2.66								
Mac	7.72	1.60	4.97	7.34							
Pol	0.11	6.34	3.09	0.34	7.86						
Rus	0.47	6.46	3.44	0.89	7.89	0.59					
Ser	3.19	3.38	0.06	2.75	4.99	3.20	3.54				
Slk	0.88	5.67	2.35	0.43	7.23	0.79	1.33	2.44			
Slv	2.96	3.93	0.35	2.50	5.60	2.97	3.33	0.42	2.17		
Sor	2.19	3.32	0.43	1.83	4.70	2.16	2.54	0.49	1.52	0.17	
Ukr	0.31	5.64	2.61	0.10	7.08	0.21	0.75	2.70	0.51	2.45	1.84

As in many other cases, Bulgarian and Macedonian stay at the periphery and display the greatest divergence, while the rest of the South Slavic languages move away from the more conservative ones. This picture is only an echo of well known classifications but it shows that the lambda has a certain power even if it is only an image of a purely structural, very immaterial and non-etymological entity.

Using the *non-significant dissimilarities* in Table 6.2 we obtain a graph displaying the situation, cf. Figure 6.2.

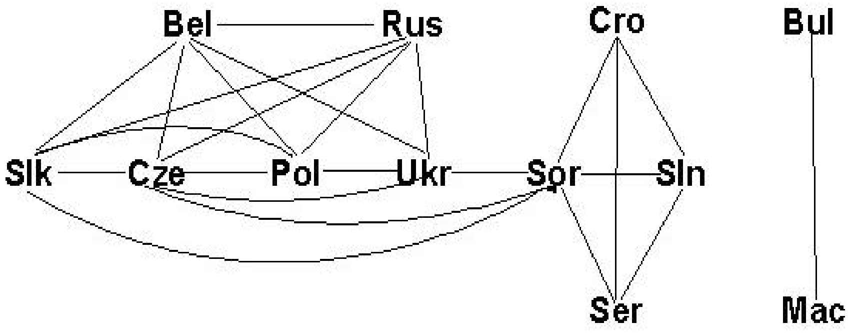


Figure 6.2. Structural similarities between Slavic languages

6.2. Different texts of the same genre

Using the tables in Chapter 4 where individual texts are presented according to their genre we summarize the results in Table 6.3 and perform some comparisons between languages. Here n = number of texts or text parts used for computing lambda, $s^2(\bar{\Lambda})$ = the variance of the mean, i.e. $s^2(\Lambda)/n$.

Again, we use the asymptotic u-test. Comparing e.g. Czech and Slovak poetry we obtain

$$u = \frac{1.8646 - 1.6505}{\sqrt{0.001437 + 0.002236}} = 3.53$$

which is highly significant. A similar result can be obtained for the comparison of prose, viz. u = 3.89. Hence either Czech writers have richer texts or Czech is more synthetic than Slovak in spite of their great similarity (e.g. in Czech even the conditional conjunction can be conjugated). However, Slovak prose does not differ significantly from Turkish prose, etc.

Table 6.3
Mean lambdas in different genres of individual languages

Language	**Genre**	***n***	$\bar{\Lambda}$	$s^2(\bar{\Lambda})$
Slovak	poetry	9	1.6506	0.002236
Slovak	prose	5	1.5071	0.002293
Czech	poetry	35	1.8646	0.001437
Czech	prose	37	1.7198	0.000699
Turkish	prose	40	1.5022	0.000966
Hawaiian	prose	38	0.6533	0.000117
Dutch	prose	5	0.9628	0.010262
Finnish	prose	5	1.5884	0.007743
English	prose	56	1.2922	0.001320
English	poetry	19	1.4276	0.003214
Italian	prose	7	1.3706	0.001722
Italian	poetry	8	1.4111	0.010295
Hungarian	prose	30	1.8444	0.000420
Hungarian	poetry	15	1.9924	0.003189
Romanian	prose	15	1.6369	0.006959
Romanian	poetry	42	1.8290	0.000689
Latin	prose	12	2.3335	0.000322
Latin	poetry	18	2.0078	0.003307
Japanese	prose	7	0.9867	0.001747
Spanish	prose	18	0.9595	0.000642

German	prose	236	1.4697	0.000158
German	poetry	27	1.6109	0.001907
Kannada	prose	5	1.9899	0.004043
Lakota	prose	4	1.2460	0.003656
Maori	prose	5	0.8954	0.004286
Marathi	prose	5	1.7139	0.012930
Marathi	poetry	11	1.8168	0.004130
Marquesan	prose	7	0.7488	0.006345
Rarotongan	prose	5	0.8836	0.002268
Russian	poetry	5	1.6492	0.009300
Russian	prose	10	1.6029	0.001829
Samoan	prose	5	0.8277	0.000495
Slovenian	prose	5	1.6109	0.006206
Tagalog	prose	3	1.3499	0.005230
French	poetry	7	1.2846	0.000269
French	prose	20	1.0889	0.000680

Comparing mean lambdas of poetry and performing the *u*-test we obtain the results presented in Table 6.4.

Table 6.4
The u-test for comparing the mean lambdas of poetry in 11 languages

	Cz	**E**	**Fr**	**G**	**Hu**	**It**	**Lt**	**Mr**	**Ro**	**Rus**
E	6.41									
Fr	14.04	2.42								
G	4.39	2.56	6.99							
Hu	1.88	7.06	12.04	5.34						
It	4.19	0.14	1.23	1.81	5.01					
Lt	2.08	7.18	12.09	5.50	0.19	5.12				
Mr	0.64	4.54	8.02	2.65	2.05	3.38	2.21			
Ro	0.77	6.43	17.59	4.28	2.62	3.99	2.83	0.18		
Rus	2.08	1.98	3.73	0.36	3.07	1.70	3.19	1.45	1.80	
Slk	3.53	3.02	7.31	0.62	4.64	2.14	4.80	2.08	3.30	0.01

Again, a classification could be performed but we simply consider only the non-significant dissimilarities and link the respective languages with an edge. We obtain a weighted graph presented in Figure 6.3.

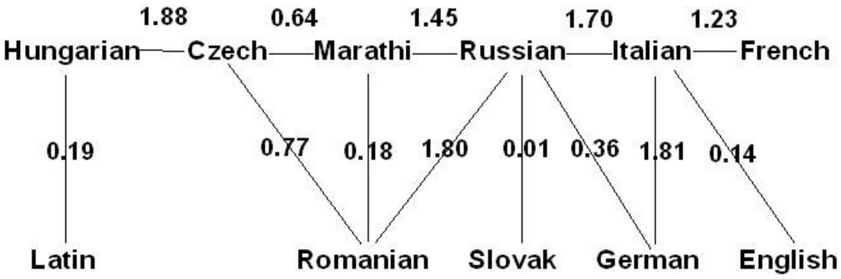


Figure 6.3. Structural similarities between poetic texts

Though poetry is not a good background for language comparison, we do not compare here morphological properties or kinships but a quite abstract property. Nevertheless, if we look at the graph, we see from left to right decreasing synthetism. Of course the sequence is not perfect and would be more complex if more languages were added.

Comparing the prose in several languages we obtain the results as shown in Table 6.5. One could, again, perform a classification or ordering according to the number of significant dissimilarities or the sums of *u*'s for individual languages or using one of about 500 taxonomic methods, but adding even one more language may change everything. Hence, we would rather draw the graph of similarities (non-significant dissimilarities) as shown in Figure 6.4

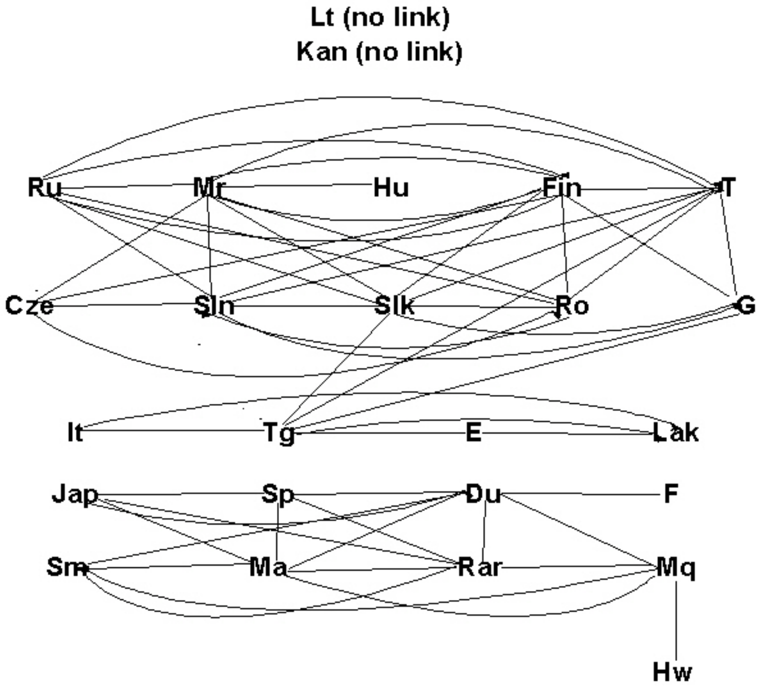


Figure 6.4. The *u*-test for comparing the mean lambdas of prose in 25 languages

The graph yields a very complex picture which is preliminarily not easy to interpret. To obtain an image which can be explicated both historically, typologically and geographically one would need a great amount of only modern texts. Nevertheless, even now, the presented “lambda-technique” shows that strongly synthetic languages tend to form a group. First of all we remark that Latin and Kannada have absolutely no similarity with the other languages; thus a classical language and a Dravidian language stand apart. Continuing, the three strongly agglutinating languages (Hungarian, Turkish, Finnish) and the strongly inflectional cognate languages (Marathi, Romanian, Russian, Czech, Slovenian, Slovak, German) belong to the main cluster. The mid cluster signalizes a development towards analytism. The four languages in it (Italian, Tagalog, English and Lakota) belong to three language families. English has no similarity with the synthetic group. The group containing Polynesian languages is rather analytic. In Japanese the frequency structure depends on the definition of the word. If postpositions are considered affixes, Japanese would be nearer to the synthetic group. French has a strong trend towards analytism – depending on word definition – and Spanish and Dutch depend here on the choice of texts: modern texts in Spanish would show a different picture. Dutch has lost a lot of Indo-European inflection.

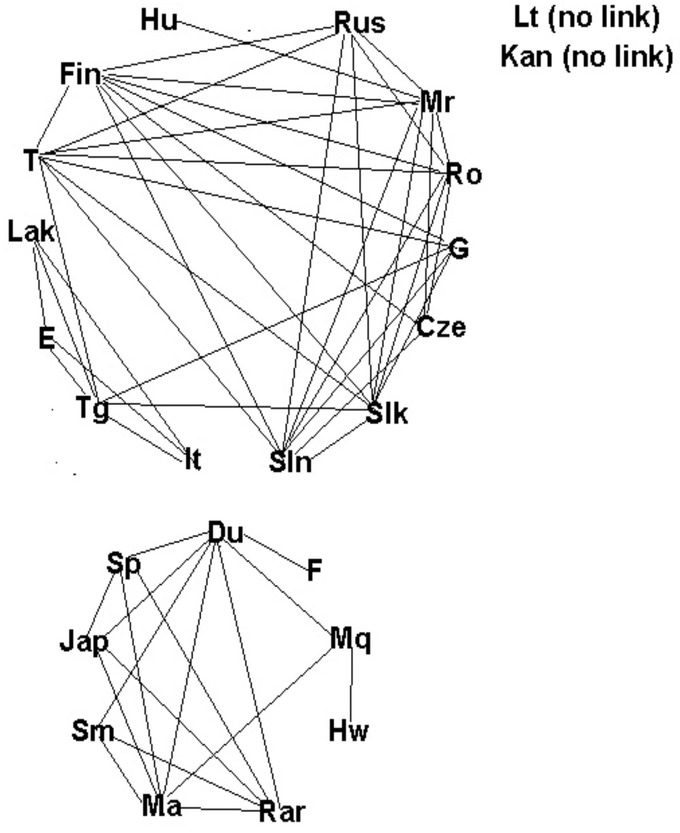


Figure 6.5. The *u*-test for comparing the mean lambdas of prose in 25 languages

Table 6.5

The *u*-test for comparing the mean lambdas of prose in 25 languages

	Cze	Du	E	Fin	F	G	Hw	Hu	It	Jap	Kan	Lak	Lt	Ma	Mr	Mq	Rar	Ro	Rus	Sm	Slk	Sln	Sp	Tg
Dutch	7.23																							
English	9.52	3.06																						
Finnish	1.43	4.66	3.11																					
French	16.99	1.21	4.55	5.44																				
German	8.54	4.97	4.62	1.34	13.15																			
Hawaiian	37.33	3.04	16.9	10.6	15.43	49.23																		
Hungarian	3.72	8.53	13.2	2.83	22.78	15.59	51.40																	
Italian	7.10	3.73	1.42	2.24	5.75	2.29	16.73	10.24																
Japanese	14.82	0.22	5.52	6.18	2.07	11.07	7.72	18.42	6.52															
Kannada	3.92	8.59	9.53	3.7	13.11	8.03	20.72	2.18	8.16	13.18														
Lakota	7.18	2.40	0.65	3.21	2.39	3.62	9.65	9.37	1.70	3.53	8.48													
Latin	19.21	13.32	25.70	8.30	39.32	39.43	80.19	17.96	21.30	29.61	5.20	17.24												
Maori	11.68	0.56	5.3	6.32	2.75	8.61	3.65	13.83	6.13	1.18	11.99	3.93	21.19											
Marathi	0.05	4.93	3.53	0.87	5.36	2.13	9.29	1.13	2.84	6.00	2.12	3.63	5.38	6.24										
Marquesan	11.57	1.66	6.21	7.07	4.06	8.94	1.19	13.32	6.92	2.64	12.18	4.97	19.41	1.42	6.95									
Rarotongan	15.35	0.71	6.82	7.04	3.78	11.90	4.71	18.53	7.71	1.63	13.93	4.71	28.49	0.15	6.74	1.45								
Romanian	0.95	5.14	3.79	0.4	6.27	1.98	11.69	2.42	2.86	6.97	3.37	3.79	8.16	6.99	0.55	7.70	7.84							
Russian	2.32	5.82	5.54	0.15	10.26	2.99	21.53	5.09	3.90	10.30	5.05	4.82	15.75	9.05	0.91	9.45	11.24	0.36						
Samoan	25.82	1.30	10.9	8.38	7.62	25.12	7.05	33.61	11.53	3.36	17.25	6.49	52.68	0.98	7.65	0.95	1.06	9.37	16.08					
Slovak	3.89	4.86	3.58	0.81	7.67	0.76	17.39	6.48	2.15	8.19	6.07	3.39	16.16	7.54	1.68	8.16	9.23	1.35	1.49	12.87				
Slovenian	1.31	5.05	3.67	0.19	6.29	1.77	12.04	2.86	2.70	7.00	3.74	3.67	8.94	6.99	0.74	7.70	7.90	0.23	0.09	9.57	1.13			
Spanish	20.76	0.03	7.51	6.87	3.56	18.04	11.11	27.15	8.46	0.56	15.05	4.37	44.25	0.91	6.48	2.52	1.41	7.77	12.94	3.91	10.11	7.87		
Tagalog	4.80	3.11	0.71	2.09	3.4	1.63	9.53	6.58	0.25	4.35	6.65	1.10	13.20	4.66	2.70	5.59	5.39	2.60	3.01	6.90	1.81	2.44	5.09	
Turkish	5.33	5.09	4.39	0.92	10.19	0.97	25.80	9.19	2.54	9.90	6.89	3.77	23.16	8.37	1.80	8.81	10.88	1.51	1.90	17.65	0.09	1.28	13.53	1.93

We are aware of the fact that this image is preliminary. It is only a first comparison of text structure on a very abstract level void of the influence of text size. However, we are sure that a thorough analysis of present day texts would bring both some corrections and strengthening of some similarities. Figure 6.4 contains an implicit interpretation and is made in this form for the sake of lucidity. The usual circular form of the same state is presented in Figure 6.5.

Usually one evaluates the quantitative properties of such graphs but it would be premature here. All we have at present is a package of intuitive hypotheses which cannot be tested because other properties with which the frequency structure could be related are not yet quantified. The research in this direction must be postponed.

7. Text development

7.1. Change of lambda

Authors writing longer works necessarily make pauses in writing. The pause may be made in the given chapter or between two chapters of a novel. The latter case is quite normal. Either the next chapter is a continuation of the previous one or a jump is made and a new vocabulary is used, as can be seen in the studies on the type-token relation. With the new chapter, the frequency structure may change. Thus each chapter is a unit of its own and the lambda can show whether the author changes his techniques or enriches the vocabulary of the text. Of course, the length of chapters may be different but lambda eliminates the size differences.

For the sake of illustration let us consider the sequence of lambdas in the Russian novel by Ostrovskij *How the steel was tempered*, taken from the *Kelih-Corpus* (2009a,b). Looking at Table 7.1, where the chapters are ordered, one can see that lambda does not have a smooth course. It oscillates around the mean $\bar{\Lambda}$ = 1.9485.

Table 7.1
Lambda in the first ten chapters of Ostrovskij´s novel
How the steel was tempered in Russian

Chapter	***N***	***L***	***Λ***	**σ(*Λ*)**
1	4107	2051	1.8043	0.000193
2	4136	2217	1.9383	0.000200
3	6323	3091	1.8583	0.000187
4	3733	2264	2.1660	0.000224
5	3769	1982	1.8810	0.000137
6	7534	3519	1.8108	0.000113
7	6019	3106	1.9505	0.000143
8	5352	2927	2.0392	0.000123
9	3291	1839	1.9657	0.000296
10	5399	2995	2.0708	0.000113

We may ask whether this oscillation is too drastic at places. To this end we perform a series of *u*-tests between neighbouring chapters and obtain the results presented in Table 7.2. For example the test for Chapter 1 and 2 yields

$$u = \frac{1.8043 - 1.9383}{\sqrt{0.000193 + 0.000200}} = -6.76$$

This result shows that there are no non-significant steps. The minus sign means increase, the + sign means decrease of lambda.

Table 7.2
u-tests between neighbouring chapters in
Ostrovskij´s *How the steel was tempered* in Russian

Chapters	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
u	-6.76	+4.07	-15.18	+15.00	+4.44	-8.73	-5.43	+3.59	-5.19

Since we have the same text in other Slavic languages, let us consider the development of lambda in them graphically. In Figure 7.1 the translations of the same text are presented.

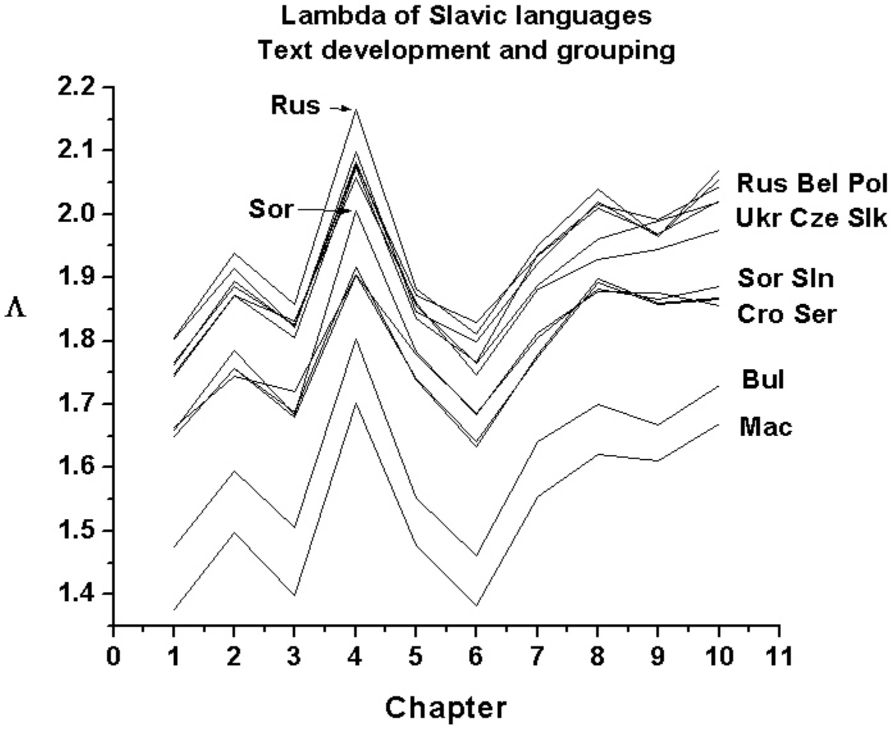


Figure 7.1. The change of lambda in 12 Slavic languages based on Ostrovskij´s novel *How the steel was tempered*

The profiles of all languages are quite similar but the heights are different. The height expresses *grosso modo* the degree of synthetism of a language and at the same time the geographic positions of Slavic languages. The Sorbian spoken in Germany has a transient position. It is possible to develop tests both for parallelism of the lines and their height. However a simple comparison will do.

7.1.1. Difference in height

In order to measure the ***difference in height*** between two languages, we simply compute the mean of the absolute differences in individual positions (here chapters), i.e.

$$(7.1)\quad DH = \frac{1}{n}\sum_{i=1}^{n} | \Lambda_i(L_1) - \Lambda_i(L_2) |,$$

where L_i (i = 1,2) is a language, and n is the number of comparisons. Consider for example the difference between Macedonian and Russian. The lambda values together with the computed absolute difference are presented in Table 7.3. The sum of the differences yields 4.1952 and the mean height difference is 4.1952/10 = 0.4195, which is the greatest among Slavic languages.

Table 7.3
Parallel lambda differences in Macedonian and Russian
in Ostrovskij´s *How the steel was tempered*.

Chapter	**Macedonian** Λ_1	**Russian** Λ_2	**Difference** $\|\Lambda_1 - \Lambda_2\|$
1	1.3770	1.8043	0.4273
2	1.4982	1.9383	0.4401
3	1.3992	1.8583	0.4591
4	1.7027	2.1660	0.4633
5	1.4771	1.8810	0.4039
6	1.3815	1.8108	0.4293
7	1.5541	1.9505	0.3964
8	1.6212	2.0392	0.4180
9	1.6100	1.9657	0.3557
10	1.6687	2.0708	0.4021

In the same way the differences between all Slavic data can be computed. The computation would yield a matrix which could be used for the classification of Slavic languages but it would yield only one more classification corroborating the other ones. Nevertheless, we present the differences in height in Table 7.4. The mean height differences can be obtained by dividing the numbers in Table 7.4. by 10. From the given data one can perform different kinds of taxonomies; here we restrict ourselves to showing the full difference of each language to the other ones. The result is presented in Table 7.5. The table presents in numerical form what can be seen in Figure 7.1: Macedonian is the most peripheral language

(= greatest sum of height differences); Slovak is the most central language (= smallest sum of height differences.)

Table 7.4
Parallel lambda differences in Slavic languages
in Ostrovskij´s *How the steel was tempered*.

	Bu	**Cr**	**Cz**	**Mac**	**Pol**	**Ru**	**Ser**	**Slk**	**Sl**	**Sor**	**Ukr**
Beloruss.	3.1138	1.4613	0.2804	3.9577	0.1982	0.2377	1.4866	0.4522	1.3106	1.2234	0.2368
Bulgarian		1.6525	2.904	0.8439	3.0616	3.3513	1.6272	2.708	1.8032	1.8904	2.956
Croatian			1.2515	2.4964	1.4091	1.6988	0.0467	1.0555	0.2275	0.3113	1.3035
Czech				3.7479	0.2134	0.4943	1.2768	0.2186	1.1008	1.0136	0.3579
Macedon.					3.9055	4.1952	2.4711	3.5519	2.6471	2.7343	3.7999
Polish						0.3341	1.4344	0.369	1.2584	1.1712	0.2126
Russian							1.7241	0.6433	1.5481	1.4609	0.4406
Serbian								1.0808	0.221	0.3122	1.3288
Slovak									0.9048	0.8176	0.347
Slovenian										0.2444	1.1528
Sorbian											1.0656

Table 7.5
Sum of height differences between Slavic languages
(referring to Ostrovskij´s *How the steel was tempered*)

Language	**Sum of height differences**
Macedonian	34.3509
Bulgarian	25.9119
Russian	16.1284
Belorussian	13.9587
Polish	13.5675
Ukrainian	13.2015
Serbian	13.0097
Croatian	12.9141
Czech	12.8681
Slovenian	12.4187
Sorbian	12.2449
Slovak	12.1489

7.1.2. Difference in profile

The ***difference in the profile*** of the sequences is the mean absolute difference between the individual slopes of the sequences. The slopes are computed individually, i.e., between the neighbouring chapters. Fortunately, the steps on the abscissa are given by natural numbers, hence the slope is always $\Lambda_{i+1} - \Lambda_i$, because $x_{i+i} - x_i = 1$. Hence the difference in profile is defined as

$$(7.2) \quad DP = \frac{1}{n-1}\sum_{i=1}^{n-1}\left[\,|\, (\Lambda_{1,i+1} - \Lambda_{1,i}) - (\Lambda_{2,i+1} - \Lambda_{2,i}) \,|\, \right].$$

For the profile difference between Macedonian and Russian we obtain

DP(Mac,Russ) = [|(1.4982 – 1.3770) – (1.9383 – 1.8043)| + |(1.3992 – 1.4982) – (1.8583 – 1.9393)| +… + [(1.6687 – 1.6100) – (2.0708 – 1.9657)|]/9 = 0.0316.

In Table 7.6 one finds the profile differences for the Ostrovskij book between all Slavic languages. Summing up the differences of each language to all the other ones one obtains the result presented in Table 7.7.

Table 7.6
Profile differences between Slavic languages concerning
Ostrovskij´s *How the steel was tempered*

	Bu	**Cr**	**Cz**	**Mac**	**Pol**	**Rus**	**Ser**	**Slk**	**Sl**	**Sor**	**Ukr**
Bel	0.2172	0.3538	0.2914	0.2722	0.2069	0.1772	0.3473	0.4354	0.4934	0.3574	0.2183
Bu		0.344	0.2518	0.088	0.2707	0.241	0.3536	0.3416	0.4808	0.2496	0.1303
Cr			0.272	0.3326	0.2797	0.4249	0.0509	0.2342	0.2776	0.329	0.2611
Cz				0.2154	0.2033	0.3666	0.2793	0.1686	0.3194	0.2248	0.1779
Mac					0.2689	0.284	0.3449	0.3016	0.4454	0.1688	0.1397
Pol						0.3033	0.2866	0.2829	0.3731	0.3213	0.1704
Rus							0.4339	0.466	0.5858	0.3688	0.2525
Ser								0.2555	0.2501	0.3161	0.3124
Slk									0.291	0.282	0.2865
Sl										0.4456	0.3987
Sor											0.2873

As can be seen, one obtains different results than with height differences. This fact is a reminder of the relativity of any classification. If one considers only one property, one can obtain a different result than with another single property. In the empirical sciences one replaces monothetic classifications with polythetic ones, but even these are merely approximations to something that should be substantiated theoretically. Today, one does not speak of the correctness of a classification but rather of its purposefulness. This is supported by the fact that the great number of numerical taxonomic methods used today may furnish quite different results. Classification is only a first ordering of the universe of discourse, and as long as it cannot be derived from a theory, it is only a report of facts.

Table 7.7
Sum of profile differences

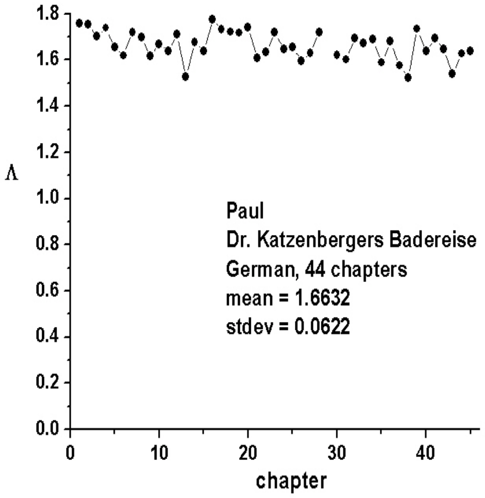
Language	***DP***
Slovenian	4.3609
Russian	3.9040
Belorussian	3.3705
Sorbian	3.3507
Slovak	3.3453
Serbian	3.2306
Croatian	3.1598
Bulgarian	2.9686
Polish	2.9671
Macedonian	2.8615
Czech	2.7705
Ukrainian	2.6351

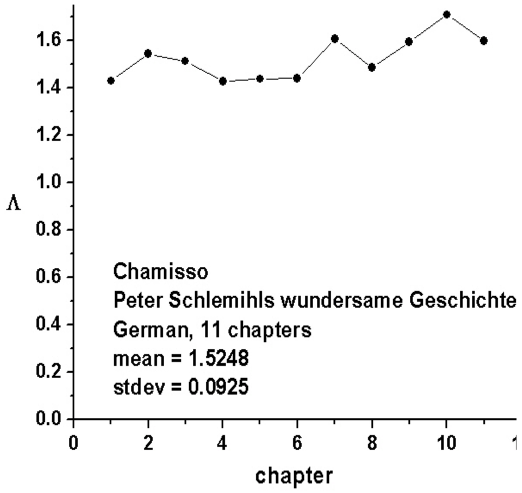
One can compare the development of lambda also in two texts which have different numbers of parts. Though such a comparison cannot be perfect, one compares at least as many parts as the shorter text has. The rest must be omitted.

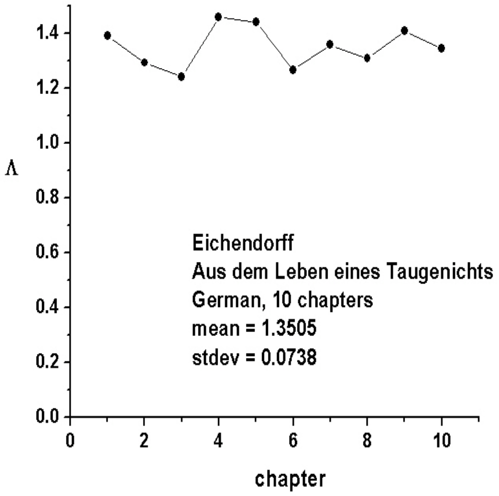
7.2. Mean sequential difference

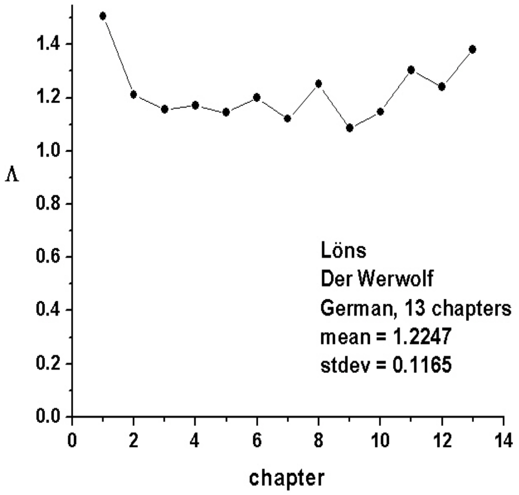
The course of lambda is not smooth. It depends on the writer, length of the work (in terms of its parts), genre, theme and surely also the historical time of its genesis. In order to illustrate this volatile property, we present graphically some results in Figure 7.2. In J. Paul´s *Dr. Katzenbergers Badereise* we left out Chapter 29 containing only 5 words, thus representing rather an anomaly.

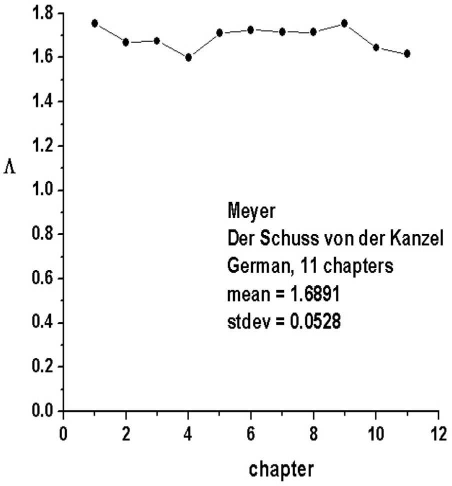
For the time being we do not see any regularity or controlling mechanism. But if one analyzed many works not very different in all the above mentioned dimensions, perhaps a pattern would appear, possibly in a form similar to the components of a Fourier sequence or some recurrent difference function.

