Studies
in
Quantitative Linguistics
8

Ioan-Iovitz Popescu
Ján Mačutek
Emmerich Kelih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

Vectors and Codes of Text

RAM - Verlag

Vectors and codes of text

by

Ioan-Iovitz Popescu

in cooperation with

Ján Mačutek
Emmerich Kelih
Radek Čech
Karl-Heinz Best
Gabriel Altmann

2010
RAM-Verlag

Studies in quantitative linguistics

Editors
Fengxiang Fan (fanfengxiang@yahoo.com)
Emmerich Kelih (emmerich.kelih@uni-graz.at)
Reinhard Köhler (koehler@uni-trier.de)
Ján Mačutek (jmacutek@yahoo.com)
Eric S. Wheeler (wheeler@ericwheeler.ca)

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, II + 161 pp.

Gedruckt mit Unterstützung der Karl-Franzens-Universität Graz

ISBN: 978-3-942303-02-6

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

Preface

The present book is a continuation of our endeavour to introduce in textology new quantitative methods and evaluate some older ones (cf. Popescu et. al. 2009, Popescu, Mačutek, Altmann 2009; Tuzzi, Popescu, Altmann 2010). We illustrated the measurements and performed evaluations of texts from many languages. Needless to say, all results ensuing from the use of vectors, codes and chains must be tested on further languages and texts. Since this is ongoing empirical research, some modifications and adaptations of the methods presented may be necessary.

Nevertheless, the more we advanced the clearer we saw the abysmal playground hidden in texts. With some progress in using and elaborating quantitative methods in text analysis some new problems and formerly unrecognized phenomena appeared, thus we were confronted not only with methodological challenges, but also with new questions and problems in linguistic text theory in general.

We restricted ourselves to formal features (frequencies, codes and chains) accessible to the cooperating linguists and avoided sociolinguistic, psycholinguistic and other problems. We nevertheless hope that the methods presented could be made useful for other investigations, too.

The book consists of nine chapters. In Chapter 1 we introduce briefly the extensive domain of possible problems concerning comparisons and research strategies. In Chapters 2 to 6 we examine different vectors of texts, show their behaviour, compare texts and languages and take a step towards capturing the text dynamics looking at it from different points of view.

In Chapters 7 and 8 we goedelize the text in one special way, show breaks in the syntactic continuity in the text and ascribe to it its binary code which can be compared and tested.

The last chapter, Chapter 9, is devoted to chaining phenomena restricted here to Belza-Skorochoďko chains, revealing many new vistas touching behaviour, perseveration, psycholinguistic and other aspects. As a matter of fact, each topic could be developed infinitely but we strived for presenting simple methods, developed tests and showed a way of ternary plotting.

We hope that other scholars will adopt the methods for different purposes and for analyzing other languages, in order to get stronger corroboration of the procedures presented.

In this place we want to express our gratitude to Claudiu Vasilescu, who patiently wrote for us dilettantes all the Excel programmes and discretely concealed his amazement about our naivety. We had suffered much more without his kind help. Radek Čech was supported by the Czech Science Foundation, grant no. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach).

I.-I. Popescu

Contents

Preface

1. Introduction 1

2. The adjusted modulus 3
2.1. German data 4
2.2. Italian data 12
2.3. Slavic data 15
2.4. General data 22

3. The vector *T* 26
3.1. Stepwise and retrospective dissimilarity 26
3.2. Dispersion 38
3.3. Randomness 40
3.4. Prospective dissimilarity 42

4. Vectorial method of text comparison 53
4.1. Comparisons of texts 53
4.2. Cross-linguistic comparison 55
4.3 Vector distance 69

5. The ternary plot 76

6. Further simple methods for measuring the dynamics 95

7. The binary code of sentence 100
7.1. Goedelization 100
2.2. Breaks in the sequence 120

8. The binary code of text 124
8.1. The classical method 124
8.2. Other methods 127
8.3. Using the binary code 130

9. Belza-Skorochod'ko chaining 135

10. Conclusions 146

Appendix I. Texts used 147
References 156
Author index 159
Subject index 160

1. Introduction

"Not everything that can be counted counts, and
not everything that counts can be counted."
A. Einstein

The comparison of vocabularies of two texts in the same language can be performed in principle in two different ways:

(A) with regard to the identity of individual words,

(B) without regard to the identity of individual words.

In case (A) there are again two possibilities:

(A-1) The vocabularies of the two texts are considered sets and these sets are compared for similarity, with the frequencies of individual words ignored.

(A-2) The frequencies of individual words are taken into account as a kind of weight, and the weights of identical words are compared. This is the most common practice in quantitative lexicology (cf. Brunet 1988; Muller 1992; Labbé C., Labbé D. 2001, 2003, 2006; Labbé D. 2007; Merriam 2002; Rudman 1998; Tuldava 1971, 1998; Viprey, Ledoux 2006). The weights (= frequencies) are usually relativized because of different text lengths.

Needless to say, an analysis of type (A) can be practised only in texts of the same language. However, general textology is interested also in possible tendencies existing in all languages and must take into account some properties of the text for whose computation the identity of words is irrelevant.[1] Thus one must go beyond the level of lexicology and consider some abstract forms formed by the words of the text. There are several possibilities here, but two of them are quite conspicuous, namely the comparison of

(B-1) the rank-frequency sequence of words which can be considered either as a distribution or as a simple sequence, or of

(B-2) the frequency spectrum of words, where the random variable X is the occurrence number (= frequency) (x = 1,2,3,…) and $f(x)$ is the number of words having frequency x. This version can be attained by a simple transformation of (B-1).

In case (B-1) one takes into account the identity of ranks, in case (B-2) one takes into account the identity of occurrences. In all B-cases one can use for comparison some non-parametric tests, e.g. the chi-square, or one can reduce the data to some moments of the distributions and perform the comparison using Ord's scheme (Ord 1972) for which only the mean, the variance and the third central moment are necessary. Ord´s scheme is represented by the vector $<I, S> = <m_2/m_1', m_3/m_2>$, where m_i are the individual moments. In this way one can transcend the material base of the text but still take into account some rather

[1] In case of comparison of vocabularies of a text and its translation in another language there is seldom a one-to-one correspondence of words.

abstract properties of (B-1) and (B-2). Here we shall present another vector which can easily be computed for any text and any variant of (B).

The above-said shows that there is not only *direct* text comparison based on word identity; as a matter of fact, texts have an infinite number of properties all of which can in principle be quantified and their numerical forms compared. Even psychological/psycholinguistic or aesthetic properties have already been quantified (cf. e.g. Paivio, Yuille, Madigan 1968; Paivio 1971). Hence, there are different aspects of research for which text comparisons are necessary. Let us mention only some of them:

(a) Text unfolding, i.e. observing the dynamics of a property in the course of text;

(b) properties of genres, i.e. observing the common features of different texts even in different languages;

(c) style identity, used also in forensic linguistics but especially in music, concerning similar technical means used in different texts of the same author;

(d) historical development of texts in a language, i.e. the change of a property in the history of written texts, beginning from simple forms up to modern novels;

(e) ontogenetic development of texts in children;

(f) the speech of individual persons in a stage play;

(g) general textology surpassing the boundaries of individual persons, languages and epochs and using rather abstract properties.

All these approaches can be combined and must lead to the establishment of a special aspect of text theory.

Our procedure is rather explorative; we bring some results but are not always able to unveil the secrets of the background mechanisms whose existence must be assumed. However, the way of their operation is far from being known or even hypothesized. We try to go new ways offering new methods important for the description of individual texts or groups of texts rather than results. The tiresome work with text processing for different evaluations must be left to interested researchers specialized in individual domains.

Methodologically, our way in the depth of the text can be described in four steps. First, we consider it a whole and process it as a whole. Only a complete text contains the complete information. In the second step we reduce it to distributions of various entities and try to model them. Here we search for the genesis of attractors without the existence of which no communication is possible. Self-regulation is an intrinsic principle of language stability and this is warranted by the existence of attractors. In the third step, we reduce the properties of a certain attractor to a vector consisting of three components, study its form and compare texts. At last, in the fourth step, we reduce a property to a single number, the binary code of text, and show its applicability to different properties. Graphically, the procedure can be presented as follows:

Text

↓

Distribution = a ranked set of numbers

↓

Vector = three numbers

↓

Binary code = a single number

The binary code, though it is only a number, can be partitioned in a sum of numbers which reveal the given special structure of the text. Its study is not very advanced but here at least the first steps are made.

2. The adjusted modulus

In this part we restrict ourselves to capturing one of the aspects B-1 (cf. Introduction). We disregard the individuality of words in texts and consider only the rank-frequency sequence of word forms. It has been shown in many publications (cf. e.g. Popescu, Mačutek, Altmann 2009) that after stating the frequencies of word forms (or of other entities) there are three clearly determinable quantities, viz. $f(1)$ — the frequency of the most frequent word, V — the vocabulary size of different forms which is identical with the greatest rank, and the fixed point h which can be computed as

$$(2.1) \quad h = \begin{cases} r, & if\ there\ is\ an\ r = f(r) \\ \dfrac{f(i)r_j - f(j)r_i}{r_j - r_i + f(i) - f(j)}, & if\ there\ is\ no\ r = f(r) \end{cases}.$$

i.e. the h-point is that point at which $r = f(r)$. 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 $r_i + 1 = r_j$.

Using these three quantities we determine the vector

$$(2.2) \quad P = \left(\frac{f(1)}{h}, \frac{V}{h}\right)$$

and compute its length or modulus in the usual way as

$$(2.3) \quad M = \left(\left(\frac{f(1)}{h}\right)^2 + \left(\frac{V}{h}\right)^2\right)^{1/2} = \frac{1}{h}\left(f(1)^2 + V^2\right)^{1/2}.$$

All quantities in P are in some way associated with text size N (they increase with increasing N) and the dependence is visible but cannot be declared as significant because of great dispersion. However, if we divide the modulus M by $\log_{10} N$, i.e.

$$(2.4) \quad A = \frac{M}{log_{10} N},$$

the dependence disappears and we obtain a relatively constant property of the text, namely the adjusted relationship of the three conspicuous points, the adjusted modulus. Of course, even this indicator displays variation but this variation is rather due to style, genre or language. Its thorough study would surely be helpful in deciphering some background mechanisms of writing. Theoretically, its sampling properties cannot be derived because the sampling distribution of *V* is not known and (preliminarily) cannot be stated, while *N* is a constant and the properties of *f*(1) and *h* are known (cf. Mačutek, Popescu, Altmann 2007). Nevertheless, for each set of texts in one language and one genre, the empirical properties, e.g. mean, standard deviation, etc. may be determined.

2.1. German data

For further processing we present some commented results in Tables 2.1 to 2.7. In Table 2.1 the data necessary for the computation of *A* are presented. The German texts were taken from the Gutenberg Project which is accessible on the Internet.

Table 2.1
The adjusted modulus *A* of 253 German texts

ID	*N*	*V*	*f*(1)	*h*	*M*	*A*
Arnim 01	7846	2221	271	33	67.80	17.41
Arnim 02	1201	564	46	13	43.53	14.13
Arnim 03	4167	1429	189	26	55.44	15.32
Busch 01	15820	4642	527	44	106.18	25.29
Chamisso 01	2210	884	82	18	49.32	14.75
Chamisso 02	1847	808	84	16	50.77	15.54
Chamisso 03	1428	630	70	14	45.28	14.35
Chamisso 04	3205	1209	123	20	60.76	17.33
Chamisso 05	2108	853	79	18	47.59	14.32
Chamisso 06	1948	801	75	17	47.32	14.39
Chamisso 07	1362	670	44	13	51.65	16.48
Chamisso 08	1870	788	80	16	49.50	15.13
Chamisso 09	1320	593	96	14	42.91	13.75
Chamisso 10	1012	536	52	11	48.96	16.29
Chamisso 11	1386	656	66	14	47.09	14.99
Droste 01	16172	4064	525	49	83.63	19.87
Droste 02	884	492	48	9.62	51.39	17.44
Droste 03	700	425	31	9	47.35	16.64
Droste 04	786	408	34	10.5	38.99	13.47
Droste 05	1274	657	51	12.5	52.72	16.98

Droste 08	965	509	39	11	46.41	15.55
Eichendorff 01	3080	1079	177	21	52.07	14.93
Eichendorff 02	4100	1287	210	25	52.16	14.44
Eichendorff 03	4342	1334	182	28	48.08	13.22
Eichendorff 04	1781	739	79	16	46.45	14.29
Eichendorff 05	1680	699	70	16	43.91	13.61
Eichendorff 06	3223	1059	130	22	48.50	13.82
Eichendorff 07	2594	932	121	20	46.99	13.76
Eichendorff 08	3987	1320	159	25	53.18	14.77
Eichendorff 09	3285	1185	155	22	54.32	15.45
Eichendorff 10	3052	1073	131	22	49.13	14.10
Goethe 01	7554	2222	318	33	68.02	17.54
Goethe 05	559	332	30	8	41.67	15.17
Goethe 09	653	379	30	9	42.24	15.01
Goethe 10	480	301	18	7	43.08	16.07
Goethe 11	468	297	18	7	42.51	15.92
Goethe 12	251	169	14	6	28.26	11.78
Goethe 14	184	129	10	5	25.88	11.43
Goethe 17	225	124	11	6	20.75	8.82
Heine 01	19522	5769	939	46.5	125.70	29.30
Heine 02	603	361	50	8.5	42.88	15.42
Heine 03	394	211	21	7	30.29	11.67
Heine 04	20107	5305	946	46.5	115.89	26.93
Heine 07	263	169	17	5	33.97	14.04
Hoffmann 01	2974	1176	95	22	53.63	15.44
Hoffmann 02	1076	534	29	11	48.62	16.04
Hoffmann 03	8163	2511	290	34	74.34	19.00
Immermann 01	28943	6397	918	63	102.58	22.99
Kafka 01	10256	2321	448	41	57.65	14.37
Kafka 02	3181	1210	159	22.5	54.24	15.49
Kafka 03	1072	513	34	12.33	41.70	13.76
Kafka 04	625	321	23	9.5	33.88	12.12
Kafka 05	247	166	14	5	33.32	13.92
Kafka 06	178	137	6	4	34.28	15.23
Kafka 07	132	89	9	3.66	24.44	11.53
Kafka 08	139	102	9	3.5	29.26	13.65
Kafka 09	596	343	25	9	38.21	13.77
Kafka 10	86	62	4	4	15.53	8.03
Kafka 11	151	104	9	4.5	23.20	10.65
Kafka 12	160	101	9	5	20.28	9.20