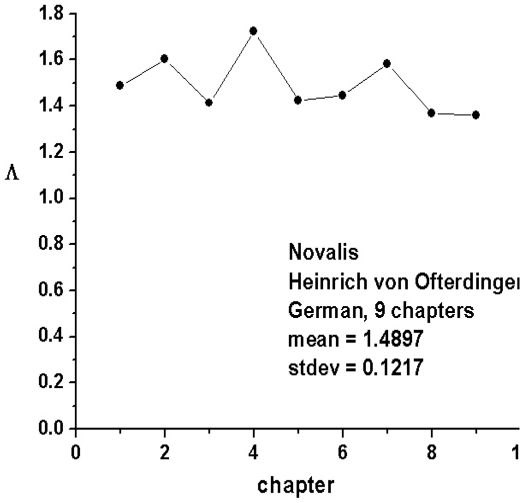

Λ
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
chapter
Paul
Dr. Katzenbergers Badereise
German, 44 chapters
mean = 1.6632
stdev = 0.0622


Λ
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
chapter
Chamisso
Peter Schlemihls wundersame Geschichte
German, 11 chapters
mean = 1.5248
stdev = 0.0925


Λ
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
chapter
Eichendorff
Aus dem Leben eines Taugenichts
German, 10 chapters
mean = 1.3505
stdev = 0.0738


Λ
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
14
chapter
Löns
Der Werwolf
German, 13 chapters
mean = 1.2247
stdev = 0.1165


Λ
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
12
chapter
Meyer
Der Schuss von der Kanzel
German, 11 chapters
mean = 1.6891
stdev = 0.0528


Λ
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
chapter
Novalis
Heinrich von Ofterdinger
German, 9 chapters
mean = 1.4897
stdev = 0.1217

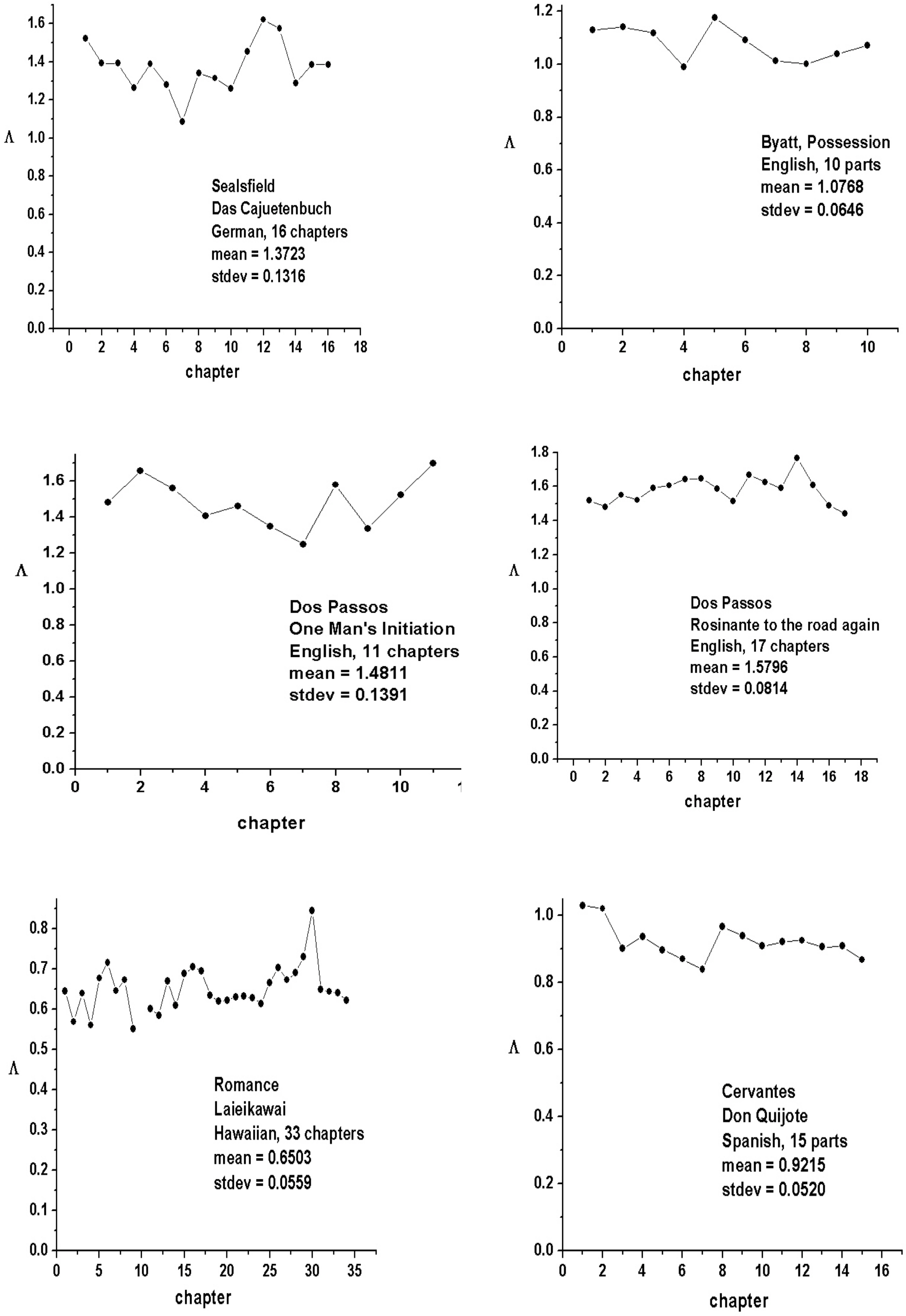


Figure 7.2. Courses of lambda in some longer works

Since lambda does not depend on *N*, the chapters of the text can – from our point of view – be considered as a series having different properties. A number of them have been presented in L. Hřebíček´s works (1997, 2000, 2007) and a number of other ones can be found in the literature (cf. e.g. Popescu, Mačutek, Altmann 2010). Here we shall stay at the lowest level and compute simply the mean sequential difference of the lambdas in one text. It is not the same as von Neumann´s mean square successive difference (von Neumann, Kent, Bellinson, Hart 1941), which can be transformed into the variance. We define the indicator of *sequential difference* (called also *roughness*) as

$$(7.3)\quad D_S = \frac{1}{n-1}\sum_{i=1}^{n-1} |\Lambda_i - \Lambda_{i+1}|.$$

It expresses the degree of frequency uniformity of a work consisting of several parts. Frequency uniformity is not the same as vocabulary similarity of parts, i.e. using the same words, but rather the uniformity of the technique of exploiting the dictionary. It is not a measure of "knowledge" of the word-stock but rather a technique of composition. The indicator D_S is akin to the Integer Lyapunov Coefficient (cf. Hřebíček 2000) which uses the logarithms of absolute differences.

Consider the sequence of lambdas in the first 10 chapters of the Russian version of Ostrovskij´s *How the steel was tempered* (Table 7.8).

Table 7.8
Lambda in the first ten chapters of Ostrovskij's novel
How the steel was tempered in Russian

Chapter	Λ	$\lvert\Lambda_i - \Lambda_{i+1}\rvert$
1	1.8043	0.1340
2	1.9383	0.0800
3	1.8583	0.3077
4	2.1660	0.2850
5	1.8810	0.0702
6	1.8108	0.1397
7	1.9505	0.0887
8	2.0392	0.0735
9	1.9657	0.1051
10	2.0708	

The sum of absolute differences is 1.2839 and D_S = 0.1427. In Table 7.9 the values of D_S in other Slavic languages (same text) and some other ones are shown.

Table 7.9
Mean sequential differences in some texts

Language	**Work**	D_S
Tamási	Szülöföldem	0.0354
Cervantes	Don Quijote	0.0376
Gárdonyi	Egri csillagok	0.0424
Meyer	Der Schuß von der Kanzel	0.0481
Hawaiian	Romance Laieikawai	0.0523
Gárdonyi	Isten rabjai	0.0545
J. Paul	Dr. Katzenbergers Badereise	0.0607
Byatt	Possession	0.0660
Dos Passos	Rosinante to the road again	0.0710
Apuleius, L.	Metamorphoses	0.0715
Slovenian	Ostrovskij, How the steel was tempered	0.0827
Chamisso	Peter Schlemihls wundersame Geschichte	0.0862
Eichendorff	Aus dem Leben eines Taugenichts	0.0959
Löns	Der Werwolf	0.1040
Slovak	Ostrovskij, How the steel was tempered	0.1042
Serbian	Ostrovskij, How the steel was tempered	0.1074
Polish	Ostrovskij, How the steel was tempered	0.1089
Tucholsky	Schloss Gripsholm	0.1103
Croatian	Ostrovskij, How the steel was tempered	0.1110
Czech	Ostrovskij, How the steel was tempered	0.1124
Dos Passos	One man´s initiation	0.1139
Sorbian	Ostrovskij, How the steel was tempered	0.1205
Sealsfield	Das Cajuetenbuch	0.1209
Ukrainian	Ostrovskij, How the steel was tempered	0.1216
Belorussian	Ostrovskij, How the steel was tempered	0.1271
Macedonian	Ostrovskij, How the steel was tempered	0.1283
Bulgarian	Ostrovskij, How the steel was tempered	0.1310
Russian	Ostrovskij, How the steel was tempered	0.1427
Novalis	Heinrich von Ofterdingen	0.1438

The Slavic languages are intermixed with the other ones but they display the tendency to having relatively great D_S. The greater the D_S, the more divergent is the frequency structure of individual parts. The simplest interpretation consists in assuming either great time gaps in writing, differences in content or strong intervention of editors. The last case is well known with the novel by N. Ostrovskij and in its translations.

No conclusions can be drawn from the order of works either about the author or about the language. Preliminarily, we scrutinize only the sequence of frequency structures of a work.

7.3. Runs

Nevertheless, there are trends in individual texts that can be discovered by other means. They will be presented in the subsequent chapters.

We mark with minus the lambda value smaller than the mean, and with plus the value greater than the mean. For the data in Table 7.8 whose mean is 1.9485 we obtain the chronological sequence of chapters

(I) $- - - + - - + + + +$

to which different tests can be applied. First we state whether there are too few or too many runs. A run is an uninterrupted sequence of equal things. In the above example there are r = 4 runs. Let the number of pluses n_A = 5 and the number of minuses n_B = 5, and $n_A + n_B = n$ which is here n = 10. The expectation of the number of runs is given as

$$(7.4)\quad E(r) = 1 + \frac{2n_A n_B}{n},$$

and the standard deviation as

$$(7.5)\quad \sigma_r = \sqrt{\frac{2n_A n_B(2n_A n_B - n)}{n^2(n-1)}}.$$

Even if the number of individual categories is small, an asymptotic normal test is quite sufficient to test the existence of a trend. For the above example we obtain

$$E(r) = 1 + \frac{2(5)5}{10} = 6,$$

$$\sigma_r = \sqrt{\frac{2(5)5[2(5)5 - 10]}{10^2(10-1)}} = 1.4907.$$

Inserting the above expressions in

$$(7.6) \quad u = \frac{r - E(r)}{\sigma_r}$$

we obtain

$$u = \frac{4-6}{1.4907} = -1.34\,.$$

This *u* is not significant and tells us that 4 runs under the above conditions are rather random.

For the translations of *How the steel was tempered* in other Slavic languages whose necessary values can be found in the Appendix we obtain the results in Table 7.10. As can be seen, the translations follow exactly the original Russian text: none of the run tests is significant and there are only two variant outcomes. At the same time the result shows that in spite of morphological differences among these languages, the typological differences are not very advanced. We can conclude that for the comparison of the vocabulary richness of groups of texts both the normalized indicator lambda and the test of runs are reliable methods.

Table 7.10
Runs in Slavic languages

Language	**runs**	n_A	n_B	n	$E(r)$	σ_r	u
Belorussian	4	4	6	10	5.8	1.4236	-1.26
Bulgarian	4	5	5	10	6	1.4907	-1.34
Croatian	4	5	5	10	6	1.4907	-1.34
Czech	4	4	6	10	5.8	1.4236	-1.26
Macedonian	4	5	5	10	6	1.4907	-1.34
Polish	4	5	5	10	6	1.4907	-1.34
Russian	4	5	5	10	6	1.4907	-1.34
Serbian	4	4	6	10	5.8	1.4236	-1.26
Slovak	4	4	6	10	5.8	1.4236	-1.26
Slovenian	4	5	5	10	6	1.4907	-1.34
Sorbian	4	5	5	10	6	1.4907	-1.34
Ukrainian	4	5	5	10	6	1.4907	-1.34

In the same way the speeches of Italian presidents with respect to the years of origin can be analyzed. Using the table in Appendix we obtain for the 60 speeches of 10 Italian presidents with mean 1.5153 the following sequence (c.f. Table 5) (+ = greater than the mean, − = smaller than the mean):

(II) +++++++++++++++++++++ − +++++++ − − − − − − − − − + − − − + − −
− − − − − − + + + + + − + + +

yielding n = 60, r = 11, n_A = 38, n_B = 22, $E(r)$ = 1 + 2(38)22/60 = 28.87, σ^2 = 2(38)22[2(38)22 – 60]/[60^2(59)] = 12.6896, σ = 3.5622, hence u = (11 – 28.87)/3.5622 = −5.02, showing that the number of runs is significantly small: there is a tendency to aggregate values greater than the mean and also those smaller than the mean. A (big) wavelike movement with respect to the mean can be observed. We can conclude that the subconscious influence of a predecessor on his successor is considerable, a fact that could have a sociological source (e.g. the president goes but his writer remains).

Table 7.11
Chronologically ordered lambdas of Italian Presidents

President	Year	Λ	President	Year	Λ	President	Year	Λ
L. Einaudi	1949	1.6928	G. Saragat	1969	1.5357	F. Cossiga	1989	1.4455
L. Einaudi	1950	1.5781	G. Saragat	1970	1.4946	F. Cossiga	1990	1.4243
L. Einaudi	1951	1.7686	G. Leone	1971	1.5970	F. Cossiga	1991	1.5976
L. Einaudi	1952	1.8488	G. Leone	1972	1.5598	O.L. Scalfaro	1992	1.3316
L. Einaudi	1953	1.7489	G. Leone	1973	1.6580	O.L. Scalfaro	1993	1.3904
L. Einaudi	1954	1.7303	G. Leone	1974	1.6160	O.L. Scalfaro	1994	1.3152
G. Gronchi	1955	1.7038	G. Leone	1975	1.5968	O.L. Scalfaro	1995	1.2787
G. Gronchi	1956	1.6672	G. Leone	1976	1.5727	O.L. Scalfaro	1996	1.4869
G. Gronchi	1957	1.6194	G. Leone	1977	1.5601	O.L. Scalfaro	1997	1.1357
G. Gronchi	1958	1.6236	S. Pertini	1978	1.3602	O.L. Scalfaro	1998	1.1550
G. Gronchi	1959	1.6718	S. Pertini	1979	1.2348	C.A. Ciampi	1999	1.4862
G. Gronchi	1960	1.6703	S. Pertini	1980	1.3086	C.A. Ciampi	2000	1.5429
G. Gronchi	1961	1.6677	S. Pertini	1981	1.2042	C.A. Ciampi	2001	1.5288
A. Segni	1962	1.5701	S. Pertini	1982	1.2590	C.A. Ciampi	2002	1.5397
A. Segni	1963	1.6008	S. Pertini	1983	1.1797	C.A. Ciampi	2003	1.5585
G. Saragat	1964	1.6580	S. Pertini	1984	1.2583	C.A. Ciampi	2004	1.5676

G. Saragat	1965	1.5736	F. Cossiga	1985	1.3665	C.A. Ciampi	2005	1.4860
G. Saragat	1966	1.6031	F. Cossiga	1986	1.4165	G. Napolitano	2006	1.5677
G. Saragat	1967	1.6120	F. Cossiga	1987	1.5774	G. Napolitano	2007	1.5878
G. Saragat	1968	1.5774	F. Cossiga	1988	1.3839	G. Napolitano	2008	1.5692

Let us consider some other texts consisting of several chapters as presented in Table 7.12. There are only 2 texts, namely one book by Dos Passos and the Hawaiian texts, in which the asymptotic normal test is significantly negative (smaller than – 2). Since here the expectation is much greater than the number of runs, there are too few runs, the work from the view of its vocabulary formation more monotonous, more compact and more uniform.

Table 7.12
Runs in some texts consisting of more parts

Author	**Title**	r	n_A	n_B	n	$E(r)$	σ_r	u
Byatt	Possession	4	5	5	10	6.0000	1.4907	-1.3416
Cervantes	Don Quijote	8	9	6	15	8.2000	1.7857	-0.1120
Chamisso	Peter Schlemihls wundersame Geschichte	6	6	5	11	6.4545	1.5588	-0.2916
Dos Passos	One Man's Initiation	4	5	6	11	6.4545	1.5588	-1.5747
Dos Passos	Rosinante to the road again	5	7	10	17	9.2353	1.9298	-2.1947
Eichendorff	Aus dem Leben eines Taugenichts	8	5	5	10	6.0000	1.4907	1.3416
Hawaiian	Laieikawai	11	20	13	33	16.7576	2.6957	-2.1358
Löns	Der Werwolf	5	8	5	13	7.1538	1.6257	-1.3248
Meyer	Der Schuss von der Kanzel	4	5	6	11	6.4545	1.5588	-1.5747
Novalis	Heinrich von Ofterdingen	7	6	3	9	5.0000	1.2247	1.6330
Paul	Dr. Katzenbergers Badereise	24	22	22	44	23.0000	3.2778	0.3051
Sealsfield	Das Cajuetenbuch	7	7	9	16	8.8750	1.8998	-0.9869
Milton	Paradise lost	2	3	1	2	2.5000	0.5000	-1.0000
Tucholsky	Schloss Gripsholm	3	4	2	6	3.6667	0.9428	-0.7071
Gárdonyi	Egri csillagok	3	3	2	5	3.4000	0.9165	-0.4364
Kafka	Betrachtung	9	18	8	10	9.8889	2.0310	-0.44

7.4. Length of phases (runs-up-and-down test)

In Table 7.11, the texts of Italian presidents, we can see that the values of lambda constitute a sequence of increasing and decreasing phases. The mean is not relevant in this case. Comparing the subsequent lambda values we can symbolize the transitions from a greater value to a lower one as minus, and the contrary case as plus. If there are two or more equal numbers one after another, one can consider them as one item. Thus we obtain

(III)
− + + − − − − − − + + − − − + + − + + − − − + − + − − − − − − + − + − + + + + − +
− + − + − − + − + + + − + + + − + + −

Here we have n = 59 signs
20 runs of length 1
5 runs of length 2
7 runs of length ≥ 3.

In order to test whether such a sequence is stationary, we compare the empirical numbers of lengths with their theoretical expected frequencies.

The expected number of runs of length d can be computed as

$$(7.7) \quad e_d = \frac{2(d^2+3d+1)(n-d-2)}{(d+3)!}\ ,$$

hence $e_1 = 2(5)(59\text{-}3)/4! = 23.33$
$e_2 = 2(11)(59\text{-}4)/5! = 10.08.$

The expected value $e_{3,4,...}$ can be computed from the sum of all, minus e_1 and e_2. The number of runs of all length follows from

$$(7.8) \quad e(d \geq 1) = 2\sum_{d=1}^{n-3}\frac{(d^2+3d+1)(n-d-2)}{(d+3)!} = 2\left(\frac{2n-7}{6}+\frac{1}{n!}\right).$$

Since the expression $1/n!$ is very small, it can be neglected and one obtains

$$e(d \geq 1) = (2n - 7)/3.$$

In our case it is

$$e(d \geq 1) = (2(59)\text{-}7)/3 = 37.$$

Hence $e(d \geq 3) = e(d \geq 1) - e_1 - e_2 = 37 - 23.33 - 10.08 = 3.59$. The observed run lengths can be compared with the expected ones using the usual chi-square test

$$(7.9)\quad X^2 = \sum_{d=1}^{d\geq 3} \frac{(f_d - e_d)^2}{e_d}.$$

For our data we obtain

$$X^2 = (20 - 23.33)^2/23.33 + (5 - 10.08)^2/10.08 + (7 - 3.59)^2/3.59 = \\ = 6.27 \sim 6.3.$$

The chi-square has 2.5 DF and yields $P(\chi^2) = 0.04$. Since this is smaller than 0.05, we reject the H_0 hypothesis and conclude that there are too many long runs; the sequence is not random. The deciphering of the background of this non-stationarity must be left to sociolinguists. We can continue to analyze the sequence from other points of view.

The phase test yields more definite results with texts partitioned in many parts. As examples we can present e.g. Dos Passos´ *Rosinante to the road again* yielding $n = 17$, $f(1) = 5$, $f(2) = 2$, $f(\geq 3) = 2$ and $X^2 = 2.07$, or Paul´s *Dr. Katzenbergers Badereise* with $n = 43$, $f(1) = 22$, $f(2) = 9$, $f(\geq 3) = 1$ yielding $X^2 = 3.10$, both non-significant, that is, the distribution of phase lengths does not display any deviation.

7.5. Length of runs

As can be seen in the sequence (II) (Italian presidents) some of the runs are very long. Does this length arise by chance or is there some tendency in the background? We shall approach the problem using an approximation. The longer the sequence (n) is, the better the approximation.

Let n_1 be the number of signs of one sort in the sequence (here, let it be "+"), n_2 the number of signs of another sort (here "–") and let s be the length of the longest run. The expected value of s in the Poisson distribution is

$$(7.10)\quad E(s) = \lambda = n_2\left(\frac{n_1}{n}\right)^s.$$

Hence the probability of exceeding can be computed as

$$(7.11)\quad P(s) = 1 - e^{-\lambda}.$$

Consider the sequence (II) in which there are $n_1 = 38$ and $n_2 = 22$, $n = 60$. The longest run is $s = 21$. Hence $\lambda = 22(38/60)^{21} = 0.001502$. Inserting this value in (7.10) we obtain

$$P(21) = 1 - e^{-0.001502} = 0.0015.$$

Since 0.0015 < 0.05, we may state than this length did not arise by chance; there is some tendency in the background. In the case of Italian presidents it is perhaps the perseveration evoked by the special opportunity or the existence of professional writers, etc.

In Table 7.13 we present some tests using partitioned texts. The fact that none of them displays a significantly long run is a sign of non-perseveration or frequent breaking of rhythm caused by non-continuous writing.

Table 7.13
Test for longest run in some texts

Author	**Text**	***n***	***n_A***	***n_B***	***s***	***P***
Paul	Dr. Katzenbergers Urlaubsreise	44	22	22	5	0,49
Dos Passos	Rosinante to the road again	17	7	10	5	0,11
Katka	Die Betrachtung	18	8	10	5	0,16
Sealsfield	Das Kajüttenbuch	28	14	14	5	0,35
Cervantes	Don Quijote	15	9	6	3	0,72
Löns	Der Werwolf	13	8	5	6	0,23

7.6. Uniformity

The smoothness of a sequence can be expressed in many different ways. Smoothness concerns the differences between subsequent elements of the sequence. Since in nature there are no absolutely smooth straight lines, curves or planes, the study of smoothness gave rise to new disciplines (cf. e.g. Tricot 1995; Falconer 1990). In linguistics it was especially L. Hřebíček (2000) who applied different procedures to linguistic phenomena.

In a previous work (cf. Popescu, Kelih et al. 2010: 95 ff.) we tried to apply very simple measures of text dynamics, namely the *non-smoothness* and the *roughness* indicators which are easy to compute. They can be applied to any numerically expressed property. Here we have to do with the indicator lambda expressing the frequency structure of vocabulary used in subsequent parts of a written work or a series of works. In the first case we merely consider the change of direction of lambda expressed by the extreme points of the sequence. In a sequence 3, 5, 4, 2 there are three extremes (3,5,2); in the sequence 5,4,3,2 there are two (5,2). The extremes are easy to count either from the tables or from a figure.

The ***non-smoothness*** can be computed as the proportion of the number of extremes (*m*) to the number of all points (*n*) in the sequence. But since the first and the last point are automatically extremes, we compute the indicator as

$$(7.12)\quad NS = \frac{m-2}{n-2}.$$

For example in Figure 7.2 we see that in the text by Dos Passos, *One Man´s Initiation* there are *n* = 11 points out of which *m* = 8 are extremes. Hence

$NS = (8 – 2)/(11 – 2) = 0.6667.$

As a matter of fact, *NS* is a simple proportion whose variance is given as

$$(7.13\ )\, Var(NS) = \frac{NS(1-NS)}{n-2} = \frac{(m-2)(n-m)}{(n-2)^3}.$$

For the above work of Dos Passos we obtain *Var*(*NS*) = (8 – 2)(11 – 8)/(11 – 2)^3 = 0.0247.

Table 7.14
Non-smoothness of vocabulary exploitation in some texts

Author	**Text**	***n***	***m***	***NS***	***VAR(NS)***
Byatt	Possession	10	6	0.5000	0.0313
Cervantes	Don Quijote	15	10	0.6154	0.0182
Chamisso	Peter Schlemihls wundersame Geschichte	11	7	0.5556	0.0274
Dos Passos	One Man's Initiation	11	8	0.6667	0.0247
Dos Passos	Rosinante to the road again	17	10	0.5333	0.0166
Eichendorff	Aus dem Leben eines Taugenichts	10	8	0.7500	0.0234
Hawaiian	Laieikawai	33	20	0.5806	0.0079
Löns	Der Werwolf	13	11	0.8182	0.0135
Meyer	Der Schuss von der Kanzel	11	8	0.6667	0.0247
Novalis	Heinrich von Ofterdingen	9	7	0.7143	0.0292
Paul	Dr. Katzenbergers Badereise	44	33	0.7381	0.0046
Sealsfield	Das Cajuetenbuch	16	10	0.5714	0.0175

The indicator *NS* can be considered a non-weighted measure of smoothness of text deployment. It takes into account only the existence of an extreme but not its value. In Table 7.14 one finds the computation of non-smoothness of some works shown in the Appendix. For graphical presentation see also Figure 7.2. The greater *NS*, the greater is the non-smoothness of the text. As can be seen in Table 7.14, there is no text with *NS* < 0.5. This is, as mentioned several times above, the consequence of time-breaks in text.

We can test whether two works or two authors differ significantly. To this end we use the normal approximation. First we compute the pooled mean *NS* as

$$(7.14)\quad \overline{NS} = \frac{m_1 - 2 + m_2 - 2}{n_1 - 2 + n_2 - 2} = \frac{m_1 + m_2 - 4}{n_1 + n_2 - 4}$$

and the variance of the difference ($NS_1 - NS_2$)

$$(7.15)\quad \mathrm{Var}(NS_1 - NS_2) = \overline{NS}(1 - \overline{NS})\left(\frac{1}{n_1 - 2} + \frac{1}{n_2 - 2}\right).$$

In order to obtain a standardized normal deviate, we set up

$$(7.16)\quad u = \frac{NS_1 - NS_2}{\sqrt{Var(NS_1 - NS_2)}} =$$

$$= \frac{(m_1 - 2)(n_2 - 2) - (m_2 - 2)(n_1 - 2)}{\sqrt{(n_1 - 2)(n_2 - 2)}\sqrt{\frac{(m_1 + m_2 - 4)(n_1 + n_2 - m_1 - m_2)}{n_1 + n_2 - 4}}}.$$

For the two books by Dos Passos in Table 7.14 we obtain

$$u = \frac{6(15) - 8(9)}{\sqrt{9(15)}\sqrt{\frac{(8 + 10 - 4)(11 + 17 - 8 - 10)}{11 + 17 - 4}}} = 0.6414$$

which indicates a non-significant difference in (non-)smoothness. An alternative approximation can be made directly using the *NS* and its variance from Table 7.14. Thus we obtain

$$u = \frac{0.6667 - 0.5333}{\sqrt{0.0247 + 0.0166}} = 0.6564$$

which is slightly greater but also not significant.

The indicator *NS* simply shows the existence of non-smoothness. Since it lies in the interval <0,1>, many extremes (= great *NS* > 0.5) indicate non-smoothness while few extremes (*NS* < 0.5) indicate smoothness. As can be seen in Table 7.14, all the texts display a kind of non-smoothness

8. Historical development

Studying the historical development of rank-frequency structuring in texts we are confronted with enormous variation in the sequence of works. Even if we take one genre, the personalities of writers are very different. Further, frequency structuring does not depend only of the theme but also on the aims, effects to be evoked, attitudes of the authors, efforts made for originality, the historical circumstances, etc. And lastly, the sample can in no case be representative of the given epoch unless it is very specific, e.g. Nobel lectures or End-of-year speeches of Italian presidents. Though ordering of lambdas according to increasing or decreasing values is possible, it does not mirror the historical order. If we scrutinize the works in historical order and the sample is large, then every year must be characterized by its mean or its variance or a confidence interval, otherwise we may obtain a very chaotic image. Besides, the picture may get still more complicated if some works were started by a writer in year X and finished ten years later, and, still worse, published after another ten years.

Two kinds of studies are possible: (a) the historical development of texts in general or partitioned in genres in a language, and (b) the historical development of a writer.

8.1. Development in a language

Historical development can proceed in three different ways. Either it is a weakly oscillating constant, or it is a weakly oscillating development attaining at times its extreme and then returning, or it is a chaotic movement. The *weakly oscillating constant* can be observed in texts of the same sort but in practice there is always a slope which itself may oscillate as time goes on. The *weakly oscillating development* can be captured by a straight line but only in a small time interval, otherwise the movement would diverge; negative values or infinity are not possible for language properties. And finally, there may be a *chaotic movement* within the general interval of lambda. Such a case can be expected not only in the sequence of different authors but also in the chapters of one and the same work.

Let us consider some examples. In Table 8.1 and Figure 8.1 the lambda of German literary works in the course of 250 years – with great gaps that can be filled up – is shown. Since we took several works of the authors, we ascribed the lambdas to their midlife years. The development can be captured by a straight line with negative slope, but it can be predicted that the slope would become zero if we add more works, as can be judged from the last two texts. Of course, this picture is not complete because the 20th century is almost completely missing. For the sake of simplicity the texts from the 21st century can be omitted for evaluation, if necessary, or the sample can be enlarged.

Table 8.1
Historical development of mean lambda in German literature

Author	**Increasing midlife year**	**mean Λ**
Lessing	1755	1.5514
Schiller	1782	1.6463
Novalis	1787	1.5108
Goethe	1791	1.4536
Paul	1794	1.6522
Hoffmann	1799	1.4418
Arnim	1806	1.3734
Chamisso	1810	1.5248
Immermann	1818	1.1151
Droste-Huelshoff	1823	1.6108
Eichendorff	1823	1.3505
Rückert	1827	1.6044
Heine	1827	1.5480
Sealsfield	1829	1.3894
Storm	1853	1.0070
Keller	1855	1.4754
Meyer	1862	1.6891
Busch	1870	1.3569
Raabe	1871	1.3081
Löns	1890	1.2247
Wedekind	1891	1.3277
Sudermann	1893	1.0216
Schnitzler	1897	1.3778
Kafka	1904	1.5153
Tucholsky	1913	1.1688
Sloggi (pseudonym)	2001	1.5208
Rieder	2001	1.3436

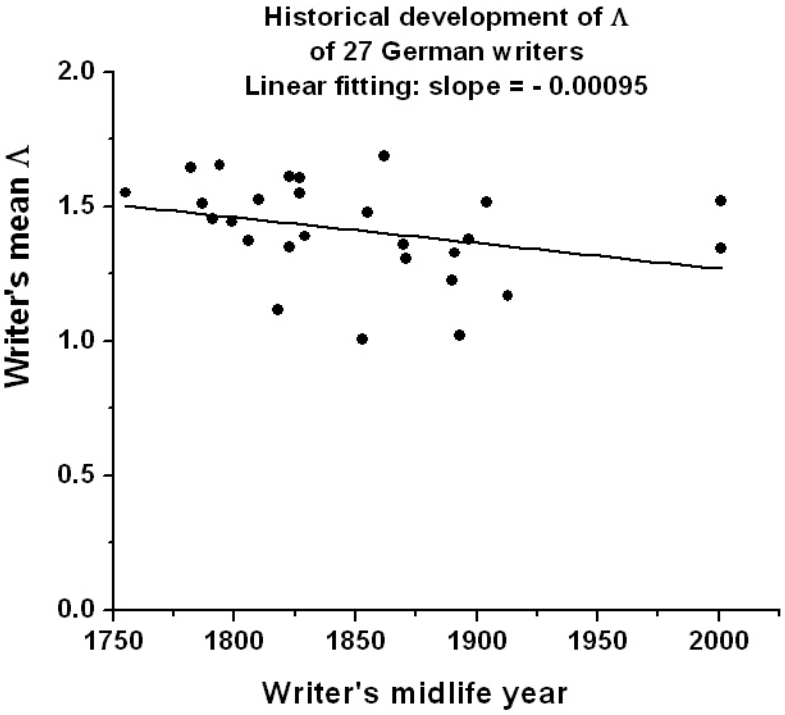


Figure 8.1. Development of lambda in German literature

In the Turkish prose of the last century we see a linear trend, too, whose slope is still nearer to zero than in German. This comparison is not quite correct because in German we took a mixed sample while in Turkish it was only prose. Nevertheless, the results presented in Table 8.2 and Figure 8.2 show the given trend. Here we took individual works and the year of their appearance. The slope is almost zero.

Table 8.2
Development of lambda in Turkish literature in the 20th century
(Data from F. Can)

	Author	Title	Increasing year	Λ
1	Rauf, M	Eylül	1901	1.2207
2	Gürpınar, H.R.	Toraman	1919	1.8453
3	Seyfettin. O.	Efruz Bey	1919	1.7303
4	Karay, R.H.	İstanbul’un Bir Yüzü	1920	1.8330
5	Güntekin, R.N.	Çalıkuşu	1922	1.3369
6	Karaosmanoğlu, Y.K.	Nur Baba	1922	1.5609

7	Adıvar, H.E.	Kalb Ağrısı	1924	1.4266
8	Enis, S.	Zaniyeler	1924	1.5730
9	Uşaklıgil, H.Z.	Kırık Hayatlar	1924	1.4013
10	Safa, P.	Peyami Safa	1925	1.4619
11	Yesari, M.	Tipi Dindi!	1933	1.5182
12	Esendal, M.Ş.	Ayaşlı İle Kiracıları	1934	1.2446
13	Uçuk, C.	Dikenli Çit	1937	1.4565
14	Uçuk, C.	Mithat Cemal Kuntay	1938	1.1931
15	Hisar, A.Ş.	Fahim Bey ve Biz	1941	1.7044
16	Ali, S.	Kürk Mantolu Madonna	1943	1.5096
17	Bilbaşar, K.	Denizin Çağırışı	1943	1.9256
18	Faik, S.	Medarı Maişet Motoru	1944	1.6989
19	Tanpınar, A.H.	Huzur	1949	1.3786
20	Akbal, O.	Garipler Sokağı	1950	1.6822
21	Kemal, O.	Cemile	1952	1.5629
22	Kemal, Y.	İnce Memed	1955	1.1144
23	Atılgan, Y.	Aylak Adam	1959	1.5823
24	Tahir, K.	Yorgun Savaşçı	1965	1.4217
25	Baykurt, F.	Tırpan	1970	1.1968
26	Buğra, T.	İbiş’in Rüyası	1970	1.4449
27	Altan, Ç.	Büyük Gözaltı	1972	1.4881
28	Atay, O.	Tutunamayanlar	1972	1.4069
29	Ağaoğlu, A.	Ölmeye Yatmak	1973	1.6191
30	Füruzan	Kırkyedililer	1974	1.6310
31	Kür, P.	Yarın Yarın	1976	1.4153
32	Edgü, F.	O; Hâkkari’de Bir Mevsim	1977	1.6215
33	İleri, S.	Ölüm İlişkileri	1979	1.6918
34	Pamuk, O.	Sessiz Ev	1983	1.2523
35	Tekin, L.	Sevgili Arsız Ölüm	1983	1.2005
36	Eroğlu, M.	Issızlığın Ortasında	1984	1.3570
37	İlhan, A.	Haco Hanım Vay!..	1984	1.6031
38	Arslanoğlu, K.	Devrimciler	1987	1.4539
39	Gürsel, N.	Boğazkesen:Fatih’in Romanı	1995	1.5486
40	Altan, A.	Kılıç Yarası Gibi	1998	1.7751

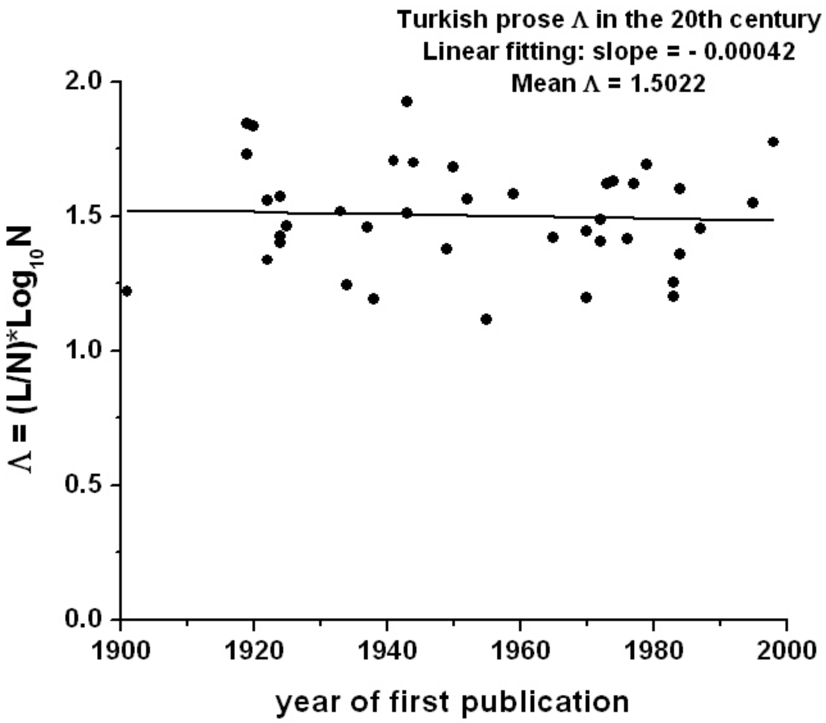


Figure 8.2. Development of lambda in Turkish prose in the 20th century (Data from F. Can)

A third example is the survey of lambdas in the Nobel lectures. The data are presented in Table 8.3 and Figure 8.3. As can be seen, the slope is slightly positive but can be changed by any new text. Evidently, it will always oscillate around zero.

Table 8.3
Lambdas of Nobel lectures

Nobelist	Year	*Λ*
Yeats, W.B.	1923	1.2152
Lewis, S.	1930	1.3312
O'Neill, E.	1936	1.2987
Buck, P.	1938	1.0426
Eliot, T.S.	1948	1.3459
Faulkner, W.	1949	1.3059
Russell, B.	1950	1.2379
Churchill, W.	1953	1.6126

Marshall, G.C.	1953	1.3031
Hemingway, E.	1954	1.3622
Steinbeck, J.	1962	1.5060
Bellow, S.	1976	1.3605
Golding, W.	1983	1.2864
Buchanan, J.M. Jr.	1986	1.2427
Gordimer, N.	1991	1.4169
Walcott, D.	1992	1.4685
Morrison, T.	1993	1.3527
Carter, J.	2002	1.5071
Pinter, H.	2005	1.3205
Lessing, D.	2007	1.0793

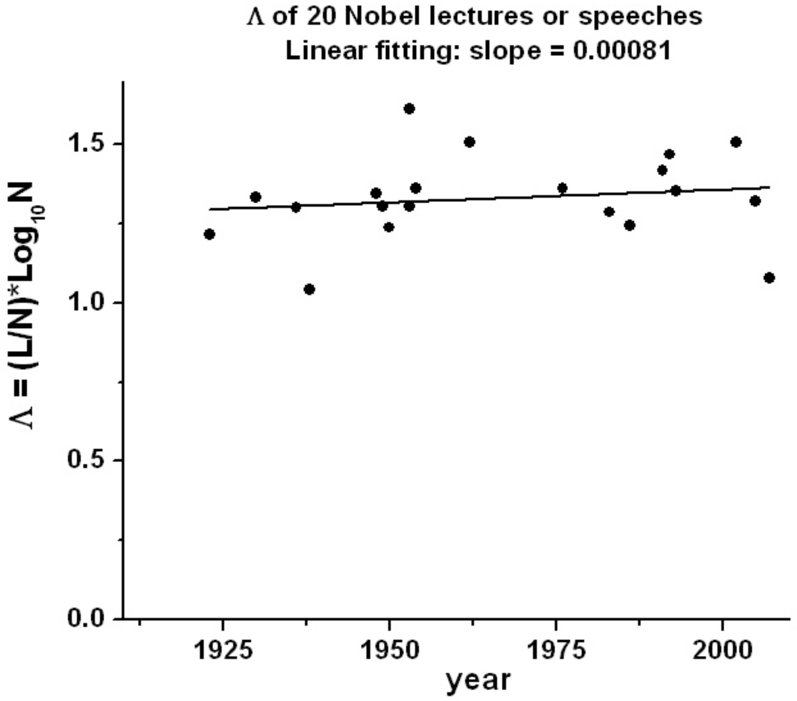


Figure 8.3. Lambdas of Nobel lectures

Thus we have three different slopes but in each of the three cases the motion is rather chaotic and cannot be captured by simple formulas. Nevertheless, one can characterize the development using the several indicators presented in previous chapters.

The *sequential difference* D_S for the three sets of data, German literature, Turkish prose and Nobel lectures, is almost equal in German and Turkish, viz.

0.2253 and 0.2238, while in the Nobel lectures it is 0.1521. That means the difference to the neighbour is greater in literature than in public speeches. This impression is corroborated also by the D_S of speeches of Italian presidents, which is 0.0776, i.e. half of the Nobel lectures. This is probably caused by the fact that in speeches of Italian presidents the same person is maximally 6 times its own neighbour and there is no need for striving for originality.

For runs of values below and above the mean (see Chapter 7.3) we obtain three quite different values: u(German) = −0.19, u(Turkish) = 0.32 and u(Nobel) = 0.00. Though none is significant, German and Turkish display contrary tendencies.

The probability of the longest run is not significant in the three samples. In German it is 0.16, and in Turkish and Nobel lectures it is equally 0.46. As shown in Chapter 7.4, this differs strongly from Italian presidential speeches where the probability of the longest run is 0.0015. The longer the longest run, the stronger is the perseveration of the given structure.

Runs-up-and-down display non-significant results for the three text groups: X^2(German) = 0.88, X^2(Turkish) = 4.27, X^2(Nobel) = 1.92.

The non-smootheness (*NS*) measured in form of the ratio of number of extremes (cf. Chapter 7.5) ($m - 2$) to all points ($n - 2$) yields the following results: *NS*(German) = 0.68, *NS*(Turkish) = 0.76, *NS*(Nobel) = 0.61. The greater the value of *NS*, the more the sequence displays a kind of regular periodicity. This phenomenon can textologically be interpreted as a subconscious deviation from the structure of predecessors, because it can hardly be assumed that a writer consciously controls the frequency structure of his text. The result can be seen without test if we compare these values with that of Italian presidents yielding *NS*(Italian) = 0.55 ($n = 60$, $m = 34$) which lies in the vicinity of the neutral point 0.5. The significance of the deviation of NS from the expectation can be tested either exactly by means of the binomial distribution or asymptotically using the u criterion. Using the exact binomial test we have the parameters $p = 0.5$, $N = n - 2$ and compute the cumulative probability from $m - 2$ to N, for example in German we have $n = 27$, $m = 19$ hence

$$P(X \geq m-2) = \sum_{x=m-2}^{n-2} \binom{n-2}{x} 0.5^{n-2} = \sum_{x=17}^{25} \binom{25}{x} 0.5^{25} = 0.0539$$

which is not significant. For the Turkish prose with $n = 40$, $m = 31$ we obtain $P(\mathrm{X} \geq 29) = 0.0008$ and for Nobel lectures with $n = 20$, $m = 13$, $P(\mathrm{X} \geq 11) = 0.2403$. Hence only the development in Turkish displays a clearly significant non-perseveration or non-conservatism feature. The picture would surely change (in whatever direction) if we took more works into consideration. In any case, there is a background movement which is partially caused by individuals but represents a collective dynamic.

8.2. Development of a writer

The development of a writer – in our view – does not mean the uses of more (or fewer) words in the course of time but rather the technique of playing with words in his texts. We assume, a writer knows the words of his language (except for scientific terminology etc.), hence we do not study the development of his vocabulary but the development of his technique of using it. Some writers use a very restricted vocabulary in all their works, while other ones control it as occasion demands.

In Table 8.4 we present the works by the Romanian writer M. Eminescu (1850-1889) in his 146 poems. They can be found on the Internet under

http://en.wikipedia.org/wiki/Mihai_Eminescu
http://www.gabrielditu.com/eminescu/contents.asp
http://www.romanianvoice.com/poezii/poeti/eminescu.php

Again, the text sizes do not play any role; one can see that e.g. *Odă în metru antic* (1883) has N = 103 and Λ = 1.6267 while a much longer work *Memento mori* (1872) with N = 9773 has Λ = 1.6175.

Table 8.4
Lambdas in the poems by M. Eminescu

First published in	**Poem title**	***N***	***V***	***L***	Λ	***Var***(Λ)
1866	De-aş avea	93	61	62.4787	1.3225	0.0068656
1866	Din străinătate	244	168	174.3565	1.7060	0.0012760
1866	La Bucovina	184	140	141.8929	1.7465	0.0014636
1866	La mormântul lui Aron Pumnul	150	116	118.8191	1.7237	0.0014240
1866	Lida	66	57	57.6503	1.5894	0.0023418
1866	Misterele nopţii	155	110	111.8929	1.5812	0.0019284
1866	O călărire în zori	346	233	245.8645	1.8042	0.0017673
1867	Ce-ţi doresc eu ţie, dulce Românie	183	127	129.7148	1.6037	0.0031009
1867	Din lyra spartă...	51	44	43.8284	1.4675	0.0027909
1867	Horia	143	119	120.9907	1.8236	0.0015187
1867	La moartea lui Heliade	332	225	231.7707	1.7600	0.0012492
1867	Nu e steluţă	54	40	39.8284	1.2778	0.0040765
1868	Amorul unei marmure	266	184	187.1290	1.7059	0.0019147
1868	La o artistă (Ca a nopţii poezie)	142	104	106.4049	1.6128	0.0025453

1868	Numai poetul	48	40	39.8284	1.3950	0.0048527
1868	Speranța	245	143	154.6554	1.5082	0.0014295
1869	Amicului F.I.	257	194	194.6569	1.8253	0.0010524
1869	Când	126	100	101.8929	1.6985	0.0007371
1869	Când marea...	114	80	81.8929	1.4776	0.0035516
1869	Când priveşti oglinda mărei	101	82	83.4787	1.6566	0.0022125
1869	Cine-i?	129	93	95.7148	1.5660	0.0035048
1869	De ce să mori tu?	266	172	177.8098	1.6209	0.0020175
1869	De-aş muri ori de-ai muri	258	168	171.1356	1.5997	0.0022058
1869	Întunericul şi poetul	249	176	180.4694	1.7367	0.0020325
1869	Junii corupți	458	309	322.7674	1.8752	0.0019875
1869	La moartea principelui Ştirbey	132	98	99.4787	1.5981	0.0034323
1869	La o artistă (Credeam ieri...)	219	152	158.7279	1.6963	0.0011282
1869	Lebăda	41	37	37.2361	1.4647	0.0048212
1869	Locul aripelor	259	173	174.8995	1.6297	0.0022923
1869	O stea pin ceruri	78	65	64.8284	1.5726	0.0023380
1869	Ondina (Fantazie)	871	535	557.6579	1.8823	0.0004409
1869	Prin nopți tăcute	48	40	39.8284	1.3950	0.0048527
1869	Unda spumă	59	47	46.8284	1.4055	0.0035713
1869	Viața mea fu ziuă	105	86	86.6569	1.6681	0.0020361
1870	Epigonii	921	565	590.0754	1.8992	0.0005702
1870	Îngere palid...	63	53	52.8284	1.5088	0.0032357
1870	La moartea lui Neamțu	245	173	175.3071	1.7095	0.0013976
1870	La Quadrat	110	79	80.8929	1.5012	0.0020147
1870	Sus în curtea cea domnească	128	104	104.6569	1.7229	0.0014895
1871	Andrei Mureşanu	2008	1011	1057.0693	1.7387	0.0003263
1871	Aveam o muză	421	281	290.0402	1.8080	0.0019170
1871	Basmul ce i l-aş spune ei	398	262	272.0975	1.7774	0.0011167
1871	Copii eram noi amandoi	375	250	265.9252	1.8253	0.0009625
1871	Frumoasă şi jună	113	82	85.3657	1.5510	0.0037088
1871	Înger de pază	91	71	72.0645	1.5514	0.0011987
1871	Iubită dulce, o, mă lasă	337	212	216.0618	1.6205	0.0017821
1871	Iubitei	416	240	248.4707	1.5643	0.0011037
1871	Mortua est!	491	295	307.1973	1.6837	0.0016838
1871	Noaptea...	177	128	130.3071	1.6550	0.0023696
1871	Replici	147	73	81.8627	1.2070	0.0072703
1871	Steaua vieții	70	55	55.2426	1.4561	0.0038160
1872	Când crivățul cu iarna...	708	420	438.8507	1.7666	0.0012523

1872	Cugetările sarmanului Dionis	571	389	406.6445	1.9632	0.0009578
1872	Demonism	882	500	524.0894	1.7502	0.0009308
1872	Doi aştri	40	39	38.4142	1.5385	0.0007810
1872	Ecò	698	442	461.6051	1.8807	0.0009960
1872	Egipetul	688	452	465.7789	1.9211	0.0004419
1872	Feciorul de imparat fara de stea	6030	2271	2445.6464	1.5332	0.0000806
1872	Memento mori	9773	3576	3961.9447	1.6175	0.0000681
1872	Miradoniz	636	377	406.0453	1.7898	0.0007029
1872	Odin şi poetul	1429	724	763.2834	1.6852	0.0004045
1873	Adânca mare...	75	62	63.0645	1.5767	0.0039859
1873	Ah, mierea buzei tale	228	144	147.5432	1.5259	0.0015717
1873	Care-i amorul meu în astă lume	213	157	158.8929	1.7369	0.0019103
1873	Cum oceanu-ntărâtat...	77	67	67.2426	1.6474	0.0014965
1873	Dacă treci râul Selenei	356	232	244.4991	1.7523	0.0011642
1873	Din Berlin la Potsdam	128	99	101.3006	1.6677	0.0038990
1873	Dumnezeu şi om	443	320	327.2346	1.9548	0.0016413
1873	Floare albastră	247	185	192.1356	1.8612	0.0023214
1873	Ghazel	331	231	235.8836	1.7957	0.0010875
1873	Inger si demon	876	520	537.7659	1.8064	0.0009795
1873	Mitologicale	681	442	466.0482	1.9389	0.0006468
1873	Murmură glasul mării	119	100	101.9907	1.7789	0.0020327
1873	O arfă pe-un mormânt	157	118	120.3071	1.6827	0.0011127
1873	Privesc oraşul furnicar	173	136	140.7559	1.8209	0.0021390
1874	Cum negustorii din Constantinopol	101	84	84.6569	1.6800	0.0021639
1874	Împarat şi proletar	1510	857	896.5804	1.8876	0.0004909
1874	În căutarea Şeherezadei	915	594	615.4903	1.9921	0.0004121
1874	Napoleon	240	169	176.8790	1.7542	0.0009666
1874	O, adevăr sublime...	334	226	235.3818	1.7786	0.0010531
1874	Pustnicul	380	270	273.9640	1.8599	0.0011154
1875	Făt-Frumos din tei	415	281	289.4209	1.8258	0.0012316
1876	Calin	2299	1123	1199.1833	1.7534	0.0002680
1876	Crăiasa din poveşti	122	94	96.9515	1.6580	0.0019899
1876	Dorința	102	85	87.8126	1.7292	0.0011998
1876	Lacul	90	70	71.0711	1.5432	0.0011759
1876	Melancolie	274	192	201.9549	1.7968	0.0013554
1876	Mureşanu	2051	961	1029.0071	1.6616	0.0002963
1876	Vis	177	138	139.8929	1.7767	0.0009436

1878	Departe sunt de tine	135	105	106.8929	1.6868	0.0029829
1878	Oricâte stele...	85	73	72.8284	1.6531	0.0012850
1878	Povestea codrului	220	168	171.7800	1.8290	0.0007176
1878	Singurătate	172	134	135.0711	1.7556	0.0014232
1879	Atat de frageda…	176	133	138.7396	1.7701	0.0009544
1879	De câte ori, iubito...	102	84	85.9907	1.6933	0.0010291
1879	Despartire	304	202	208.7707	1.7051	0.0016393
1879	Foaia veştedă (dupa Lenau)	115	99	100.0645	1.7931	0.0020503
1879	Freamăt de codru	179	143	144.4853	1.8185	0.0008707
1879	Pajul Cupidon...	148	115	118.7800	1.7418	0.0032239
1879	Pe aceeaşi ulicioară...	138	103	104.4853	1.6202	0.0022833
1879	Revedere	141	102	103.0711	1.5711	0.0024387
1879	Rugăciunea unui dac	357	253	259.4127	1.8549	0.0013102
1879	Sonete	265	194	196.3071	1.7951	0.0012290
1879	Stelele-n cer	91	76	76.6569	1.6503	0.0011559
1880	Dintre sute de catarge	50	41	41.6503	1.4153	0.0071176
1880	O, mamă...	140	98	99.8929	1.5313	0.0031582
1881	Scrisoarea I	1272	707	743.5881	1.8148	0.0005579
1881	Scrisoarea II	696	423	442.3648	1.8067	0.0012462
1881	Scrisoarea III	2278	1146	1236.8800	1.8230	0.0002243
1881	Scrisoarea IV	1256	699	749.9394	1.8504	0.0006690
1881	Scrisoarea V	1027	550	581.7540	1.7059	0.0004247
1882	Nu mă-nțelegi	384	257	261.7793	1.7618	0.0013245
1883	Adio	159	111	114.6410	1.5872	0.0040790
1883	Când amintirile...	97	80	80.6569	1.6520	0.0023051
1883	Ce e amorul?	124	94	97.9274	1.6533	0.0040566
1883	Ce te legeni...	102	76	79.3657	1.5629	0.0031921
1883	Criticilor mei	130	91	91.2426	1.4837	0.0019737
1883	Cu mâne zilele-ți adaogi...	141	105	106.4787	1.6230	0.0024824
1883	De-oi adormi (variantă)	122	105	105.2426	1.7998	0.0021823
1883	De-or trece anii...	87	63	64.8929	1.4467	0.0039308
1883	Din valurile vremii...	152	104	105.8929	1.5200	0.0024687
1883	Glossa	380	191	200.2494	1.3595	0.0017594
1883	Iar când voi fi pământ (variantă)	131	106	107.4787	1.7371	0.0014690
1883	Iubind în taină...	87	77	76.8284	1.7128	0.0019761
1883	La mijloc de codru...	55	35	41.6363	1.3175	0.0033965
1883	Lasă-ți lumea...	225	167	170.5432	1.7829	0.0010596
1883	Luceafărul	1737	820	885.3917	1.6514	0.0003643

1883	Mai am un singur dor	125	103	103.2426	1.7319	0.0015466
1883	Nu voi mormânt bogat (variantă)	113	99	99.6569	1.8106	0.0018141
1883	Odă în metru antic	103	83	83.2426	1.6267	0.0028468
1883	Pe langa plopii fara soti	199	138	140.7213	1.6256	0.0020005
1883	Peste vârfuri	47	39	40.0645	1.4254	0.0032489
1883	S-a dus amorul	219	152	155.5432	1.6623	0.0028969
1883	Se bate miezul nopții...	45	40	39.8284	1.4632	0.0033581
1883	Și dacă...	53	37	37.2426	1.2116	0.0058106
1883	Somnoroase păsărele...	55	46	46.2426	1.4633	0.0024934
1883	Te duci...	84	68	72.9112	1.6703	0.0036142
1883	Trecut-au anii	88	74	74.2426	1.6405	0.0012177
1883	Veneția (de Gaetano Cerri)	79	71	70.8284	1.7013	0.0022928
1884	Din noaptea	68	57	56.8284	1.5314	0.0028826
1885	Sara pe deal	156	128	129.4787	1.8203	0.0025494
1886	La steaua	71	62	61.8284	1.6121	0.0026994
1887	De ce nu-mi vii	123	82	85.7800	1.4575	0.0036578
1887	Kamadeva	81	70	70.2426	1.6550	0.0013843
1887	Povestea teiului	390	261	271.1328	1.8013	0.0010717
1887	Venere şi Madona	393	247	256.5522	1.6936	0.0013175

Looking at Figure 8.4 where the lambdas of his individual works are shown we see again the good quality of the normalization. The minimum lambda is 1.2070, the maximum never surpasses 2.00 – it is 1.9921 in *În căutarea Şeherezadei*. The mean lambda of Eminescu's 146 poems in Table 8.4 is 1.6685. Nevertheless, each year displays a relatively great dispersion. Hence the linear regression yielding Λ = 2.8926 – 0.00065(year) is not significant; the technique of the author does not change. In this form the first parameter does not say anything. But even if we take the mean lambdas from Table 8.5 and replace the years by ordinal numbers (1,2,…,21), we obtain a non-significant regression coefficient 0.0018; however the first coefficient is 1.6590 representing the (unweighted) mean of means.

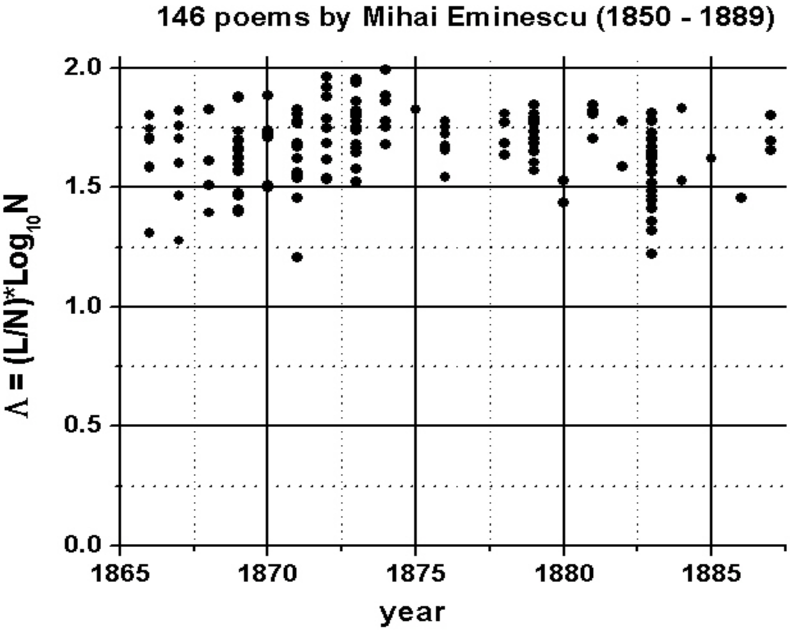


Figure 8.4. Lambda of 146 works by M. Eminescu in historical succession

Taking means of individual years we obtain the results in Table 8.5.

Table 8.5
Mean lambdas of M. Eminescu´s works

Year of first publication	Annual mean Λ
1866	1.6348
1867	1.5865
1868	1.5555
1869	1.6316
1870	1.6656
1871	1.6203
1872	1.7446
1873	1.7533
1874	1.8254
1875	1.8258
1876	1.6844
1878	1.7270
1879	1.7222

1880	1.4840
1881	1.7975
1882	1.7792
1883	1.5856
1884	1.5314
1885	1.8318
1886	1.6207
1887	1.6519

The graphical presentation is shown in Figure 8.5.

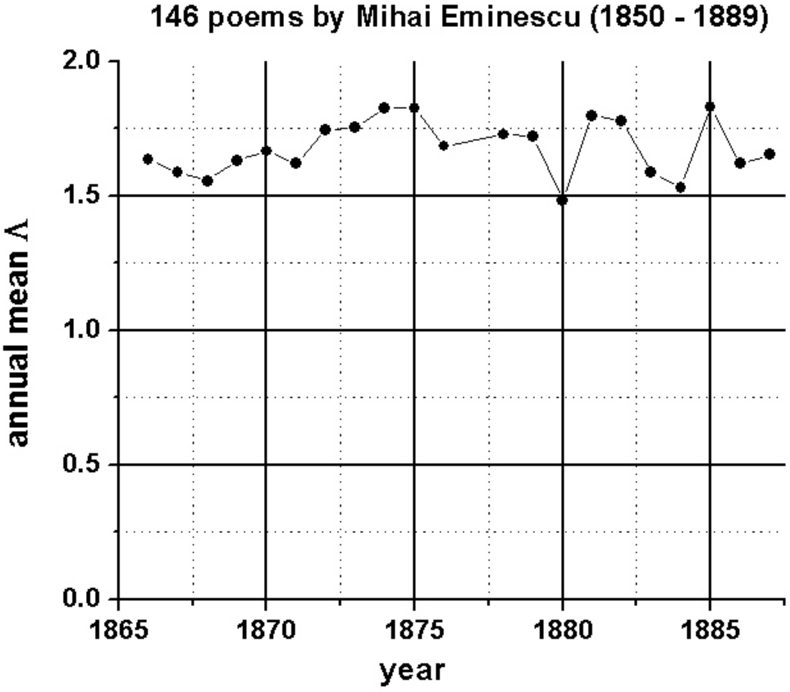


Figure 8.5. Yearly mean lambdas of M. Eminescu

Again, it would be possible to apply all the indicators we used in previous chapters. Since we are now engaged with means, computing the unweighted general mean for Eminescu´s 21 creative years (from Table 8.5) yields $\bar{\Lambda}$ = 1.6790. Comparing this general mean with individual means we obtain the sequence

– – – – – – +++++++ – ++ – – + – +

containing 11 cases of + (n_A = 11), 10 cases of − (n_B = 10), and r = 8. Performing the u-test from Chapter 7.2 we obtain u = −3.81 showing that M. Eminescu had a rather smooth development characterized by two long periods at the beginning which caused the significantly small number of runs, and then, from 1880 on, a very unsettled movement. One can say that he abandoned the acquired frequency structures and sought new ways of expression. It must be remarked that the partitioning in years is merely a convention but for more detailed analysis it would be necessary to take into account the exact date of appearance of each poem.

Runs express only the motion but not the weight of every step. Hence we compute the mean *sequential difference* according to formula (7.3) and obtain D_S(Eminescu) = 0.0995. Compared with the dynamics of German, Turkish and Nobel-lecture texts it is very small but it is greater than that of Italian presidents (0.0776). We must consider this result as a component of Eminescu´s personal style.

As to runs-up-and-down, the chi-square test yields X^2 = 1.05, which with P = 0.59 is not significant. Hence the sequence can be considered stationary.

As to the length of runs, we have n_A = 11, n_B = 10, n = 21, hence λ = $10(11/21)^7$ = 0.1522, and $P(s)$ = 1 – exp(-0.1522) = 0.14, which is not significant. Hence the longest run could arise by chance.

In Table 8.5 we find n = 21 points out of which m = 13 are extremes. Thus the non-smootheness is NS = 11/19 = 0.5789. Comparing it with the expectation 0.5 we obtain $u = (0.5789 – 0.5)/\sqrt{0.1190} = 0.72$, a non-significant difference, i.e. Eminescu has a rather stationary, conservative lambda-structure.

Having these results for M. Eminescu, the research could continue in three directions: (i) studying further properties in Eminescu´s work and associate them with the above ones and (ii) scrutinize other Romanian writers and compare their individual development with that of Eminescu or, lastly, with the general movement in Romanian poetry. Needless to say, (iii) comparisons with the same phenomenon in other languages would, perhaps, reveal some aspects of the dynamics of literary creation.

9. Child language development

Since the vocabulary of children increases in the course of years, it is to be expected that the lambda will increase, too. However, the testing of this hypothesis is not easy if we consider the number of children in different cultures, in different ages and in different personal milieus. That means, the homogeneity of texts would be especially important. In different circumstances quite different means of lambda may develop. Since our sources are scarce, we analyze at least the available ones that can be found on the Internet. In Table 9.1 (reproducing Table 6e of the Appendix) we bring some English stories narrated by children using as source (1) http://www.goodnightstories.com/stories.htm and written by children using as source (2) http://www.kids-space.org/. The beginning of this research can be offered by the data in Table 9.1.

Table 9.1
Stories told or written by children in English
(G = gender, m = masculine, f = feminine)

Author	**Age**	**G**	**Text**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Bernice	6	f	Fairies school	266	147	152.2978	1.3884	0.001852
Emily	6	f	My friend Jeffrey	112	70	71.8995	1.3155	0.003848
Eva	6	f	The little fish	180	91	98.8041	1.2379	0.004023
Leon	6	m	The dragon	71	42	48.8965	1.2749	0.006820
Tomin	6	m	Dream	109	56	61.1701	1.1434	0.006363
Anna	7	f	The Fairies	123	87	88.0711	1.4964	0.002741
Natalia	7	f	The beautiful female frog	136	82	85.2334	1.3371	0.004339
Olivia	7	f	The owls that are red	134	79	87.8902	1.3952	0.004150
Jordan	7	m	My unicorn	39	27	28.0645	1.1449	0.010573
Banner	7	m	Sam the Robo	75	46	46.6569	1.1665	0.005933
Emma	8	f	Pat's sleepover	145	79	81.3071	1.2120	0.006331
Sophia	8	f	The flying horse	51	37	38.9907	1.3055	0.007460
Ibrahim	8	m	The Sneaky Mouse	211	86	102.0259	1.1239	0.004776
Ibrahim	8	m	The Three Cats	281	111	120.1450	1.0470	0.001135
Steven	8	m	My first time visiting space	212	105	110.7055	1.2148	0.004038
Lorelei	9	f	The giraffe story	241	122	125.1421	1.2369	0.002296
Lorelei	9	f	The life story of Bea The Cran	258	128	135.0459	1.2623	0.003318

Michaela	9	f	The ants first home	142	86	87.4853	1.3260	0.003330
Colin	9	m	Cars	125	65	68.1356	1.1430	0.001686
Jack	9	m	The lost giraffe	104	65	69.0552	1.3393	0.002126
Kate	10	f	Destroer	91	57	60.9275	1.3116	0.003178
Micaela	10	f	Fairyland	1941	509	583.1789	0.9879	0.000611
Olivia	10	f	The_magic_bag	221	101	104.5563	1.1091	0.002307
Jordan	10	m	The wolf who would fly	148	86	88.7214	1.3010	0.003164
Raven	10	m	The big whale	84	57	58.4787	1.3396	0.006769
Bethany	11	f	The prince and princess	418	176	196.4680	1.2320	0.002410
Kaia	11	f	The hungry bear	187	108	114.4062	1.3899	0.001479
Maria	11	f	Dark fairy	185	92	98.9823	1.2130	0.004841
Toni	11	m	The rift	339	193	201.2885	1.5024	0.001310
Trevor	11	m	Flying in the sky	81	57	57.6569	1.3585	0.005270
Lizzie	12	f	I won't grow up	357	165	172.9090	1.2364	0.001610
Lyndsay	12	f	Fantasy	328	152	160.8352	1.2337	0.003570
Lyndsay	12	f	The Jensons New Year	1538	457	532.4480	1.1033	0.000746
Graham	12	m	Kielbasa	408	172	205.5659	1.3154	0.002758
Christy	13	f	The Pear	1367	451	534.2903	1.2256	0.000812
Jammie	13	f	Long Wait	243	136	146.8503	1.4417	0.002788
Jesse	13	f	The Light Festival in which Orli was conquered by the Light Angel	945	327	379.8462	1.1960	0.000881
Stacey	13	m	The Vain Rabbits and the Cunning Bear	360	170	179.0468	1.2714	0.003206
Vincent	13	m	The Unicorn's battle	171	101	111.7702	1.4595	0.003470

Collecting the individual texts according to age and gender but not differentiating between narrated and written texts we obtain the numbers presented in Table 9.2. The means of lambdas are not very stable because the number of observations is small, but the table allows us to make a first glance at a future research domain.

Though our data for English are not sufficient for drawing conclusions, nevertheless, we may report at least some preliminary observations. The girls began at a higher level than the boys but in the course of seven years the level did not change. The girls come back to the level they began with. The linear regression of Λ against age yields $\Lambda_f = 1.3933 - 0.0125(\text{age})$ where the regression parameter

Table 9.2
Lambdas in English stories told or written by children classified according to age and gender

Age	f	m	mean f	mean m
6	1.3884, 1.3155, 1.2379	1.2749, 1.1434	1.3139	1.2092
7	1.4964, 1.3371, 1.3952	1.1449, 1.1665	1.4096	1.1557
8	1.2120, 1.3055	1.1239, 1.0470, 1.2148	1.2588	1.1286
9	1.2369, 1.2623, 1.3260	1.1430, 1.3393	1.2751	1.2412
10	1.3116, 0.9879, 1.1091	1.3010, 1.3396	1.1362	1.3203
11	1.2320, 1.3899, 1.2130	1.5024, 1.3585	1.2783	1.4305
12	1.2364, 1.2337, 1.1033	1.3154	1.1911	1.3154
13	1.2256, 1.4417, 1.3260	1.2714, 1.4595	1.3311	1.3655

meter is not significantly different from zero; it is even slightly negative (in our data). The boys began at a lower level, but slowly developed a non-standard treatment of frequencies and arrived at the level of girls. The linear regression is Λ_m = 0.9454 + 0.0343(age). Here, too, the regression coefficient is not different from zero but evidently the boys, perhaps, attain a certain level of courage to form the text more freely or purposefully. This can be either a sign of stronger future development of the boys or a sign of slower development in the past than that of the girls. According to our data the development lines of both genders cross at the age of 9.57.

The two sequences are shown in Figure 9.1.