Kafka 13	232	150	9	6	25.04	10.59
Kafka 14	142	104	11	3	34.86	16.20
Kafka 15	189	136	7	4.5	30.26	13.29
Kafka 16	255	177	10	6	29.55	12.28
Kafka 17	111	80	11	3	26.92	13.16
Kafka 18	61	48	3	2.5	19.24	10.78
Kafka 19	41	33	3	2	16.57	10.27
Kafka 20	1402	539	74	14.75	36.89	11.72
Kafka 21	610	364	18	9.5	38.36	13.77
Kafka 22	2129	887	89	18.33	48.63	14.61
Kafka 23	255	153	13	6	25.59	10.63
Kafka 24	584	276	25	8.5	32.60	11.79
Kafka 25	3414	1214	104	23	52.98	14.99
Kafka 26	134	98	7	3.5	28.07	13.20
Kafka 27	428	240	14	8	30.05	11.42
Kafka 28	470	272	13	8	34.04	12.74
Keller 01	25625	5516	1399	59	96.45	21.88
Keller 02	301	196	20	5	39.40	15.90
Keller 03	13149	3512	724	43	83.39	20.25
Keller 04	1896	897	103	15	60.19	18.36
Lessing 01	114	78	7	4	19.58	9.52
Lessing 02	208	141	13	4	35.40	15.27
Lessing 03	61	48	4	2.5	19.27	10.79
Lessing 04	47	41	2	2	20.52	12.27
Lessing 05	182	120	7	4.5	26.71	11.82
Lessing 06	362	227	13	7	32.48	12.69
Lessing 07	231	161	9	4	40.31	17.06
Lessing 08	74	64	4	2	32.06	17.15
Lessing 09	327	193	24	6	32.41	12.89
Lessing 10	254	154	12	6	25.74	10.71
Löns 01	1672	706	95	15	47.49	14.73
Löns 02	2988	928	141	23	40.81	11.74
Löns 03	4063	1162	172	26	45.18	12.52
Löns 04	3713	1081	167	24	45.58	12.77
Löns 05	4676	1235	254	28	45.03	12.27
Löns 06	4833	1364	244	29	47.78	12.97
Löns 07	7743	1862	414	36	52.99	13.62
Löns 08	6093	1724	328	31	56.61	14.96
Löns 09	9252	2126	453	39	55.74	14.05
Löns 10	6546	1736	274	35	50.21	13.16

Löns 11	4102	1294	217	27	48.60	13.45
Löns 12	4432	1318	221	26	51.40	14.10
Löns 13	1361	556	60	14	39.94	12.75
Meyer 01	1523	801	56	14	57.35	18.02
Meyer 02	573	331	26	8	41.50	15.05
Meyer 03	1052	551	46	11	50.27	16.63
Meyer 04	2550	1142	79	18	63.60	18.67
Meyer 05	1249	658	47	12	54.97	17.75
Meyer 06	833	471	34	10	47.22	16.17
Meyer 07	1229	652	47	13	50.28	16.28
Meyer 08	1028	556	43	11	50.70	16.83
Meyer 09	776	441	40	9	49.20	17.03
Meyer 10	940	493	41	11	44.97	15.13
Meyer 11	2398	1079	88	17	63.68	18.84
Novalis 01	2894	1129	139	21	54.17	15.65
Novalis 02	3719	1487	208	22	68.25	19.12
Novalis 03	5321	1819	233	25	73.35	19.69
Novalis 04	2777	1282	130	18	71.59	20.79
Novalis 05	8866	2769	473	35	80.26	20.33
Novalis 06	4030	1467	178	23	64.25	17.82
Novalis 07	1744	792	77	16	49.73	15.34
Novalis 08	2111	816	75	17	48.20	14.50
Novalis 09	8945	2681	442	32	84.91	21.49
Novalis 10	5367	1939	238	26	75.14	20.15
Novalis 11	1358	646	83	11.66	55.86	17.83
Novalis 12	4430	1697	195	24	71.17	19.52
Novalis 13	1080	514	58	12.33	41.95	13.83
Paul 01	854	487	37	10	48.84	16.66
Paul 02	383	255	14	6	42.56	16.48
Paul 03	520	311	26	8	39.01	14.36
Paul 04	580	354	21	8	44.33	16.04
Paul 05	1331	677	44	12	56.54	18.10
Paul 06	526	305	16	8	38.18	14.03
Paul 07	508	316	15	7	45.19	16.70
Paul 08	402	248	22	6	41.50	15.93
Paul 09	1068	547	37	10	54.82	18.10
Paul 10	1558	778	53	13	59.98	18.79
Paul 11	2232	1027	84	15	68.70	20.51
Paul 12	620	365	25	8	45.73	16.38
Paul 13	1392	652	40	13	50.25	15.98

Paul 14	1400	714	49	14	51.12	16.25
Paul 15	1648	793	65	15	53.04	16.49
Paul 16	320	223	12	5	44.66	17.83
Paul 17	1844	897	73	15	60.00	18.37
Paul 18	870	489	42	11	44.62	15.18
Paul 19	1236	676	38	13	52.08	16.84
Paul 20	2059	1011	78	16	63.38	19.13
Paul 21	3955	1513	172	24	63.45	17.64
Paul 22	478	302	15	7	43.20	16.12
Paul 23	656	386	26	9	42.99	15.26
Paul 24	1465	730	80	13	56.49	17.84
Paul 25	588	361	18	8	45.18	16.31
Paul 26	1896	887	61	15	59.27	18.08
Paul 27	749	410	26	9	45.65	15.88
Paul 28	241	172	8	5	34.44	14.46
Paul 29	1825	872	68	14	62.47	19.16
Paul 30	388	238	17	6	39.77	15.36
Paul 31	1630	753	72	14	54.03	16.82
Paul 32	163	119	6	4	29.79	13.47
Paul 33	596	355	23	8	44.47	16.02
Paul 34	5	5	1	1	5.10	7.30
Paul 35	1947	897	82	17	52.98	16.11
Paul 36	425	253	15	7	36.21	13.78
Paul 37	368	239	12	6	39.88	15.54
Paul 38	1218	636	40	12	53.10	17.21
Paul 39	388	248	13	7	35.48	13.70
Paul 40	1370	655	53	14	46.94	14.96
Paul 41	1032	546	43	11	49.79	16.52
Paul 42	1546	731	50	13	56.36	17.67
Paul 43	4148	1591	152	26	61.47	16.99
Paul 44	1881	896	66	15	59.90	18.29
Paul 45	2723	1102	155	18	61.82	18.00
Paul 46	3095	1276	99	21	60.94	17.46
Paul 47	516	319	19	8	39.95	14.73
Paul 48	1200	604	50	13	46.62	15.14
Paul 49	562	336	19	8	42.07	15.30
Paul 50	430	255	23	7	36.58	13.89
Paul 51	3222	1323	116	20	66.40	18.93
Paul 52	1731	815	71	15	54.54	16.84
Paul 53	1839	864	75	14	61.95	18.98

Paul 54	6644	2417	245	30	80.98	21.19
Paul 55	7854	2680	321	33	81.79	21.00
Paul 56	963	482	47	10	48.43	16.23
Pseudonym 01	728	363	30	10	36.42	12.73
Pseudonym 02	612	326	23	9	36.31	13.03
Raabe 01	13045	3003	691	45	68.48	16.64
Raabe 02	3173	962	134	23	42.23	12.06
Raabe 03	2690	950	135	21	45.69	13.32
Raabe 04	6253	2110	282	30	70.96	18.69
Raabe 05	5087	1801	196	26	69.68	18.80
Rieder 01	1161	510	36	12	42.61	13.90
Rieder 02	1231	472	55	13	36.55	11.83
Rückert 01	141	97	10	4	24.38	11.34
Rückert 02	327	202	9	7	28.89	11.49
Rückert 03	152	107	8	4	26.82	12.29
Rückert 04	721	412	22	9	45.84	16.04
Rückert 05	212	145	10	5	29.07	12.50
Schnitzler 01	2793	961	109	19.5	49.60	14.39
Schnitzler 02	1936	825	59	17	48.65	14.80
Schnitzler 03	801	410	28	11	37.36	12.87
Schnitzler 04	2489	870	135	20.67	42.59	12.54
Schnitzler 05	2123	822	110	17.67	46.93	14.11
Schnitzler 06	1539	668	50	14.5	46.20	14.49
Schnitzler 07	5652	1451	259	31.25	47.17	12.57
Schnitzler 08	1711	666	63	14.62	45.76	14.15
Schnitzler 09	6552	1993	207	31.73	63.15	16.55
Schnitzler 10	1349	629	49	14.5	43.51	13.90
Schnitzler 11	1595	723	97	15	48.63	15.18
Schnitzler 12	6173	1476	400	31	49.33	13.01
Schnitzler 13	1184	544	44	13	41.98	13.66
Schnitzler 14	3900	1309	139	25.5	51.62	14.38
Sealsfield 01	1352	600	45	13	46.28	14.78
Sealsfield 02	4663	1825	142	27	67.80	18.48
Sealsfield 03	3238	1197	114	21	57.26	16.31
Sealsfield 04	3954	1399	161	24	58.68	16.31
Sealsfield 05	3187	1079	96	22	49.24	14.05
Sealsfield 06	2586	1010	67	20	50.61	14.83
Sealsfield 07	2939	1035	75	20	51.89	14.96
Sealsfield 08	4865	1333	138	27	49.63	13.46
Sealsfield 09	7259	2295	263	31	74.52	19.30

Sealsfield 10	4838	1620	138	26	62.53	16.97
Sealsfield 11	3785	1265	98	26	48.80	13.64
Sealsfield 12	3019	1191	95	20	59.74	17.17
Sealsfield 13	2370	1071	89	17	63.22	18.73
Sealsfield 14	2744	1198	82	19	63.20	18.38
Sealsfield 15	4786	1545	164	27	57.54	15.64
Sealsfield 16	4497	1602	137	26	61.84	16.93
Sealsfield 17	6705	2273	192	30	76.04	19.87
Sealsfield 18	4162	1252	285	24	53.50	14.78
Sealsfield 19	5626	1653	171	29	57.30	15.28
Sealsfield 20	8423	2735	273	35	78.53	20.01
Sealsfield 21	6041	2040	220	29	70.75	18.71
Sealsfield 22	5748	1655	157	29	57.33	15.25
Sealsfield 23	1752	799	80	14	57.36	17.68
Sealsfield 24	1696	753	68	14	54.00	16.72
Sealsfield 25	1368	704	40	12	58.76	18.74
Sealsfield 26	1517	679	44	15	45.36	14.26
Sealsfield 27	4195	1516	179	24	63.61	17.56
Sealsfield 28	1515	586	70	15	39.34	12.37
Storm 01	38306	6233	1292	76	83.76	18.27
Sudermann 01	11437	2427	507	43	57.66	14.21
Tucholsky 01	8544	2449	351	35	70.69	17.98
Tucholsky 02	7106	1935	207	35	55.60	14.44
Tucholsky 03	9699	2502	336	38	66.43	16.66
Tucholsky 04	7415	1968	214	35	56.56	14.61
Tucholsky 05	4823	1399	174	28	50.35	13.67
Wedekind 01	4035	1336	122	26	51.60	14.31
Wedekind 02	6040	1731	179	31	56.14	14.85
Wedekind 03	7402	1934	276	34	57.46	14.85
Wedekind 04	1297	646	44	13	49.81	16.00
Wedekind 05	1935	580	89	19	30.88	9.40
Wedekind 06	5955	1689	249	34	50.21	13.30
Wedekind 07	605	341	22	9	37.97	13.65
Wedekind 08	2033	855	87	17	50.55	15.28

The titles of the individual texts are shown in Appendix I. The texts are of fictional or poetic character. We assume that A is related to some other textual properties but the limitations on the size of the present report do not allow us to set up hypotheses.

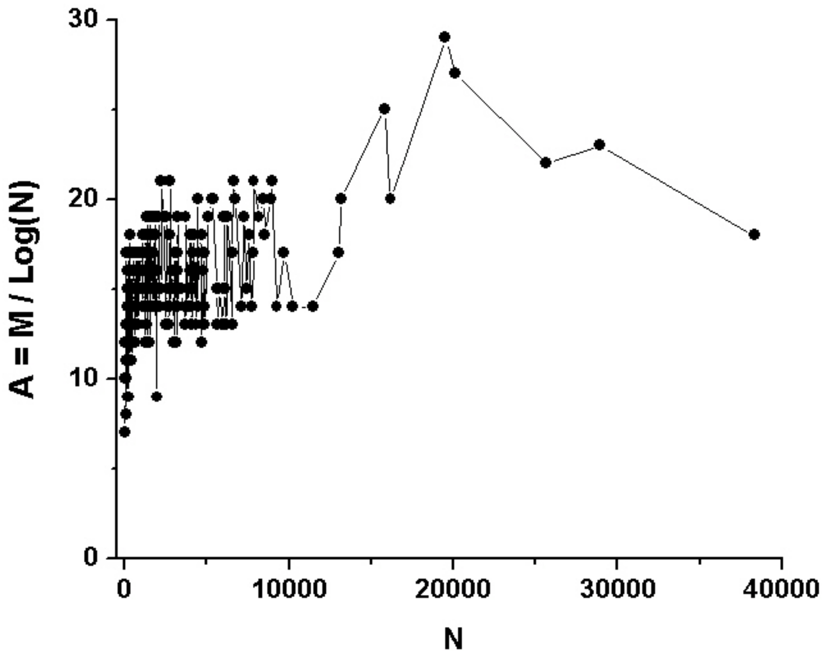


Figure 2.1. The adjusted modulus in 253 German texts

If ordered according to *N*, the adjusted modulus *A* yields for German texts a very compact picture with a small number of real outliers which may be caused by stylistic peculierities. The mean is μ_G = 15.43, *sd* = 2.9302. In any case it would be possible to set up 95% or 99% confidence intervals in order to study the mechanisms in texts outside of this interval. It can be expected that press texts or scientific texts will be quite different.

2.2. Italian data

Let us consider now texts of the same genre, namely the end-of-year speeches of Italian presidents (cf. Tuzzi, Popescu, Altmann 2009a,b). Here, not only the genre but also the content concerns the same universe of discourse. Though the interests and views of individual presidents must necessarily differ and change, they speak about present-day problems of Italy. The results are presented in Table 2.2.