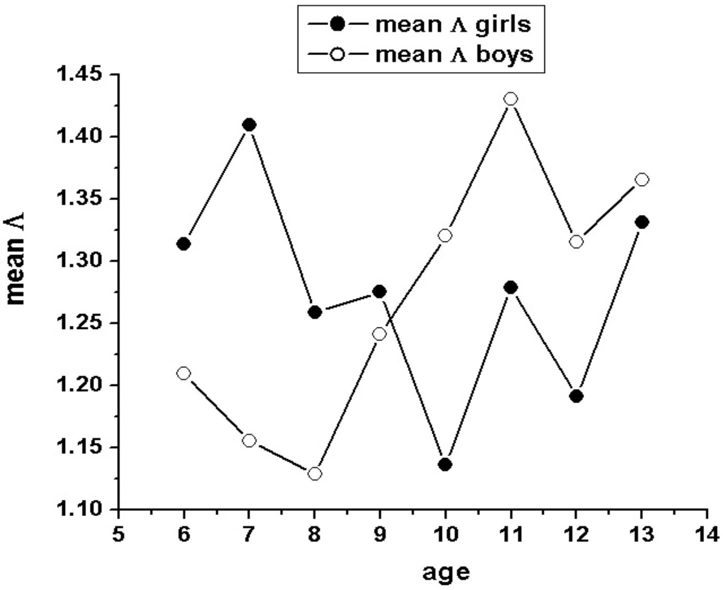


Figure 9.1. Children development of lambda in English

The data for Czech children presented in Appendix (Table 4e) are given in the form of school classes, e.g. class 2 means children between 7 – 8 years. For the sake of simplicity we shall take the mean age, 7.5. The results are shown in Table 9.3

Table 9.3
Lambdas in Czech stories written by children according to age and gender

Age	f	m	mean f	mean m
7.5	0.9520, 1.2348, 0.9565, 1.2287, 1.1594	1.1047, 1.0320, 1.2457, 1.1846, 1.2513	1.1063	1.1637
8.5	1.4267, 1.4917, 1.6241, 1.2947, 1.6562, 1.4000, 1.6096	1.4124, 1.4783, 1.7528, 1.5095, 1.4553, 1.4584	1.5004	1.5111
9.5	1.5511, 1.5902, 1.5843, 1.5957, 1.6131	1.5550, 1.5068, 1.4109, 1.4850	1.5869	1.4894
10.5	1.5811, 1.5809, 1.3461, 1.6278, 1.6524, 1.5918	1.4987, 1.6087	1.5634	1.5537
11.5	1.5436, 1.6989, 1.6962, 1.2919, 1.6198	1.4600	1.5701	1.4600
12.5	1.7992, 1.7197, 1.6319, 1.5534	1.6726, 1.6751	1.6761	1.6739
13.5	1.7039, 1.8366, 1.7842	1.5742, 1.7011	1.7749	1.6377
14.5	1.5805, 1.7328, 1.9103, 1.6883, 1.7389, 1.6392, 1.5983	1.8058, 18512, 1.6961, 1.8685	1.6951	1.8053

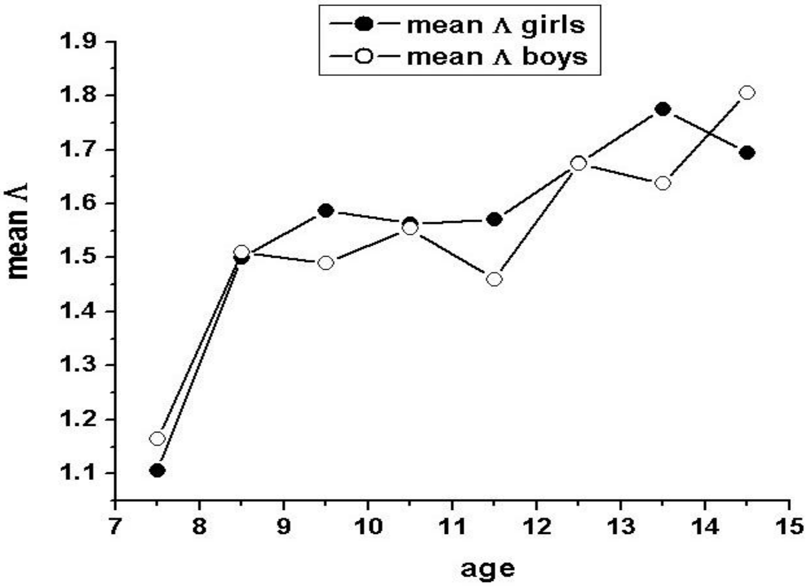


Figure 9.2. Children development of lambda in Czech

Though the data are very scarce, we can suppose that boys and girls have approximately the same development. Because of the different representation of ages, a comparison with English is not possible. The Czech example can be considered representative, the English one not yet. This is also caused by the fact that in the English data we have both narrated and written texts. It may be supposed that deeper research will discover a difference. We hope that future investigations will bring more light in this domain. The number of texts from different languages must be increased in order to obtain more reliable results.

10. Pathological texts

Reading pathological texts necessarily evokes the question “what is text?” Omitting hypertexts one can say that it is something ordered linearly. In our case this “something” is usually words or signs or phrases or sentences belonging to a certain language. But broadly speaking, any linear sequence of things of the same kind may represent a text, e.g. a row of houses, the human life (which is linear in time), a melody, etc. (cf. Altmann 2009). It is a question of definition, not a problem of reality. In textology, one considers text a written or spoken enunciation understandable to at least a group of people, that is, written in a certain language. However, Dadaistic texts break even this presupposition and mix language with articulated but nonsense words, destroy syntax and create senseless sentences in a given language or produce even texts (= linear sequences of written symbols) void of any language. The extent of this mélange is not determined. And even the literary evaluation of these texts differs from usual norms. Nevertheless, if the text entities are identified, it is always possible to compute the lambda indicator. In Table 10.1 we ordered 34 Dada texts by increasing lambda because the Dadaistic world view is (intentionally, pathologically or effectually) reversed, hence low-lambda poems should mean "beautiful" poems and vice versa. A “best-dada” is, perhaps, a text differing most drastically from “normal” texts and “worst-dada” is a text of high literary value. So, at first sight one can see that "the best" Dada-poems belong to the famous German painter Kurt Schwitters while "the worst" Dada-poem, *Gesang an die Welt*, by George Grosz (Λ = 1.9849) becomes in fact the top German poem found so far (with lambda higher than Λ = 1.8825 of *Elegie 12* by Goethe – not given in the Appendix). The 34 Dadaistic poems analyzed in Table 10.1 can be found on the Internet at http://members.peak.org/~dadaist/English/Graphics/ poems.html.

The very low lambdas at the beginning of Table 10.1 are caused by abnormal word rank-frequencies. In pathological texts even parabolic Zipf's curves have been reported (see Piotrowska, Piotrowska 2004), a phenomenon not known in “normal” texts. The first ten poems in Table 10.1 differ drastically from texts written in the same language (German, English), while those written in non-language (e.g. the Simultangedicht “kaa gee dee” by Schwitters, No. 4 in Table 10.1) cannot be compared with anything. If we do not ascribe the Dada-poems to a certain language and consider the written sequences as words, then computing their “vocabulary richness” using transformation (1.11) yields quite “normal” results from R_1 = 0.3960 up to R_1 = 0.8737. That means, it is not the “vocabulary” alone that distinguishes these texts from “normal” ones.

Table 10.1
Lambda values of Dadaistic poems

	Author	**Poem Title**	***N***	***V***	***L***	***Λ***
1	Schwitters, K.	What a b what a b what a beauty	85	5	19.4707	0.4420
2	Schwitters, K.	Zwölf	43	11	12.8929	0.4898
3	Schwitters, K.	Simultangedicht "kaa gee dee"	43	16	17.4787	0.6640
4	Schwitters, K.	Das Urgebet der Scholle	27	13	14.4721	0.7672
5	Schwitters, K.	Cigarren	34	19	19.2426	0.8668
6	Schwitters, K.	Sie puppt mit Puppen	75	31	35.8771	0.8970
7	Hülsenbeck, R.	Die Primitiven	12	10	10.2361	0.9205
8	Arp, H.	Sekundenzeiger	78	25	39.7103	0.9633
9	Schwitters, K.	Unsittliches I-Gedicht	13	12	11.4142	0.9781
10	Ball, H.	Seepferdchen und Flugfische	57	32	32.2426	0.9932
11	Schwitters, K.	Perhaps Strange	81	37	43.5680	1.0265
12	Schwitters, K.	A. M.	57	36	36.2426	1.1164
13	Schwitters, K.	Gedicht	19	18	17.4142	1.1720
14	Schwitters, K.	Die zute Tute	43	32	31.8284	1.2091
15	Ernst, M.	Die Wasserprobe	43	32	32.6503	1.2403
16	Schwitters, K.	An Anna Blume	216	103	116.5193	1.2593
17	Schwitters, K.	Seenot	35	29	28.8284	1.2718
18	Schwitters, K.	So, so!	39	32	32.2426	1.3154
19	Ball, H.	Gadji beri bimba	117	74	76.4049	1.3506
20	Arp, H.	Opus Null	265	138	147.7548	1.3511
21	Ball, H.	Karawane	47	41	40.8284	1.4525
22	Ball, H.	Totentanz	160	102	108.1538	1.4899
23	Herzfelde, W.	Trauerdiriflog	55	48	47.4142	1.5003
24	Soupault, P.	Tomatenblüten	54	45	47.5373	1.5251
25	Hennings, E.	Morfin	68	56	57.8863	1.5600
26	Hennings, E.	Ätherstrophen	75	62	62.6569	1.5665
27	Hülsenbeck, R.	DaDa Schalmei	132	95	97.8191	1.5715
28	Soupault, P.	Westwego	154	109	114.1701	1.6217
29	Hennings, E.	Nach dem Cabaret	74	66	65.8284	1.6628

30	Hennings, E.	Tänzerin	77	68	69.0645	1.6921
31	Baargeld, J.T.	Bimmelresonnanz II	65	60	61.5765	1.7174
32	Herzfelde, W.	Das Dadalyripipidon	99	88	88.2426	1.7788
33	Hennings, E.	Gesang zur Dämmerung	100	91	91.6503	1.8330
34	Grosz, G.	Gesang an die Welt	352	263	274.3623	1.9849

While in “normal” texts one can venture some predictions, in “abnormal” texts everything is open and the analysis yields a diagnosis *a posteriori*. As an example we present the results of the analysis of the Romanian poem *Magie neagră* (1991) (“Black witchcraft”) written by a schizophrenic patient Tiberius who committed suicide. A purely formal analysis shows that the mean lambda significantly differs from that of Romanian poems and the sequence of lambdas has an expressed course. In Figure 10.1 the decreasing lambda is evident and can be captured even by a straight. We obtain Λ = 1.6084 – 0.0332(page) with both the *t*-test for the regression coefficient and the *F*-test for regression being significant. The determination coefficient is merely $R^2 = 0.50$. It would be more appropriate to use a decreasing wave function but without a prev-

Table 10.2
Page-wise analysis of a non-dada text

	N	***V***	***L***	***Λ***
Whole text	879	400	420.1586	1.4072
page 1	52	47	46.4142	1.5317
page 2	103	77	77.6569	1.5176
page 3	88	68	68.6569	1.5171
page 4	90	75	75.2426	1.6338
page 5	94	66	66.6569	1.3992
page 6	95	64	64.6569	1.3460
page 7	75	53	53.6569	1.3415
page 8	74	47	48.9907	1.2375
page 9	60	52	51.8284	1.5360
page 10	77	54	54.2426	1.3289
page 11	60	36	37.4787	1.1107
			average =	1.4091
			stdev =	0.1545

ious hypothesis which can be uttered only by a psychiatrist or an *a posteriori* interpretation of a specialist it would be mere speculation. At the present state of affairs it cannot be said what kind of surprise can be expected in the next text.

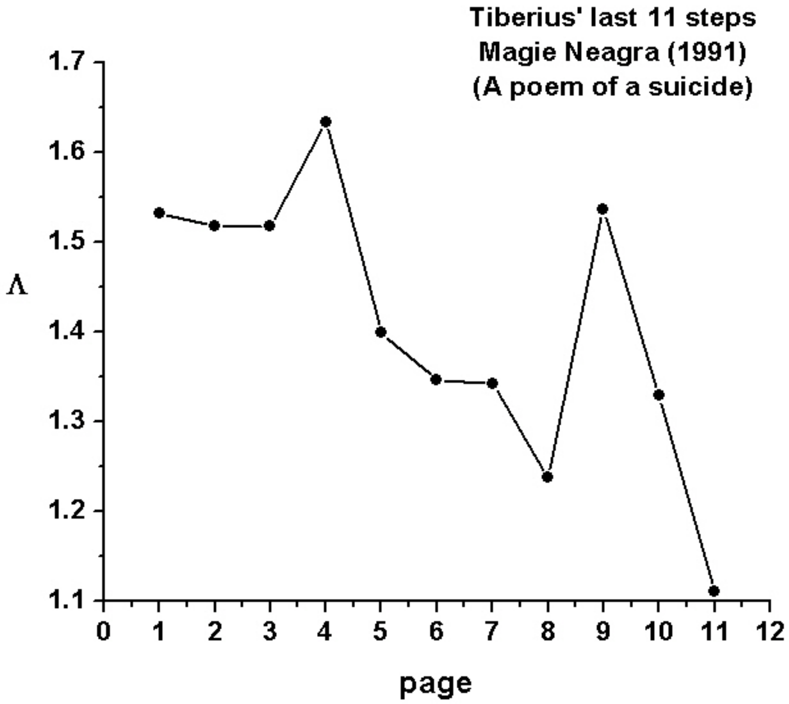


Figure 10.1. A poem of suicide by a schizophrenic patient

We can consider these types of texts as a special genre in which an inner stimulus leads to deviations of different kinds. This whole domain must be left to specialists in literary history and psychiatry. Nevertheless, lambda is an expression of a measurable property and in the future, perhaps, the individual values or courses of functions will be attributable to special psychic states.

11. Conclusions

The lambda structure of a text is an enciphered quantitative concept subsuming different text properties. It expresses the internal structure of the rank-frequency distribution of words but can be used for any finite set of linguistic entities. Its advantage is the normalization, i.e. the elimination of the influence of text size on the frequency structure. Further, it has a good testability for differences between texts and can be used for expressing the development of a writer or of a genre.

It captures the (ir)regularities in the frequency distribution of words and is associated with the *h*-point whose special function expresses the vocabulary richness, i.e. a simple transformation of lambda yields the vocabulary richness based on the *h*-point. Its relationship to other text properties must still be investigated; but since no property is isolated, it is to be expected that they will be discovered soon.

The study of lambda may help to describe the movement of text organization both in individual texts, e.g. novels, and in the history of the given genre. It could turn out to be useful for the demarcation of genres, hence the development of texts from primitive forms to highly structured journalistic ones can also be scrutinized using it.

Since it can differentiate writers, it can differentiate also styles, but in order to do it, the style must be defined and captured quantitatively. A mere name for a style is not sufficient. While genre can mostly be localized – press texts appear in newspapers, scientific articles in respective journals, etc. – style is positioned at a higher level because even texts of the same genre can be differentiated both as to formal features and the way of writing (serious, ironic, humoristic, informal, etc.). Of course, all stylistic concepts concerning non-formal properties are intuitive, not strictly defined and not quantified.

The study of lambda opens some new vistas, namely the quantitative study of the development of children's ability to treat language, the study of abnormal texts created consciously by violating to some aspects of the standard language, and finally, the study of texts created by some psychic disturbance. Last but not least, it opens a way for theoretical advancement one of whose aspects will be indicated here. Consider the data in Table 8.4 concerning the complete work of the Romanian poet M. Eminescu. Every poet has his own ductus and style – this fact is easier visible in painting and audible in music – but at the same time he tries to be original in every new work. Thus there are two antagonistic forces influencing his creations: the force of perseveration and the force of innovation. Their clash may result in peculiar formations characteristic for their regularity. Here we show only the above mentioned case. If for each poem one computes its lambda-difference to all other poems, one obtains the *u*-tests signalizing dissimilarity ($|u| > 1.96$) or non-significant dissimilarity (= similarity) if $|u| \leq 1.96$. For the given writer the number of similarities is characteristic both for each text separately and for the writer as a whole. For the sake of simplicity we consider *y*, the number of similarities, a continuous variable. We first compute for each lambda (text) all its similarities and set up the hypothesis that the relative rate of change of *y*, i.e. *dy/y* is proportional to a function of innovation minus a function of perseveration. The equation can be written as follows

$$(11.1)\quad \frac{dy}{y} = \left(\frac{b}{x-m} - \frac{c}{M-x} \right) dx\,.$$

Here $x = 1/\Lambda$ is lambda reciprocal, *b,c* are parameters and *m, M* can be considered the corresponding minimum and maximum of *x* respectively. Solving this equation we obtain

$$(11.2)\quad y = a(x-m)^b (M-x)^c$$

which is the well known doubly truncated beta-function. For the sake of simpler graphical presentation we consider $x = 1/\Lambda$ and compute the parameters iteratively. We obtain the results presented in Table 11.1 and graphically in Figure 11.1.

Table 11.1 (lambdas taken from Table 8.4)
The relationship between $x = 1/\Lambda$ and the number of similarities (degrees) of a text in Eminescu´s poems (*x* increasing)

Lambda	*x*	*y*		**Lambda**	*x*	*y*		**Lambda**	*x*	*y*
1.9921	0.5019828	3		1.7502	0.5713633	70		1.6202	0.6172078	75
1.9632	0.5093725	8		1.7465	0.5725737	77		1.6175	0.618238	47
1.9548	0.5115613	12		1.7418	0.5741187	96		1.6128	0.6200397	76
1.9389	0.5157564	12		1.7387	0.5751423	61		1.6121	0.6203089	80
1.9211	0.5205351	12		1.7371	0.5756721	80		1.6037	0.623558	82
1.8992	0.5265375	22		1.7369	0.5757384	90		1.5997	0.6251172	74
1.8876	0.5297733	25		1.7367	0.5758047	90		1.5981	0.6257431	82
1.8823	0.5312649	25		1.7319	0.5774005	87		1.5894	0.6291682	72
1.8807	0.5317169	34		1.7292	0.5783021	81		1.5872	0.6300403	83
1.8752	0.5332765	42		1.7237	0.5801474	82		1.5812	0.6324311	64
1.8612	0.5372878	51		1.7229	0.5804167	82		1.5767	0.6342361	79
1.8599	0.5376633	40		1.7128	0.5838393	85		1.5726	0.6358896	64
1.8549	0.5391126	44		1.7095	0.5849664	82		1.5711	0.6364967	65
1.8504	0.5404237	37		1.706	0.5861665	80		1.566	0.6385696	74
1.8290	0.5467469	47		1.7059	0.5862008	85		1.5643	0.6392636	58

1.8258	0.5477051	51		1.7059	0.5862008	69		1.5629	0.6398362	69
1.8253	0.5478551	49		1.7051	0.5864759	84		1.5514	0.6445791	54
1.8253	0.5478551	49		1.7013	0.5877858	91		1.5510	0.6447453	68
1.8236	0.5483659	54		1.6985	0.5887548	70		1.5432	0.6480041	52
1.823	0.5485464	42		1.6963	0.5895184	73		1.5385	0.6499838	47
1.8209	0.5491790	60		1.6936	0.5904582	74		1.5332	0.6522306	41
1.8203	0.5493600	64		1.6933	0.5905628	72		1.5314	0.6529973	61
1.8185	0.5499038	49		1.6868	0.5928385	92		1.5313	0.6530399	61
1.8148	0.5510249	46		1.6852	0.5934014	63		1.5259	0.6553509	48
1.8106	0.5523031	63		1.6837	0.5939300	79		1.5200	0.6578947	53
1.808	0.5530973	63		1.6827	0.5942830	73		1.5088	0.6627784	53
1.8067	0.5534953	61		1.6800	0.5952381	79		1.5082	0.663042	44
1.8064	0.5535872	55		1.6703	0.5986948	96		1.5012	0.6661338	47
1.8042	0.5542623	62		1.6681	0.5994844	78		1.4837	0.6739907	40
1.8013	0.5551546	59		1.6677	0.5996282	98		1.4776	0.6767731	47
1.7998	0.5556173	72		1.6623	0.6015761	85		1.4675	0.6814310	41
1.7968	0.556545	61		1.6616	0.6018296	60		1.4647	0.6827337	49
1.7957	0.5568859	60		1.6580	0.6031363	75		1.4633	0.6833869	39
1.7951	0.557072	61		1.6566	0.603646	77		1.4632	0.6834336	42
1.7931	0.5576934	72		1.6550	0.6042296	79		1.4575	0.6861063	41
1.7898	0.5587216	59		1.6550	0.6042296	67		1.4561	0.686766	41
1.7829	0.560884	65		1.6533	0.6048509	93		1.4467	0.6912283	41
1.7789	0.5621451	76		1.6531	0.6049241	65		1.4254	0.7015575	34
1.7786	0.56224	68		1.6520	0.6053269	78		1.4153	0.706564	43
1.7774	0.5626196	69		1.6514	0.6055468	59		1.4055	0.7114906	28
1.7767	0.5628412	68		1.6503	0.6059504	65		1.3950	0.7168459	32
1.7701	0.5649398	68		1.6474	0.6070171	66		1.3950	0.7168459	32
1.7666	0.5660591	74		1.6405	0.6095703	65		1.3595	0.7355645	18
1.7618	0.5676013	72		1.6297	0.6136099	73		1.3225	0.7561437	21
1.7600	0.5681818	72		1.6267	0.6147415	76		1.3175	0.7590133	18
1.7556	0.5696058	73		1.6256	0.6151575	71		1.2778	0.7825951	11
1.7542	0.5700604	70		1.6230	0.6161429	75		1.2116	0.8253549	8
1.7534	0.5703205	62		1.6209	0.6169412	74		1.2070	0.8285004	9
1.7523	0.5706785	72		1.6205	0.6170935	71				

The determination coefficient $R^2 = 0.86$ is sufficiently high. The F-test is highly significant.

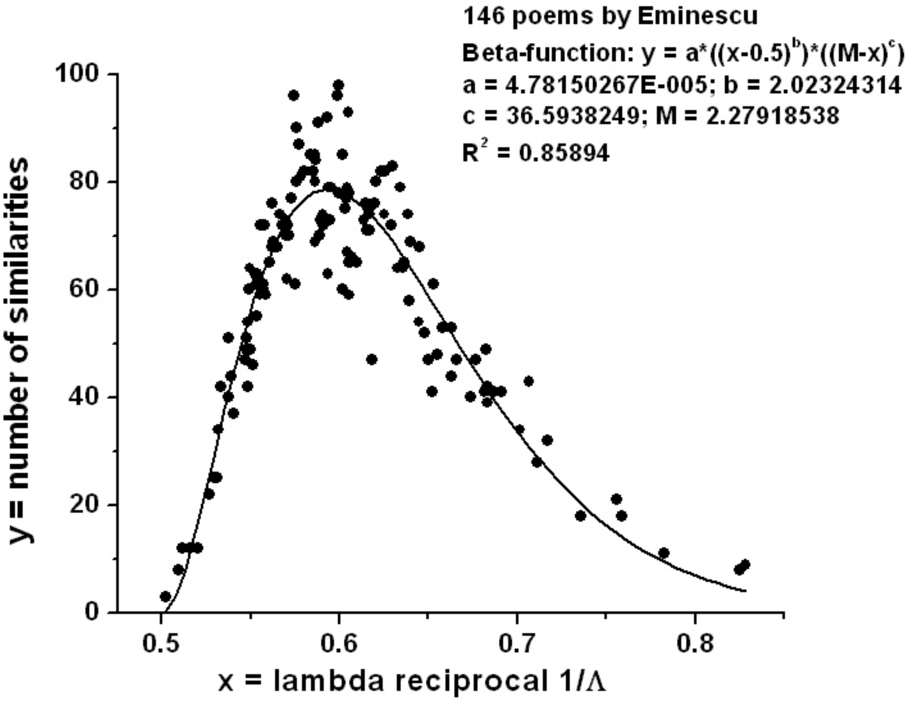


Figure 11.1. Graphical presentation of <1/Λ, degrees>

In subsequent research one may scrutinize whether all writers follow this regularity, whether there are differences between languages and genres, whether the course of the curve has commonalities with the development of the writer, whether it holds only for different texts or also for chapters of a novel, etc.

In the present book, we tried to show the usefulness of the strictly defined concept of lambda, a normalized indicator capturing the complete rank-frequency distribution, its appearance in texts and some of its relationships. Its use opens an infinite domain of research in quantitative textology.

References

Altmann, G. (1978). Zur Anwendung der Quotiente in der Textanalyse. *Glottometrika 1, 91-106.* Bochum: Brockmeyer.

Altmann, G. (2006). Fumndamentals of quantitative linguistics. In: Genzor, J., Bucková, M. (eds.), *Favete linguis: 15-27.* Bratislava: Slovak Academic Press.

Altmann, G. (2009). Texte und Theorien. In: Delcourt, Ch., Hug, M. (eds.), *Mélanges offerts à Charles Muller: 37-45.* Paris: Conseil International de la Langue Français.

Andersen, S., Altmann, G. (2006). Information content of words in texts. In: Grzybek, P. (ed.), *Contributions to the Science of Text and Language. Word length studies and related issues: 91-115*. Dordrecht: Springer.

Baayen, R.H. (1989). *A corpus-based approach to morphological productivity. Statistical analysis and psycholinguistic interpretation.* Diss. Amsterdam: Free University.

Bernet, Ch. (1988). Faits lexicaux. Richesse du vocabulaire. In: Thoiron, P. et al. (eds.), *Etudes sur la richesse et la structure lexicale: 1-11* Paris: Champion.

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux: structure et evolution.* Genève: Slatkine.

Bunge, M. (1963). *The myth of simplicity. Problems of scientific philosophy*. Englewood Cliffs, N.J.: Prentice.-Hall

Carroll, J.B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice Hall.

Cossette, A. (1994). *La richesse lexicale et sa mesure.* Geneva-Paris: Slatkine-Champion.

Covington, M.A., McFall, J.D. (2010). Cutting the Gordian Knot: The Moving Average Type-Token Ratio (MATTR*). Journal of Quantitative Linguistics 17(2), 94-100.*

Ejiri, K., Smith, A.E. (1993). Proposal of a new ´constraint measure´ for text. In: Köhler, R., Rieger, B.B. (eds.), *Contributions to quantitative linguistics: 195-211*. Dordrecht: Kluwer.

Falconer, K.J. (1990). *Fractal geometry. Mathematical foundations and Applications*. Chichester: Wiley.

Guiraud, P. (1954). *Les catactères stitistiques du vocabulaire.* Paris: Presses Universitaires de France.

Guiraud, P. (1959). *Problèmes et methods de la statistique linguistique*. Dordrecht: Reidel.

Herdan, G. (1960). *Type-token mathematics.* The Hague: Mouton.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. New York: Springer.

Hess, C.E., Sefton, K.M., Landry, R.G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research 29, 129-134.*

Hess, C.E., Sefton, K.M., Landry, R.G. (1989). The reliability of type-token ratios for the oral language of school age children. *Journal of Speech and Hearing Research 32, 536-540.*

Holmes, D.I. (1991). Vocabulary richness and the Prophetic Voice. *Literary and Linguistic Computing 6(4), 259-268.*

Holmes, D.I. (1992). A stylometric analysis of Mormon Scripture and related texts. *Journal of the Royal Statsitical Society A 155(1), 91-120.*

Holmes, D.I., Forsyth, R.S. (1995). The Federalist revised: new directions in authorship attribution. *Literary and Linguistic Computing 10(2), 111-127.*

Honore, T. (1979). Some simple measures of richness of vocabulary. *ALLC Bulletin 7, 172-177.*

Hřebíček, L. (1997). *Lectures on Text Theory***.** Prague: Oriental Institute**.**

Hřebíček, L. (2000). *Variation in sequences***.** Prague: Oriental Institute**.**

Hřebíček, L. (2007). *Text in Semantics.* Prague: Oriental Institute**.**

Kelih, E. (2009a). Slawisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas´ stal´ (KZS)". In: Kelih, E., Levickij, V., Altmann, G. (eds.), *Methods of text analysis: 106-124*. Černivci: ČNU.

Kelih, E. (2009b). Preliminary analysis of a Slavic parallel corpus. In: Levická, J., Garabík, R. (eds.), *NLP, Corpus Linguistics, Corpus Based Grammar Research. Fifth International Conference Smolenice, Slovakia, 25-27 November 2009. Proceedings: 175-183.* Bratislava: Tribun.

Köhler, R., Galle, M. (1993). Dynamic aspects of text characteristics. In: Hřebíček, L., Altmann, G. (eds.), *Quantitative text analysis: 46-53.* Trier: WVT.

Kuraszkiewicz, W. (1963). *La richesse du vocabulaire dans quelques grands textes polonais en vers.* Wroclaw: Ossolineum.

Malvern, D., Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversuity. *Language Testing 19, 85-104.*

Martel, P. (1986). Richesse lexicale et variable sociologiques. In: *Méthodes quantitatives et informatique dans l´étude des texts. Colloque international de CNRS à Université de Nice, 5-8 juin 1985: 599-608.* Paris: Champion

Martynenko, G. (2010). Measuring lexical richness and its harmony. In: Grzybek, P., Kelih, E., Mačutek, J. (eds.*), Text and language: 125-132.* Wien: Praesens.

Menard, N. (1983). *Mesure de la richesse lexicale*. Paris: Slatkine.

Muller, Ch. (1968). Mesure de la richesse lexicale. *Travaux de linguistique et de literature 6, 73-84.*

Muller, Ch. (1971). Sur la mesure de la richesse lexicale. Théorie et experiences, homage à René Michéa. *Études de linguistique appliqué 74-87.*

Müller, D. (2002). Computing the type token relation from the *a priori* distribution of types. *Journal of Quantitative Linguistics 9, 193-214.*

Müller, W. (1971). Wortschatzumfang und Textlänge. *Muttersprache 81(4), 266-276.*

Nešitoj, V.V. (1975). Dlina teksta i ob´em slovarja. Pokazateli leksičeskogo bogatstva teksta. In: *Metody izučenija leksiki: 110-118.* Minks: BGU.

Neumann, J.v., Kent, R.H., Bellinson, H.R., Hart, B.I. (1941). The mean square successive difference. *Annals of Mathematical Statistics 12, 153-162.*

Orlov, J.K. (1983). Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Guiter, H., Arapov, M.V. (eds.), *Studies on Zipf´s law: 154-233.* Bochum: Brockmeyer.

Orlov, J.K., Boroda, M.G., Nadarejšvili, I.Š. (1982). *Sprache, Text, Kunst. Quantitative Analysen.* Bochum: Brockmeyer.

Panas, E. (2001). The generalized Torquist: Specification and estimation of a new vocabulary text-size function. *Journal of Quantitative Linguistics 8, 233-252.*

Piotrowska, W., **Piotrowska, X.** (2004). Statistical parameters in pathological texts. *Journal of Quantitative Linguistics 11(1-2), 133-140.*

Popescu, I.-I. (2007). Text ranking by the weight of highly frequent words. In: Grzybek, P., Köhler, R. (eds.), *Exact methods in the study of language and text: 555-565.* Berlin-New York: de Gruyter.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N. (2009). *Word frequency studies.* Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Altmann, G., Köhler, R. (2009). Zipf´s law – another view. *Quality and Quantity 44(4), 713-731.*

Popescu, I.-I., Kelih, E., Mačutek, J., Čech, R., Best, K.-H., Altmann, G. (2010). *Vectors and codes of text.* Lüdenscheid: RAM.

Popescu, I.-I., Mačutek. J., Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM.

Popescu, I.-I., Mačutek, J., Altmann, G. (2010). Word forms, style and typology. *Glottotheory 3(1), 89-96*

Ratkowsky, D.A., Hantrais, L. (1975). Tables for comparing the richness and structure of vocabulary in texts of different length. *Computers and Humanities 9, 69-75.*

Ratkowsky, D.A., Halstead, M.H., Hantrais, L. (1980). Measuring vocabulary richness in literary works. A new proposal and a re-assessment of some earlier measures. *Glottometrika 2, 125-147.*

Richards, B. (1987). Type/token ratios: what do they really tell us? *Journal of Child Language 14, 201-209.*

Schach, E. (1987). Empirische Eigenschaften der TTR bei ausgewählten Texten. In: Wagner, K.R. (ed.), *Wortschatz-Erwerb: 102-114.* Bern: Lang.

Serant, D., Thoiron, P. (1988). Richesse lexicale et topographie des formes répétées. In: Thoiron, P. et al. (ed.), *Éudes sur la richesse et la structure lexicale: 125-129.* Paris-Genève: Champion; Slatkine.

Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistical Association 70, 542-547.*

Sichel, H.S. (1986). Word frequency distributions and type-token characteristics. *The Mathematical Scientist 11, 45-72.*

Skinner, B.F. (1939). The alliteration in Shakespeare´s sonnets: A study in literary behaviour. *The Psychological Record 3, 186-192.*

Skinner, B.F. (1957). *Verbal behaviour.* Acton: Copley.

Suprun, A.E. (1979). K količestvennoj ocenke leksičeskogo bogatstva teksta. *Naučnye doklady vyssej školy/Filologičeskie nauki 1, 44-48.*

Tešitelová, M. (1972). On the so-called vocabulary richness. *Prague Studies in Mathematical Linguistics 3, 103-120.*

Thoiron, P. (1986). Indice de diversité et mesure de la richesse lexicale. In: *Méthodes quantitatives et informatique dans l´étude des texts. Colloque international de CNRS à Université de Nice, 5-8 juin 1985: 831-840.* Paris: Champion

Thoiron, P. (1988). Richesse lexicale et classement des texts. In: Thoiron, P. et al. (eds.)*, Études sur la richesse et la structure lexicale: 141-163.* Paris-Genève: Champion; Slatkine.

Thoiron, P., Labbé, D., Serant, D. (eds.) (1988). *Études sur la richesse et la structure lexicale.* Paris-Genève: Champion; Slatkine.

Tricot, C. (1995). *Curves and fractal dimension.* New York: Springer.

Tuldava, J. (1977). O kvantitativnych charakteristikach bogatstva leksičeskogo sostava chudožestvennych tekstov. *Acta et Commentationes Universitatis Tartuensis 437,159-175.*

Tuldava, J. (1995). On the relation between text length and vocabulary size. In: Tuldava, J., *Methods in quantitative linguistics: 131-150.* Trier: WVT.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts.* Lüdenscheid: RAM.

Tweedie, F., Baayen, R.H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities 32, 323-352.*

Wachal, R.S., Spreen, O. (1973). Some measures of lexical diversity in aphasic and normal language performance. *Language and Speech 16, 169-181.*

Weizman, M. (1971). How useful is the logarithmic type.-token ratio? *Journal of Linguistics 7, 237-243.*

Woronczak, J. (1965). Metody obliczania wskaźników bogactwa slownikowego tekstów. In: Mayenowa, M.R. (ed.), *Poetyka i matematyka: 145-163.* Warszawa: PIW.