Table 2.2
The adjusted modulus of 60 Italian presidential End-of-year-speeches

ID	***N***	***V***	***f*(1)**	***h***	***M***	***A***
1949Einaudi	194	140	10	5	28.07	12.27
1950Einaudi	150	105	9	4	26.35	12.11
1951Einaudi	230	169	9	5	33.85	14.33
1952Einaudi	179	145	7	4	36.29	16.11
1953Einaudi	190	143	8	4	35.81	15.71
1954Einaudi	260	181	12	5	36.28	15.02
1955Gronchi	388	248	16	7	37.31	14.41
1956Gronchi	665	374	29	8	46.89	16.61
1957Gronchi	1130	549	65	12	46.07	15.09
1958Gronchi	886	460	41	11	41.98	14.24
1959Gronchi	697	388	33	9	43.27	15.22
1960Gronchi	804	434	41	10	43.59	15.00
1961Gronchi	1252	622	67	13	48.12	15.54
1962Segni	738	381	35	10	38.26	13.34
1963Segni	1057	527	46	12	45.37	15.00
1964Saragat	465	278	21	8	34.85	13.06
1965Saragat	1052	510	52	12	43.97	14.55
1966Saragat	1200	597	44	13	47.89	15.55
1967Saragat	1056	526	51	11	48.04	15.89
1968Saragat	1173	562	56	13	43.44	14.15
1969Saragat	1583	692	86	15	46.49	14.53
1970Saragat	1929	812	85	17	49.48	15.06
1971Leone	262	168	12	5	33.69	13.93
1972Leone	767	394	32	10	41.61	14.42
1973Leone	1250	616	67	12	51.64	16.67
1974Leone	801	426	32	9	47.47	16.35
1975Leone	1328	632	63	13	48.86	15.64
1976Leone	1366	649	52	13	50.08	15.97
1977Leone	1604	717	80	14	51.53	16.08
1978Pertini	1492	603	53	14	42.24	13.31
1979Pertini	2311	800	70	18	44.61	13.26
1980Pertini	1360	535	50	14	39.08	12.47
1981Pertini	2819	911	96	20	45.80	13.28
1982Pertini	2486	854	90	19	45.20	13.31
1983Pertini	3746	1149	118	24	48.82	13.66

1984Pertini	1340	514	42	14	37.75	12.07
1985Cossiga	2359	859	118	17	51.00	15.12
1986Cossiga	1348	561	65	14	40.34	12.89
1987Cossiga	2092	904	109	15	60.70	18.28
1988Cossiga	2384	875	123	19	46.51	13.77
1989Cossiga	1912	778	85	17	46.04	14.03
1990Cossiga	3345	1222	155	20	61.59	17.48
1991Cossiga	418	241	22	7	34.57	13.19
1992Scalfaro	2774	978	118	18	56.29	16.35
1993Scalfaro	2942	1074	129	19	58.16	16.77
1994Scalfaro	3606	1190	171	21	57.25	16.09
1995Scalfaro	4233	1341	180	23	59.71	16.46
1996Scalfaro	2085	866	88	16	54.40	16.39
1997Scalfaro	5012	1405	167	28	51.45	13.91
1998Scalfaro	3995	1175	137	24	50.34	13.98
1999Ciampi	1941	831	66	17	50.52	15.37
2000Ciampi	1844	822	70	16	51.56	15.79
2001Ciampi	2098	898	89	18	50.13	15.09
2002Ciampi	2129	909	96	17	53.77	16.16
2003Ciampi	1565	718	63	14	51.48	16.12
2004Ciampi	1807	812	76	15	54.37	16.69
2005Ciampi	1193	538	54	13	42.71	13.88
2006Napolitano	2204	929	125	17	56.81	16.99
2007Napolitano	1792	793	101	16	49.96	15.36
2008Napolitano	1713	775	75	15	51.91	16.05

In the Italian texts the mean A is μ_A = 14.92 and the standard deviation is sd = 1.4234, both smaller than in German texts. As can be seen in Figure 2.2, the indicator A is almost constant, with deviations caused by style, not by text size. Not even historically, i.e. ordering the Presidents chronogically, a trend can be observed. This is the first hint concerning the impact of theme on the repetitive strutucre of words.

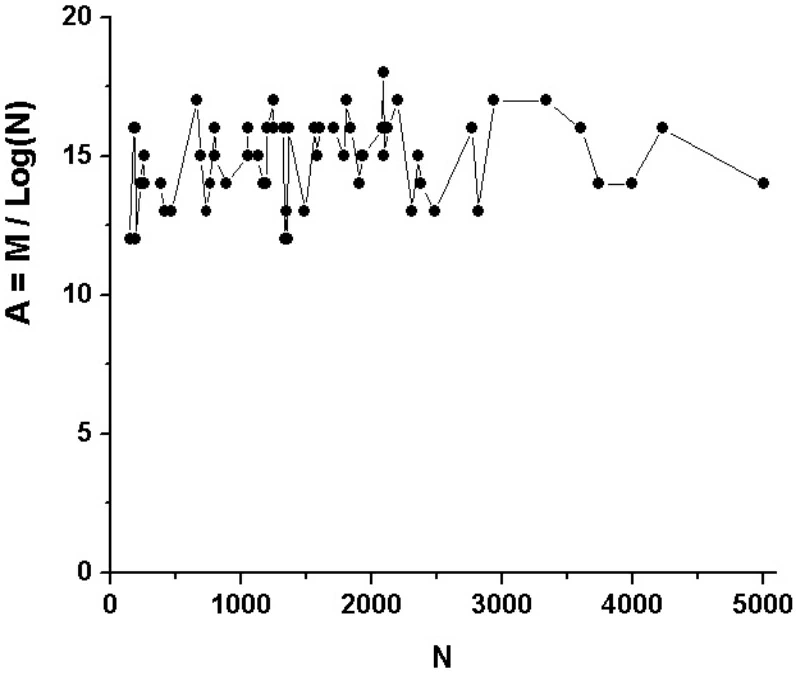


Figure 2.2. Adjusted modulus of 60 Italian presidential End-of-year Speeches

2.3. Slavic data

Now we compare the adjusted modulus in 12 Slavic languages based on the translation of the same text from Russian (N. Ostrovskij, “How the steel was tempered”). The modulus has been computed for each of the first ten chapters separately. For the computations, E. Kelih´s (2009) special corpus has been used. The results are presented in Table 2.3 and Figure 2.3. Here and below Bel = Belorussian, Bul = Bulgarian, Cro = Croatian, Cze = Czech, Mac = Macedonian, Pol = Polish, Rus = Russian, Ser = Serbian, Slk = Slovak, Sln = Slovenian, Sor = Sorbian, Ukr = Ukrainian.

Table 2.3
Adjusted modulus in 12 Slavic languages (same text)

ID	***N***	***V***	***f*(1)**	***h***	***M***	***A***
Bel 01	4145	1916	175	19	101.26	27.99
Bel 02	4177	2079	153	17	122.62	33.87
Bel 03	6367	2863	219	24	119.64	31.45
Bel 04	3791	2116	129	17.33	122.33	34.18

Bel 05	3791	1854	125	18.5	100.44	28.07
Bel 06	7547	3347	186	25	134.09	34.58
Bel 07	6063	2953	158	24	123.22	32.57
Bel 08	5362	2783	146	22.2	125.53	33.66
Bel 09	3312	1776	94	18.24	97.50	27.70
Bel 10	5319	2814	147	21.33	132.11	35.46
Bul 01	4653	1709	194	23	74.78	20.39
Bul 02	4734	1913	170	21.62	88.83	24.17
Bul 03	7224	2581	273	28	92.69	24.02
Bul 04	4305	2007	155	20	100.65	27.70
Bul 05	4277	1706	150	23	74.46	20.51
Bul 06	8673	2979	280	31	96.52	24.51
Bul 07	6992	2729	289	25	109.77	28.55
Bul 08	6242	2591	235	22	118.26	31.16
Bul 09	3787	1663	129	20.25	82.37	23.02
Bul 10	6278	2633	260	23.33	113.41	29.86
Cro 01	4582	1900	192	21	90.94	24.84
Cro 02	4689	2096	174	20	105.16	28.65
Cro 03	7160	2888	281	27	107.47	27.88
Cro 04	4316	2149	149	19.21	112.14	30.85
Cro 05	4255	1881	183	19.5	96.92	26.71
Cro 06	8553	3222	366	28.5	113.78	28.94
Cro 07	6841	2958	247	24	123.68	32.25
Cro 08	6075	2845	229	22	129.74	34.29
Cro 09	3760	1795	183	19.33	93.34	26.11
Cro 10	6184	2823	254	22.66	125.08	32.99
Cze 01	3925	1773	180	20.33	87.66	24.39
Cze 02	4381	2109	183	17	124.52	34.20
Cze 03	6670	2904	309	25	116.82	30.55
Cze 04	3920	2111	183	16	132.43	36.86
Cze 05	3852	1854	163	19	97.96	27.32
Cze 06	8117	3369	329	28.5	118.77	30.38
Cze 07	6390	2945	254	24	123.16	32.36
Cze 08	5738	2805	216	21.33	131.89	35.09
Cze 09	3451	1820	142	17	107.38	30.35
Cze 10	5736	2891	219	20	144.96	38.57
Mac 01	4810	1636	193	23.62	69.74	18.94
Mac 02	4898	1836	184	22.25	82.93	22.47
Mac 03	7470	2456	283	30.73	80.45	20.77
Mac 04	4424	1937	157	21.5	90.39	24.79
Mac 05	4425	1667	155	23.66	70.76	19.41
Mac 06	8914	2842	316	31.5	90.78	22.98

Mac 07	7153	2606	314	26	100.96	26.19
Mac 08	6414	2484	282	25	100.00	26.27
Mac 09	3850	1610	146	23.5	68.79	19.19
Mac 10	6461	2536	325	24.6	103.93	27.28
Pol 01	4348	1970	160	18.8	105.13	28.90
Pol 02	4368	2149	149	19	113.38	31.15
Pol 03	6694	2995	227	24.24	123.91	32.39
Pol 04	4003	2200	131	16	137.74	38.24
Pol 05	3997	1962	138	18.33	107.30	29.79
Pol 06	7937	3481	273	23	151.81	38.93
Pol 07	6348	3061	196	21	146.06	38.41
Pol 08	5753	2928	172	19	154.37	41.06
Pol 09	3501	1855	113	17.5	106.20	29.96
Pol 10	5786	2970	165	20	148.73	39.53
Rus 01	4107	1907	169	20	95.72	26.49
Rus 02	4136	2088	152	18.5	113.16	31.29
Rus 03	6323	2909	213	24.8	117.61	30.94
Rus 04	3733	2157	127	17	127.10	35.58
Rus 05	3769	1882	125	19	99.27	27.76
Rus 06	7534	3369	183	26.33	128.14	33.05
Rus 07	6019	2972	164	24	124.02	32.81
Rus 08	5352	2814	140	20.75	135.78	36.42
Rus 09	3291	1761	99	18.25	96.65	27.48
Rus 10	5399	2853	169	23.5	121.62	32.58
Ser 01	4579	1899	191	20.73	92.07	25.15
Ser 02	4656	2082	173	20	104.46	28.48
Ser 03	7093	2852	273	27	106.11	27.56
Ser 04	4290	2129	142	20	106.69	29.37
Ser 05	4241	1877	184	19	99.26	27.36
Ser 06	8566	3237	373	28	116.37	29.59
Ser 07	6816	2941	246	25	118.05	30.79
Ser 08	6029	2823	224	21.5	131.72	34.84
Ser 09	3749	1787	184	18.5	97.11	27.17
Ser 10	6208	2816	263	22.5	125.70	33.14
Slk 01	4275	1895	185	21.5	88.56	24.39
Slk 02	4325	2068	183	20	103.80	28.55
Slk 03	6496	2864	289	27.5	104.67	27.45
Slk 04	3885	2087	162	16.5	126.87	35.34
Slk 05	3862	1862	163	20.25	92.30	25.73
Slk 06	8021	3292	328	27	122.53	31.38
Slk 07	6337	2937	231	25.25	116.68	30.69
Slk 08	5781	2771	222	24	115.83	30.79

Slk 09	3412	1757	144	18.5	95.29	26.97
Slk 10	5699	2818	206	22	128.43	34.20
Sln 01	5209	1955	409	24	83.22	22.39
Sln 02	5199	2098	372	22	96.85	26.06
Sln 03	7971	2944	604	28	107.33	27.51
Sln 04	4787	2199	306	21	105.72	28.73
Sln 05	4720	1929	386	24	81.97	22.31
Sln 06	9546	3354	730	32	107.27	26.95
Sln 07	7520	3038	498	26	118.41	30.55
Sln 08	6822	2955	429	27	110.59	28.85
Sln 09	4075	1874	258	21	90.08	24.95
Sln 10	6797	2920	457	26	113.67	29.66
Sor 01	4851	1976	237	22	90.46	24.54
Sor 02	4812	2152	209	21	102.96	27.96
Sor 03	7395	2942	312	26	113.79	29.41
Sor 04	4483	2261	224	20	113.60	31.11
Sor 05	4272	1950	174	20.33	96.30	26.52
Sor 06	8795	3444	346	28.67	120.73	30.61
Sor 07	7058	3075	282	23.5	131.40	34.14
Sor 08	6316	2917	231	21.5	136.10	35.81
Sor 09	3850	1902	136	18.5	103.07	28.75
Sor 10	6648	2997	260	25	120.33	31.48
Ukr 01	4119	1895	120	19	99.94	27.65
Ukr 02	4160	2078	99	18	115.58	31.93
Ukr 03	6282	2877	140	22.67	127.06	33.45
Ukr 04	3764	2127	80	16.75	127.07	35.54
Ukr 05	3755	1864	89	17.33	107.68	30.12
Ukr 06	7542	3309	160	25	132.51	34.18
Ukr 07	5999	2949	157	23	128.40	33.99
Ukr 08	5362	2809	114	20.33	138.28	37.08
Ukr 09	3278	1796	82	16.4	109.63	31.18
Ukr 10	5351	2821	139	21	134.50	36.07

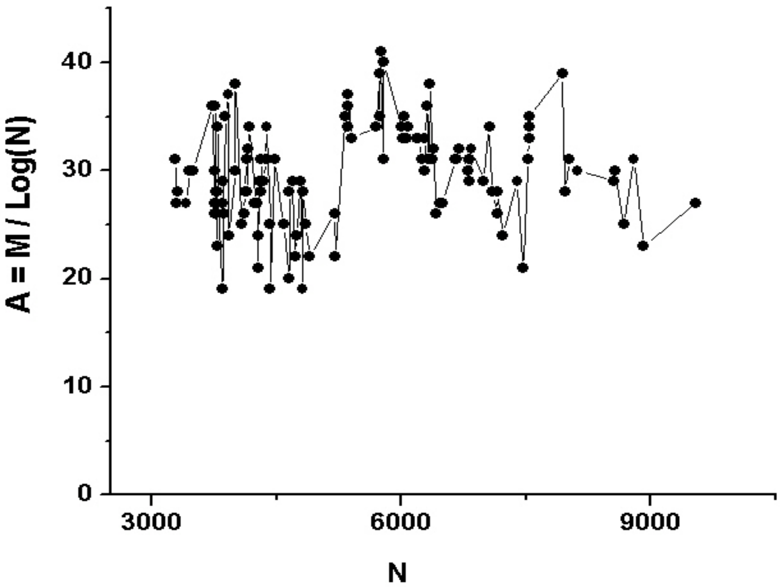


Figure 2.3. Adjusted modulus in 12 Slavic languages

Again, the adjusted modulus *A* is constant, but Figure 2.3 presents both a mixture of languages and a mixture of chapters, hence the dispersion is very great. In order to disentangle the oscillation, in Table 2.4 the values of *A* are presented chapterwise and languagewise.

Table 2.4
Indicator *A* of ten chapters of the same text in 12 Slavic languages

Chapter	Bel	Bul	Cro	Cze	Mac	Pol	Rus	Ser	Slk	Sln	Sor	Ukr
1	27.99	20.39	24.84	24.39	18.94	28.90	26.49	25.15	24.39	22.39	24.54	27.65
2	33.87	24.17	28.65	34.20	22.47	31.15	31.29	28.48	28.55	26.06	27.96	31.93
3	31.45	24.02	27.88	30.55	20.77	32.39	30.94	27.56	27.45	27.51	29.41	33.45
4	34.18	27.70	30.85	36.86	24.79	38.24	35.58	29.37	35.34	28.73	31.11	35.54
5	28.07	20.51	26.71	27.32	19.41	29.79	27.76	27.36	25.73	22.31	26.52	30.12
6	34.58	24.51	28.94	30.38	22.98	38.93	33.05	29.59	31.38	26.95	30.61	34.18
7	32.57	28.55	32.25	32.36	26.19	38.41	32.81	30.79	30.69	30.55	34.14	33.99
8	33.66	31.16	34.29	35.09	26.27	41.06	36.42	34.84	30.79	28.85	35.81	37.08
9	27.70	23.02	26.11	30.35	19.19	29.96	27.48	27.17	26.97	24.95	28.75	31.18
10	35.46	29.86	32.99	38.57	27.28	39.53	32.58	33.14	34.20	29.66	31.48	36.07

If one plots the individual points in the given order, one obtains the results as shown in Figure 2.4.