Yule, G.U. (1944). *The statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

Appendix: All data

Table 1: Belorussian

Translation (1950) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4145	1916	2066.6239	1.8036	0.000328
2	4177	2079	2207.9228	1.9140	0.000153
3	6367	2863	3049.7554	1.8221	0.000125
4	3791	2116	2223.9183	2.0994	0.000262
5	3791	1854	1954.6261	1.8452	0.000213
6	7547	3347	3500.7174	1.7987	0.000122
7	6063	2953	3082.5568	1.9232	0.000200
8	5362	2783	2902.4616	2.0187	0.000167
9	3312	1776	1849.5612	1.9658	0.000353
10	5319	2814	2936.1131	2.0567	0.000252

Table 2a: Bulgarian

Translation (1976) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4653	1709	1872.2601	1.4758	0.000152
2	4734	1913	2055.0948	1.5955	0.000140
3	7224	2581	2819.1777	1.5059	0.000201
4	4305	2007	2135.4096	1.8026	0.000148
5	4277	1706	1827.2504	1.5513	0.000230

6	8673	2979	3219.7384	1.4620	0.000133
7	6992	2729	2984.4337	1.6410	0.000254
8	6242	2591	2796.4254	1.7003	0.000160
9	3787	1663	1765.8570	1.6685	0.000266
10	6278	2633	2860.8613	1.7307	0.000120

Table 2b: Bulgarian private letters (2005)

Name	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Boris 2	761	400	428.4493	1.6222	0.001690
Ceneva1	352	201	205.3782	1.4858	0.002612
Ceneva2	515	285	289.7990	1.5260	0.001816
Janko1	483	286	297.0321	1.6506	0.001243
Janko 3	406	238	247.2991	1.5889	0.001660

Table 3: Croatian

Translation (1945) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4582	1900	2064.7868	1.6498	0.000256
2	4689	2096	2244.0413	1.7569	0.000128
3	7160	2888	3136.0351	1.6884	0.000126
4	4316	2149	2274.8955	1.9160	0.000176
5	4255	1881	2038.2506	1.7383	0.000129
6	8553	3222	3550.8479	1.6325	0.000102

7	6841	2958	3172.7856	1.7787	0.000104
8	6075	2845	3046.4531	1.8973	0.000143
9	3760	1795	1955.6023	1.8595	0.000262
10	6184	2823	3048.0405	1.8687	0.000116

Table 4a: Czech poetry

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Blatný, I.	Verše I (1948)	26253	8588	9621.8897	1.6197	0.000040
Blatný, I.	Verše II (1933 – 1953)	25494	9007	10077.9504	1.7419	0.000031
Bondy, E.	Básnické dílo V (1991)	20772	6869	7626.2192	1.5851	0.000073
Diviš, I.	Noé vypouští krkavce (1975)	7870	4655	4854.7688	2.4034	0.000115
Diviš, I.	Češi pod Hauscaránem (1998)	6645	3283	3587.6026	2.0645	0.000100
Dryje, F.	Mrdat (1998)	9608	4253	4463.1333	1.8500	0.000111
Holan, V.	První testament (1940)	5286	2521	2709.4732	1.9080	0.000140
Holan, V.	Spisy I (1965)	19425	7794	8626.2336	1.9044	0.000068
Holan, V.	Toskána (1962)	5543	2651	2794.7592	1.8877	0.000147
Holan, V.	Propast propasti (1982)	23995	6754	7662.5059	1.3987	0.000062
Holub, M.	Časoprostor (1958 – 1998)	5064	2491	2694.9286	1.9715	0.000139
Hrubín, F.	Zpívám (1930s – 1971)	6802	2774	2998.8271	1.6898	0.000140
Jakoubek, J.	Balady, romance a hněvance (1992)	24508	7596	8422.1613	1.5084	0.000058
Jirous, I.M.	Magorova summa (1998)	29842	10232	11137.949	1.6701	0.000050
Jirousová, V.	Co je tu, co tu není (1970s – and 1980s)	8692	3802	4125.597	1.8699	0.000110

Kainar, J.	Synkopy (1930s – 1971)	6796	3059	3293.5380	1.8575	0.000125
Kaprál, Z.	Plané palposty (1998)	5052	2766	2881.4802	2.1120	0.000138
Kotrlá, I.	Února (1992)	7415	3537	3811.9716	1.9896	0.000086
Kovtun, J.	Hřbet velryby (1995)	11305	4954	5392.7512	1.9335	0.000066
Král, P.	Chiméry a exil (1998)	15595	7126	7439.0518	2.0001	0.000052
Mikulášek, O.	Verše (1930s – 1990s)	17827	6424	7158.2734	1.7070	0.000084
Motýl, P.	Lahve z ubytovny (2000)	9234	4131	4480.9079	1.9243	0.000194
Nápravník, M.	Vůle k noci (1959)	9354	4774	5104.6938	2.1672	0.000084
Nuska, B.	Okamžiky (1998)	13675	5399	6295.6049	1.9042	0.000080
Orten, J.	Knihy veršů (1930s – early of the 1940s)	26242	7891	8949.6461	1.5071	0.000062
Reynek, B.	Svěcení (second half of the 20th century)	6010	3002	3196.8868	2.0102	0.000103
Seifert, J.	Třeba vám nesu růže (the 1920s – the 1980s)	25426	8678	9668.7819	1.6752	0.000053
Suchý, J.	Růže růžová (1971)	9428	3238	3576.8931	1.5079	0.000095
Topol, J.	Básně (1997)	24924	10177	10926.9805	1.9300	0.000047
Topol, J.	Miluju tě k zbláznění (1988)	12803	4894	5374.6524	1.7243	0.000144
Trojak, B.	Pan Twardowski (1998)	2585	1606	1668.8487	2.2032	0.000305
Trojak, B.	Kuním štětcem (1996)	1571	1035	1071.0131	2.1789	0.000312
Vodseďálek, I.	Bloudění (1962 – 1968)	8579	3942	4176.8397	1.9151	0.000132
Vokolek, V.	Tanec bludných kořenů (1940s – 1980s)	10310	4771	5153.7313	2.0062	0.000064

Table 4b: Czech

Translation (1948) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	3925	1773	1928.8121	1.7661	0.000199
2	4381	2109	2267.9280	1.8851	0.000174
3	6670	2904	3182.1657	1.8244	0.000130
4	3920	2111	2272.3298	2.0829	0.000153
5	3852	1854	1992.5997	1.8548	0.000221
6	8117	3369	3664.2640	1.7648	0.000131
7	6390	2945	3170.4844	1.8882	0.000165
8	5738	2805	2994.6971	1.9617	0.000174
9	3451	1820	1940.3427	1.9892	0.000152
10	5736	2891	3083.2654	2.0204	0.000103

Table 4c: Czech prose

Author	**Text. year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Bondy, E.	Leden na vsi (1977)	23643	7433	8432.7212	1.5600	0.000065
Fejt, V.	Věž (2000)	15818	5831	6365.5766	1.6900	0.000070
Fischl, V.	Strýček Bosko (1993)	15738	4256	4941.7011	1.3179	0.000104
Flos, F.	Lovci kožešin (1970)	4568	2161	2335.0989	1.8707	0.000295
Fučík, J.	Reportáž psaná na oprátce (1945)	23815	7305	8186.8762	1.5047	0.000046
Hájek, P.	Kráska a netvor (1998)	11695	5337	5790.5904	2.0143	0.000067
Hovorka, J.	Okarína do re mi (1992)	23402	8014	9032.8855	1.6865	0.000040

Hrabal, B.	Blitzkrieg (1952)	522	323	341.9900	1.7805	0.001175
Hrabal, B.	Česká rapsodie (1992)	1989	906	1022.4929	1.6966	0.000494
Hrabal, B.	Expozé panu ministru (1968)	1044	638	684.17	1.9783	0.000494
Hrabal, B.	Kůň truhláře Bárty (1991)	1500	736	805.3042	1.7045	0.000454
Hrabal, B.	Květnové idy (1992)	1196	648	722.2958	1.8580	0.000378
Hrabal, B.	Lednová povídka (1952)	984	543	586.22	1.7831	0.000731
Hrabal, B.	Modrý pokoj (1991)	1668	848	911.0045	1.7598	0.000516
Hrabal, B.	Modrý pondělí (1993)	2520	1240	1344.4094	1.8141	0.000369
Hrabal, B.	Pogrom (1991)	1839	942	1060.6844	1.8835	0.000304
Hrabal, B.	Pohádka o zlaté Praze (1990)	1448	793	840.7820	1.8358	0.000686
Hrabal, B.	Praha, město utajených infarktů (1991)	2004	985	1088.7907	1.7943	0.000278
Hrabal, B.	Protokol (1952)	999	556	626.98	1.8826	0.000481
Hrabal, B.	Únorová povídka (1952)	2858	1274	1432.06	1.7317	0.000311
Hrabal, B.	Veselé vánoce (1992)	1264	683	763.8220	1.8748	0.000381
Hrabal, B.	Zavražděný kohout (1993)	1435	710	787.4061	1.7313	0.000622
Jedlička, J.	Kde život náš je v půli se svou poutí (1966)	24666	9718	10945.2796	1.9489	0.000043
Karel, M.	Gypsová dáma (1967)	17654	6182	6702.1943	1.6123	0.000094
Klíma, L.	Sus triumfans (1920s)	10073	3885	4241.0792	1.6854	0.000080
Klíma, L.	Melia (1920s)	8264	3085	3371.0090	1.5979	0.000155
Körner, V.	Adelheid (1967)	24943	7539	8520.3150	1.5019	0.000037
Krupička, J.	Stará pevnost (2001)	29947	9802	10952.08849	1.6371	0.000037
Páral, V.	Veletrh splněných přání (1964)	21275	7632	8534.2035	1.736	0.000050
Pecka, K.	Pasáž (1995)	28758	9267	10209.57056	1.583	0.000038
Škvorecký, J.	Eva byla nahá (1996)	13106	5632	6071.8926	1.9076	0.000043

Správcová, B.	Spravedlnost (2000)	22041	7039	7816.0331	1.5402	0.000044
Uhde, M.	Modrý anděl (1979)	5459	1899	2073.5795	1.4198	0.000207
Vávra, V.	Muž v jiných končinách světa (1992)	23528	7613	8493.5334	1.5782	0.000038
Viewegh, M.	Názory na vraždu (1990)	25476	8954	9607.0179	1.6616	0.000046
Vodňanský, J.	Velký dračí propadák aneb Král v kukani (1997)	16145	5545	6048.5864	1.5766	0.000091
Weil, J.	Žalozpěv za 77 297 obětí (1958)	3565	1755	1898.9984	1.8921	0.000153

Table 4d: Czech scientific texts

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Fučíková, T.	Vnitřní lékařství: Imunologie (2002)	11816	3988	4267.771	1.4709	0.000078
Kořenský, J.	Člověk, řeč, poznání (2004)	23077	6453	7119.041	1.3460	0.000061
Mokrejš, A.	Hermeneutické pojetí zkušenosti (1996)	15344	3854	4617.261	1.2596	0.000080
Mráz, M.	Smyslové vnímání a čas v Aristotelově filosofii (1999)	7275	2184	2375.041	1.2608	0.000127
Pudil, R	Akutní koronární syndromy (2003)	16816	4161	4598.038	1.1554	0.000121
Racek, J	Volné radikály a antioxidanty v klinické praxi (2001)	11861	3629	3950.439	1.3569	0.000107
Rubeš, Z.	Logika pro humanitní obory (2002)	17768	4229	4877.204	1.1665	0.000066
Sousedík, S.	René Descartes a české baroko (1996)	11996	4224	4495.088	1.5285	0.000106
Viktorinová, M.	Kopřivka a angioedém (2001)	11911	3755	4094.19	1.4010	0.000088
Zbořil, V.	Kortikosteroidy v léčbě nespecifických střevních zánětů (2001)	17221	4828	5359.19	1.3183	0.000085

Table 4e. Czech prosaic texts written by children
School journals “Červotoč” and “Dušan”
http://sedmikraska.cz/dilna/cervotoc.php **and** http://zs.staravesno.indos.cz/dusan.htm

Author	**Grade**	**Gender**	**Text**	***N***	***V***	***L***	***Λ***	***Var(Λ)***
Hanka	2th	f	Hrála jsem si	16	12	12.6503	0.9520	0.002677
Barbora	2th	f	O Vánocích	22	20	20.2361	1.2348	0.002822
Eliška	2th	f	O zimních prázdninách	13	10	11.1623	0.9565	0.005874
Karolinka	2th	f	O prázdninách	26	21	22.5765	1.2287	0.001627
Natálie	2th	f	O prázdninách	21	19	18.4142	1.1594	0.005842
Adam	2th	m	O zimních prázdninách	19	17	16.4142	1.1047	0.006661
Jan	2th	m	Byl jsem bruslit	15	12	13.1623	1.0320	0.005030
Kuba	2th	m	O Vánocích	22	21	20.4142	1.2457	0.001769
Lukáš	2th	m	O zimních prázdninách	22	20	19.4142	1.1846	0.005492
Pepa	2th	m	O zimních prázdninách	30	26	25.4142	1.2513	0.003593
Ineska	3th	f	Za sto let	97	69	69.6569	1.4267	0.002305
Klára	3th	f	Život v budoucnosti	116	81	83.8191	1.4917	0.002233
Laura	3th	f	Dinosauři	83	70	70.2426	1.6241	0.001333
Alena	3th	f	Můj plyšový mazlíček	45	35	35.2426	1.2947	0.007407
Markéta	3th	f	Můj plyšový mazlíček	158	114	119.0160	1.6562	0.002325
Nikola	3th	f	Můj plyšový mazlíček	67	48	51.3658	1.4000	0.002370
Sára	3th	f	Můj plyšový mazlíček	100	79	80.4787	1.6096	0.003071
Tomáš	3th	m	Můj plyšový mazlíček	56	45	45.2426	1.4124	0.005351
David	3th	m	Raketa	113	78	81.3658	1.4783	0.003709
Janek	3th	m	Neuvěřitelný robot	76	71	70.8284	1.7528	0.001527

Metoděj	3th	m	Domorodý člověk	102	76	76.6568	1.5095	0.002519
Petr	3th	m	Superprášek	50	43	42.8284	1.4553	0.004569
Vojta	3th	m	Doba ledová	53	45	44.8284	1.4584	0.004191
Anička	4th	f	Minulost a budoucnost	269	169	171.7213	1.5511	0.001842
Kamila	4th	f	Budoucnost	124	90	94.1942	1.5902	0.000818
Míša	4th	f	Budoucnost	127	82	95.6410	1.5843	0.001545
Monika	4th	f	Budoucnost	70	58	60.5373	1.5957	0.002333
Vendy	4th	f	Minulost a budoucnost	66	55	58.5132	1.6131	0.002581
Adam	4th	m	Budoucnost	80	62	65.3658	1.5550	0.003245
Adam	4th	m	Sci-fi	212	135	137.3137	1.5068	0.001866
Kryštof	4th	m	Černé okénko	126	81	84.6410	1.4109	0.004394
Martin	4th	m	Lev	65	53	53.2426	1.4850	0.005777
Kateřina	5th	f	Andy	232	143	155.0696	1.5811	0.001253
Natálie	5th	f	Oksana	153	102	110.7122	1.5809	0.002014
Tereza	5th	f	Dmitro	169	96	102.1131	1.3461	0.004809
Klára	5th	f	Tisíc a jeden recept	168	121	122.8929	1.6278	0.001617
Klára	5th	f	Víte, že	141	106	108.4049	1.6524	0.002574
Markéta	5th	f	Pohroma na silnici	250	162	165.9574	1.5918	0.001386
Honza	5th	m	Jak chytit zloděje	169	108	113.5843	1.4974	0.002603
Florian	5th	m	Ráchel	109	85	86.0645	1.6087	0.003599
Andrea	6th	f	Cesta časem	153	95	108.1029	1.5436	0.003129
Irena	6th	f	Já a první žárovka	390	245	255.7156	1.6989	0.001015
Karolína	6th	f	Duch	274	187	190.6476	1.6962	0.001922
Veronika	6th	f	Zážitek ze soustředění	162	92	94.7214	1.2919	0.001114

Zuzana	6th	f	Moje procházka	159	111	116.9919	1.6198	0.002185
Patrik	6th	m	V budoucnosti aut	115	80	81.4787	1.4600	0.003429
Anežka	7th	f	Oheň	231	170	175.8379	1.7992	0.001375
Ela	7th	f	Výlet do ZOO	413	264	271.4947	1.7197	0.000828
Gabriela	7th	f	Podařená dovolená	188	127	134.9019	1.6319	0.002468
Míša	7th	f	Stanování	130	90	95.5280	1.5534	0.002050
Radek	7th	m	Zážitek z prázdnin	131	96	103.4877	1.6726	0.002894
Uhlobaron	7th	m	Boj o postup	102	84	85.0645	1.6751	0.003992
Eva	8th	f	Vánoce	167	123	128.0160	1.7039	0.001245
Lucka	8th	f	Procházka zimní krajinou	315	237	244.1783	1.9366	0.001055
Marfuška	8th	f	Výlet 8. tříd	325	225	230.8445	1.7842	0.001517
Kuba	8th	m	Silvestr	194	132	133.4853	1.5742	0.002155
Victor	8th	m	Steel drum	67	63	62.4142	1.7011	0.001109
Anežka	9th	f	Cizí kultura	293	185	187.7213	1.5805	0.001601
Daniela	9th	f	Myšlení pro budoucí svět	490	304	315.6167	1.7328	0.000477
Kateřina	9th	f	A nebylo dřív líp?	337	241	254.6929	1.9103	0.002149
Tereza	9th	f	Malý hudební výlet	169	127	128.0711	1.6883	0.001650
Aneta	9th	f	Co mi dala ZŠ	321	208	222.6892	1.7389	0.001723
Katka	9th	f	Jak nás škola připravila	280	184	187.5498	1.6392	0.001369
Kristýna	9th	f	Už od června	222	144	151.2280	1.5983	0.003447
Radek	9th	m	Kolektiv třídy	179	142	143.4787	1.8058	0.001693
Jan	9th	m	Cizí kultura	163	134	136.4049	1.8512	0.002041
Martin	9th	m	Cukrářka Mařena	992	500	527.4829	1.6961	0.000838
Vojtěch	9th	m	Spirituály	309	227	231.8771	1.8685	0.002226

Table 5: Dutch prose

Author	Text, year	*N*	*V*	*L*	*Λ*	*Var*(*Λ*)
Conscience, H.	Siska van Roosemael (1912)	14268	2204	2685.1083	0.7818	0.000107
Couperus, L.	Jan en Florence (1917)	22075	4779	5607.4437	1.1034	0.000074
Couperus, L.	Extaze (1892)	30743	3870	4688.8152	0.6845	0.000058
Couperus, L.	Reis-impressies (1894)	18998	4340	5473.5074	1.2327	0.000060
Reyneke van Stuwe, J.	Het vroolijke leven (the first half of the 20th century)	14841	3168	3599.0264	1.0116	0.000115

Table 6a: English poetry

Mostly from http://www.gutenberg.org/browse/languages/en

Author	Text, year	*N*	*V*	*L*	*Λ*	*Var*(*Λ*)
Browning, R.	Christmas-Eve and Easter-Day (1850)	9476	2777	3300.9452	1.3852	0.000151
Browning, R.	Dramatic Romances, 1st half (1845)	17318	4086	5028.8130	1.2308	0.000092
Browning, R.	Dramatic Romances, 2nd half (1845)	17367	4202	5324.6508	1.2999	0.000077
Browning, R.	Browning's Shorter Poems I (1899)	21103	4529	5676.5509	1.1632	0.000078
Browning, R.	Browning's Shorter Poems II (1899)	14582	3394	4165.1018	1.1893	0.000100
Byron, Lord	Poetical Works I. 1st half (1898)	27785	5084	6448.8631	1.0314	0.000073
Byron, Lord	Poetical Works I. 2nd half (1898)	27761	6047	7194.6282	1.1516	0.000064
Dos Passos, J.	A Pushcart at the Curb. Nights at Bassano (1922)	3838	1423	1780.7655	1.6630	0.000134
Dos Passos, J.	A Pushcart at the Curb. On foreign travel (1922)	1746	634	789.5402	1.4661	0.000350

Dos Passos, J.	A Pushcart at the Curb. Phases of the moon (1922)	2734	1063	1324.9526	1.6655	0.000332
Dos Passos, J.	A Pushcart at the Curb. Quai de la Tournelle (1922)	1796	782	933.7530	1.6919	0.000157
Dos Passos, J.	A Pushcart at the Curb. Vagones de tercera (1922)	2363	967	1207.1920	1.7234	0.000160
Dos Passos, J.	A Pushcart at the Curb. Winter in castle (1922)	4023	1472	1884.5598	1.6885	0.000125
Eliot, T.S.	Poems (1920)	6535	2419	2881.8693	1.6825	0.000168
Eliot, T.S.	The Waste Land (1922)	3036	1184	1367.8999	1.5690	0.000404
Ginsberg, A.	Howl (1956)	2925	1313	1511.0000	1.7905	0.000196
Hopkins, G.M.	Poems (1876 – 1889)	10939	3168	3582.8966	1.3229	0.000175
Poe, E.A.	Poems of later life (1845 – 1849)	5838	1606	2006.8899	1.2947	0.000185

Table 6b: English prose

Mostly from http://www.gutenberg.org/browse/languages/en

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Ackroyd, P.	Hawksmoor. Part I. 1st half (1985)	25764	4172	5676.0000	0.9717	0.000075
Ackroyd, P.	Hawksmoor. Part I. 2nd half (1985)	25514	4138	5485.0000	0.9473	0.000078
Ackroyd, P.	Hawksmoor. Part II. 1st half (1985)	25758	4114	5571.0000	0.9540	0.000069
Ackroyd, P.	Hawksmoor. Part II. 2nd half (1985)	25699	3904	5301.0000	0.9097	0.000058
Amis, M.	Time’s Arrow. Part 1 (1991)	27880	5105	6643.0000	1.0592	0.000062
Amis, M.	Time’s Arrow. Parts 2 and 3 (1991)	19427	4519	5691.0000	1.2562	0.000056
Barnes, J.	Flaubert's Parrot 1st third (1984)	21046	4552	5699.5647	1.1708	0.000068
Barnes, J.	Flaubert's Parrot 2nd third (1984)	21074	4488	5625.1861	1.1541	0.000084
Barnes, J.	Flaubert's Parrot 3rd third (1984)	21075	4145	5148.4002	1.0563	0.000081

Bradbury, M.	The history man. 1st third (1975)	32314	4810	6812.3029	0.9507	0.000037
Byatt, A.S.	Possession. 1st tenth (1990)	19473	4363	5125.0000	1.1289	0.000083
Byatt, A.S.	Possession. 2nd tenth (1990)	19541	4285	5197.0000	1.1411	0.000107
Byatt, A.S.	Possession. 3rd tenth (1990)	19431	4228	5069.0000	1.1187	0.000076
Byatt, A.S.	Possession. 4th tenth (1990)	18918	3679	4379.0000	0.9901	0.000061
Byatt, A.S.	Possession. 5th tenth (1990)	19835	4573	5426.0000	1.1757	0.000087
Byatt, A.S.	Possession. 6th tenth (1990)	19406	4145	4938.0000	1.0911	0.000077
Byatt, A.S.	Possession. 7th tenth (1990)	19429	3773	4584.0000	1.0117	0.000089
Byatt, A.S.	Possession. 8th tenth (1990)	19403	3691	4528.0000	1.0007	0.000048
Byatt, A.S.	Possession. 9th tenth (1990)	19355	3925	4692.0000	1.0392	0.000094
Byatt, A.S.	Possession. 10th tenth (1990)	18817	3765	4712.0000	1.0703	0.000092
Dos Passos, J.	One Man's Initiation. I (1920)	1680	667	771,3743	1,4809	0,000325
Dos Passos, J.	One Man's Initiation. II (1920)	1040	486	571,1291	1,6568	0,000548
Dos Passos, J.	One Man's Initiation. III (1920)	2043	778	961,3707	1,5577	0,000345
Dos Passos, J.	One Man's Initiation. IV (1920)	4423	1339	1706,1992	1,4064	0,000253
Dos Passos, J.	One Man's Initiation. V (1920)	3394	1159	1402,6038	1,4591	0,000295
Dos Passos, J.	One Man's Initiation. VI (1920)	5006	1388	1824,5594	1,3484	0,000237
Dos Passos, J.	One Man's Initiation.VII (1920)	5565	1488	1855,1828	1,2486	0,000342
Dos Passos, J.	One Man's Initiation.VIII (1920)	307	179	194,7869	1,5781	0,001721
Dos Passos, J.	One Man's Initiation. IX (1920)	4010	1192	1487,0521	1,3362	0,000251
Dos Passos, J.	One Man's Initiation. X (1920)	395	206	231,5652	1,5222	0,001445
Dos Passos, J.	One Man's Initiation. XI (1920)	181	127	136,1265	1,6980	0,002861

Dos Passos, J.	Rosinante to the road again. I (1920 – 1922)	2376	947	1067.3320	1.5165	0.000308
Dos Passos, J.	Rosinante to the road again. II((1920 – 1922)	4141	1397	1696.0692	1.4815	0.000195
Dos Passos, J.	Rosinante to the road again. III (1920 – 1922)	4459	1500	1894.4124	1.5504	0.000189
Dos Passos, J.	Rosinante to the road again.IV (1920 – 1922)	1498	630	717.2578	1.5205	0.000753
Dos Passos, J.	Rosinante to the road again.V (1920 – 1922)	3808	1418	1693.7422	1.5926	0.000166
Dos Passos, J.	Rosinante to the road again. VI (1920 – 1922)	362	205	227.2881	1.6065	0.001584
Dos Passos, J.	Rosinante to the road again. VII (1920 – 1922)	1934	799	967.1798	1.6435	0.000399
Dos Passos, J.	Rosinante to the road again. VIII (1920 – 1922)	784	424	446.1811	1.6472	0.000572
Dos Passos, J.	Rosinante to the road again. IX (1920 – 1922)	2276	925	1076.1253	1.5873	0.000235
Dos Passos, J.	Rosinante to the road again. X (1920 – 1922)	1130	509	560.2528	1.5137	0.000684
Dos Passos, J.	Rosinante to the road again. XI (1920 – 1922)	2958	1168	1422.6268	1.6693	0.000161
Dos Passos, J.	Rosinante to the road again. XII (1920 – 1922)	2796	1068	1319.9250	1.6270	0.000461
Dos Passos, J.	Rosinante to the road again.XIII (1920 – 1922)	1033	478	544.7293	1.5894	0.000637
Dos Passos, J.	Rosinante to the road again. XIV (1920 – 1922)	2610	1138	1350.5169	1.7679	0.000177
Dos Passos, J.	Rosinante to the road again. XV (1920 – 1922)	991	479	532.3813	1.6095	0.000563
Dos Passos, J.	Rosinante to the road again. XVI (1920 – 1922)	5475	1788	2181.2887	1.4894	0.000224
Dos Passos, J.	Rosinante to the road again. XVII (1920–1922)	2662	979	1119.5872	1.4406	0.000322
Milton, J.	Paradise Lost. Books I to III (1667)	19574	4329	5147.0000	1.1286	0.000071
Milton, J.	Paradise Lost. Books IV to VI (1667)	21438	4386	5175.0000	1.0455	0.000077
Milton, J.	Paradise Lost. Books VII to IX (1667)	18765	3886	4631.0000	1.0545	0.000062
Milton, J.	Paradise Lost. Books X to XII (1667)	20219	4236	5026.0000	1.0703	0.000109
Wells, H.G.	The Invisible Man. 1st half (1897)	24431	3885	5486.0000	0.9853	0.000073

Wells, H.G.	The Invisible Man. 2nd half (1897)	24454	3811	5366.0000	0.9630	0.000087
Yeats, W.B.	The Celtic Twilight. 1st half (1893)	19877	3105	4225.0000	0.9137	0.000069
Yeats, W.B.	The Celtic Twilight. 2nd half (1893)	19868	2694	3983.0000	0.8618	0.000096

Table 6c: English Nobel lectures

From http://nobelprize.org/nobel_prizes/literature/laureates/

Author	**Text, year**	***N***	***V***	***L***	***Λ***	**_Var_(_Λ_)**
Buchanan Jr., J.M.	Nobel Prize for Economy (lecture) (1986)	4622	1232	1567.3100	1.2427	0.000179
Bellow, S.	Nobel Prize for Literature (lecture) (1976)	4760	1495	1760.8600	1.3605	0.000238
Yeats, W.B.	Nobel Prize for Literature (lecture) (1923)	3806	1115	1291.7841	1.2152	0.000255
Lewis, S.	Nobel Prize for Literature (lecture) (1930)	5004	1597	1800.7000	1.3312	0.000212
O'Neill, E.	Nobel Prize for Literature (banquet speech) (1936)	536	236	255.0642	1.2987	0.001581
Buck, P.	Nobel Prize for Literature (lecture) (1938)	9082	1821	2392.2778	1.0426	0.000352
Eliot, T.S.	Nobel Prize for Literature (banquet speech) (1948)	1304	491	563.3743	1.3459	0.000595
Faulkner, W.	Nobel Prize for Literature (banquet speech) (1949)	553	238	263.2980	1.3059	0.002882
Russell, B.	Nobel Prize for Literature (lecture) (1950)	5690	1571	1875.8123	1.2379	0.000222
Churchill, W.	Nobel Prize for Literature (banquet speech) (1953)	426	248	261.2670	1.6126	0.000995
Hemingway, E.	Nobel Prize for Literature (banquet speech) (1954)	332	175	179.3782	1.3622	0.002241
Steinbeck, J.	Nobel Prize for Literature (banquet speech) (1962)	962	428	485.6392	1.5060	0.000446
Golding , W.	Nobel Prize for Literature (lecture) (1983)	4501	1369	1584.9300	1.2864	0.000201
Gordimer, N.	Nobel Prize for Literature (lecture) (1991)	3772	1261	1494.3228	1.4169	0.000271

Walcott , D.	Nobel Prize for Literature (lecture) (1992)	6331	1956	2445.6682	1.4685	0.000146
Morrison, T.	Nobel Prize for Literature (lecture) (1993)	2971	1017	1157.2155	1.3527	0.000414
Pinter, H.	Nobel Prize for Literature (lecture) (2005)	4921	1490	1759.9814	1.3205	0.000336
Lessing, D.	Nobel Prize for Literature (lecture) (2007)	4974	1241	1452.1722	1.0793	0.000251
Marshall, G.C.	Nobel Prize for Peace (lecture) (1953)	3247	1001	1204.9100	1.3031	0.000368
Carter, J.	Nobel Prize for Peace (lecture) (2002)	2330	939	1042.8500	1.5071	0.000279
Feynman, R.P.	Nobel Prize for Physics (lecture) (1965)	11126	1662	2391.6931	0.8698	0.000179

Table 6d: English scientific texts

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Bernstein, R.J.	Rorty’s Inspirational Liberalism (2003)	6673	1853	2165.708	1.2412	0.000146
Bybee, J.	From usage to grammar: The mind’s response to repetition (2006)	9533	1722	2215.870	0.9249	0.000120
Gutting, G.	Rorty’s Critique of Epistemology (2003)	7928	1701	2124.218	1.0447	0.000215
Hudson, R.	Why education needs linguistics (and vice versa) (2004)	11418	2197	2734.978	0.9719	0.000108
Chomsky, N.	Three Factors in Language Design (2005)	10364	2263	2879.978	1.1158	0.000080
MacWhinney, B.	The Emergence of Grammar from Perspective (2005)	9574	1986	2722.391	1.1320	0.000094
Rouse, J.	From Realism or Antirealism to Science as Solidarity (2003)	8812	1972	2419.967	1.0834	0.000152
Schoenemann, T.	Syntax as an Emergent Characteristic of the Evolution of Semantic Complexity (1999)	17531	2579	3380.855	0.8184	0.000090
Taylor, Ch.	Rorty and philosophy (2003)	10302	2173	2689.938	1.0478	0.000111
Williams, M.	Rorty on Knowledge and Truth (2003	7990	1943	2349.956	1.1478	0.000097

Table 6e: English stories told or written by children

http://www.goodnightstories.com/stories.htm (2010) and http://www.kids-space.org/ (2010)

Author	Age	Gender	Text	N	V	L	Λ	$Var(\Lambda)$
Bernice	6	f	Fairies school	266	147	152.2978	1.3884	0.001852
Emily	6	f	My Friend Jeffrey	112	70	71.8995	1.3155	0.003848
Eva	6	f	The little Fish	180	91	98.8041	1.2379	0.004023
Leon	6	m	The dragon	71	42	48.8965	1.2749	0.006820
Tomin	6	m	Dream	109	56	61.1701	1.1434	0.006363
Anna	7	f	The Fairies	123	87	88.0711	1.4964	0.002741
Natalia	7	f	The beautiful female frog	136	82	85.2334	1.3371	0.004339
Olivia	7	f	The owls that are red	134	79	87.8902	1.3952	0.004150
Jordan	7	m	My Unicorn	39	27	28.0645	1.1449	0.010573
Banner	7	m	Sam the Robo	75	46	46.6569	1.1665	0.005933
Emma	8	f	Pat's Sleepover	145	79	81.3071	1.2120	0.006331
Sophia	8	f	The flying horse	51	37	38.9907	1.3055	0.007460
Ibrahim	8	m	The Sneaky Mouse	211	86	102.0259	1.1239	0.004776
Ibrahim	8	m	The Three Cats	281	111	120.1450	1.0470	0.001135
Steven	8	m	My first time visiting space	212	105	110.7055	1.2148	0.004038
Lorelei	9	f	The giraffe story	241	122	125.1421	1.2369	0.002296
Lorelei	9	f	The Life Story Of Bea The Cran	258	128	135.0459	1.2623	0.003318
Michaela	9	f	The Ants First Home	142	86	87.4853	1.3260	0.003330
Colin	9	m	Cars	125	65	68.1356	1.1430	0.001686
Jack	9	m	The Lost Giraffe	104	65	69.0552	1.3393	0.002126

Kate	10	f	Destroer	91	57	60.9275	1.3116	0.003178
Micaela	10	f	Fairyland	1941	509	583.1789	0.9879	0.000611
Olivia	10	f	The_magic_bag	221	101	104.5563	1.1091	0.002307
Jordan	10	m	The Wolf Who Could fly	148	86	88.7214	1.3010	0.003164
Raven	10	m	The big whale	84	57	58.4787	1.3396	0.006769
Bethany	11	f	The prince and princess	418	176	196.4680	1.2320	0.002410
Kaia	11	f	The hungry bear	187	108	114.4062	1.3899	0.001479
Maria	11	f	Dark fairy	185	92	98.9823	1.2130	0.004841
Toni	11	m	The Rift	339	193	201.2885	1.5024	0.001310
Trevor	11	m	Flyind in the Sky	81	57	57.6569	1.3585	0.005270
Lizzie	12	f	I Won't Grow Up	357	165	172.9090	1.2364	0.001610
Lyndsay	12	f	Fantasy	328	152	160.8352	1.2337	0.003570
Lyndsay	12	f	The Jensons New Year	1538	457	532.4480	1.1033	0.000746
Graham	12	m	Kielbasa	408	172	205.5659	1.3154	0.002758
Christy	13	f	The Pear	1367	451	534.2903	1.2256	0.000812
Jammie	13	f	Long Wait	243	136	146.8503	1.4417	0.002788
Jesse	13	f	The Light Festival in which Orli was conquered by the Light Angel	945	327	379.8462	1.1960	0.000881
Stacey	13	m	The Vain Rabbits and the Cunning Bear	360	170	179.0468	1.2714	0.003206
Vincent	13	m	The Unicorn's battle	171	101	111.7702	1.4595	0.003470