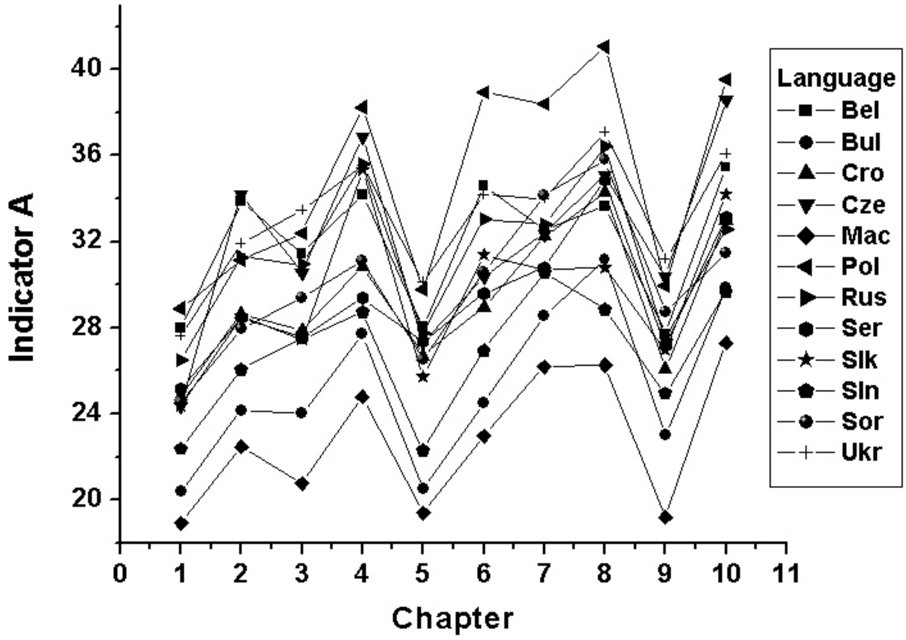


Figure 2.4. Indicator *A* in 10 chapters of the same text in 12 Slavic languages

It can easily be seen that the curves are almost parallel. This agreement can be due to the same content, and the difference in level can be due to (morphological, morphosyntactic) differences between Slavic languages. If we take the means of *A* and the standard deviations of individual languages, we obtain the results presented in Table 2.5. The 95% confidence interval for the mean of each language is computed as

$$\bar{A} \pm 1.96s/\sqrt{12} = \bar{A} \pm 1.96 s_{\bar{A}}$$

Table 2.5
Some sampling properties of Slavic languages

Language	$\bar{A}$	s	$s_{\bar{A}}$	Interval	
Pol	34.836	4.783	1.381	32.129	37.543
Ukr	33.119	2.912	0.841	31.471	34.767
Cze	32.007	4.328	1.249	29.558	34.456
Bel	31.953	2.986	0.862	30.263	33.643
Rus	31.440	3.362	0.970	29.538	33.342
Sor	30.033	3.378	0.975	28.122	31.944
Slk	29.549	3.569	1.030	27.530	31.568
Cro	29.351	3.144	0.907	27.572	31.130

Ser	29.345	2.924	0.844	27.691	30.999
Sln	26.796	2.873	0.829	25.171	28.421
Bul	25.389	3.759	1.085	23.262	27.516
Mac	22.829	3.184	0.919	21.027	24.631

The situation is presented graphically in Figure 2.5.

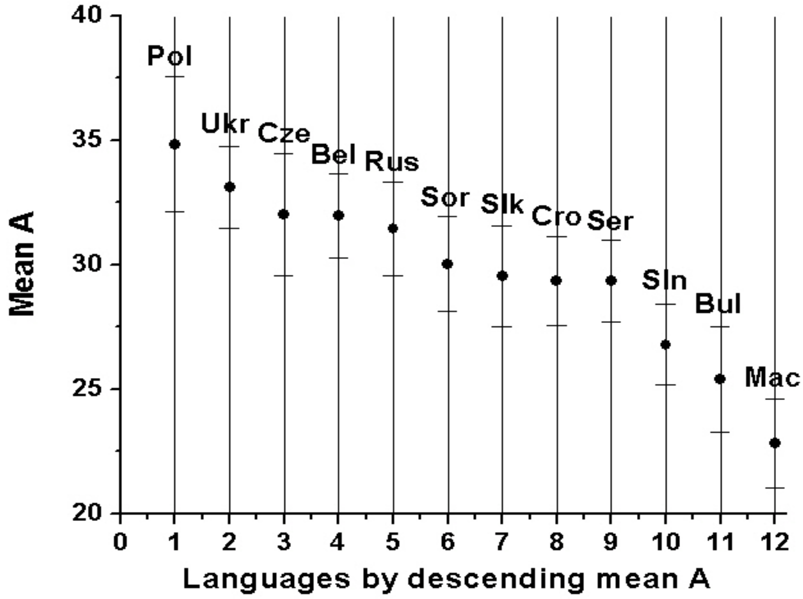


Figure 2.5. Mean *A* and its interval in 12 Slavic languages

As can be seen, Figure 2.5 displays a slightly different ordering of Slavic languages, a status quo from the textological point of view which need not agree with historical or geographical facts. Nevertheless a rough geographical (areal) order of the Slavic languages can be obtained: It starts with a “mixture” of East- and West Slavic languages, and ends with South Slavic languages (starting with Croatian). May be – this must be investigated more systematically – this ordering is the result of some morphologial and morphosyntactical characteristics, which are roughly in some relation to the degree of analytism/synthetism in these languages. For a more systematic study a deeper analysis of indicators in parallel corpora is required. At the same time it shows that classification is not a plain, unequivocal partition of the universe of discourse in classes; it is rather a play with rough sets (cf. Dubois, Prade 1990; Pawlak 1991; Bazan, Szczuka, Wojna, Wojnarski 2004).

Here we considered a translation from Russian, but it does not mean that Russian is in the mid of other languages exactly for this reason. If one would translate a work from Macedonian, one would probably obtain the same “rough” image.

2.4. General data

Though data from other languages at our disposal are not always representative, we compute the adjusted modulus in order to see the approximate location of a language. The data are presented in Table 2.6. The basic data were taken from Popescu et al. (2009)

Table 2.6
The adjusted modulus in 14 languages
(E = English, H = Hungarian, Hw = Hawaiian, In = Indonesian, Kn = Kannada, Lk = Lakota, Lt = Latin, M = Maori, Mq = Marquesan, Mr = Marathi, R = Romanian, Rt = Rarotongan, Sm = Samoan, T = Tagalog)

ID	*N*	*V*	*f*(1)	*h*	*M*	*A*
E 01	2330	939	126	16	59.213	17.585
E 02	2971	1017	168	22	46.854	13.491
E 03	3247	1001	229	19	54.045	15.391
E 04	4622	1232	366	23	55.879	15.247
E 05	4760	1495	297	26	58.624	15.941
E 07	5004	1597	237	25	64.580	17.457
E 13	11265	1659	780	41	44.713	11.035
H 01	2044	1079	225	12	91.851	27.745
H 02	1288	789	130	8	99.955	32.141
H 03	403	291	48	4	73.733	28.301
H 04	936	609	76	7	87.675	29.507
H 05	413	290	32	6	48.627	18.589
Hw 03	3507	521	277	26	22.695	6.402
Hw 04	7892	744	535	38	24.115	6.188
Hw 05	7620	680	416	38	20.978	5.404
Hw 06	12356	1039	901	44	31.256	7.638
In 01	376	221	16	6	36.930	14.341
In 02	373	209	18	7	29.968	11.653
In 03	347	194	14	6	32.417	12.761
In 04	343	213	11	5	42.657	16.825
In 05	414	188	16	8	23.585	9.012
Kn 003	3188	1833	74	13	141.115	40.278
Kn 004	1050	720	23	7	102.910	34.063
Kn 005	4869	2477	101	16	154.941	42.019
Kn 006	5231	2433	74	20	121.706	32.729
Kn 011	4541	2516	63	17	148.046	40.481
Lk 01	345	174	20	8	21.893	8.627
Lk 02	1633	479	124	17	29.105	9.059

Lk 03	809	272	62	12	23.248	7.995
Lk 04	219	116	18	6	19.565	8.359
Lt 01	3311	2211	133	12	184.583	52.439
Lt 02	4010	2334	190	18	130.096	36.106
Lt 03	4931	2703	103	19	142.366	38.551
Lt 04	4285	1910	99	20	95.628	26.330
Lt 05	1354	909	33	8	113.700	36.307
Lt 06	829	609	19	7	87.042	29.824
M 01	2062	398	152	18	23.669	7.141
M 02	1175	277	127	15	20.315	6.617
M 03	1434	277	128	17	17.950	5.686
M 04	1289	326	137	15	23.574	7.580
M 05	3620	514	234	26	21.721	6.104
Mq 01	2330	289	247	22	17.281	5.132
Mq 02	457	150	42	10	15.577	5.856
Mq 03	1509	301	218	14	26.547	8.351
Mr 001	2998	1555	75	14	111.201	31.983
Mr 018	4062	1788	126	20	89.622	24.835
Mr 026	4146	2038	84	19	107.354	29.675
Mr 027	4128	1400	92	21	66.810	18.478
Mr 288	4060	2079	84	17	122.394	33.918
R 01	1738	843	62	14	60.377	18.635
R 02	2279	1179	110	16	74.008	22.041
R 03	1264	719	65	12	60.161	19.396
R 04	1284	729	49	10	73.064	23.504
R 05	1032	567	46	11	51.715	17.160
R 06	695	432	30	10	43.304	15.237
Rt 01	968	223	111	14	17.793	5.959
Rt 02	845	214	69	13	17.296	5.909
Rt 03	892	207	66	13	16.713	5.665
Rt 04	625	181	49	11	17.047	6.097
Rt 05	1059	197	74	15	14.029	4.638
Sm 01	1487	267	159	17	18.280	5.762
Sm 02	1171	222	103	15	16.315	5.317
Sm 03	617	140	45	13	11.312	4.054
Sm 04	736	153	78	12	14.311	4.992
Sm 05	447	124	39	11	11.817	4.459
T 01	1551	611	89	14	44.103	13.823
T 02	1827	720	107	15	48.527	14.878
T 03	2054	645	128	19	34.609	10.448

In order to get some lucidity in this set of data, we present the means of individual languages as shown in Table 2.7. The results obtained above are included in the table.

Table 2.7
Means of the adjusted modulus for 28 languages

Language	***A***
Samoan	4.917
Rarotongan	5.654
Hawaiian	6.408
Marquesan	6.446
Maori	6.626
Lakota	8.510
Indonesian	12.918
Tagalog	13.049
Italian	14.920
English	15.164
German	15.430
Romanian	19.329
Macedonian	22.829
Bulgarian	25.389
Slovenian	26.796
Hungarian	27.257
Marathi	27.778
Serbian	29.345
Croatian	29.351
Slovak	29.549
Sorbian	30.033
Russian	31.440
Belorussian	31.953
Czech	32.007
Ukrainian	33.119
Polish	34.836
Latin	36.593
Kannada	37.914

A preliminary look at the table shows that the languages are again ordered according to the extent of synthetism. Examination of other texts will surely change the order but in general the mean adjusted modulus is a fuzzy measure of analytism/synthetism (cf. also Popescu et al. 2009)

The *A*-value of Italian is that of Presidential addresses. Results from other Italian texts show that it is higher and approaches Romanian. In the same way, our *A*-value of German is 15.43, but if we consider some of Goethe´s poetic works (“Der Gott und die Bajadere”; “Elegie”-s 2,5,13,15,19 and “Erlkönig”) we obtain a mean A = 13.45. These are, actually, symptoms of the fact that the indicator can be used as a characteristic of genre.

Examinations in this direction will never be finished. Adding further texts could enable us to interpret the indicator stylistically, typologically etc. but this must be left to specialists. Even the problem of influence of one language on another with writers writing with the same skill in two languages can be examined by this method. Further the speech of individual persons in a drama and the relationship of the adjusted modulus to the role of the given persons, etc. can be analysed this way. The results in Table 2.7 are merely the first illustrative step.

3. The vector T

3.1. Retrospective dissimilarity: stepwise and cumlative

The length of the vector and its adjustment is not the only property discriminating texts. A number of other possibilities have been shown in Popescu et al. (2009), Popescu, Mačutek, Altmann (2009). Here we shall develop the evaluation of the rank-frequency sequence using the same quantities as in Chapter 2, namely V, $f(1)$ and h, but this time we set up a vector with three components (cf. Tuzzi, Popescu, Altmann 2010), viz.

(3.1) $T(V, f(1), h) = T(x,y,z)$,

where $x = V$, $y = f(1)$, $z = h$ are its Cartesian components. In order to compare two texts or two parts of a text, we may compute the cosine of the angle between two vectors T_1 and T_2 in the usual way as

(3.2) $$cos\,\tau_{12} = \frac{\mathbf{T}_1 \cdot \mathbf{T}_2}{|\mathbf{T}_1||\mathbf{T}_2|} = \frac{(x_1x_2 + y_1y_2 + z_1z_2)}{\sqrt{x_1^2 + y_1^2 + z_1^2}\sqrt{x_2^2 + y_2^2 + z_2^2}}$$

from which the angle in radians is obtained by taking the *arccos* function, i.e.

(3.3) $$\tau_{12} = arccos(cos\,\tau_{12})\,.$$

Since $\tau = 0$ means perfect similarity[1] and $\tau = \pi/2$ maximal dissimilarity[2], τ is a *measure of dissimilarity*. The greater τ, the greater is the dissimilarity of the vectors of compared texts or text parts.[3]

This measure can be transformed to a normalized similarity measure in different ways but here we leave it in its elementary form. The derivation of an asymptotic test is associated with the difficulty of treating V as a variable. Perhaps the use of order statistics could be of help. But even in that case, the test

---

[1] Actually, $\tau = 0$ means that the considered vectors are collinear, hence their coordinates are in the same ratio $x_1/x_2 = y_1/y_2 = z_1/z_2 = constant$ or, in other words, the corresponding rank-frequencies are fully similar (identical for *constant* = 1).

[2] In this extreme case, $\tau = \pi/2$, we would have complete orthogonality. However, the actual limit will never reach this ideal limit inasmuch as the ranks, by definition, are always positive integers, that is T vectors belong to the first quadrant of the considered Cartesian coordinate system.

[3] Similarly, we can introduce the angle u between text vectors $P(x = V/h,\ y = f_1/h)$ as well, which is computed as $cos\,u_{12} = (x_1x_2 + y_1y_2)/((x_1^2 + y_1^2)^{1/2}(x_2^2 + y_2^2)^{1/2})$. Actually, it can be easily shown that $u \approx \tau$ for V and/or f_1 much greater than h, as is the case in actual texts.

would be very involved. In the present chapter no tests will be performed; we consider only the dynamics of the angle.

For the sake of illustration we consider the rank-frequency distributions of word forms in the first two chapters of A.v.Chamisso's *Peter Schlemihls wundersame Geschichte* (1814). In Table 3.1 one can see that the first two chapters have the vectors

chapter 1 = (884, 82, 18),
chapter 2 = (808, 84, 16).