Table 7: Finnish prose

http://www.gutenberg.org/browse/languages/fi

Author	**Text, year**	N	V	L	Λ	$Var(\Lambda)$
Aho, J.	Omatunto (1914)	31024	8803	9937.9244	1.4388	0.000041
Aho, J.	Juha (1911)	32773	9295	10320.3612	1.4220	0.000041
Pakkala, T.	Lapsuuteni muistoja (1895)	15266	6301	6839.8739	1.8745	0.000055
Pakkala, T.	Lapsuuteni muistoja (1885)	20832	7259	8246.0399	1.7095	0.000061
Pakkala, T.	Lapsia (1895)	18023	5486	6341.2497	1.4974	0.000060

Table 8a: French poetry

Mostly from http://www.gutenberg.org/browse/languages/fr

Author	**Text, year**	N	V	L	Λ	$Var(\Lambda)$
Apollinaire, G.	Alcools (1898 - 1912)	15688	4471	4962.9520	1.3273	0.000090
Baudelaire, Ch.	Les Fleurs du Mal (1857)	23025	5799	6633.5079	1.2568	0.000065
Boileau, N.	Le Lutrin (1674)	9849	2841	3111.4290	1.2616	0.000132
Fontaine,. J. de La	Fables (1668)	9372	2765	3033.9196	1.2858	0.000133
Prudhomme, S.	Les vaines tendresses (1875)	10624	3161	3403.5041	1.2899	0.000198
Rimbaud, A.	Poésies complètes (1895)	12978	3874	4243.7846	1.3450	0.000107
Verlaine, P.	Oeuvres Complètes, Vol.1. 1st half (1902)	21494	5369	6312.1518	1.2723	0.000061
Verlaine, P.	Oeuvres Complètes, Vol.1. 2nd half (1902)	21434	5148	6012.1951	1.2149	0.000085

Table 8b: French prose

Mostly from http://www.gutenberg.org/browse/languages/fr

Author	**Text, year**	N	V	L	Λ	$Var(\Lambda)$
Balzac, H. de	Eugenie Grandet 1st third (1833)	21354	4798	5761.3524	1.1681	0.000078
Balzac, H. de	Eugenie Grandet 2nd third (1833)	21203	4379	5155.0616	1.0519	0.000098
Balzac, H. de	Eugenie Grandet 3rd third (1833)	21305	4135	4968.8188	1.0095	0.000126
Flaubert, G.	Un coeur simple (1877)	11117	3313	3723.1635	1.3550	0.000133
Flaubert, G.	Hérodias (1877)	9596	3208	3567.6454	1.4805	0.000105
Hugo, V.	Les misérables, I. Livre 1 (1862)	22128	4754	5640.5563	1.1076	0.000062
Hugo, V.	Les misérables. I. Livre 2 (1862)	21464	4016	4798.7695	0.9685	0.000111
Hugo, V.	Les misérables. I. Livres 3. 4. 5 (1862)	30983	6444	7717.3690	1.1187	0.000046
Hugo, V.	Les misérables. I. Livres 6. 7 (1862)	30501	5035	5984.9564	0.8799	0.000065
Hugo, V.	Les misérables. I. Livre 8 (1862)	6821	1764	1994.0454	1.1208	0.000242
Maupassant, G. de	Boule de Suif.1st half (1880)	18292	4580	5272.3026	1.2285	0.000087
Maupassant, G. de	Boule de Suif.2nd half (1880)	18168	4169	4735.9729	1.1103	0.000093
Musset, A. de	Oeuvres Complètes. 7. Merle blanc (1842)	9303	2598	2903.8387	1.2388	0.000154
Musset, A. de	Oeuvres Complètes. 7. Mimi Pinson (1845)	10540	2681	3065.8756	1.1702	0.000152
Musset, A. de	Oeuvres Complètes. 7. La mouche (1853)	11459	2698	3112.3533	1.1025	0.000173
Musset, A. de	Oeuvres Complètes. 7. Croisilles (1839)	10328	2410	2789.6109	1.0842	0.000193
Musset, A. de	Oeuvres Complètes. 7. Pierre et Camille (1844)	15602	3183	3760.2337	1.0106	0.000097
Prévost, Abbé	Manon Lescaut. Part 1 (1753)	31412	4773	6399.6731	0.9162	0.000057
Prévost, Abbé	Manon Lescaut. Part 2 (1753)	25910	4365	5634.9308	0.9598	0.000080

Table 9a: German poetry

Mostly from http://www.gutenberg.org/browse/languages/de

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Arp, H.	Sekundenzeiger (2010) http://www.balloon-painting.de/arp.htm	78	25	39.7103	0.9633	0.018363
Droste-Huelshoff, A.	Das Fegefeuer (1842)	702	425	444.0768	1.8049	0.000604
Droste-Huelshoff, A.	Der Fundator (1842)	791	408	429.7014	1.5829	0.001150
Droste-Huelshoff, A.	Der Geierpfiff (1842)	965	509	534.5509	1.6532	0.000841
Droste-Huelshoff, A.	Der Tod des Erzbischofs Engelbert (1842)	889	492	528.0074	1.7481	0.000634
Droste-Huelshoff, A.	Die Schwestern (1842)	1286	657	692.0214	1.6867	0.000492
Goethe, J.W.v	Der Erlkönig (1782)	225	124	127.9574	1.3377	0.003839
Goethe, J.W.v	Der Gott und die Bajadere (1797)	559	332	351.4129	1.7271	0.000713
Goethe, J.W.v	Elegie 2 (1788-1790)	251	251	175.4127	1.677	0.001566
Goethe, J.W.v	Elegie 5 (1788-1790)	184	184	132.5432	1.6314	0.002501
Goethe, J.W.v	Elegie 13 (1788-1790)	480	480	309.8352	1.7307	0.001221
Goethe, J.W.v	Elegie 15 (1788-1790)	468	468	306.8045	1.7505	0.000952
Goethe, J.W.v	Elegie 19 (1788-1790)	653	653	398.4337	1.7175	0.001298
Heine, H.	Belsazar (1815)	263	169	178.7186	1.6445	0.001261
Heine, H.	Die Heimkehr - Die Wallfahrt nach Kevlaar (1823–1824)	394	211	222.1861	1.4637	0.000998
Heine, H.	Die Heimkehr – Götterdämmerung (1823 – 1824)	603	361	400.2184	1.8453	0.001099
Moericke. E.	Peregrina (Aus: Maler Nolten 2010) http://www.thokra.de/html/morike_4.html#p91	593	378	385.0871	1.8008	0.001539

Rückert. F.	Amor ein Besenbinder (2010) http://www.thokra.de/html/ruckert.html	327	202	204.7213	1.5743	0.002217
Rückert. F.	Barbarossa (idem)	141	97	101.6018	1.5487	0.002018
Rückert. F.	Der Frost (idem)	152	107	109.7148	1.5749	0.002202
Rückert. F.	Die goldne Hochzeit (idem)	721	412	422.8997	1.6763	0.000919
Rückert. F.	Erscheinung der Schnitterengel (idem)	212	145	149.1942	1.6371	0.002896
Schiller. F.	Das Lied von der Glocke (1799)	2063	1029	1106.219	1.7773	0.001566
Schiller. F.	Winternacht (1796)	294	216	218.7213	1.8363	0.001263
Schwitters. K.	Die Puppen (1944)	75	31	35.8771	0.8970	0.013904

Table 9b. German prose

Mostly from http://www.gutenberg.org/browse/languages/de

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Arnim, L.A.	Der tolle Invalide auf dem Fort Ratonneau (1818)	7846	2221	2448.1279	1.2152	0.000142
Arnim, L.A.	Des ersten Bergmanns ewige Jugend (1810)	1201	564	594.5277	1.5245	0.000835
Arnim, L.A.	Frau von Saverne (1817)	4167	1429	1588.3089	1.3797	0.000291
Busch, W.	Eduards Traum (1891)	15820	4642	5111.7925	1.3569	0.000123
Chamisso, A.	Peter Schlemihls wundersame Geschichte I (1813)	2210	884	944.0268	1.4286	0.000603
Chamisso, A.	Peter Schlemihls wundersame Geschichte II (1813)	1847	808	871.9499	1.5421	0.000526
Chamisso, A.	Peter Schlemihls wundersame Geschichte III (1813)	1428	630	683.8122	1.5107	0.000751
Chamisso, A.	Peter Schlemihls wundersame Geschichte IV (1813)	3205	1209	1304.8611	1.4273	0.000270
Chamisso, A.	Peter Schlemihls wundersame Geschichte V (1813)	2108	853	910.8642	1.4362	0.000780
Chamisso, A.	Peter Schlemihls wundersame Geschichte VI (1813)	1948	801	853.3242	1.4410	0.000318
Chamisso, A.	Peter Schlemihls wundersame Geschichte VII (1813)	1362	670	697.6565	1.6054	0.000742

Chamisso, A.	Peter Schlemihls wundersame Geschichte VIII (1813)	1870	788	847.9470	1.4836	0.000552
Chamisso, A.	Peter Schlemihls wundersame Geschichte IX (1813)	1320	593	672.5900	1.5901	0.000887
Chamisso, A.	Peter Schlemihls wundersame Geschichte X (1813)	1012	536	574.5767	1.7062	0.000671
Chamisso, A.	Peter Schlemihls wundersame Geschichte XI (1813)	1386	656	705.0680	1.5982	0.000667
Droste-Huelshoff, A.	Die Judenbuche (1842)	16172	4064	4527.6626	1.1783	0.000108
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 1 (1826)	3080	1079	1227.5503	1.3904	0.000171
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 2 (1826)	4100	1287	1466.4345	1.2922	0.000325
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 3 (1826)	4342	1334	1482.2452	1.2418	0.000273
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 4 (1826)	1781	739	799.2915	1.4589	0.000609
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 5 (1826)	1680	699	749.9189	1.4397	0.001009
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 6 (1826)	3223	1059	1162.5893	1.2655	0.000483
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 7 (1826)	2594	932	1031.0925	1.3570	0.000710
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 8 (1826)	3987	1320	1446.7247	1.3065	0.000218
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 9 (1826)	3285	1185	1315.3763	1.4081	0.000412
Eichendorff, J.F.	Aus dem Leben eines Taugenichts 10 (1826)	3052	1073	1177.7687	1.3447	0.000360
Goethe, J.W.v	Reineke Fuchs. 1st half (1794)	19556	3948	4802.8200	1.0539	0.000078
Goethe, J.W.v	Reineke Fuchs. 2nd half (1794)	22093	4280	5312.1731	1.0446	0.000049
Goethe, J.W.v	Die neue Melusine (1807 – 1808)	7554	2222	2502.0181	1.2845	0.000217
Heine, H.	Die Harzreise (1824)	19522	5769	6648.1871	1.4611	0.000061
Heine, H.	Ideen. Das Buch Le Grand (1827)	20107	5305	6192.1506	1.3253	0.000083
Hoffmann, E.T.A.	Der Sandmann - Clara an Nathanael (1817)	1076	534	548.7188	1.5461	0.000693
Hoffmann, E.T.A.	Der Sandmann - Nathanael an Lothar (1817)	2974	1177	1247.4645	1.4569	0.000454
Hoffmann, E.T.A.	Der Sandmann - Nathanael an Lothar (1817)	8163	2511	2759.1983	1.3223	0.000182
Immermann, K.L.	Der Karneval und die Somnabule (1830)	28943	6397	7233.6006	1.1151	0.000051
Kafka, F.	Betrachtung - Das Gassenfenster (1913)	111	80	86.1473	1.5874	0.002177
Kafka, F.	Betrachtung - Das Unglück des Junggesellen (1913)	139	102	105.7800	1.6308	0.001382
Kafka, F.	Betrachtung - Der Ausflug ins Gebirge (1913)	132	89	93.1876	1.4971	0.003893

Kafka, F.	Betrachtung - Der Fahrgast (1913)	232	150	152.7213	1.5572	0.003075
Kafka, F.	Betrachtung - Der Kaufmann (1913)	596	343	357.8019	1.6661	0.001603
Kafka, F.	Betrachtung - Der Nachhauseweg (1913)	151	104	107.1290	1.5459	0.002495
Kafka, F.	Betrachtung - Der plötzliche Spaziergang (1913)	247	166	172.6663	1.6726	0.001593
Kafka, F.	Betrachtung - Die Abweisung (1913)	189	136	137.8929	1.6609	0.003117
Kafka, F.	Betrachtung - Die Bäume (1913)	41	33	32.8284	1.2913	0.003848
Kafka, F.	Betrachtung - Die Vorüberlaufenden (1913)	160	101	103.7213	1.4288	0.003083
Kafka, F.	Betrachtung - Entlarvung eines Bauernfängers (1913)	625	321	332.4679	1.4873	0.001146
Kafka, F.	Betrachtung - Entschlüsse (1913)	178	137	138.0711	1.7456	0.002317
Kafka, F.	Betrachtung - Kinder auf der Landstraße (1913)	1072	513	532.2830	1.5046	0.001110
Kafka, F.	Betrachtung - Kleider (1913)	142	104	110.8907	1.6808	0.001593
Kafka, F.	Betrachtung - Unglücklichsein (1913)	1402	539	596.3584	1.3385	0.000978
Kafka, F.	Betrachtung - Wunsch. Indianer zu werden (1913)	61	48	47.8284	1.3998	0.003397
Kafka, F.	Betrachtung - Zerstreutes Hinausschaun (1913)	86	62	62.2426	1.4001	0.003767
Kafka, F.	Betrachtung - Zum Nachdenken (1913)	255	177	180.5432	1.7039	0.002295
Kafka, F.	Das Schweigen der Sirenen (1917)	428	240	245.7121	1.5107	0.001642
Kafka, F.	Der Geier (1920)	255	153	158.7055	1.4978	0.002041
Kafka, F.	Die Sorge des Hausvaters (1919)	470	272	276.8902	1.5742	0.001707
Kafka, F.	Ein Bericht für eine Akademie (1917)	3181	1210	1342.6652	1.4784	0.000352
Kafka, F.	Ein Brudermord (1917)	610	364	371.3689	1.6957	0.001046
Kafka, F.	Ein Hungerkünstler (1922)	3414	1214	1289.6148	1.3347	0.000414
Kafka, F.	Ein Landarzt (1918)	2129	887	956.4639	1.4952	0.000697
Kafka, F.	In der Strafkolonie (1919)	10256	2321	2717.1024	1.0626	0.000137
Kafka, F.	Nachts (1920)	134	98	100.4049	1.5938	0.002791
Kafka, F.	Vor dem Gesetz (1915)	584	276	289.9742	1.3736	0.001269
Keller, G.	Das Tanzlegendchen (1871)	1896	897	86.1473	1.5874	0.000380
Keller, G.	Romeo und Julia auf dem Dorfe (1856)	25625	5516	105.7800	1.6308	0.000064

Keller, G.	Spiegel. das Kätzchen (1855)	13149	3512	93.1876	1.4971	0.000116
Keller, G.	Vom Fichtenbaum (about 1860)	301	196	152.7213	1.5572	0.001133
Krummacher, F.A.	Das Krokodil (Fabel) http://gutenberg.spiegel.de/?id=12&xid=625&kapitel=36&cHash=8cc889f2ec2 (2010)	499	280	303.2544	1.6397	0.001661
Lessing, G.E.	Der Besitzer des Bogens (1759)	114	78	79.8929	1.4415	0.004375
Lessing, G.E.	Der Esel mit dem Löwen (1759)	61	48	48.2426	1.4120	0.002134
Lessing, G.E.	Der Fuchs (1759)	47	41	40.4142	1.4378	0.001884
Lessing, G.E.	Der Knabe und die Schlange (1759)	231	161	164.6410	1.6846	0.001163
Lessing, G.E.	Der Rangstreit der Tiere (1759)	327	193	209.7546	1.6130	0.001551
Lessing, G.E.	Die Erscheinung (1759)	208	141	148.4877	1.6548	0.003740
Lessing, G.E.	Die Furien (1759)	182	120	121.4853	1.5086	0.001989
Lessing, G.E.	Jupiter und das Schaf (1759)	362	227	231.7858	1.6383	0.001615
Lessing, G.E.	Minerva (1759)	74	64	64.6503	1.6331	0.001967
Lessing, G.E.	Zeus und das Pferd (1759)	254	154	158.8836	1.5043	0.001860
Lichtwer, M.G.	Die Rehe (Fabel) http://gutenberg.spiegel.de/?id=5&xid=625&kapitel=49&cHash=8cc889f2ec2#gb_found (2010)	518	292	299.2412	1.568	0.001415
Löns, H.	Der Werwolf - 1. Die Haidbauern (1910)	1672	706	781.8373	1.5072	0.000619
Löns, H.	Der Werwolf - 2. Die Mansfelder (1910)	2988	928	1041.7551	1.2117	0.000447
Löns, H.	Der Werwolf - 3. Die Braunschweiger (1910)	4063	1162	1303.2939	1.1576	0.000457
Löns, H.	Der Werwolf - 4. Die Weimaraner (1910)	3713	1081	1217.5676	1.1706	0.000454
Löns, H.	Der Werwolf - 5. Die Marodebruede (1910)	4676	1235	1457.4724	1.1439	0.000460
Löns, H.	Der Werwolf - 6. Die Bruchbauern (1910)	4833	1364	1572.5177	1.1987	0.000361
Löns, H.	Der Werwolf - 7. Die Wehrwoelfe (1910)	7743	1862	2231.6345	1.1208	0.000186
Löns, H.	Der Werwolf - 8. Die Schnitter (1910)	6093	1724	2015.0224	1.2517	0.000258
Löns, H.	Der Werwolf - 9. Die Kirchenleute (1910)	9252	2126	2530.6862	1.0849	0.000165

Löns, H.	Der Werwolf - 10. Die Hochzeiter (1910)	6546	1736	1968.2937	1.1474	0.000165
Löns, H.	Der Werwolf - 11. Die Kaiserlichen (1910)	4102	1294	1481.1298	1.3046	0.000340
Löns, H.	Der Werwolf - 12. Die Schweden (1910)	4432	1318	1506.6884	1.2397	0.000248
Löns, H.	Der Werwolf - 13. Die Haidbauern (1910)	1361	556	599.5646	1.3806	0.000883
Meyer, C.F.	Der Schuss von der Kanzel 1 (1877)	1523	801	840.4087	1.7563	0.000536
Meyer, C.F.	Der Schuss von der Kanzel 2 (1877)	573	331	346.8974	1.6698	0.001094
Meyer, C.F.	Der Schuss von der Kanzel 3 (1877)	1052	551	582.8191	1.6742	0.000725
Meyer, C.F.	Der Schuss von der Kanzel 4 (1877)	2550	1142	1197.3296	1.5995	0.000438
Meyer, C.F.	Der Schuss von der Kanzel 5 (1877)	1292	658	690.3727	1.6625	0.000728
Meyer, C.F.	Der Schuss von der Kanzel 6 (1877)	833	471	491.7594	1.7242	0.000756
Meyer, C.F.	Der Schuss von der Kanzel 7 (1877)	1229	652	682.7688	1.7164	0.000352
Meyer, C.F.	Der Schuss von der Kanzel 8 (1877)	1028	556	584.7997	1.7134	0.000590
Meyer, C.F.	Der Schuss von der Kanzel 9 (1877)	776	441	470.8607	1.7535	0.001455
Meyer, C.F.	Der Schuss von der Kanzel 10 (1877)	940	493	519.5186	1.6432	0.000721
Meyer, C.F.	Der Schuss von der Kanzel 11 (1877)	2398	1079	1145.9911	1.6152	0.000337
Novalis, H.O.	Heinrich von Ofterdingen - Die Erfuellung (1802)	5367	1939	2144.3593	1.4902	0.000237
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 1 (1802)	2894	1129	1243.2165	1.4870	0.000391
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 2 (1802)	3719	1487	1668.7753	1.6021	0.000264
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 3 (1802)	5321	1819	2018.2224	1.4132	0.000226
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 4 (1802)	2777	1282	1388.6005	1.7219	0.000278
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 5 (1802)	8866	2769	3197.8734	1.4239	0.000138
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 6 (1802)	4030	1467	1617.2371	1.4468	0.000333
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 7 (1802)	1744	792	850.9498	1.5816	0.000626
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 8 (1802)	2111	816	869.3676	1.3691	0.000574
Novalis, H.O.	Heinrich von Ofterdingen - Die Erwartung 9 (1802)	8945	2681	3081.9564	1.3615	0.000139
Novalis, H.O.	Hyazinth und Rosenblütchen (1802)	1358	646	713.8339	1.6468	0.000569
Novalis, H.O.	Neue Fragmente - Sophie (1802)	4430	1697	1860.9322	1.5318	0.000175

Novalis, H.O.	Neue Fragmente - Traktat vom Licht (1802)	1080	514	556.7288	1.5637	0.000774
Paul, J.	1. Dr. Katzenbergers Badereise (1809)	854	487	512.3110	1.7586	0.000619
Paul, J.	2. Reisezwecke (1809)	383	255	260.7121	1.7584	0.000489
Paul, J.	3. Ein Reisegefaehrte (1809)	520	311	326.3134	1.7044	0.000882
Paul, J.	4. Bona (1809)	580	354	365.6117	1.7420	0.001032
Paul, J.	5. Herr von Niess (1809)	1331	677	705.3279	1.6556	0.000937
Paul, J.	6. Fortsetzung der Abreise (1809)	526	305	312.5991	1.6171	0.001605
Paul, J.	7. Fortgesetzte Fortsetzung der Abreise (1809)	508	316	322.6383	1.7185	0.000965
Paul, J.	8. Beschluss der Abreise (1809)	402	248	262.0725	1.6978	0.000888
Paul, J.	9. Halbtagfahrt nach St. Wolfgang (1809)	1068	547	570.4589	1.6177	0.000624
Paul, J.	10. Mittags-Abenteuer (1809)	1558	778	814.0288	1.6681	0.000489
Paul, J.	11. Wagen-Sieste (1809)	2232	1027	1092.4338	1.6390	0.000438
Paul, J.	12. die Avantuere (1809)	620	365	379.8378	1.7107	0.000853
Paul, J.	13. Theodas ersten Tages Buch (1809)	1392	652	676.3218	1.5274	0.000674
Paul, J.	14. Missgeburten-Adel (1809)	1400	714	745.9797	1.6764	0.000436
Paul, J.	15. Hasenkrieg (1809)	1648	793	840.0866	1.6399	0.000633
Paul, J.	16. Ankunft-Sitzung (1809)	320	223	227.4760	1.7808	0.000919
Paul, J.	17. Blosse Station (1809)	478	302	308.8269	1.7311	0.000913
Paul, J.	18. Maennikes Seegefecht (1809)	656	386	400.6074	1.7202	0.000938
Paul, J.	19. Mondbelustigungen (1809)	1465	730	794.6886	1.7173	0.000551
Paul, J.	20. Zweiten Tages Buch (1809)	588	361	369.6391	1.7409	0.001096
Paul, J.	21. Hemmrad der Ankunft im Badeorte (1809)	1896	887	929.6366	1.6072	0.000476
Paul, J.	22. Niessiana (1809)	749	410	425.7108	1.6338	0.001255
Paul, J.	23. Ein Brief (1809)	241	172	173.8995	1.7188	0.002002
Paul, J.	24. Mittagtischreden (1809)	1825	872	920.9306	1.6457	0.000346
Paul, J.	25. Musikalisches Deklamatorium (1809)	388	238	247.7186	1.6528	0.000938
Paul, J.	26. Neuer Gastrollenspieler (1809)	1630	753	810.0167	1.5963	0.000699

Paul, J.	27. Nachtrag (1809)	163	119	120.0711	1.6296	0.002288
Paul, J.	28. Darum (1809)	596	355	368.7674	1.7171	0.001430
Paul, J.	30. Tischgebet und Suppe (1809)	1947	897	959.5221	1.6211	0.000512
Paul, J.	31. Aufdeckung und Sternbedeckung (1809)	425	253	259.1263	1.6026	0.001269
Paul, J.	32. Erkennszene (1809)	368	239	243.3716	1.6969	0.001017
Paul, J.	33. Abendtisch-Reden ueber Schauspiele (1809)	1218	636	660.4759	1.6732	0.000513
Paul, J.	34. Brunnen-Beaengstigungen (1809)	388	248	252.7858	1.6866	0.002216
Paul, J.	35. Theodas Brief an Bona (1809)	1370	655	693.7838	1.5885	0.000934
Paul, J.	36. Herzens-Interim (1809)	1032	546	574.9571	1.6790	0.000857
Paul, J.	37. Neue Mitarbeiter an allem (1809)	1546	731	763.6143	1.5752	0.000506
Paul, J.	38. Wie Katzenberger … (1809)	3095	1276	1351.0030	1.5237	0.000483
Paul, J.	39. Doktors Hoehlen-Besuch (1809)	516	319	329.6120	1.7328	0.001692
Paul, J.	40. Theodas Hoehlen-Besuch (1809)	1200	604	638.4810	1.6383	0.000624
Paul, J.	41. Drei Abreisen (1809)	562	336	345.9730	1.6928	0.000696
Paul, J.	42. Theodas kuerzeste Nacht der Reise (1809)	430	255	268.9611	1.6472	0.001325
Paul, J.	43. Praeliminar-Frieden ... (1809)	3222	1323	1413.4771	1.5390	0.000379
Paul, J.	44. Die Stuben-Treffen (1809)	1731	815	869.5594	1.6267	0.000688
Paul, J.	45. Ende der Reisen und Noeten (1809)	1839	864	921.6598	1.6361	0.000386
Paul, J.	I. Die Kunst. einzuschlafen (1809)	4148	1591	1713.6207	1.4946	0.000247
Paul, J.	I. Huldigungpredigt (1809)	1844	897	951.6060	1.6853	0.000431
Paul, J.	I. Wuensche fuer Luthers Denkmal (1809)	6644	2417	2625.1154	1.5103	0.000178
Paul, J.	II. Das Glueck (1809)	1881	896	943.1259	1.6418	0.000614
Paul, J.	II. Ueber Charlotte Corday (1809)	7854	2680	2960.5063	1.4682	0.000166
Paul, J.	II. Ueber Hebels alemannische Gedichte (1809)	870	489	520.0381	1.7570	0.000881
Paul, J.	III. Die Vernichtung (1809)	2723	1102	1235.6792	1.5588	0.000461
Paul, J.	III. Polymeter (1809)	963	482	516.2581	1.5995	0.000765
Paul, J.	III. Rat zu urdeutschen Taufnamen (1809)	1236	676	520.0381	1.7496	0.000598

Paul, J.	IIII. Dr. Fenks Leichenrede (1809)	2059	1011	1067.8164	1.7185	0.000281
Paul, J.	V. Ueber den Tod nach dem Tode (1809)	3955	1513	1658.7034	1.5086	0.000370
Raabe, W.	Deutscher Mondschein (1873)	6253	2110	2355.0997	1.4297	0.000218
Raabe, W.	Ein Besuch (1884)	2690	950	1060.4001	1.3520	0.000453
Raabe, W.	Eine Silvester-Stimmung (1878)	3173	962	1069.6008	1.1803	0.000355
Raabe, W.	Im Siegeskranze (1866)	13045	3003	3638.3903	1.1478	0.000141
Raabe, W.	Theklas Erbschaft (1865)	5087	1801	1964.3420	1.4312	0.000217
Rieder, E.	Brief an einen Toten (2010)	1231	472	511.0053	1.2828	0.000804
Rieder, E.	Liebe Mutter (2010)	1161	510	531.7592	1.4037	0.000743
Schnitzler, A.	Albine (about 1900)	1936	825	2355.0997	1.4297	0.000881
Schnitzler, A.	Amerika (about 1900)	801	410	1060.4001	1.3520	0.001304
Schnitzler, A.	Das Schicksal (about 1900)	6552	1993	1069.6008	1.1803	0.000218
Schnitzler, A.	Der Andere (about 1900)	2489	870	3638.3903	1.1478	0.000585
Schnitzler, A.	Der Fürst ist im Hause (about 1900)	1711	666	1964.3420	1.4312	0.000843
Schnitzler, A.	Der Sohn (1899)	2793	961	511.0053	1.2828	0.000362
Schnitzler, A.	Die Braut (about 1900)	2123	822	531.7592	1.4037	0.000576
Schnitzler, A.	Die Frau des Weisen (1898)	5652	1451	2355.0997	1.4297	0.000248
Schnitzler, A.	Die Toten schweigen (about 1900)	6173	1476	1060.4001	1.3520	0.000210
Schnitzler, A.	Er wartet auf den vazierenden Gott (about 1900)	1184	544	1069.6008	1.1803	0.000731
Schnitzler, A.	Erbschaft (about 1900)	1539	668	3638.3903	1.1478	0.000681
Schnitzler, A.	Frühlingsnacht im Seziersaal (about 1900)	1595	723	1964.3420	1.4312	0.000779
Schnitzler, A.	Mein Freund Ypsilon (about 1900)	3900	1309	511.0053	1.2828	0.000458
Schnitzler, A.	Welch eine Melodie (about 1900)	1349	629	531.7592	1.4037	0.000680
Sealsfield, C.	Das Cajuetenbuch - Callao 1825 (1843)	8423	2735	2966.4563	1.3825	0.000193
Sealsfield, C.	Das Cajuetenbuch - Das Diner (1843)	1368	704	730.4652	1.6746	0.000608
Sealsfield, C.	Das Cajuetenbuch - Das Paradies der Liebe (1843)	1515	586	636.4313	1.3360	0.000534
Sealsfield, C.	Das Cajuetenbuch - Der Abend (1843)	1517	679	705.5967	1.4796	0.000667

Sealsfield, C.	Das Cajuetenbuch - Der Fluch Kishogues (1843)	4162	1252	1507.9459	1.3113	0.000353
Sealsfield, C.	Das Cajuetenbuch - Der Kapitaen (1843)	5626	1653	1788.8537	1.1924	0.000286
Sealsfield, C.	Das Cajuetenbuch - Die Fahrt und die Kajuete (1843)	4195	1516	1664.9112	1.4378	0.000280
Sealsfield, C.	Das Cajuetenbuch - Die Praerie am Jacinto (1843)	1352	600	628.7768	1.4561	0.000816
Sealsfield, C.	Das Cajuetenbuch - Ein Morgen im Paradiese (1843)	1752	799	860.9260	1.5939	0.000486
Sealsfield, C.	Das Cajuetenbuch - Havanna 1816 (1843)	6041	2040	2223.9184	1.3920	0.000193
Sealsfield, C.	Das Cajuetenbuch - Sehr Seltsam! (1843)	5748	1655	1775.8324	1.1615	0.000160
Sealsfield, C.	Das Cajuetenbuch - Selige Stunden (1843)	1696	753	803.4532	1.5299	0.000602
Sealsfield, C.	Das Cajuetenbuch 1 (1843)	4663	1825	1936.3384	1.5234	0.000272
Sealsfield, C.	Das Cajuetenbuch 2 (1843)	3238	1197	1283.8108	1.3918	0.000308
Sealsfield, C.	Das Cajuetenbuch 3 (1843)	3954	1399	1529.5646	1.3915	0.000339
Sealsfield, C.	Das Cajuetenbuch 4 (1843)	3187	1079	1148.6162	1.2626	0.000414
Sealsfield, C.	Das Cajuetenbuch 5 (1843)	2586	1010	1053.2213	1.3899	0.000503
Sealsfield, C.	Das Cajuetenbuch 6 (1843)	2939	1035	1085.5876	1.2811	0.000575
Sealsfield, C.	Das Cajuetenbuch 7 (1843)	4865	1333	1435.1893	1.0877	0.000223
Sealsfield, C.	Das Cajuetenbuch 8 (1843)	7259	2295	2518.9217	1.3398	0.000178
Sealsfield, C.	Das Cajuetenbuch 9 (1843)	4838	1620	1725.6404	1.3143	0.000267
Sealsfield, C.	Das Cajuetenbuch 10 (1843)	3785	1265	1332.5020	1.2597	0.000361
Sealsfield, C.	Das Cajuetenbuch 11 (1843)	3019	1191	1261.5481	1.4541	0.000426
Sealsfield, C.	Das Cajuetenbuch 12 (1843)	2370	1071	1138.8995	1.6217	0.000411
Sealsfield, C.	Das Cajuetenbuch 13 (1843)	2744	1198	1257.1705	1.5753	0.000414
Sealsfield, C.	Das Cajuetenbuch 14 (1843)	4786	1545	1675.8163	1.2885	0.000326
Sealsfield, C.	Das Cajuetenbuch 15 (1843)	4497	1602	1706.8173	1.3865	0.000264
Sealsfield, C.	Das Cajuetenbuch 16 (1843)	6705	2273	2429.4817	1.3865	0.000245
Sloggi (pseudonym)	Eine kleine Geschichte mit der Zeit (2010)	728	363	381.0986	1.4983	0.001410
Sloggi (pseudonym)	Taumelnde Realitaet (2010)	612	326	338.5495	1.5416	0.001323
Storm, T.	Der Schimmelreiter. 1st half (1888)	19181	4006	4588.4182	1.0245	0.000106

Storm, T.	Der Schimmelreiter 2nd half (1888)	19119	3848	4419.0859	0.9896	0.000124
Sudermann, H.	Die Reise nach Tilsit (1917)	11437	2427	2879.0828	1.0216	0.000134
Tucholsky, K.	Schloss Gripsholm 1 (1931)	8544	2449	2757.3764	1.2689	0.000235
Tucholsky, K.	Schloss Gripsholm 2 (1931)	7106	1935	2100.4428	1.1385	0.000204
Tucholsky, K.	Schloss Gripsholm 3 (1931)	9699	2502	2790.4180	1.1470	0.000200
Tucholsky, K.	Schloss Gripsholm 4 (1931)	7415	1968	2138.9185	1.1164	0.000281
Tucholsky, K.	Schloss Gripsholm 5 (1931)	4823	1399	1537.4552	1.1742	0.000257
Wedekind, F.	Der Verführer (about 1900)	2033	855	920.8471	1.4984	0.000612
Wedekind, F.	Frühlingsstürme (about 1900)	5955	1689	1901.3018	1.2052	0.000295
Wedekind, F.	Mine-Haha I (1903)	4035	1336	1427.9679	1.2761	0.000386
Wedekind, F.	Mine-Haha II (1903)	6040	1731	1872.0151	1.1719	0.000232
Wedekind, F.	Mine-Haha III (1903)	7402	1934	2167.9088	1.1333	0.000216
Wedekind, F.	Mine-Haha IV (1903)	1297	646	676.1460	1.6228	0.000826
Wedekind, F.	Rabbi Esra (about 1900)	1935	580	645.2454	1.0960	0.000504
Wedekind, F.	Silvester (about 1900)	605	341	352.1927	1.6194	0.001516