Table 3.1
Word-form rank-frequency distributions in Chamisso's *Peter Schlemihl*

ID	*N*	*V*	*f*(1)	*h*
Chamisso 01	2210	884	82	18
Chamisso 02	1847	808	84	16
Chamisso 03	1428	630	70	14
Chamisso 04	3205	1209	123	20
Chamisso 05	2108	853	79	18
Chamisso 06	1948	801	75	17
Chamisso 07	1362	670	44	13
Chamisso 08	1870	788	80	16
Chamisso 09	1320	593	96	14
Chamisso 10	1012	536	52	11
Chamisso 11	1386	656	66	14

Inserting these numbers in formula (3.2) we obtain

$$\cos\tau = \frac{884(808) + 82(84) + 18(16)}{\sqrt{884^2 + 82^2 + 18^2}\sqrt{808^2 + 84^2 + 16^2}} = 0.9999383327$$

from which

τ radians = arccos(0.9999383327) = 0.01110566827

whose rounded form is shown in Table 3.2.

This approach allows us to study a text composed of different parts to answer the following questions:

1. How does the (dis)similarity of individual parts develop compared with the first part? Here we compare each part of the text with its beginning part yielding the text special initiating dynamics. The example in Table 3.1 yields the

dissimilarities against chapter 1 as given in Table 3.2 and presented by Figure 3.1. We can call this view *stepwise (dis)similarity with retrospective view.*

The given text is constructed almost like a classical drama: at the beginning, the development of the text is straightforward, i.e., the dissimilarities are small; then a conflict appears, attains a crisis and is solved in a catharsis.

Table 3.2
Dissimilarities of chapters against chapter 1

Chapter	τ
1	0
2	0.0111
3	0.0182
4	0.0097
5	0.0008
6	0.0012
7	0.0269
8	0.0087
9	0.0681
10	0.0042
11	0.0078

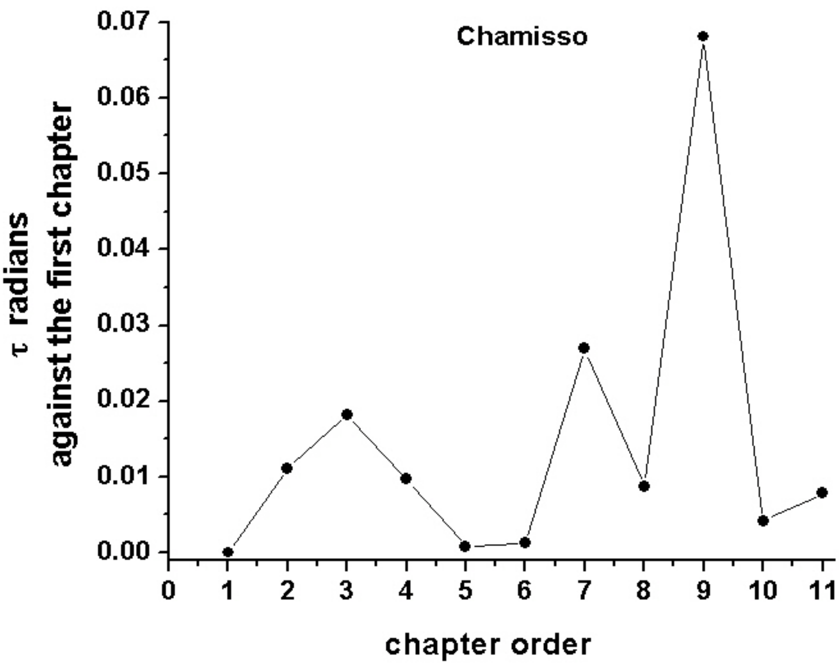


Figure 3.1. The τ angle of consecutive chapters with the first one in Chamisso's *Peter Schlemihl*

Of course other interpretations are possible and a future study will show them. One can also imagine that Chamisso made a "pause" (break) of whatever kind between some chapters or some kind of "Stilbruch", e.g. changing of the style appeared. In such a pause, the rhythms acquired by the Skinner effect faded out and the new chapter began so to say with *tabula rasa* of frequencies in the brain of the writer. Thus, one can associate this approach with the contents of the text or with the psychological state of the writer. The differentiation can be made only on the basis of modern texts whose authors could be interviewed but we do not have this possibility. We can call this view *stepwise dissimilarity.*

2. What is the relationship/(dis)similarity of the stepwise sums of next parts of the text to the first part? Here the chapters are added and compared with the beginning chapter. This view can be called *cumulative (dis)similarity with retrospective view.*

In order to illustrate this case we take the same text and show the development in Table 3.3 and Figure 3.2.

Table 3.3
Cumulative word-form rank-frequency distributions
in Chamisso's *Peter Schlemihl*

Chap. 1 *versus*	***N***	***V***	***f*(1)**	***h***	***cos τ***	***τ*** radians
1	2210	884	82	18	1	0
1+2	4055	1435	166	25	0.9997	0.0229
1+2+3	5487	1773	236	28	0.9992	0.0401
1+…+4	8693	2480	359	35	0.9987	0.0516
1+…+5	10806	2877	428	40	0.9985	0.0556
1+…+6	12755	3211	503	44	0.998	0.0632
1+…+7	14118	3460	547	44	0.9979	0.0648
1+…+8	15989	3754	627	48	0.9973	0.0734
1+…+9	17310	3979	723	49	0.9962	0.0876
1+…+10	18323	4219	775	50	0.9960	0.0896
1+…+11	19710	4446	841	51	0.9955	0.0949

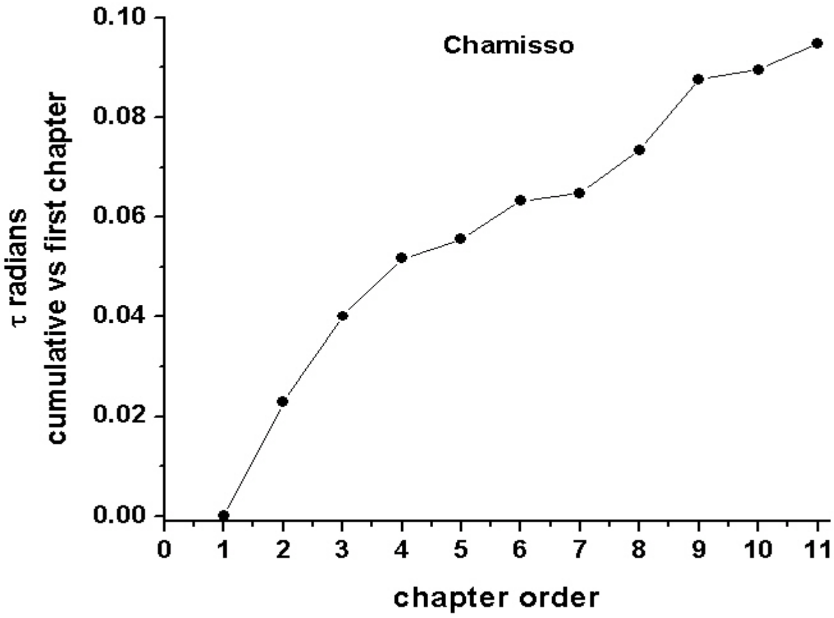


Figure 3.2. The τ angle of cumulative word-form rank-frequency distributions in Chamisso's *Peter Schlemihl*

Evidently, the cumulative comparison displays a relatively smooth increase of dissimilarity.

In Table 3.4 one can find the retrospective stepwise and cumulative dissimilarities of eleven works of German writers, namely

Novalis, *Heinrich von Ofterdingen* (1802)
Jean Paul, *Dr. Katzenbergers Badereise* (1809)
A.v. Chamisso, *Peter Schlemihls wundersame Geschichte* (1814)
E.Th.A. Hoffmann, *Der Sandmann* (1817)
J.v. Eichendorff, *Aus dem Leben eines Taugenichts* (1826)
Ch. Sealsfield, *Das Kajütenbuch* (1841)
C.F. Meyer, *Der Schuß von der Kanzel* (1877)
F. Wedekind, *Mine-Haha* (1901)
H. Löns, *Der Werwolf* (1910)
F. Kafka, *Betrachtung* (1913)
K. Tucholsky, *Schloss Gripsholm* (1931).

In Table 3.4 we ordered the works according to the number of chapters in order to save space, but special problems require different ordering.

Table 3.4
Development of retrospective dissimilarities in 11 German texts
(*S* = stepwise, *C* = cumulative)

Part	Hoffmann		Wedekind		Tucholsky	
	S	*C*	*S*	*C*	*S*	*C*
1	0	0	0	0	0	0
2	0.0264	0.0025	0.0121	0.0297	0.0360	0.0091
3	0.0348	0.0431	0.0507	0.0719	0.0089	0.0322
4			0.0231	0.0722	0.0342	0.0423
5					0.0195	0.0505

Part	Novalis		Eichendorff		Meyer	
	S	*C*	*S*	*C*	*S*	*C*
1	0	0	0	0	0	0
2	0.0169	0.0319	0.0008	0.0394	0.0109	0.0029
3	0.0069	0.0497	0.0270	0.0543	0.0137	0.0148
4	0.0219	0.0563	0.0561	0.0595	0.0019	0.0197
5	0.0471	0.0934	0.0629	0.0655	0.0017	0.0219
6	0.0034	0.1052	0.0405	0.0751	0.0044	0.0233
7	0.0256	0.1094	0.0335	0.0882	0.0033	0.0276
8	0.0309	0.1110	0.0427	0.0964	0.0077	0.0327
9	0.0414	0.1322	0.0325	0.1028	0.0209	0.0378
10	0.0052	0.1429	0.0411	0.1113	0.0140	0.0426
11					0.0117	0.0466

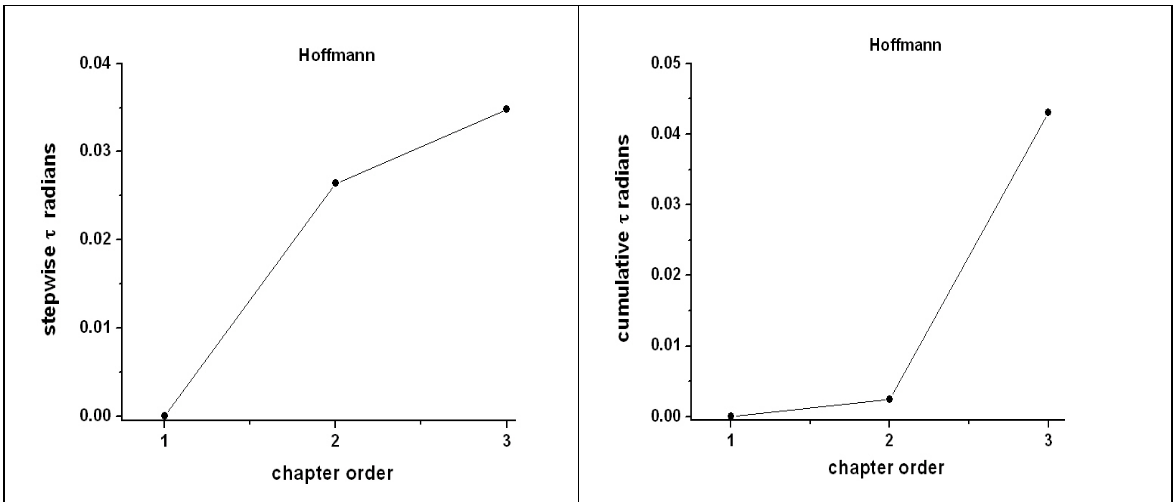
Part	Chamisso		Löns		Kafka	
	S	*C*	*S*	*C*	*S*	*C*
1	0	0	0	0	0	0
2	0.0111	0.0229	0.0174	0.0373	0.0077	0.0050
3	0.0182	0.0401	0.0132	0.0619	0.0189	0.0094
4	0.0097	0.0516	0.0195	0.0894	0.0230	0.0080
5	0.0008	0.0556	0.0691	0.1313	0.0385	0.0085
6	0.0012	0.0632	0.0432	0.1597	0.0241	0.0080
7	0.0269	0.0648	0.0850	0.2012	0.0069	0.0156
8	0.0087	0.0734	0.0543	0.2203	0.0404	0.0230

9	0.0681	0.0876	0.0762	0.2458	0.0278	0.0192
10	0.0042	0.0896	0.0228	0.2545	0.0340	0.0219
11	0.0078	0.0949	0.0324	0.2655	0.0171	0.0220
12			0.0324	0.2783	0.0395	0.0271
13			0.0266	0.2802	0.0173	0.0282
14					0.0139	0.0279
15					0.0716	0.0333
16					0.0282	0.0341
17					0.0438	0.0341
18					0.0703	0.0352

Part	**Sealsfield**		**Paul**				
	S	*C*	*S*	*C*	Part	*S*	*C*
1	0	0	0	0	29	0.0049	0.0757
2	0.0074	0.0096	0.0212	0.0028	30	0.0065	0.0768
3	0.0205	0.0200	0.0092	0.0128	31	0.0196	0.0812
4	0.0400	0.0306	0.0167	0.0152	32	0.0286	0.0812
5	0.0139	0.0308	0.0113	0.0192	33	0.0113	0.0816
6	0.0088	0.0336	0.0241	0.0194	34	0.0154	0.0857
7	0.0034	0.0399	0.0284	0.0184	35	0.0181	0.0863
8	0.0283	0.0536	0.0132	0.0241	36	0.0261	0.0865
9	0.0401	0.0635	0.0086	0.0284	37	0.0131	0.0870
10	0.0116	0.0664	0.0087	0.0318	38	0.0247	0.0869
11	0.0027	0.0707	0.0083	0.0393	39	0.0050	0.0894
12	0.0068	0.0745	0.0076	0.0408	40	0.0028	0.0908
13	0.0099	0.0778	0.0146	0.0432	41	0.0080	0.0923
14	0.0087	0.0788	0.0074	0.0461	42	0.0199	0.0958
15	0.0312	0.0863	0.0062	0.0502	43	0.0044	0.0970
16	0.0118	0.0915	0.0221	0.0503	44	0.0640	0.1032
17	0.0126	0.0968	0.0066	0.0518	45	0.0044	0.1054
18	0.1490	0.0991	0.0100	0.0544	46	0.0170	0.1059
19	0.0285	0.0999	0.0197	0.0527	47	0.0068	0.1077
20	0.0262	0.1059	0.0048	0.0532	48	0.0196	0.1083
21	0.0334	0.1107	0.0377	0.0643	49	0.0157	0.1096
22	0.0202	0.1120	0.0263	0.0643	50	0.0128	0.1126

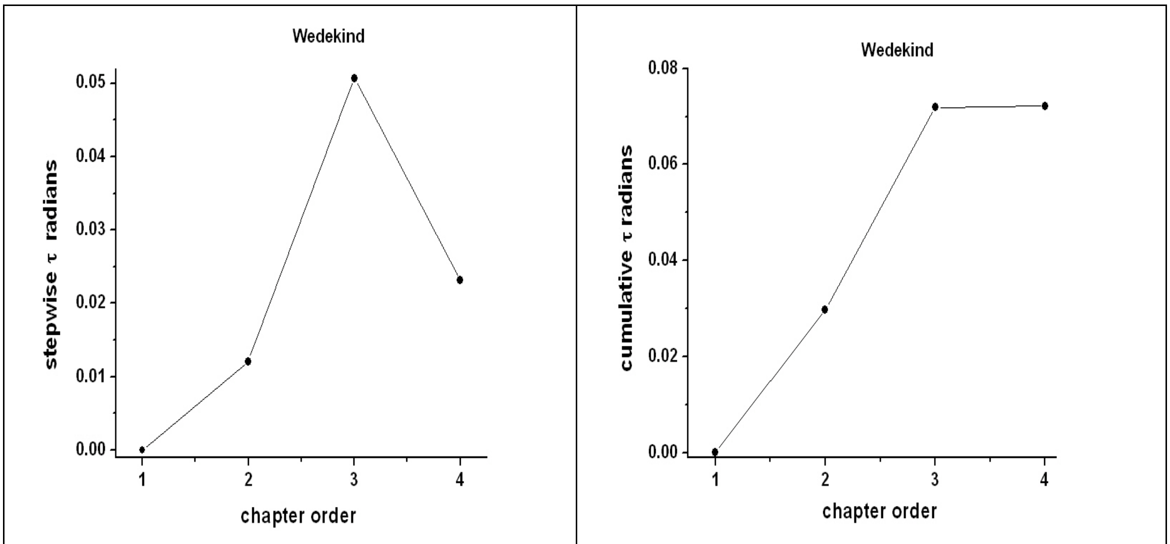
23	0.0253	0.1130	0.0090	0.0651	51	0.0113	0.1146
24	0.0155	0.1134	0.0334	0.0710	52	0.0116	0.1169
25	0.0187	0.1129	0.0261	0.0715	53	0.0265	0.1195
26	0.0102	0.1128	0.0080	0.0717	54	0.0441	0.1264
27	0.0431	0.1158	0.0126	0.0731	55	0.0214	0.1278
28	0.0442	0.1173	0.0306	0.0732			

In Figure 3.3 one can see the graphic presentation of the eleven texts.



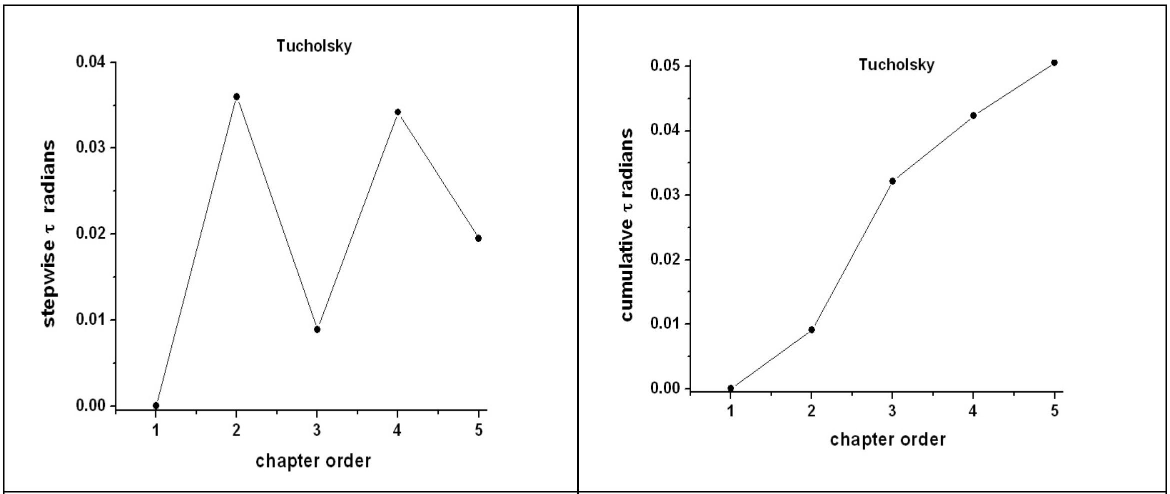
Hoffmann. stepwise	Hoffmann. cumulative

Figure 3.3a. Hoffmann



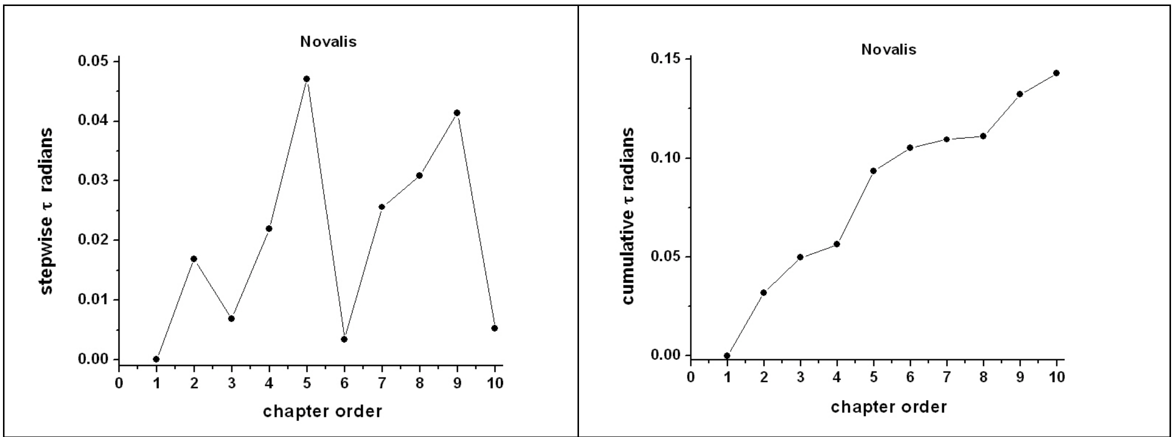
Wedekind stepwise	Wedekind cumulative

Figure 3.3b. Wedekind



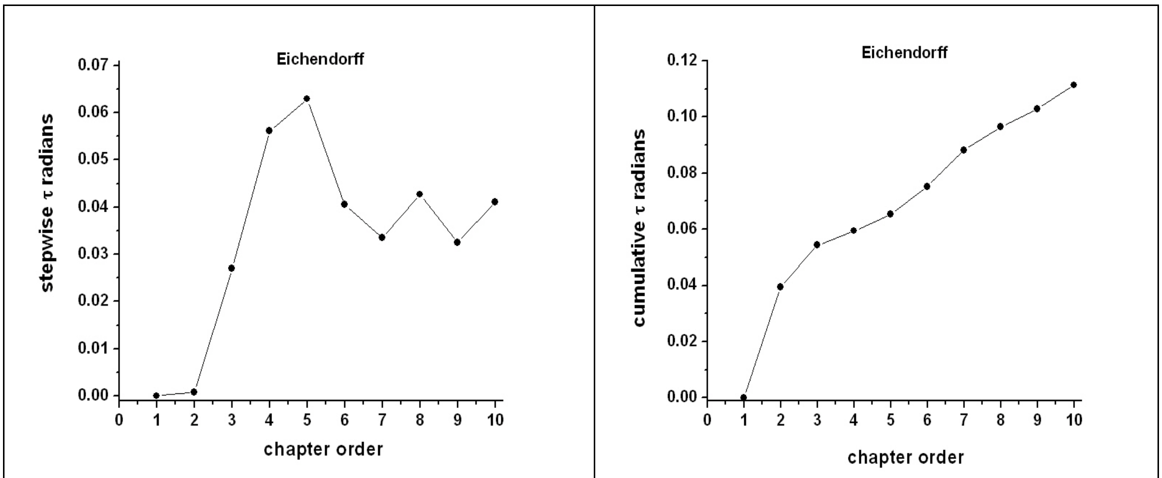
Tucholsky stepwise	Tucholsky cumulative

Figure 3.3c. Tucholsky



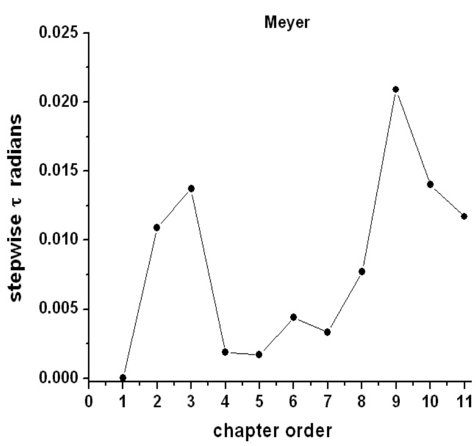
Novalis stepwise	Novalis cumulative

Figure 3.3d. Novalis

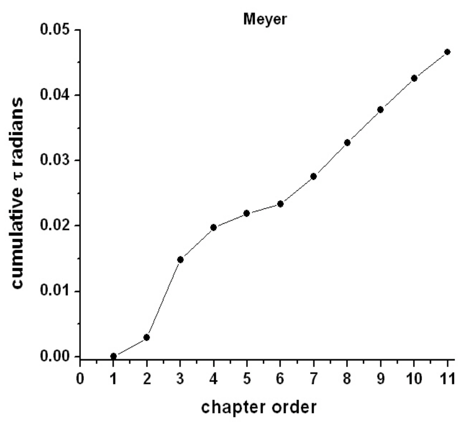


Eichendorff stepwise	Eichendorff cumulative

Figure 3.3e. Eichendorff

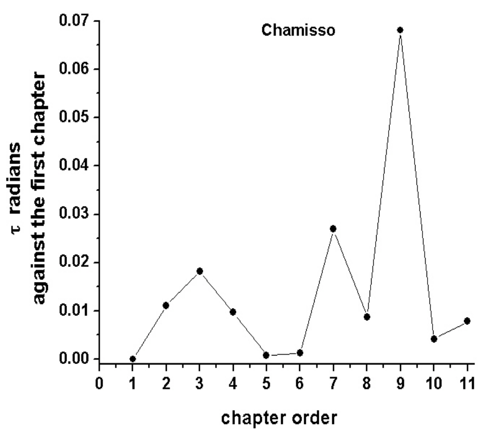


Meyer stepwise

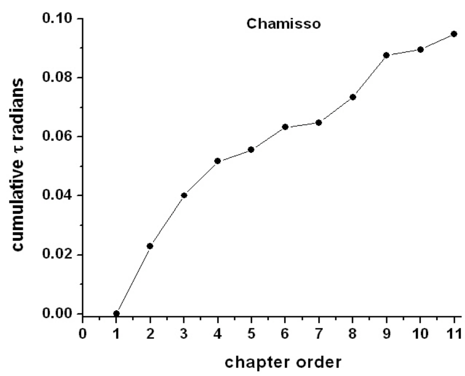


Meyer cumulative

Figure 3.3f. Meyer

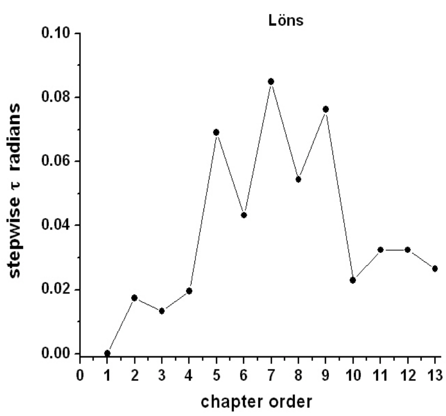


Chamisso stepwise

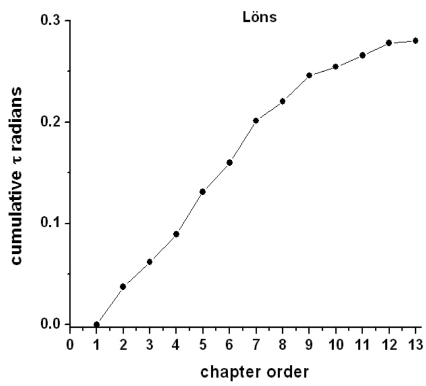


Chamisso cumulative

Figure 3.3g. Chamisso

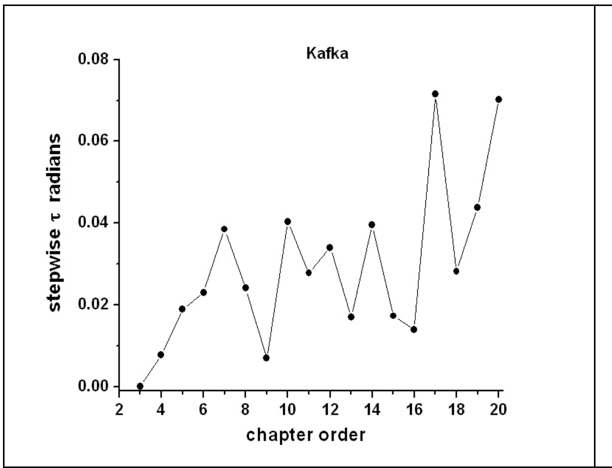


Löns stepwise

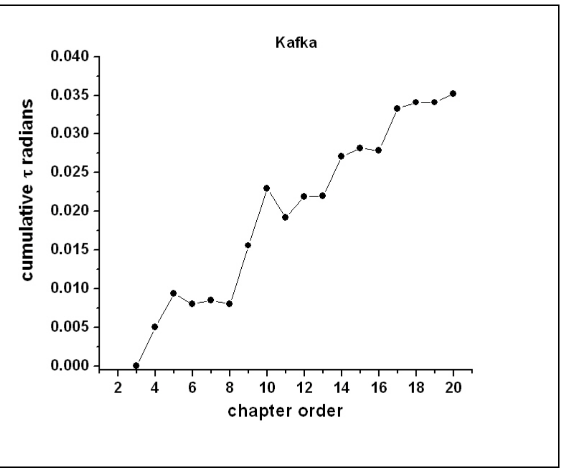


Löns cumulative

Figure 3.3h. Löns

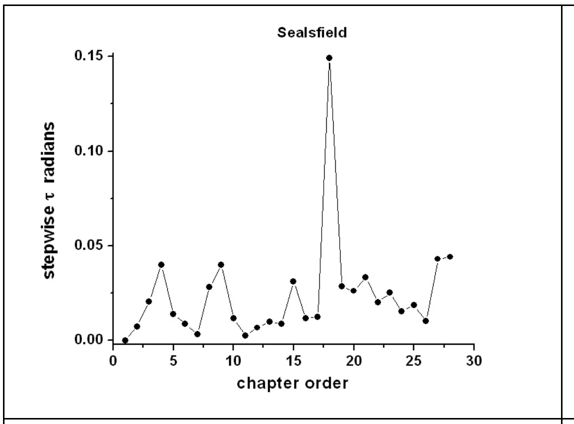


Kafka stepwise

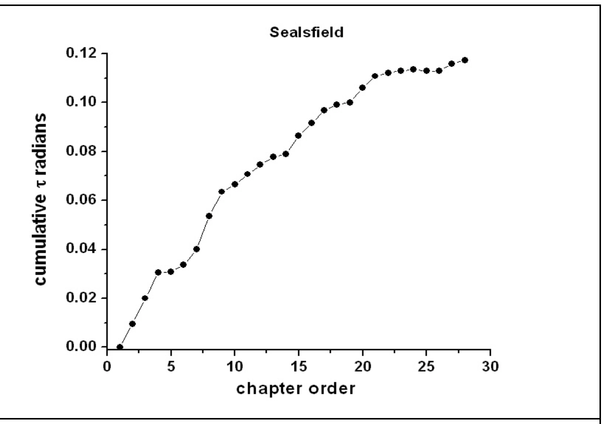


Kafka cumulative

Figure 3.3i. Kafka

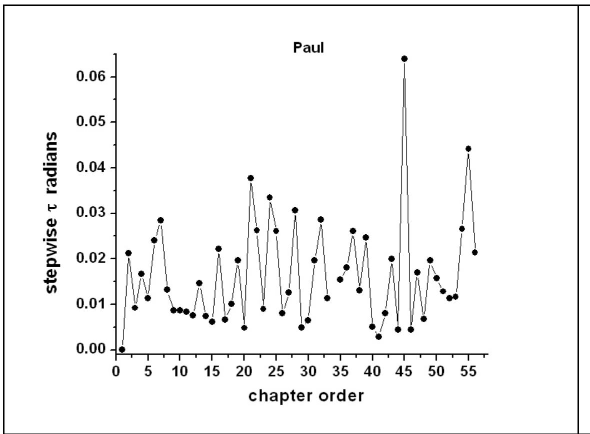


Sealsfield stepwise

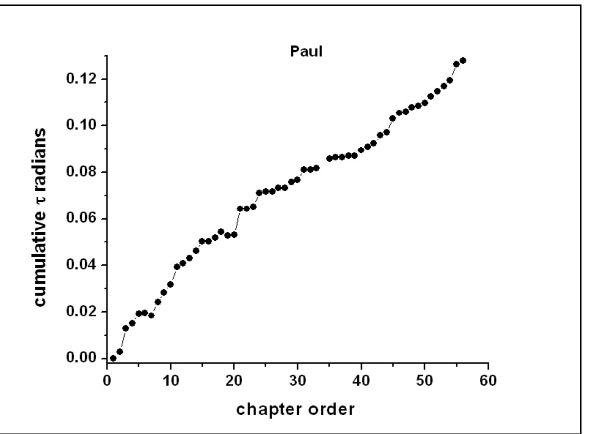


Sealsfield cumulative

Figure 3.3j. Sealsfield



Paul stepwise



Paul cumulative

Figure 3.3k. Paul

Figure 3.3. The stepwise and cumulative course of retrospective dissimilarity in 11 German texts

As can be seen, the *retrospective stepwise* figures display very different forms. We may ask whether there is something common in these figures or whether one can at least propose a comparative measure or, finally, what changes in the course of time and in the course of increasing the number of chapters. To this end one may take some indicators from chaos theory.