Table 10: Hawaiian prose

http://www2.hawaii.edu/~kroddy/moolelo/kawelo/mokuna1.htm
http://www.gutenberg.org/files/13603/13603.txt

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Kawelo	Mokuna I (1909)	3507	521	764.2747	0.7725	0.000409
Kawelo	Mokuna II (1909)	7892	744	1228.9000	0.6068	0.000201
Kawelo	Mokuna III (1909)	7620	680	1047.4841	0.5336	0.000271
Kawelo	Mokuna IV (1909)	12356	1039	1876.6807	0.6215	0.000076

Laieikawai	Introduction (1917. anonymous romance)	506	135	155.8483	0.8337	0.003169
Laieikawai	I. The birth of the Princess	2348	321	449.3212	0.6446	0.000661
Laieikawai	II. The flight to Paliuli	2788	322	459.8502	0.5684	0.000815
Laieikawai	III. Kauakahialii meets the Princess	2535	348	476.0875	0.6392	0.000585
Laieikawai	IV. Aiwohikupua goes to woo the Princess	3191	387	509.6389	0.5600	0.000501
Laieikawai	V. The boxing match with Cold-nose	1832	309	379.6029	0.6768	0.001046
Laieikawai	VI. The house thatched with bird feathers	1996	323	432.8242	0.7156	0.000503
Laieikawai	VII. The Woman of the Mountain	1296	219	268.6868	0.6453	0.001451
Laieikawai	VIII. The refusal of the Princess	1658	266	346.3423	0.6725	0.000755
Laieikawai	IX. Aiwohikupua deserts his sisters	3273	357	513.2585	0.5512	0.000505
Laieikawai	XI. Abandoned in the forest	1547	234	291.3328	0.6006	0.000489
Laieikawai	XII. Adoption by the Princess	1742	251	314.2530	0.5847	0.001111
Laieikawai	XIII. Hauailiki goes surf riding	1617	268	337.5305	0.6698	0.000771
Laieikawai	XIV. The stubbornness of Laieikawai	2105	302	385.4147	0.6085	0.000538
Laieikawai	XV. Aiwohikupua meets the guardians of Paliuli	1132	219	254.7657	0.6873	0.002218
Laieikawai	XVI. The Great Lizard of Paliuli	1199	227	274.5380	0.7050	0.001600
Laieikawai	XVII. The battle between the Dog and the Lizard	1313	232	292.2356	0.6940	0.000865
Laieikawai	XVIII. Aiwohikupua's marriage with the Woman of the Mountain	2045	301	391.9715	0.6346	0.000838
Laieikawai	XIX. The rivalry of Hina and Poliahu	1840	260	348.7167	0.6187	0.000895
Laieikawai	XX. A suitor is found for the Princess	1951	280	368.8892	0.6221	0.001163
Laieikawai	XXI. The Rascal of Puna wins the Princess	1664	242	325.2627	0.6296	0.000983
Laieikawai	XXII. Waka's revenge	2245	306	423.4071	0.6320	0.000612

Laieikawai	XXIII. The Puna Rascal deserts the Princess	2222	298	416.3870	0.6272	0.000645
Laieikawai	XXIV. The marriage of the chiefs	1731	245	328.0722	0.6137	0.000866
Laieikawai	XXV. The Seer finds the Princess	1976	292	398.6615	0.6649	0.000769
Laieikawai	XXVI. The Prophet of God	1953	300	416.5577	0.7019	0.000593
Laieikawai	XXVII. A journey to the Heavens	1738	269	360.7592	0.6725	0.001041
Laieikawai	XXVIII. The Eyeball-of-the-Sun	2435	365	496.3962	0.6904	0.000582
Laieikawai	XXIX. The warning of vengeance	1245	228	293.6043	0.7299	0.000956
Laieikawai	XXX. The coming of the Beloved	1423	291	380.7179	0.8436	0.000648
Laieikawai	XXXI. The Beloved falls into sin	2046	305	400.6518	0.6483	0.000691
Laieikawai	XXXII. The Twin Sister	1740	260	345.3927	0.6433	0.000583
Laieikawai	XXXIII. The Woman of Hana	2099	314	404.8231	0.6407	0.001159
Laieikawai	XXXIV. The Woman of the Twilight	2062	284	386.6041	0.6214	0.000978

Table 11a. Hungarian poetry

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Arany, J.	Elbeszélő költemények. part 1 (1847-1887)	28453	12001	13633.3263	2.1342	0.000034
Arany, J.	Elbeszélő költemények. part 2 (1847-1887)	29715	11595	13326.1928	2.006	0.000029
Arany, J.	Elbeszélő költemények. part 3 (1847-1887)	24334	9496	11056.2758	1.9928	0.000036
Arany, J.	Elbeszélő költemények. part 4 (1847-1887)	30840	11540	12911.1265	1.8793	0.000037
Arany, J.	Toldi trilogia (1846)	10261	4433	5063.7828	1.9796	0.000107
Arany, J.	Versek. part 1 (1847-1882)	26851	10469	12140.1694	2.0024	0.000041
Arany, J.	Versek. part 2 (1847-1882)	21084	8575	9958.6577	2.0424	0.000044

Arany, J.	Versek. part 3 (1847-1882)	19151	8038	9109.5670	2.037	0.000068
Arany, J.	Zsengék. magánérdekű és Aranynak tulajdonított versek (second half of the 19th century)	7178	3859	4261.1645	2.289	0.000124
Árpas, R.	Forgószélben http://mek.niif.hu/07500/07519/07519.htm (2010)	11818	6441	7405.9822	2.5518	0.000037
Petőfi, S.	Poems (1845)	19641	6860	8114.9306	1.7738	0.000085
Petőfi, S.	Poems (1846)	19363	6770	7953.0191	1.7608	0.000048
Petőfi, S.	Poems (1947)	35543	11208	13335.5192	1.7074	0.000028
Petőfi, S.	Az apostol (1848)	13444	4923	5783.7459	1.7761	0.000076
Petőfi, S.	János Vitéz (1844)	8528	3668	4237.4818	1.9532	0.000124

Table 11b: Hungarian prose

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Bársony, I.	Délibáb. part 1 (1827)	25680	9008	11190.9068	1.9216	0.000043
Bársony, I.	Délibáb. part 2 (1828)	25075	8505	10769.4868	1.8894	0.000036
Bródy, S.	Az ezüst kecske. part 1 (1897)	25938	8728	10481.9326	1.7837	0.000062
Bródy, S.	Az ezüst kecske. part 2 (1897)	20048	6888	8197.4572	1.7591	0.000046
Gárdonyi, G.	Egri csillagok. part 1 (1899-1901)	15700	5371	7146.0094	1.9098	0.000091
Gárdonyi, G.	Egri csillagok. part 2 (1899-1901)	24458	7489	10011.1795	1.7963	0.000047
Gárdonyi, G.	Egri csillagok. part 3 (1899-1901)	26219	8158	10646.8380	1.7943	0.000040
Gárdonyi, G.	Egri csillagok. part 4 (1899-1901)	32863	9609	13188.9568	1.8140	0.000027
Gárdonyi, G.	Egri csillagok. part 5 (1899-1901)	32706	9629	13389.5101	1.8483	0.000038
Gárdonyi, G.	Isten Rabjai. part 1 (1908)	33662	9463	12979.2532	1.7456	0.000041

Gárdonyi, G.	Isten Rabjai. part 2 (1908)	30483	9393	12130.4555	1.7844	0.000042
Gárdonyi, G.	Isten Rabjai. part 3 (1908)	24500	8103	10351.9107	1.8545	0.000034
Karinthy, F.	Utazás a koponyám körül. 1st quarter (1937)	11143	4861	5605.4861	2.0358	0.000069
Karinthy, F.	Utazás a koponyám körül. 2nd quarter (1937)	11113	4724	5396.3350	1.9646	0.000108
Karinthy, F.	Utazás a koponyám körül. 3rd quarter (1937)	11153	4752	5411.3701	1.9638	0.000071
Karinthy, F.	Utazás a koponyám körül. 4th quarter (1937)	11101	4778	5469.2347	1.9931	0.000092
Karinthy, F.	Utazás. part 1 (first half of the 20th century)	21193	8471	9865.2233	2.0138	0.000045
Karinthy, F.	Utazás. part 2 (first half of the 20th century)	23103	9041	10452.1178	1.9742	0.000046
Surányi, M.	Az örvény (1924)	16192	5442	6741.2629	1.7525	0.000052
Tamási, A.	Szülőföldem. part 1 (1939)	12209	4549	5500.6524	1.8412	0.000078
Tamási, A.	Szülőföldem. part 2 (1939)	10560	4164	5028.3953	1.9159	0.000106
Tamási, A.	Szülőföldem. part 3 (1939)	10990	4288	5214.2142	1.9173	0.000091
Tamási, A.	Szülőföldem. part 4 (1939)	8567	3454	4109.8502	1.8867	0.000121
Tornyai, J.	Gyere. part 1 (first half of the 20th century)	26985	7919	9742.4336	1.5998	0.000071
Tornyai, J.	Gyere. part 2 (first half of the 20th century)	29002	8061	9984.5268	1.5363	0.000052
Wesselényi, M.	Balítéletekről 1st third (1831)	13164	5328	5929.1904	1.8554	0.000085
Wesselényi, M.	Balítéletekről 2nd third (1831)	13233	5253	5897.8381	1.8370	0.000048

Wesselényi, M.	Balítéletekről 3rd third (1831)	12980	5054	5624.3502	1.7823	0.000088
Wesselényi, M.	Balítéletekről. part 1 (1831)	18367	6852	7733.3812	1.7954	0.000082
Wesselényi, M.	Balítéletekről. part 2 (1831)	20795	7535	8489.4509	1.7628	0.000048

Table 11c: Hungarian newspaper texts

Text	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
A nominalizmus forradalma (2006) http://www.mno.hu/portal/printable?contentID=380289&sourceType=MNO	1288	789	907.1794	2.1904	0.000122
Egyre több ids húzza az igát (2006)	936	609	674.0608	2.1398	0.000226
Kunczekolbász http://www.mno.hu/portal/380473 (2006)	413	290	314.3958	1.9914	0.000780
Népszavazás előtt http://www.mno.hu/portal/380285 (2006)	403	291	332.4396	2.1491	0.000474
Orbán Viktor beszéde az Astoriánál http://www.magyarnemzet.com/portal/380017 (2006)	2044	1079	1288.8270	2.0874	0.000192

Table 12a: Indonesian newspaper texts

Text	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Assagaf-Ali Baba Jadi Asisten (21 Sep 2006)	376	221	228.4947	1.5649	0.001672
BRI Siap Cetak Miliarder Dalam (21 Oct 2006)	373	209	218.6208	1.5073	0.001964
Pelni Jamin Tiket Tidak Habis (18 Oct 2006)	414	188	195.6488	1.2367	0.002675
Pemerintah Andalkan Hujan (10 Oct 2006)	343	213	217.3651	1.6067	0.001486
Pengurus PSM Terbelah (9 Oct 2006)	347	194	199.8510	1.4631	0.001295

Table 13a: Italian poetry

Mostly from http://www.gutenberg.org/browse/languages/it

Author	Text	N	V	L	Λ	$Var(\Lambda)$
Dante, A.	Divina Commedia. I. Inferno (1472)	22934	3604	4854.604	0.9230	0.000098
Dante, A.	Divina Commedia. II. Purgatorio (1472)	15015	3940	4622.157	1.2857	0.000130
Dante, A.	Divina Commedia. III. Paradiso (1472)	32064	6782	8013.304	1.1261	0.000063
Leopardi, G.	Canti (1831)	854	483	534.3341	1.8342	0.000583
Pellico, S.	Poesie Inedite. I. 1st half (1837)	18573	5645	6531.362	1.5012	0.000133
Pellico, S.	Poesie Inedite. I. 2nd half (1837)	18497	5617	6446.307	1.4871	0.000066
Pellico, S.	Poesie Inedite. II. 1st half (1837)	19537	6096	7215.482	1.5847	0.000095
Pellico, S.	Poesie Inedite. II. 2nd half (1837)	19466	5914	7019.343	1.5467	0.000048

Table 13b: Italian prose

Mostly from http://www.gutenberg.org/browse/languages/it

Author	Text	N	V	L	Λ	$Var(\Lambda)$
Amicis, E. de	Il cuore (1886)	1129	512	537.4872	1.4533	0.000525
Amicis, E. de	Le tre capitali (1898)	27622	6687	7813.855	1.2564	0.000057
Amicis, E. de	Ricordi di Londra (1874)	23223	6155	7149.417	1.3441	0.000061
Capuana, L.	Idealismo e cosmopolitismo. In: Gli 'ismi' contemporanei (1898)	10740	3165	3543.502	1.3300	0.000140
Capuana, L.	Romanzi e novelle. In: Gli 'ismi' contemporanei (1898)	19512	5140	5872.518	1.2913	0.000065
Capuana, L.	Varieta. In: Gli 'ismi' contemporanei (1898)	22738	6707	7579.668	1.4523	0.000037

Deledda, G.	Canne al vento (1913)	3258	1237	1329.653	1.4337	0.000586
Giacomo, S.	Mattinate Napoletane (1887)	21992	5421	6108.924	1.2062	0.000050
Manzoni, A.	I promessi sposi (1827)	6064	2203	2426.397	1.5136	0.000195
Pellico, S.	Le mie prigioni (1832)	11760	3667	4007.01	1.3869	0.000088

Table 13c: Italian End-of-Year Presidential speeches

Author	**Text**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Einaudi, L.	End-of-Year Presidential speech (1949)	194	140	143.5432	1.6928	0.001487
Einaudi, L.	End-of-Year Presidential speech (1950)	150	105	108.7800	1.5781	0.003510
Einaudi, L.	End-of-Year Presidential speech (1951)	230	169	172.2333	1.7686	0.001979
Einaudi, L.	End-of-Year Presidential speech (1952)	179	145	146.8929	1.8488	0.001530
Einaudi, L.	End-of-Year Presidential speech (1953)	190	143	145.8191	1.7489	0.002234
Einaudi, L.	End-of-Year Presidential speech (1954)	260	181	186.2913	1.7303	0.002064
Gronchi, G.	End-of-Year Presidential speech (1955)	388	248	255.3558	1.7038	0.001467
Gronchi, G.	End-of-Year Presidential speech (1956)	665	374	392.7529	1.6672	0.000765
Gronchi, G.	End-of-Year Presidential speech (1957)	1130	549	599.3767	1.6194	0.000731
Gronchi, G.	End-of-Year Presidential speech (1958)	886	460	488.0442	1.6236	0.000978
Gronchi, G.	End-of-Year Presidential speech (1959)	697	388	409.8201	1.6718	0.000776
Gronchi, G.	End-of-Year Presidential speech (1960)	804	434	462.2283	1.6703	0.000865
Gronchi, G.	End-of-Year Presidential speech (1961)	1252	622	674.0538	1.6677	0.000608
Segni, A.	End-of-Year Presidential speech (1962)	738	381	404.0091	1.5701	0.000629

Segni, A.	End-of-Year Presidential speech (1963)	1057	527	559.5185	1.6008	0.001093
Saragat, G.	End-of-Year Presidential speech (1964)	465	278	289.0321	1.6580	0.001090
Saragat, G.	End-of-Year Presidential speech (1965)	1052	510	547.7775	1.5736	0.001012
Saragat, G.	End-of-Year Presidential speech (1966)	1200	597	624.7671	1.6031	0.000875
Saragat, G.	End-of-Year Presidential speech (1967)	1056	526	562.9810	1.6120	0.000941
Saragat, G.	End-of-Year Presidential speech (1968)	1173	562	602.8260	1.5774	0.000804
Saragat, G.	End-of-Year Presidential speech (1969)	1583	692	759.8210	1.5357	0.000422
Saragat, G.	End-of-Year Presidential speech (1970)	1929	812	877.5755	1.4946	0.000932
Leone, G.	End-of-Year Presidential speech (1971)	262	168	173.0226	1.5970	0.001280
Leone, G.	End-of-Year Presidential speech (1972)	767	394	414.7079	1.5598	0.001091
Leone, G.	End-of-Year Presidential speech (1973)	1250	616	669.2188	1.6580	0.000680
Leone, G.	End-of-Year Presidential speech (1974)	801	426	445.7840	1.6160	0.000799
Leone, G.	End-of-Year Presidential speech (1975)	1328	632	678.9746	1.5968	0.000665
Leone, G.	End-of-Year Presidential speech (1976)	1366	649	685.1578	1.5727	0.000484
Leone, G.	End-of-Year Presidential speech (1977)	1604	717	780.7230	1.5601	0.000499
Pertini, S.	End-of-Year Presidential speech (1978)	1492	603	639.4469	1.3602	0.000791
Pertini, S.	End-of-Year Presidential speech (1979)	2311	800	848.3508	1.2348	0.000688
Pertini, S.	End-of-Year Presidential speech (1980)	1360	535	567.9546	1.3086	0.001089
Pertini, S.	End-of-Year Presidential speech (1981)	2819	911	983.9384	1.2042	0.000525
Pertini, S.	End-of-Year Presidential speech (1982)	2486	854	921.7382	1.2590	0.000481
Pertini, S.	End-of-Year Presidential speech (1983)	3746	1149	1236.6461	1.1797	0.000318
Pertini, S.	End-of-Year Presidential speech (1984)	1340	514	539.1823	1.2583	0.000695

Cossiga, F.	End-of-Year Presidential speech (1985)	2359	859	955.7467	1.3665	0.000532
Cossiga, F.	End-of-Year Presidential speech (1986)	1348	561	610.0912	1.4165	0.000915
Cossiga, F.	End-of-Year Presidential speech (1987)	2092	904	993.7626	1.5774	0.000438
Cossiga, F.	End-of-Year Presidential speech (1988)	2384	875	976.9096	1.3839	0.000543
Cossiga, F.	End-of-Year Presidential speech (1989)	1912	778	842.2127	1.4455	0.000594
Cossiga, F.	End-of-Year Presidential speech (1990)	3345	1222	1351.7941	1.4243	0.000372
Cossiga, F.	End-of-Year Presidential speech (1991)	418	241	254.7695	1.5976	0.002019
Scalfaro, O.L.	End-of-Year Presidential speech (1992)	2774	978	1072.8016	1.3316	0.000464
Scalfaro, O.L.	End-of-Year Presidential speech (1993)	2942	1074	1179.3043	1.3904	0.000410
Scalfaro, O.L.	End-of-Year Presidential speech (1994)	3606	1190	1333.2622	1.3152	0.000268
Scalfaro, O.L.	End-of-Year Presidential speech (1995)	4233	1341	1492.5157	1.2787	0.000346
Scalfaro, O.L.	End-of-Year Presidential speech (1996)	2085	866	934.0381	1.4869	0.000594
Scalfaro, O.L.	End-of-Year Presidential speech (1997)	5012	1405	1538.4429	1.1357	0.000273
Scalfaro, O.L.	End-of-Year Presidential speech (1998)	3995	1175	1281.1874	1.1550	0.000332
Ciampi, C.A.	End-of-Year Presidential speech (1999)	1941	831	877.3226	1.4862	0.000439
Ciampi, C.A.	End-of-Year Presidential speech (2000)	1844	822	871.2039	1.5429	0.000540
Ciampi, C.A.	End-of-Year Presidential speech (2001)	2098	898	965.5417	1.5288	0.000422
Ciampi, C.A.	End-of-Year Presidential speech (2002)	2129	909	984.9410	1.5397	0.000517
Ciampi, C.A.	End-of-Year Presidential speech (2003)	1565	718	763.4969	1.5585	0.000816
Ciampi, C.A.	End-of-Year Presidential speech (2004)	1807	812	869.7050	1.5676	0.000527
Ciampi, C.A.	End-of-Year Presidential speech (2005)	1193	538	576.2236	1.4860	0.000687
Napolitano, G.	End-of-Year Presidential speech (2006)	2204	929	1033.5266	1.5677	0.000590

Napolitano, G.	End-of-Year Presidential speech (2007)	1792	793	874.5688	1.5878	0.000476
Napolitano, G.	End-of-Year Presidential speech (2008)	1713	775	831.2543	1.5692	0.000687

Table 14: Japanese prose

Author	Text, year	*N*	*V*	*L*	Λ	$Var(\Lambda)$
Ariyoshi, S.	Sumi (Ink Stick) (1961)	15315	2537	3477.0279	0.9502	0.000083
Hori, T.	Utsukushii Mura (Beautiful Village) (1933)	25011	2902	4768.7768	0.8386	0.000048
Kaiko, K.	Panikku (Panic) (1957)	21250	3553	4782.7314	0.9740	0.000053
Kobayashi, H.	Guzo suhai (Idolatry) (1950)	5613	1276	1621.3952	1.0830	0.000355
Miura, T.	Kikyo (Going home) (1962)	13138	2382	3014.3885	0.9450	0.000100
Oe, K.	Shiiku (Prize Stock) (1957)	22467	3182	4845.7574	0.9386	0.000063
Okamoto, K.	Pari Sai (Quatorze Juillet) (1938)	17700	3530	4909.2397	1.1782	0.000056

Table 15a: Kannada Social Sciences

Author	Text, Year	*N*	*V*	*L*	Λ	$Var(\Lambda)$
Pradhana Gurudhat	Aadalitha Bashe Kelavu Vicharagalu pp. 71–92 (1984)	3188	1833	1891.1101	2.0783	0.000102
Pradhana Gurudhat	Aadalitha Bashe Kelavu Vicharagalu pp. 93–103 (1984)	1050	720	733.2554	2.1098	0.00031
Nagappa, T.R.	Vayskara Shikshana mathu swayam seve pp. 1–15 (1988)	4869	2477	2558.4278	1.9376	0.000107

Nagappa, T.R.	Vayskara Shikshana mathu swayam seve pp. 16–42 (1988)	5231	2433	2484.4061	1.7661	0.000201
Murthy, D.N.S.	Shreshta arthashasthagnayaru pp. 3–53 (1990)	4541	2516	2557.6910	2.0599	0.000117
Pn. Keshava Murthy Galigali	Ashtavarga Paddhathi. pp. 2–164 (about 1985)	4141	1842	1877.3766	1.6399	0.000192
Shivshankar aralimatti	Asian Kredegalu pp. 1–12 (1983)	1302	807	829.2109	1.9836	0.000365
Om Prakash	Asthama pp. 1–42 (1982)	4735	2356	2428.2787	1.8848	0.000163
Om Prakash	Asthama pp. 43–73 (1982)	4316	2122	2224.6369	1.8737	0.000186

Table 15b: Kannada Commerce

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Poojara, B.C., Balavantha. M.U.	Lekka Parishodhaneya moolathatvagalu mattu Aacharane pp. 1–29 (1987)	3713	1664	1791.8498	1.7227	0.000188
Poojara, B.C., Balavantha. M.U.	Lekka Parishodhaneya moolathatvagalu mattu Aacharane pp. 96–126 (1987)	4508	1738	1810.2104	1.4673	0.000176
Basava, K.D.	Lekka Parishodhana Shasthra pp. 1–25 (1984)	1787	833	868.1568	1.5799	0.000535
Basava, K.D.	Lekka Parishodhana Shasthra pp. 28–40 (1984)	4556	1755	1891.0226	1.5185	0.000246

Girija shankar	Lekka Parishodhana Shasthra pp. 43–52 (1986)	1455	790	825.4514	1.7944	0.000457
Girija shankar	Lekka Parishodhana Shasthra pp. 53–80 (1986)	4554	1764	1924.8015	1.5463	0.000265
Girija shankar	Lekka Parishodhana Shasthra pp. 81–109 (1986)	4685	1738	1850.7375	1.4501	0.000173
Venkatarao, D.D.	Banking Siddhantha pp. 1–24 (1989)	4499	2005	2149.2828	1.7452	0.000172
Venkatarao, D.D.	Banking Siddhantha pp. 35–61 (1989)	4672	1920	2013.4462	1.5814	0.000188

Table 16: Lakota tape-record texts

Text, year	N	V	L	Λ	$Var(\Lambda)$
Bean, grass, and fire (1995)	219	116	125.5587	1.3418	0.001940
Iktomi meets the prairie chicken (1994)	1633	479	579.9693	1.1411	0.000562
Iktomi meets two women and Iya (1994)	809	272	317.6293	1.1417	0.001200
The fly on the window (1994)	345	174	184.7719	1.3592	0.001993

Table 17a: Latin poetry

http://www.thelatinlibrary.com/

Author	**Text, year**	N	V	L	Λ	$Var(\Lambda)$
Catullus, Gaius	all poems	13248	6046	6198.0461	1.9285	0.000083
Horatius, Flaccus	Carmina. Liber I-IV	13731	7236	7687.8752	2.3167	0.000041
Horatius, Flaccus	Epistles. Liber I-II	9900	5191	5561.1152	2.2445	0.000048
Horatius, Flaccus	Odes & Epodes	16232	8243	8765.7753	2.2737	0.000057
Horatius, Flaccus	Sermones	14242	6580	6813.3861	1.9871	0.000063
Horatius, Flaccus	Sermones. Liber 1. Sermo 1	829	609	620.5317	2.1846	0.000727
Martialis, Marcus	Epigrammata	1354	909	930.4721	2.1521	0.000363
Ovidius, Publius	Ars amatoria. liber primus	4931	2703	2782.7051	2.084	0.000110
Ovidius, Publius	Ars amatoriae. Liber I-III	14895	6087	6370.3444	1.7847	0.000061
Ovidius, Publius	Fasti	31267	10276	10991.2121	1.5801	0.000048
Silius Italicus, Tiberius	Punicorum Libri I - V	22972	9254	9745.2763	1.8501	0.000041
Silius Italicus, Tiberius	Punicorum Libri VI - X	22346	8909	9484.5611	1.846	0.000039
Silius Italicus, Tiberius	Punicorum Libri XI - XIII	14616	6634	6996.4873	1.9936	0.000047
Silius Italicus, Tiberius	Punicorum Libri XIV - XVII	18363	7805	8242.8892	1.914	0.000038
Vergilius, Publius	Aeneidos I - VI	30604	10755	11744.2459	1.7214	0.000038
Vergilius, Publius	Aeneidos VII - XII	33165	11191	12238.3370	1.6682	0.000031
Vergilius, Publius	Georgicon	14201	6771	7305.0448	2.136	0.000053
Vergilius, Publius	Georgicon liber primus	3311	2211	2328.4387	2.4754	0.000089

Table 17b: Latin prose

http://www.thelatinlibrary.com/

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Apuleius, Lucius	Fables. Book 1	4010	2334	2501.9964	2.2481	0.000192
Apuleius, Lucius	Metamorphoses Liber I	4015	2335	2504.4040	2.2478	0.000241
Apuleius, Lucius	Metamorphoses Liber II	4817	2797	3039.8954	2.3241	0.000145
Apuleius, Lucius	Metamorphoses Liber III	4104	2446	2569.1645	2.2619	0.000166
Apuleius, Lucius	Metamorphoses Liber IV	5158	3162	3325.1570	2.3933	0.000142
Apuleius, Lucius	Metamorphoses Liber V	4605	2695	2874.4212	2.2866	0.000134
Apuleius, Lucius	Metamorphoses Liber VI	4513	2777	2951.1815	2.3898	0.000186
Apuleius, Lucius	Metamorphoses Liber VII	4089	2618	2737.1158	2.4176	0.000179
Apuleius, Lucius	Metamorphoses Liber VIII	4825	3005	3137.5412	2.3953	0.000153
Apuleius, Lucius	Metamorphoses Liber IX	6524	3765	3954.4263	2.3121	0.000115
Apuleius, Lucius	Metamorphoses Liber X	5709	3456	3616.5138	2.3797	0.000094
Apuleius, Lucius	Metamorphoses Liber XI	4940	3007	3137.4113	2.3459	0.000093

Table17 c: Latin history and philosophy

http://www.thelatinlibrary.com/

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Caesar, Julius	De Bello Gallico	20463	5654	6103.9776	1.2859	0.000063
Descartes, René	Meditationes de prima philosophia	19252	4239	4610.5597	1.0261	0.000071

Tacitus, Publius	Germania & Agricola	12802	5798	6274.3184	2.0129	0.000051
Tacitus, Publius	Historiae. Liber I	11824	5347	5814.6421	2.0028	0.000076
Tacitus, Publius	Historiae. Liber II	12228	5711	6121.2842	2.0461	0.000062

Table 17 d: Latin rhetorics

http://www.thelatinlibrary.com/

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Cicero, Marcus	Academica	23180	6002	6553.1748	1.234	0.000082
Cicero, Marcus	Brutus	24979	6854	7726.2041	1.3602	0.000090
Cicero, Marcus	Orationes	13186	4513	4730.4500	1.4779	0.000086
Cicero, Marcus	Post reditum in senatu oratio	4285	1910	1983.2745	1.6808	0.000291
Cicero, Marcus	Pro Flacco Oratio	10876	4111	4303.1976	1.5971	0.000077
Quintilianus, Marcus	Oratoriae liber decimus	12140	4437	4813.8218	1.6196	0.000075

Table 18: Macedonian

Translation (1988) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4810	1636	1798.7271	1.3770	0.000308
2	4898	1836	1988.6609	1.4982	0.000116
3	7470	2456	2698.4967	1.3992	0.000094

4	4424	1937	2066.1179	1.7027	0.000248
5	4425	1667	1792.7489	1.4771	0.000264
6	8914	2842	3117.5172	1.3815	0.000114
7	7153	2606	2883.9482	1.5541	0.000142
8	6414	2484	2731.3657	1.6212	0.000076
9	3850	1610	1728.7413	1.6100	0.000227
10	6461	2536	2829.5471	1.6687	0.000106

Table 19: Maori folk narratives

Text, year	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
A Tawhaki. te Tohunga Rapu Tuna (1955)	1434	277	384.6240	0.8466	0.000866
Ka Kimi a Maui i Ona Matua (1954)	3620	514	715.1782	0.7031	0.000486
Ka pu te Ruha ka Hao te Rangatahi(about 1955)	1289	326	444.2921	1.0720	0.000689
Ko te Paamu Tuatahi Whakatiputipu Kau a te Maori (1953)	1175	277	386.0068	1.0086	0.001098
Maori Nga Mahi a Nga Tupuna (1953)	2062	398	526.9199	0.8469	0.000783

Table 20a: Marathi poetry

Author	**Text**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Dasbodh.	chapters 1 – 5 (http://sanskritdocuments.org/marathi/index.html#Dasbodh)	26859	9709	9998.7072	1.6488	0.000042

Dasbodh.	chapters 6 - 10 (http://sanskritdocuments.org/marathi/index.html#Dasbodh)	28759	8577	8809.7783	1.3659	0.000058
Dasbodh.	chapters 11 - 13 (http://sanskritdocuments.org/marathi/index.html#Dasbodh)	10440	4343	4429.77	1.7052	0.000099
Dasbodh.	chapters 14 - 17 (http://sanskritdocuments.org/marathi/index.html#Dasbodh)	15864	6316	6543.6662	1.7326	0.000074
Dasbodh.	chapters 18 - 20 (http://sanskritdocuments.org/marathi/index.html#Dasbodh)	10388	4307	4424.4645	1.7107	0.000096
Dnyaneshwar, Sant	Dnyaneshvari. chapters 1-6 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	24613	10338	10644.9446	1.8991	0.000044
Dnyaneshwar, Sant	Dnyaneshvari. chapters 7-11 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	28794	11814	12074.6543	1.87	0.000037
Dnyaneshwar, Sant	Dnyaneshvari. chapters 12-13 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	16086	7405	7603.0667	1.9882	0.000052
Dnyaneshwar, Sant	Dnyaneshvari. chapters 14-15 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	12081	5865	5960.7394	2.0141	0.000121
Dnyaneshwar, Sant	Dnyaneshvari. chapters 16-17 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	10960	5715	5810.7951	2.1418	0.000108

Dnyaneshwar, Sant	Dnyaneshvari. chapter18 (http://sanskritdocuments.org/marathi/index.html#Dnyaneshwari)	22018	9429	9675.5616	1.9084	0.000053

Table 20 b: Marathi aesthetics

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Kanchan Ganekar	Nath ha majha pp. 1–17 (1989).	4146	2038	2098.9331	1.8314	0.000212
Vishram Bedekar	Ek jhaad aani don pakshi pp. 1–18 (1992)	5504	2911	2971.9077	2.0198	0.000142
Deshpanday, Pu.L.	Maithr pp. 1–15 (1989)	5105	2617	2682.5179	1.9484	0.000167
Sunitha Deshpanday	Aahe Manohar Tari pp. 1–16 (about 1990)	5195	2382	2452.4149	1.754	0.000198
Lakshman Gayakwad	Uchalya pp. 1–16 (1990)	4339	2217	2266.7549	1.9002	0.000185
Sou Veena Gavankar	Ek Hota Kabir pp. 1–118 (1988)	3489	1865	1885.4064	1.9144	0.000312
Anil Avachat	Swathavishayi pp. 1–6 (1990)	1862	1115	1131.321	1.9868	0.000256
Kanchan Ganekar	Nath ha maja pp. 18–34 (1989)	4205	2070	2142.3856	1.8463	0.000195
Deshpandya, P.L.	Maithr pp. 86–107 (about 1990)	5218	2877	2933.3632	2.0898	0.000196
Sou Veena Gavankar	Ek hota Kabir pp. 119–207 (1988)	3356	1962	1987.3053	2.0879	0.000174

Tanle 20 c: Marathi social sciences

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
M.T.	Jeevanache Rahasay Hasthasamudrik pp. 5–89 (1991)	4693	1947	2057	1.6096	0.000189
V.A.P.	Thode Adbut Thode goodh pp. 5–48 (1986)	3642	1831	1872	1.831	0.000271
Shri V.A.P.	Phalajyotishateel shankasamadhaan pp. 4–26 (1993)	4170	1853	1898	1.6475	0.000193