The *retrospective cumulative view* is relatively simple; the sequences are – except for some cases like Kafka – relatively smooth because the accumulation effaces the weight of individual chapters. Ignoring Hoffmann and Wedekind whose texts have too few parts it can easily be shown that all retrospective cumulative dissimilarities can be captured by the power function $y = ax^b$, where x is the chapter. The results are presented in Table 3.5.

Table 3.5
Cumulative dissimilarities ordered according to exponent b

Author	**Year**	***a***	***b***	R^2
Sealsfield	1841	0.0120	0.7094	0.96
Paul	1809	0.0059	0.7568	0.99
Eichendorff	1826	0.0197	0.7597	0.95
Chamisso	1814	0.0148	0.7864	0.96
Novalis	1802	0.0181	0.9088	0.95
Löns	1910	0.0293	0.9203	0.96
Kafka	1913	0.0024	0.9463	0.96
Meyer	1877	0.0037	1.0634	0.97
Tucholsky	1931	0.0089	1.1004	0.90

All regressions are satisfactory and testify to the fact that the “whole“ of the texts moves very regularly away from the base formed by the first chapter (or first part of text). In the majority of texts the dynamics is quite monotonous as can be seen in the right parts of Figure 3.3. As is well known, for $0 < b < 1$ the function is concave, for $b = 1$ it is a straight line, and for $b > 1$ it is convex, hence this property must correlate with some other text properties which must first be defined qualitatively.

Though a certain regularity may be observed also in the relationship between the year of origin and the parameter b, there are preliminarily too few data and no clear-cut hypothesis to be tested, hence the problem must be postponed until dozens of texts have been analyzed.

Actually, the exponent b is significantly higher than the value of 1/2 as required by the two-dimensional random walk model predicting that the root-mean-square distance after n unit steps equals to $n^{1/2}$ (see http://mathworld.wolfram.com/RandomWalk2-Dimensional.html). In other words, this means that the writer advances not quite at random but rather target-oriented, systematically trying to communicate something. His trials are not chaotic, isotropic, but anisotropic, polarized, biased by his internal tension. He is creating. This manifests itself in the rank-tau plane as a departure of the distribution from the straight line and the development of an upward concavity. At the same time the mean normalized τ becomes mostly lower than 1/2, as it will be shown in 3.3 below.

However, in contrast, the *stepwise* dissimilarity of individual chapters in their relation to the first chapter, as shown in the left parts of Figure 3.3, displays a very irregular shape. Nevertheless, even the given irregularity can be characterized in some way. In what follows, we present some possibilities. We present both static and dynamic characterizations.

3.2. Dispersion

The usual way of measuring the variability of data is the variance or the standard deviation. The familiar formula of the variance is

$$(3.4) \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \, ,$$

where n is the number of values x_i and $\bar{x}$ is their mean. Its root is the standard deviation. Example: the text by Hoffmann has three stepwise τ values: (0.0000, 0.0264, 0.0348) whose mean is 0.0204, hence

$$s^2 = [(0.0000 - 0.0204)^2 + (0.0264 - 0.0204)^2 + (0.0348 - 0.0204)^2]/2 = \\ = 0.00032976,$$

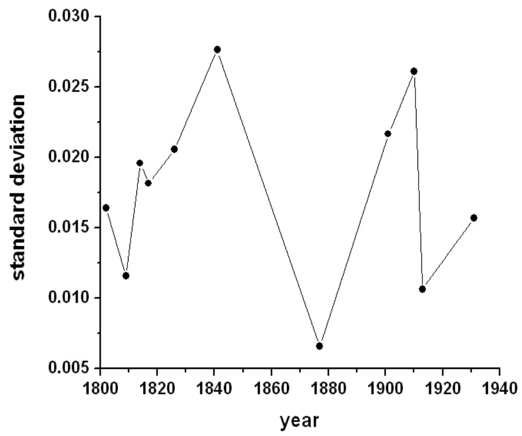
from which

$$s = 0.0182.$$

The values of standard deviations of all texts are presented in Table 3.6. The historical course of standard deviation is not regular as can be seen in the Figure in Table 3.6. Nevertheless, standard deviation is a characteristic of the formation of dissimilarity in texts. The greater the s.d., the greater are the dissimilarity jumps in the text, that is, either there are great differences in the content of chapters, or the text has been written with breaks or, finally, there was less spontaneity in writing it. Thus dissimilarities studied here can represent even a picture of the degree of spontaneity.

Table 3.6
Standard deviations of stepwise dissimilarities

Author	year	s
Novalis	1802	0.016412
Paul	1809	0.011575
Chamisso	1814	0.019571
Hoffmann	1817	0.018147
Eichendorff	1826	0.020539
Sealsfield	1841	0.027639
Meyer	1877	0.006579
Wedekind	1901	0.021657
Löns	1910	0.026113
Kafka	1913	0.010629
Tucholsky	1931	0.015657



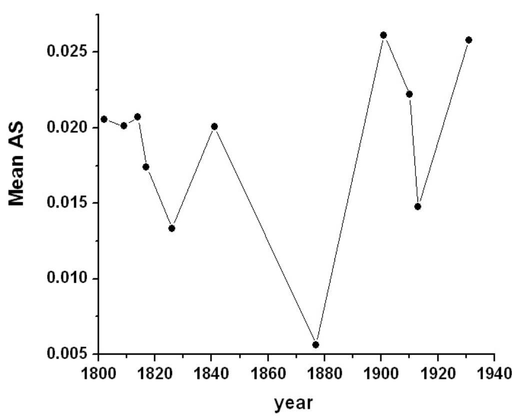
Using the mean absolute sequential difference between subsequent τ-angles defined as

$$(3.5) \quad AS = \frac{1}{n-1}\sum_{i=1}^{n-1}|\tau_i - \tau_{i+1}|,$$

one does not obtain clearer historical results as shown in Table 3.7. The time need not play any crucial role, more important are the differences in style. Neither the ordering of texts according to the number of chapters yields any smooth result.

Table 3.7
Mean absolute sequential difference

Author	year	*AS*
Novalis	1802	0.020568
Paul	1809	0.020125
Chamisso	1814	0.020699
Hoffmann	1817	0.017400
Eichendorff	1826	0.013347
Sealsfield	1841	0.020069
Meyer	1877	0.005626
Wedekind	1901	0.026123
Löns	1910	0.022200
Kafka	1913	0.014784
Tucholsky	1931	0.025781



Evidently there are no dispersion tendencies if we restrict the examination to these writers.

Standard deviation is a static measure while *AS* yields a dynamic perspective of the text.

3.3. Randomness

Chaos and randomness are concepts opposite to structure and order. But the space in between is full of concepts like stability, volatility, deterministic chaos, fractals, dimensions, self-similarity, etc. taken from the dictionary of modern mathematics. In concrete cases one must choose a simple way to show whether the results are due to chance or display some (significant) tendency. Usually, one takes recourse to statistics whose shortcomings – especially in the classical domain – are well known. Too small and too large samples furnish distorted results.

One of the ways to show that stepwise dissimilarity as presented in Table 3.4 tending to a certain structuring may be described as follows. If we normalize the stepwise dissimilarities by dividing them by their greatest value, we obtain normalized τ-angles in interval <0, 1>. If they are distributed randomly, then they follow the uniform distribution whose probability function is $f(x) = 1$ for $0 \le x \le 1$. The mean of the uniform distribution can be computed by integration as $m_1' = \int_0^1 x dx = \frac{1}{2}$ = 0.5, the second raw moment is $m_2' = \int_0^1 x^2 dx = \frac{1}{3}$ = 0.3333 and the variance is $m_2 = m_2' - m_1'^2 = 1/12 = 0.0833$. The standard deviation is then $\sqrt{0.0883} = 0.2887$ and the standard deviation of the mean is $0.2887/\sqrt{n}$. The hypothesis of the existence of structuring is thus easily testable by an asymptotic normal test, but we see that everything depends on sample size n appearing in the standard deviation of the mean. Nevertheless, even if the departure from 0.5 is not significant – because of sample size – it carries some information about structuring.

Let us exemplify the idea using the data of Chamisso in Table 3.4. In Table 3.8 one finds the original τ-angles in the second column. In the fourth column they are simply re-ranked according to size and in the fifth column all τ-angles are divided by the greatest one, namely 0.0681.

Table 3.8
Normalized stepwise τ-angles with Chamisso

Chapter	$\boldsymbol{\tau}$	**rank** $\boldsymbol{x}$	**ranked by** $\boldsymbol{\tau}$	**Normalized** $\boldsymbol{\tau}$
1	0.0000	1	0.0681	1.0000
2	0.0111	2	0.0269	0.3957
3	0.0182	3	0.0182	0.2681
4	0.0097	4	0.0111	0.1632
5	0.0008	5	0.0097	0.1422
6	0.0012	6	0.0087	0.1275
7	0.0269	7	0.0078	0.1151
8	0.0087	8	0.0042	0.0620
9	0.0681	9	0.0012	0.0179
10	0.0042	10	0.0008	0.0110
11	0.0078	11	0.0000	0.0000
Mean normalized $\boldsymbol{\tau}$ = 0.2093				

The mean of normalized τ-angles in the last column is 0.2093, a value which is “very far” from the expectation 0.5 and signalizes a kind of structuring. In order to test the significance of structuring, we set up the t-test. Since there are 11 chapters, we obtain $s_{\bar{x}}$ = 0.2887/√11 = 0.0870, hence

$$t = \frac{0.2093 - 0.5}{0.0870} = -3.34$$

which is highly significant because for the two-sided t with n-1 = 10 degrees of freedom, P(3.34) < 0.01. Thus in Chamisso there is a kind of structuring. It can easily be seen that with n = 3 the result would not be significant. As a matter of fact, this kind of structuring is possible only when the text increases.

All mean normalized τ-angles are presented in Table 3.9 where they are ordered according to the number of chapters. Optically, the greater the sample size, the more structure can be found, the more the normalized mean τ moves away from randomness as can be seen in Figure 3.4. However, not all are significantly different form the expectation. Significant structuring can be found only with Chamisso, Sealsfied and Paul.

Table 3.9
Mean normalized τ-angles and the normal test

Writer	**No of chapters, *n***	**Normalized mean τ**	***t***
Hoffmann	3	0.5867	0.52
Wedekind	4	0.4232	-0.53
Tucholsky	5	0.5479	0.37
Novalis	10	0.4234	-0.84
Eichendorff	10	0.5364	0.40
Chamisso	11	0.2093	-3.34
Meyer	11	0.3927	-1.23
Löns	13	0.4453	-0.68
Kafka	18	0.4054	-1.39
Sealsfield	28	0.1611	-6.21
Paul	55	0.2543	-6.31

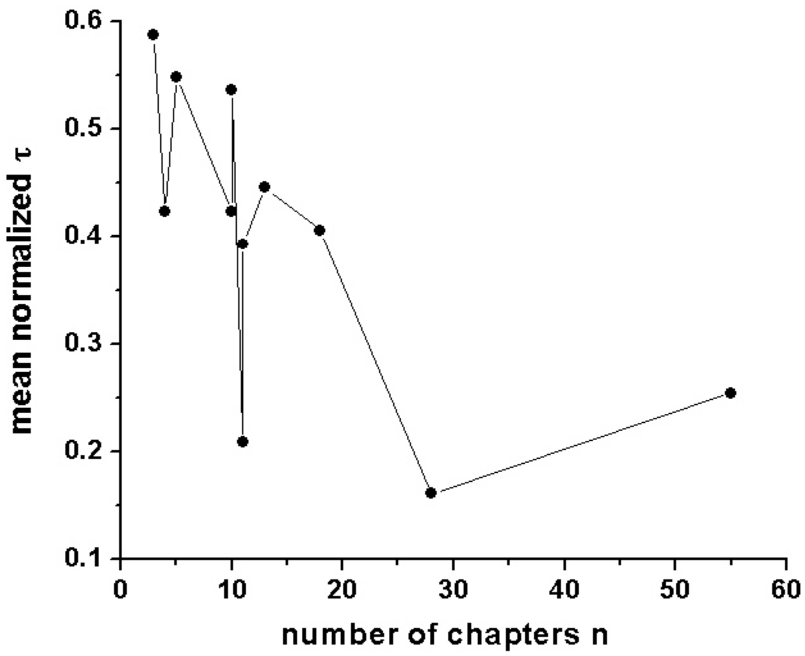


Figure 3.4. Deviations of normalized mean τ from 0.5 with increasing n

3.4. Prospective dissimilarity

There is still another view signalizing the change of the regime if a new chapter is added. In this case we cumulate the first i chapters and compare their common value with that of chapter i+1. This variant can be called *cumulative dissimilarity with prospective view*. In practice, first the i chapters are considered a whole, the rank-frequency sequence is determined, then the quantities V, $f(1)$, h are computed and compared with the respective quantities of chapter i+1.