Pandy, V.L.	Thumcha chehara thumche yaktimatv pp. 9–89 (1990)	4062	1788	1890	1.6794	0.000262
V.A.P.	Thode Adbut Thode goodh pp. 49–74 (1986)	3943	1725	1765	1.6095	0.000302
V.D.P.	Phaljothishateel shankasamadhaan pp. 27–55 (1993)	3846	1793	1830	1.7054	0.000228
Pandy, V.L.	Thumcha chehara thumcha yakthimathv pp. 90–113 (1990)	4099	1703	1821	1.6046	0.000288
V.A.P.	Thode Adbut Thode goodh pp. 91–119 (1986)	4142	1872	1920	1.6768	0.000265
V.D.P.	Phaljothishateel shankasamadhaan pp. 56–85 (1993)	4255	1731	1785	1.5225	0.000192

Table 20 d: Marathi natural, physical and professional sciences

Author	Text, year	N	V	L	Λ	$Var(\Lambda)$
Joshi, B.P.	Nisar Sheti pp. 77-97 (1991)	2998	1555	1612.4259	1.8700	0.000249
Patil, A.V.	Roopvatika Sangapon pp. 146–161 (1984)	2922	1186	1239.4511	1.4701	0.000384
Phadnis, B.A.	Galithachi Pike pp. 67–83 (1988)	4140	1731	1778.1314	1.5535	0.000319
Dashputhre, G.B.	Vanshi Aani Vanvignan pp. 29–133 (1986)	6304	2451	2736.8659	1.6496	0.000194
Pandurang, Sarala	Bhaajipala Utpadan pp. 1–32 (1988)	4957	2029	2180.1586	1.6252	0.000207
Chaudri	Limbuvargiy Phaljhaade pp. 6–35 (1984)	3735	1503	1603.1125	1.5333	0.000265
Vedhprakash Patil	Anjirachi Lagvad pp. 5–33 (1990)	3162	1262	1322.2586	1.4636	0.00037
R.P.S	Zamitil paani va shodh vaapar pp. 1–109 (1987)	5477	1807	1966.5139	1.3423	0.000239
Shivaji Tombre	Bhaat–Sheti pp. 1–206 (1990)	6206	2387	2451.367	1.4982	0.000173
Patil, N.D., Dhume, G.R.	Pik Vadisaati Khathe pp. 15–164 c	5394	1650	1835.2537	1.2697	0.000267

Table 20 e: Marathi commerce

Author	Text, year	*N*	*V*	*L*	*Λ*	*Var*(*Λ*)
Sarangar	Rashtriy Uthpann pp. 1-104 (1985)	4128	1400	1467.6548	1.2855	0.000342
Kulkarni, G.A.	Mansi Arbhaat Aani Chillar pp. 9–26 (about 1990)	5191	2386	2445.2921	1.7501	0.000288
Madhu Nene	Bankingshi Manthri. pp. 1-62 (about 1990)	3424	1412	1469.7413	1.5172	0.00022
Madhu Nene	Bankingshi Manthri. pp. 82–127 (about 1990)	4078	1607	1648.1363	1.4592	0.000219
Madhu Nene	Bankingshi Manthri. pp. 128–168 (about 1990)	4186	1458	1500.9675	1.2987	0.000237
Madhu Nene	Bankingshi Manthri. pp. 169–207 (about 1990)	3549	1628	1695.9765	1.6965	0.000326
Aravind Ladhar	Sheharbaazarshi Manthri pp. 1–13 (1986)	2946	1547	1578.209	1.8585	0.000242
And Madhav Saraf	Shearbaazar ek alibabachi guha pp. 1–49 (1990)	3372	1523	1568.0713	1.6406	0.000247
Muthali Desai Daasthane	Jagahtheel pramukh Arthvyavastha pp. 1–62 (1986)	4843	1702	1867.3479	1.4209	0.000222
And Madhav Saraf	Shearbaazar ek alibabachi guha pp. 50–113 (1990)	3601	1719	1766.5612	1.7447	0.00024

Table 20 f: Marathi official and media

Author	Text, year	*N*	*V*	*L*	*Λ*	*Var*(*Λ*)
Madhav Gadkari	Chaupher pp. 1–14 (1988)	4060	2079	2141.015	1.9029	0.000178
Madhav Gadkari	Choupher Vol 3 pp. 4–42 (1989)	4831	2312	2398.7868	1.8293	0.000147
Arun Tikker	Samaaj-Spandane pp. 8–54 (1990)	4025	2319	2343.3835	2.0987	0.000247
Madhav Gadkari	Chaupher pp. 15–27 (1988)	3954	1957	2020.1657	1.8378	0.000186
Madhav Gadkari	Chaupher pp. 43–84 (1988)	4765	2197	2261.6036	1.7457	0.000190
Arun Tikker	Samaaj-Spandane pp. 62-91 (1990)	3337	2006	2031.2669	2.1447	0.000151

Madhav Gadkari	Chaupher pp. 28–40 (1988)	3825	1931	1995.9269	1.8695	0.000233
Madhav Gadkari	Chaupher pp. 94–170 (1988)	4895	2322	2394.9731	1.8053	0.000209
Madhav Gadkari	Chaupher pp. 41–53 (1988)	3836	1970	2040.0411	1.906	0.000150
Madhav Gadkari	Chaupher pp. 201–215 (1988)	4605	2278	2341.6148	1.8627	0.000144

Table 21. Marquesan folklore texts

Text	***N***	***V***	***L***	$\boldsymbol{\Lambda}$	***Var*(**$\boldsymbol{\Lambda}$**)**
Haneamotua (about 1975)	6659	519	970.2173	0.5571	0.000243
Ka´akai o Te Henua ´Enana (about 1975)	457	150	178.5891	1.0395	0.001634
Kopuhoroto'e (1964)	2330	289	506.9815	0.7327	0.000476
Potateuatahi (about 1975)	3315	330	576.5810	0.6123	0.000777
Taheta (about 1975)	3306	362	559.1658	0.5952	0.000305
Te Hakamanu (1990)	1509	301	500.3687	1.0540	0.000669
Tekao a´akakai no Hakatauniua (about 1975)	5102	434	895.3082	0.6506	0.000408

Table 22: Polish

Translation (1974) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	$\boldsymbol{\Lambda}$	***Var*(**$\boldsymbol{\Lambda}$**)**
1	4348	1970	2106.5431	1.7627	0.000384
2	4368	2149	2273.5401	1.8948	0.000164
3	26694	2995	3190.7453	1.8235	0.000106

4	34003	2200	2310.5224	2.0793	0.000114
5	43997	1962	2077.0745	1.8717	0.000154
6	57937	3481	3723.6237	1.8295	0.000113
7	66348	3061	3228.6791	1.9341	0.000081
8	75753	2928	3074.0904	2.0091	0.000085
9	83501	1855	1945.2144	1.9692	0.000145
10	95786	2970	3108.5103	2.0213	0.000119

Table 23: Rarotongan prose

Author	**Text, Year**	***N***	***V***	***L***	**Λ**	**$Var(\Lambda)$**
Atama, Herekaiura	Legends from the Atolls: Ko Paraka e te Kehe (1983)	845	214	264.7548	0.9170	0.001261
Kauraka, Kauraka	Legends from the Atolls: Ko Tamaro e ana uhi (1983)	892	207	255.8581	0.8463	0.000965
Nikoro, Kaimaria	Legends from the Atolls: Akamaramaanga (1983)	968	223	315.9121	0.9745	0.000986
Piniata, Temu	Legends from the Atolls: Te toa ko Herehuaroa e Araitetonga (1983)	1059	197	250.6854	0.7161	0.001092
Puroku, Tepania	Legends from the Atolls: Te toa ko Teikapongi (1983)	625	181	215.5819	0.9644	0.002063

Table 24a: Romanian poetry

Mostly from http://www.romanianvoice.com/poezii/ (2010)(Text pooling is marked by asterisk)

Author	**Text, year**	***N***	***V***	***L***	**Λ**	**$Var(\Lambda)$**
Alecsandri, V.	Pasteluri (1862-187…)*	1174	687	715.6874	1.8713	0.000700
Alecsandri, V.	Dan căpitan de plai (1874)*	2905	1312	1408.5212	1.6791	0.000278
Alecsandri, V.	Doine (1842-1852)*	2145	974	1049.9596	1.6307	0.000400

Arghezi, T.	Cuvinte potrivite (1927)*	1554	912	969.9613	1.9920	0.000250
Arghezi, T.	Flori de mucigai (1931)*	396	260	281.7660	1.8483	0.001045
Bacovia, G.	selected poems http://www.romanianvoice.com/poezii/poeti/bacovia.php (2010)*	4776	1782	1968.2219	1.5162	0.000239
Barbu, I.	Joc secund (1930)*	1047	724	755.6382	2.1795	0.000699
Barbu, I.	Uvedenrode (1925 – 1926)*	1599	954	1006.3966	2.0165	0.000314
Barbu, I.	Isarlâk (1925)*	1432	850	908.8840	2.0031	0.000326
Blaga, L.	Paşii profetului (1921)	1242	685	741.5832	1.8475	0.000635
Blaga, L.	Laudă somnului (1929)	573	368	377.5246	1.8172	0.000754
Blandiana, A.	Arhitectura valurilor (1990)	1014	543	577.6261	1.7124	0.001035
Cărtărescu, M.	selected poems (http://agonia.ro/index.php/author/0011211/type/poetry/Poezie (2010)*	14700	5251	5986.4269	1.6971	0.000035
Coşbuc, G.	Nunta Zamfirei (1893)	793	448	481.7138	1.7612	0.001067
Dinescu, M.	Moartea citeşte ziarul (1989)*	1267	780	833.5361	2.0413	0.000775
Dinescu, M.	O beţie cu Marx (1989)*	1161	724	751.4132	1.9836	0.000602
Doinaş, Şt.A.	selected poems http://www.romanianvoice.com/poezii/poeti/doinas.php (2010)*	2647	1406	1486.4588	1.9221	0.000322
Doinaş, Şt.A.	Orologiul de gheaţă (idem)	435	307	312.6143	1.8962	0.000872
Eminescu, M.	În căutarea Şeherezadei (1874)	915	594	615.4903	1.9921	0.000412
Eminescu, M.	Cugetările sărmanului Dionis (1872)	571	389	406.6445	1.9632	0.000958
Eminescu, M.	Dumnezeu şi om (1873)	443	320	327.2346	1.9548	0.001641

Eminescu, M.	Mitologicale (1873)	681	442	466.0482	1.9389	0.000647
Eminescu, M.	Egipetul (1872)	688	452	465.7789	1.9211	0.000442
Eminescu, M.	Epigonii (1870)	921	565	590.0754	1.8992	0.000570
Eminescu, M.	Împarat şi proletar (1874)	1510	857	896.5804	1.8876	0.000491
Eminescu, M.	Ondina (Fantazie) (1869)	871	535	557.6579	1.8823	0.000441
Eminescu, M.	Ecò (1872)	698	442	461.6051	1.8807	0.000996
Eminescu, M.	Junii corupţi (1869)	458	309	322.7674	1.8752	0.001987
Eminescu, M.	Floare albastră (1873)	247	185	192.1356	1.8612	0.002321
Goga, O.	Clăcaşii (1905)	685	445	469.3452	1.9429	0.000576
Goga, O.	Noapte (1905)	343	243	252.9549	1.8697	0.000895
Labiş, N.	Confesiuni (1958)	589	394	412.9235	1.9420	0.000560
Labiş, N.	Marină (1958)	280	211	216.9422	1.8960	0.001348
Labiş, N.	Primele Iubiri (1956)	3156	1596	1705.3761	1.8908	0.000197
Labiş, N.	Moartea căprioarei (1956)	499	329	342.3532	1.8511	0.000981
Paunescu, A.	Totuşi. iubirea (1983)	187	122	129.3899	1.5719	0.003296
Păunescu, A.	Analfabeţilor (1990)	273	163	170.3558	1.5202	0.002972
Păunescu, A.	Repetabila povară (1974)	497	256	271.8936	1.4751	0.001290
Sorescu, M.	selected poems http://www.romanianvoice.com/poezii/poeti/sorescu.php (2010)*	4865	1832	1987.5761	1.5063	0.000211
Stănescu, N.	Un pământ numit România (1969)	627	361	395.9719	1.7666	0.001036
Topârceanu, G.	Rapsodii de toamnă (1919)	590	398	418.1699	1.9639	0.001079
Topârceanu, G.	Rapsodii de primavară (1928)	295	211	223.6638	1.8726	0.001512

Table 24b: Romanian prose

Author	**Text, year**	N	V	L	Λ	$Var(\Lambda)$
Caragiale, I.L.	Boborul (1896)	1538	811	849.3408	1.7600	0.000268
Caragiale, I.L.	Moftangii (1893)	1624	845	919.5822	1.8180	0.000560
Caragiale, I.L.	Moşii (1901)	359	329	335.3137	2.3865	0.000294
Caragiale, I.L.	Tempora (1900)	1160	586	625.7439	1.6531	0.000383
Cărtărescu, M.	De ce iubim femeile (2004)	674	352	397.2781	1.6673	0.001764
Cărtărescu, M.	Pomul se cunoaşte după roade (2006)	1188	611	670.0817	1.7343	0.000705
Creangă, I.	Capra cu trei iezi (1875)	2462	899	1003.3207	1.3820	0.000366
Liiceanu, G.	Ce înseamnă a fi european în estul postbelic? (1990)	3254	1346	1455.8387	1.5715	0.000192
Patapievici, H.-R.	Cei care urăsc (1999)	1312	683	735.2628	1.7473	0.000564
Pleşu, A.	Minima Moralia. VIII. Darul lacrimilor (1994)	1853	999	1039.2401	1.8328	0.000650
Popescu, C.T.	Antichrista http://www.gandul.info/puterea-gandului/antichrista-5771956 (2010)	688	430	447.4437	1.8454	0.001035
Preda, M.	Cel mai iubit dintre pământeni. Part 8 (1980)	37905	7680	9098.3991	1.0990	0.000041
Rebreanu, L.	Pădurea spânzuraților (1922)	25314	5595	6427.7090	1.1181	0.000052
Rebreanu, L.	Răscoala (1932)	14416	3989	4478.3811	1.2920	0.000062
Sadoveanu,.M.	Baltagul. Chapter I (1930)	2219	994	1091.5570	1.6460	0.000276

Table 25a: Russian poetry

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Pushkin, A.	Boris Godunov (1825)	14012	4420	5005.5312	1.4813	0.000079
Pushkin, A.	Captain's daughter (1836)	33361	9273	10529.7731	1.4277	0.000040
Pushkin, A.	Dubrovsky (1841)	20527	6894	7731.1695	1.6242	0.000058
Pushkin, A.	Eugene Onegin (1833)	25918	9660	10708.6356	1.8236	0.000040
Pushkin, A.	Ruslan i Lyudmila (1820)	12005	5053	5560.1518	1.8894	0.000108

Table 25b: Russian prose

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Ostrovskij, N.	How the steel was tempered. Chapter 1 (1932-1934)	4107	1907	2050.6664	1.8043	0.000193
Ostrovskij, N.	How the steel was tempered. Chapter 2 (1932-1934)	4136	2088	2216.7094	1.9383	0.000200
Ostrovskij, N.	How the steel was tempered. Chapter 3 (1932-1934)	6323	2909	3091.3609	1.8583	0.000187
Ostrovskij, N.	How the steel was tempered. Chapter 4 (1932-1934)	3733	2157	2263.6271	2.1660	0.000224
Ostrovskij, N.	How the steel was tempered. Chapter 5 (1932-1934)	3769	1882	1982.3710	1.8810	0.000137
Ostrovskij, N.	How the steel was tempered. Chapter 6 (1932-1934)	7534	3369	3518.8363	1.8108	0.000113
Ostrovskij, N.	How the steel was tempered. Chapter 7 (1932-1934)	6019	2972	3106.2350	1.9505	0.000143
Ostrovskij, N.	How the steel was tempered. Chapter 8 (1932-1934)	5352	2814	2927.0623	2.0392	0.000123
Ostrovskij, N.	How the steel was tempered. Chapter 9 (1932-1934)	3291	1761	1839.2573	1.9657	0.000296
Ostrovskij, N.	How the steel was tempered. Chapter 10 (1932-1934)	5399	2853	2995.4695	2.0708	0.000113
Pelevin, V.	Buben verchnego mira http://pelevin.nov.ru/rass/pe-bubv/1.html (2010)	3853	1792	1909.086	1.7767	0.000268

Tolstoy, L.	Boyhood (1854)	23126	6945	7967.887	1.5036	0.000046
Tolstoy, L.	Father Sergius (1898)	13538	4350	5066.62	1.5462	0.000074
Tolstoy, L.	Hadji Murat (1912)	34900	9457	11286.8	1.4692	0.000034
Tolstoy, L.	Metel’ (1856)	753	422	441.0422	1.685	0.000357
Tolstoy, L.	Sevastopol sketches (1955-1956)	35245	9577	11190.16	1.4437	0.000036
Tolstoy, L.	The death of Ivan Illych (1887)	17389	5016	5948.699	1.4506	0.000060
Turgenev, I.	Bežin lug http://quanta-textdata.uni-graz.at/ (2010)	6025	2536	2731.763	1.7138	0.000162
Dostoevskij, F.	Prestuplenie i nakazanie (p. I. ch. 1) (1866)	2595	1240	1356.702	1.785	0.000315
Gogol’, N.	Portret (1835)	17205	6073	6722.044	1.6549	0.000070

Table 26: Samoan prose

Source	**Text**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Tala o le Vavau.	O le mea na maua ai le ava. pp. 15 – 16 (1987)	1487	267	403.1697	0.8601	0.000631
Tala o le Vavau.	O le tala ia Sina ma lana tuna. pp. 17 – 19 (1987)	1171	222	303.9178	0.7964	0.000608
Tala o le Vavau.	O le tala ia Tamafaiga. pp. 49 – 52 (1987)	617	140	168.3850	0.7615	0.001115
Tala o le Vavau.	O le faalemigao. pp. 91 – 92 (1987)	736	153	214.1733	0.8343	0.001181
Tala o le Vavau.	O upu faifai ma le gaoi. p. 95 (1987)	447	124	149.4900	0.8863	0.002234

Table 27: Serbian

Translation (1949) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	Λ	***Var***(Λ)
1	4579	1899	2063.2763	1.6495	0.000204
2	4656	2082	2229.9239	1.7567	0.000151
3	7093	2852	3092.0841	1.6787	0.000150
4	4290	2129	2248.0887	1.9035	0.000218
5	4241	1877	2034.8342	1.7405	0.000185
6	8566	3237	3574.2307	1.6410	0.000103
7	6816	2941	3156.3392	1.7752	0.000110
8	6029	2823	3019.1269	1.8930	0.000158
9	3749	1787	1948.7604	1.8578	0.000202
10	6208	2816	3052.3141	1.8649	0.000158

Table 28a: Slovak poetry

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Belák, B.	Sihote. sihote (2008)	18052	6713	7224.2944	1.7034	0.000035
Buzássy, J.	V hline. v dreve. v piesni (1985)	14183	5443	5854.4083	1.7138	0.000058
Klas, T.	Hlas kňaza Gorazda (1999)	18102	5629	6379.8275	1.5006	0.000088
Richter, M.	Anjel s čiernym perím (2000)	14860	6366	6724.4670	1.8879	0.000078
Rúfus, M.	Dielo I (second half of the 20th century)	19881	7425	8055.2792	1.7416	0.000048

Rúfus, M.	Dielo II (second half of the 20th century)	11713	4556	4988.0799	1.7327	0.000103
Rúfus, M.	Dielo III (second half of the 20th century)	16002	5405	5933.4254	1.5589	0.000086
Sandtner, Š.	Ja som brána (2002)	14885	4512	5110.7771	1.4327	0.000100
Sandtner, Š.	Malí adorátori (1988)	12173	4198	4718.7741	1.5837	0.000063

Table 28b: Slovak prose

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Johanides, J.	Dívaj sa do modrých očí Londýna (2000)	22606	7637	8312.4671	1.6011	0.000034
Johanides, J.	Hmla našej trpezlivosti (2005)	18870	6215	6748.0917	1.5291	0.000079
Rakús, S.	Žobráci (1976)	25389	7082	7986.7272	1.3856	0.000050
Žarnay, J.	Časolet (1991)	27059	8852	9853.8771	1.6141	0.000050
Rakús, S.	Mačacia krajina (1985)	25364	7273	8094.0591	1.4055	0.000063

Table 28c: Slovak

Translation (1966) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4275	1895	2053.0342	1.7437	0.000205
2	4325	2068	2227.7413	1.8728	0.000236
3	6496	2864	3120.0108	1.8312	0.000145

4	3885	2087	2226.4162	2.0570	0.000142
5	3862	1862	2001.9842	1.8593	0.000306
6	8021	3292	3585.7280	1.7454	0.000116
7	6337	2937	3137.1683	1.8821	0.000134
8	5781	2771	2964.7250	1.9293	0.000145
9	3412	1757	1878.9532	1.9456	0.000226
10	5699	2818	2997.1106	1.9752	0.000164

Table 29a: Slovenian prose
http://quanta-textdata.uni-graz.at/ (2010)

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Cankar, I.	V temi	756	457	493.7225	1.8799	0.000564
Grum, S.	Vrata	1371	603	651.0865	1.4898	0.000563
Jurčič, J.	Sosedov sin (ch. I)	1966	907	990.9388	1.6601	0.000322
Kočevar, F.	Grof in menih	3491	1102	1404.1293	1.4250	0.000198
Levstik, F.	Zveženj	5588	2223	2385.3489	1.5996	0.000227

Table 29b: Slovenian

Translation (1966) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	5209	1955	2332.3012	1.6642	0.000143
2	5199	2098	2440.4324	1.7443	0.000174
3	7971	2944	3512.5388	1.7193	0.000261
4	4787	2199	2477.4815	1.9046	0.000137
5	4720	1929	2286.1240	1.7795	0.000165
6	9546	3354	4044.1405	1.6860	0.000137
7	7520	3038	3500.7281	1.8045	0.000106
8	6822	2955	3350.3265	1.8829	0.000106
9	4075	1874	2105.3648	1.8652	0.000171
10	6797	2920	3345.4684	1.8863	0.000152

Table 30: Sorbian

Translation (1960) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4851	1976	2184.2683	1.6596	0.000143
2	4812	2152	2333.5261	1.7857	0.000107
3	7395	2942	3219.3664	1.6843	0.000121
4	4483	2261	2460.8567	2.0045	0.000208

5	4272	1950	2098.9758	1.7838	0.000272
6	8795	3444	3752.3726	1.6828	0.000124
7	7058	3075	3325.8326	1.8136	0.000119
8	6316	2917	3119.5664	1.8771	0.000094
9	3850	1902	2014.7692	1.8763	0.000223
10	6648	2997	3228.2020	1.8563	0.000246

Table 31: Spanish prose
http://www.gutenberg.org/browse/languages/es

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Cervantes, M. de	Don Quijote I. Cap. 1 to 5 (1605)	11339	2369	2876.1276	1.0284	0.000168
Cervantes, M. de	Don Quijote I. Cap. 6 to 10	12090	2456	3019.0098	1.0194	0.000141
Cervantes, M. de	Don Quijote I. Cap. 11 to 20	33462	5017	6655.0132	0.8999	0.000034
Cervantes, M. de	Don Quijote I. Cap. 21 to 25	23533	3807	5037.8701	0.9359	0.000067
Cervantes, M. de	Don Quijote I. Cap. 26 to 30	24925	3773	5080.8472	0.8962	0.000069
Cervantes, M. de	Don Quijote I. Cap. 31 to 36	29691	4102	5771.2944	0.8694	0.000042
Cervantes, M. de	Don Quijote I. Cap. 37 to 44	33132	4399	6139.7123	0.8377	0.000048
Cervantes, M. de	Don Quijote I. Cap. 45 to 52	25353	4321	5553.2824	0.9647	0.000063
Cervantes, M. de	Don Quijote II. Cap. 1 to 10 (1615)	26598	4361	5640.4182	0.9383	0.000057
Cervantes, M. de	Don Quijote II. Cap. 11 to 20	30068	4721	6095.4096	0.9078	0.000052
Cervantes, M. de	Don Quijote II. Cap. 21 to 30	28503	4603	5888.5312	0.9204	0.000081

Cervantes, M. de	Don Quijote II. Cap. 31 to 40	26159	4321	5474.8305	0.9246	0.000061
Cervantes, M. de	Don Quijote II. Cap. 41 to 50	30866	4769	6221.6856	0.9049	0.000059
Cervantes, M. de	Don Quijote II. Cap. 51 to 60	29919	4669	6065.7842	0.9075	0.000033
Cervantes, M. de	Don Quijote II. Cap. 61 to 74	32653	4875	6271.0772	0.8669	0.000052
Cervantes, M. de	La Gitanilla. In: Novelas y teatro	12690	2857	3459.2561	1.1186	0.000098
Cervantes, M. de	La ilustre fregona. In: Novelas y teatro	11745	2457	3029.6657	1.0498	0.000084
Picón, J.O.	Lázaro (1882)	32534	7281	9239.2220	1.2814	0.000025

Table 32: Swedish prose

http://www.gutenberg.org/browse/languages/sv

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Strömberg, S.	Göteborgsflickor och andra historier (1919)	20162	4489	5182.3251	1.1064	0.000096
Strömberg, S.	Krigskorrespondenter och andra lögnare (1922)	24486	5120	5941.9357	1.0650	0.000073
Strömberg, S.	Baron Olson och andra historier (1919)	32740	7018	8261.7437	1.1394	0.000035
Zetterström, H.	Ada (1921)	24612	4353	5365.8795	0.9574	0.000090
Zetterström, H.	Anna-Clara och Hennes Bröder En Bok om Barn (1921)	22657	3797	4839.6939	0.9303	0.000079

Table 33a. Tagalog poetry

http://www.gutenberg.org/browse/languages/tl

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
De los Reyes, I.	Dalagang Marmol. Kay Liwayway (1912)	134	103	108.3254	1.7195	0.002863
Tolentino, A.	Dakilang Asal I. Paunawa (1907)	145	84	91.3324	1.3614	0.002496
Tolentino, A.	Dakilang Asal II. Katungkulan (1907)	381	205	219.6276	1.4878	0.000528
Tolentino, A.	Dakilang Asal III. Sa Paglilinis (1907)	193	133	141.3782	1.6742	0.003043
Tolentino, A.	Dakilang Asal IV. Sa Pagbibihis (1907)	361	185	200.7906	1.4225	0.003296
Tolentino, A.	Dakilang Asal V. Sa Pakikipanayam (1907)	337	181	191.8635	1.4391	0.001027
Tolentino, A.	Dakilang Asal VI Sa Mga Pagdalaw (1907)	1242	476	538.8603	1.3424	0.001096
Tolentino, A.	Dakilang Asal VII. Sa Mga Piging (1907)	289	146	155.3472	1.3228	0.002585
Tolentino, A.	Dakilang Asal VIII. Sa Mga Laro (1907)	146	100	101.8995	1.5106	0.003186
Tolentino, A.	Dakilang Asal IX. Sa Lansangan (1907)	669	293	318.0323	1.3432	0.001849

Table 33b. Tagalog prose

. http://www.seasite.niu.edu/. Tagalog/tagalog_short_stories_fs.htm
http://www.gutenberg.org/browse/languages/tI

Author	**Text, year**	***N***	***V***	***L***	Λ	***Var***(Λ)
Hernandez, A.V.	Limang Alas	1827	720	807.4634	1.4416	0.000580
Hernandez, A.V.	Magpinsan	1551	611	680.9899	1.4009	0.000726

Rosales, A.B.L.	Kristal Na Tubig	2054	645	748.5015	1.2072	0.000376
De los Reyes, I.	Dalagang Marmol (1912)	8768	2096	2600.5365	1.1694	0.000132
De los Reyes, I.	Dalagang Marmol I (1912)	1013	436	491.8435	1.4593	0.000606
De los Reyes, I.	Dalagang Marmol II (1912)	1418	573	641.0163	1.4247	0.000507
De los Reyes, I.	Dalagang Marmol III (1912)	2427	872	1013.0679	1.4130	0.000331
De los Reyes, I.	Dalagang Marmol V (1912)	1287	482	549.6678	1.3281	0.000502
De los Reyes, I.	Dalagang Marmol VI (1912)	1269	521	571.6024	1.3979	0.000502
De los Reyes, I.	Dalagang Marmol VII (1912)	719	331	367.6721	1.4608	0.001659
De los Reyes, I.	Dalagang Marmol VIII (1912)	634	305	335.1565	1.4813	0.001690
Almario, R.	Ang Mananayaw (1910)	6635	1731	2094.7002	1.2066	0.000091
Arsciwals, J.L.	Isa Pang Bayani (1915)	10900	2077	2692.2937	0.9972	0.000098
Nanong, B.B.	Hiwaga ng Pagibig (1922)	29625	3736	5481.8987	0.8274	0.000057
Rizal, J.	Ang Liham (1889)	3351	1163	1395.1505	1.4677	0.000170
Pascual, A.	Masakím (1910)	4329	1261	1527.0039	1.2827	0.000190

Table 34: Turkish prose

Word rank-frequencies by courtesy of Professor Fazli Can. canf@cs.bilkent.edu.tr

Author	**Text, year**	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
Adıvar, H.E.	Kalb Ağrısı (1924)	55980	14608	16820.1985	1.4266	0.000028
Ağaoğlu, A.	Ölmeye Yatmak (1973)	86291	25307	28305.7667	1.6191	0.000018
Akbal, O.	Garipler Sokağı (1950)	25200	8848	9631.5995	1.6822	0.000050

Ali, S.	Kürk Mantolu Madonna (1943)	39967	11547	13110.9317	1.5096	0.000043
Altan, A.	Kılıç Yarası Gibi (1998)	52393	18090	19707.0588	1.7751	0.000025
Altan, Ç.	Büyük Gözaltı (1972)	44023	12685	14107.9427	1.4881	0.000028
Arslanoğlu, K.	Devrimciler (1987)	74365	19904	22195.4630	1.4539	0.000028
Atay, O.	Tutunamayanlar (1972)	159707	37660	43183.7665	1.4069	0.000010
Atılgan, Y.	Aylak Adam (1959)	35929	11450	12479.9845	1.5823	0.000046
Baykurt, F.	Tırpan (1970)	88866	19730	21491.0765	1.1968	0.000019
Bilbaşar, K.	Denizin Çağırışı (1943)	29755	11508	12807.7712	1.9256	0.000038
Buğra, T.	İbiş'in Rüyası (1970)	48925	13851	15074.5378	1.4449	0.000024
Edgü, F.	O; Hâkkari'de Bir Mevsim (1977)	26140	8638	9595.4343	1.6215	0.000050
Enis, S.	Zaniyeler (1924)	45845	13458	15470.4241	1.5730	0.000033
Eroğlu, M.	Issızlığın Ortasında (1984)	71667	17791	20030.0797	1.3570	0.000018
Esendal, M.Ş.	Ayaşlı İle Kiracıları (1934)	47551	11424	12653.6256	1.2446	0.000056
Faik, S.	Medarı Maişet Motoru (1944)	36074	12010	13448.4841	1.6989	0.000044
Füruzan.	Kırkyedililer (1974)	116862	35097	37611.5735	1.6310	0.000013
Güntekin, R.N.	Çalıkuşu (1922)	90601	20836	24434.8558	1.3369	0.000018
Gürpınar, H.R.	Toraman (1919)	34479	12904	14021.9387	1.8453	0.000031
Gürsel, N.	Boğazkesen:Fatih'in Romanı (1995)	72947	20648	23229.5315	1.5486	0.000025
Hisar, A.Ş.	Fahim Bey ve Biz (1941)	34511	11611	12961.6988	1.7044	0.000034
İleri, S.	Ölüm İlişkileri (1979)	52805	17224	18916.1780	1.6918	0.000027
İlhan, A.	Haco Hanım Vay!.. (1984)	107624	31313	34287.7302	1.6031	0.000012
Karaosmanoğlu, Y.K.	Nur Baba (1922)	34605	10421	11900.1014	1.5609	0.000042

Karay, R.H.	İstanbul’un Bir Yüzü (1920)	39275	14258	15670.4547	1.8330	0.000033
Kemal, O.	Cemile (1952)	28250	9361	9919.3294	1.5629	0.000039
Kemal, Y.	İnce Memed (1955)	86996	17113	19627.4407	1.1144	0.000022
Kür, P.	Yarın Yarın (1976)	83845	21168	24102.0778	1.4153	0.000025
Pamuk, O.	Sessiz Ev (1983)	79223	18121	20251.2739	1.2523	0.000018
Rauf, M.	Eylül (1901)	68897	14823	17383.2003	1.2207	0.000028
Safa, P.	Peyami Safa. Sözde Kızlar (1925)	46211	13116	14482.3570	1.4619	0.000033
Seyfettin, O.	Efruz Bey (1919)	41825	14359	15659.4774	1.7303	0.000032
Tahir, K.	Yorgun Savaşçı (1965)	100328	26589	28518.5043	1.4217	0.000017
Tanpınar, A.H.	Huzur (1949)	98661	23419	27234.1092	1.3786	0.000017
Tekin, L.	Sevgili Arsız Ölüm (1983)	59010	12932	14848.9327	1.2005	0.000031
Uçuk, C.	Dikenli Çit (1937)	33752	9900	10856.0371	1.4565	0.000050
Uçuk, C.	Mithat Cemal Kuntay. Üç İstanbul (1938)	129541	27264	30232.1207	1.1931	0.000015
Uşaklıgil, H.Z.	Kırık Hayatlar (1924)	83450	19471	23760.7968	1.4013	0.000021
Yesari, M.	Tipi Dindi! (1933)	45874	13816	14940.7754	1.5182	0.000026

Table 35: Ukrainian

Translation (1974) of N. Ostrovskij´s *How the steel was tempered*

Chapter	***N***	***V***	***L***	***Λ***	***Var*(*Λ*)**
1	4119	1895	1991.2560	1.7475	0.000355
2	4160	2078	2151.7731	1.8720	0.000112

3	6282	2877	2986.4524	1.8056	0.000121
4	3764	2127	2184.1712	2.0749	0.000123
5	3755	1864	1928.8385	1.8362	0.000141
6	7542	3309	3435.3176	1.7662	0.000075
7	5999	2949	3076.5542	1.9376	0.000169
8	5362	2809	2897.3817	2.0152	0.000152
9	3278	1796	1856.3038	1.9909	0.000139
10	5351	2821	2932.7495	2.0435	0.000122

Author index

Subject index