Table 3.10
Prospective dissimilarity in Chamisso´s *Peter Schlemihl*

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	884	82	18	2	808	84	16	0.9999	0.0111
1+2	1435	166	25	3	630	70	14	1.0000	0.0066
1+2+3	1773	236	28	4	1209	123	20	0.9995	0.0309
1+…+4	2480	359	35	5	853	79	18	0.9987	0.0519
1+…+5	2877	428	40	6	801	75	17	0.9985	0.0548
1+…+6	3211	503	44	7	670	44	13	0.9960	0.0900
1+…+7	3460	547	44	8	788	80	16	0.9984	0.0561
1+…+8	3754	627	48	9	593	96	14	0.9999	0.0118
1+…+9	3979	723	49	10	536	52	11	0.9965	0.0834
1+…+10	4219	775	50	11	656	66	14	0.9966	0.0819

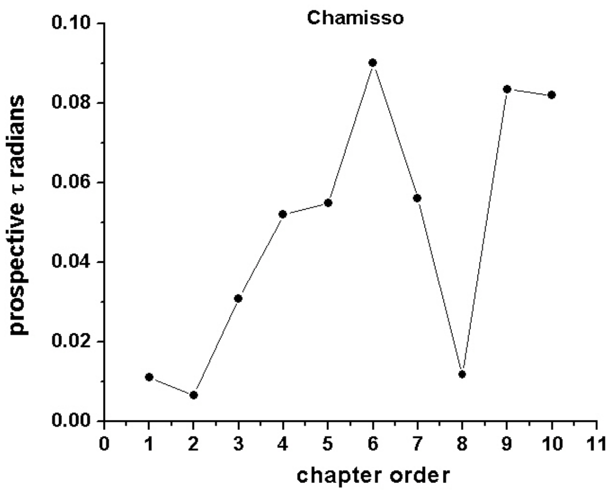


Figure 3.5. Prospective dissimilarity in Chamisso´s Peter Schlemihl

Table 3.11
Prospective dissimilarity in Eichendorff

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	1079	177	21	2	1287	210	25	1.0000	0.0008
1 + 2	1891	387	32	3	1334	182	28	0.9978	0.0664
1 + 2+ 3	2583	569	42	4	739	79	16	0.9939	0.1105
1 + …+ 4	2872	648	44	5	699	70	16	0.9925	0.1223
1 + …+ 5	3096	718	46	6	1059	130	22	0.9944	0.1059
1 + …+ 6	3482	843	51	7	932	121	20	0.9941	0.1086
1 + …+ 7	3765	964	55	8	1320	159	25	0.9915	0.1308
1 + …+ 8	4242	1123	59	9	1185	155	22	0.9917	0.1288
1 + …+ 9	4706	1278	63	10	1073	131	22	0.9897	0.1439

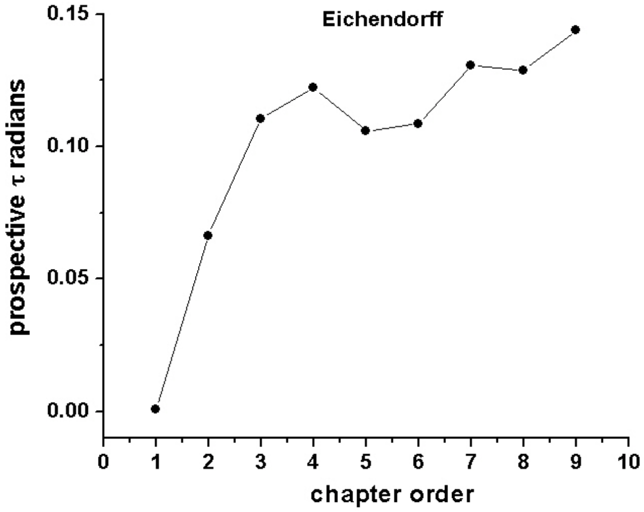


Figure 3.6. Prospective dissimilarity in Eichendorff

Table 3.12
Prospective dissimilarity in Kafka

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
3	513	34	12	4	321	23	10	10.000	0.0077
3+4	715	44	16	5	166	14	5	0.9997	0.0239
3+4+5	821	62	18	6	137	6	4	0.9995	0.0324
3+…+6	901	66	18	7	89	9	4	0.9994	0.0347
3+…+7	933	69	19	8	102	9	4	0.9998	0.0198

3+…+8	974	71	19	9	343	25	9	10.000	0.0067
3+…+9	1171	94	20	10	62	4	4	0.9988	0.0498
3+…+10	1103	98	21	11	104	9	5	0.9997	0.0242
3+…+11	1234	104	21	12	101	9	5	0.9995	0.0327
3+…+12	1272	111	22	13	150	9	6	0.9994	0.0353
3+…+13	1341	117	23	14	104	11	3	0.9998	0.0217
3+…+14	1384	128	23	15	136	7	5	0.9990	0.0440
3+…+15	1444	135	23	16	177	10	6	0.9992	0.0409
3+…+16	1533	143	25	17	80	11	3	0.9988	0.0484
3+…+17	1559	154	25	18	48	3	3	0.9987	0.0509
3+…+18	1576	157	25	19	33	3	2	0.9990	0.0453
3+…+19	1587	158	25	20	539	74	15	0.9992	0.0389

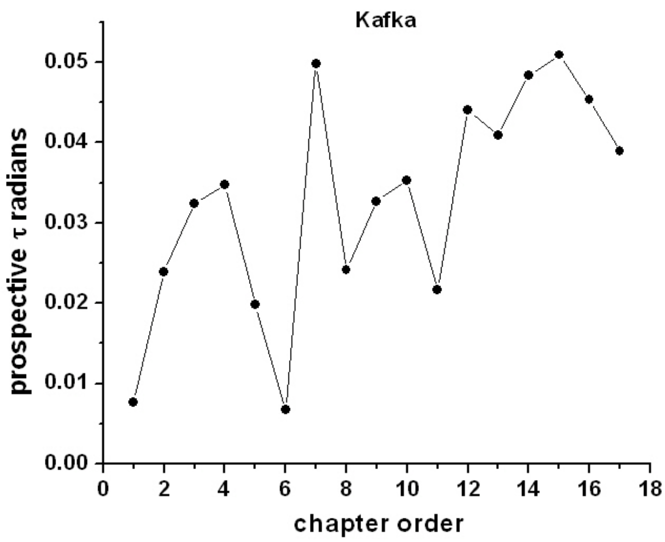


Figure 3.7. Prospective dissimilarity in Kafka

Table 3.13
Prospective dissimilarity in Löns

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	706	95	15	2	928	141	23	0.9998	0.0174
1+2	1366	236	28	3	1162	172	26	0.9997	0.0242
1+2+3	2059	408	38	4	1081	167	24	0.9991	0.0425
1+…+4	2535	575	43	5	1235	254	28	0.9998	0.0210
1+…+5	3055	829	51	6	1364	244	29	0.9961	0.0881

1+…+6	3551	1073	59	7	1862	414	36	0.9972	0.0747
1+…+7	4273	1487	69	8	1724	328	31	0.9892	0.1469
1+…+8	4911	1815	74	9	2126	453	39	0.9896	0.1441
1+…+9	5687	2268	85	10	1736	274	35	0.9752	0.2230
1+…+10	6216	2542	89	11	1294	217	27	0.9754	0.2221
1+…+11	6541	2759	92	12	1318	221	26	0.9730	0.2331
1+…+12	6820	2980	95	13	556	60	14	0.9540	0.3046

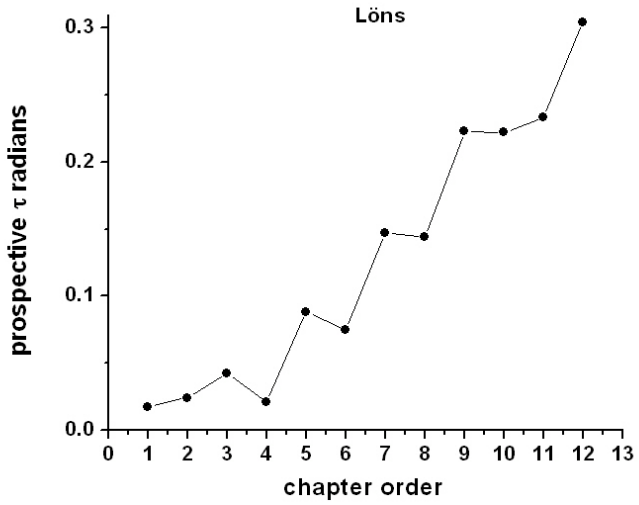


Figure 3.8. Prospective dissimilarity in Löns

Table 3.14
Prospective dissimilarity in Meyer

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	801	56	14	2	331	26	8	0.9999	0.0109
1+2	1027	74	16	3	551	46	11	0.9999	0.0122
1+2+3	1398	118	20	4	1142	79	18	0.9999	0.0152
1+…+4	2206	197	29	5	658	47	12	0.9998	0.0185
1+…+5	2558	234	33	6	471	34	10	0.9998	0.0209
1+…+6	2789	259	35	7	652	47	13	0.9998	0.0219
1+…+7	3151	306	37	8	556	43	11	0.9998	0.0212
1+…+8	3415	349	38	9	441	40	9	0.9999	0.0147
1+…+9	3624	389	38	10	493	41	11	0.9996	0.0267
1+…+10	3828	430	40	11	1079	88	17	0.9995	0.0309

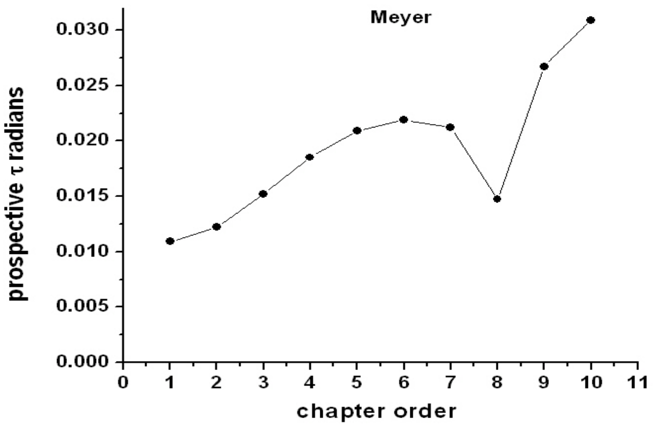


Figure 3.9. Prospective dissimilarity in Meyer

Table 3.15
Prospective dissimilarity in Novalis

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	1129	139	21	2	1487	208	22	0.9999	0.0169
1 + 2	2235	347	30	3	1819	233	25	0.9996	0.0266
1 + 2 + 3	3345	580	38	4	1282	130	18	0.9975	0.0707
1 +…+ 4	3941	710	41	5	2769	473	35	1.0000	0.0093
1 +…+ 5	5404	1183	53	6	1467	178	23	0.9955	0.0949
1 +…+ 6	5883	1361	58	7	792	77	16	0.9915	0.1308
1 +…+ 7	6098	1438	60	8	816	75	17	0.9902	0.1404
1 +…+ 8	6320	1501	62	9	2681	442	32	0.9976	0.0698
1 +…+ 9	7472	1943	71	10	1939	238	26	0.9913	0.1323

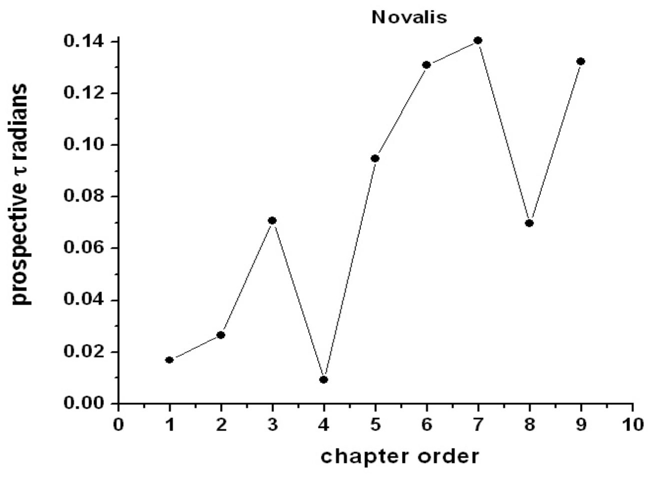


Figure 3.10. Prospective dissimilarity in Novalis

Table 3.16
Prospective dissimilarity in Paul

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	487	37	10	2	255	14	6	0.9998	0.0212
1 + 2	656	51	12	3	311	26	8	1.0000	0.0094
1 + 2+ 3	868	77	16	4	354	21	8	0.9996	0.0295
1 + …+ 4	1079	98	18	5	677	44	12	0.9997	0.0257
1 + …+ 5	1505	142	22	6	305	16	8	0.9991	0.0432
1 + …+ 6	1661	157	24	7	316	15	7	0.9989	0.0474
1 + …+ 7	1814	169	25	8	248	22	6	0.9999	0.0113
1 + …+ 8	1925	191	26	9	547	37	10	0.9995	0.0317
1 + …+ 9	2203	228	28	10	778	53	13	0.9994	0.0353
1 + …+ 10	2629	281	32	11	1027	84	15	0.9997	0.0250
1 + …+ 11	3182	365	38	12	365	25	8	0.9989	0.0469
1 + …+ 12	3354	390	40	13	652	40	13	0.9985	0.0551
1 + …+ 13	3624	430	42	14	714	49	14	0.9987	0.0502
1 + …+ 14	3941	479	44	15	793	65	15	0.9992	0.0399
1 + …+ 15	4327	544	47	16	223	12	5	0.9974	0.0722
1 + …+ 16	4394	553	48	17	897	73	15	0.9990	0.0444
1 + …+ 17	4877	621	50	18	489	42	11	0.9991	0.0428
1 + …+ 18	5103	663	51	19	676	38	13	0.9973	0.0736
1 + …+ 19	5423	695	53	20	1011	78	16	0.9987	0.0508
1 + …+ 20	5960	766	54	21	1513	172	24	0.9999	0.0161
1 + …+ 21	6704	938	59	22	302	15	7	0.9959	0.0905
1 + …+ 22	6809	953	60	23	386	26	9	0.9973	0.0732
1 + …+ 23	6954	979	60	24	730	80	13	0.9995	0.0320
1 + …+ 24	7213	1059	61	25	361	18	8	0.9953	0.0969
1 + …+ 25	7311	1077	62	26	887	61	15	0.9970	0.0781
1 + …+ 26	7685	1134	63	27	410	26	9	0.9964	0.0843
1 + …+ 27	7785	1160	63	28	172	8	5	0.9946	0.1036
1 + …+ 28	7823	1166	63	29	872	68	14	0.9975	0.0706
1 + …+ 29	8137	1234	65	30	238	17	6	0.9967	0.0810
1 + …+ 30	8189	1251	65	31	753	72	14	0.9984	0.0573
1 + …+ 31	8409	1323	66	32	119	6	4	0.9941	0.1088
1 + …+ 32	8436	1327	66	33	355	23	8	0.9957	0.0925
1 + …+ 33	8548	1348	67	35	897	82	17	0.9978	0.0662
1 + …+ 35	8833	1431	70	36	253	15	7	0.9947	0.1033
1 + …+ 36	8891	1446	71	37	239	12	6	0.9937	0.1124
1 + …+ 37	8936	1455	71	38	636	40	12	0.9951	0.0992
1 + …+ 38	9151	1495	72	39	248	13	7	0.9938	0.1114

1 + …+ 39	9213	1504	72	40	655	53	14	0.9966	0.0822
1 + …+ 40	9392	1557	74	41	546	43	11	0.9963	0.0866
1 + …+ 41	9565	1600	74	42	731	50	13	0.9952	0.0980
1 + …+ 42	9775	1650	75	43	1591	152	26	0.9974	0.0725
1 + …+ 43	10452	1802	79	44	896	66	15	0.9952	0.0976
1 + …+ 44	10758	1868	81	45	1102	155	18	0.9994	0.0334
1 + …+ 45	11233	2023	81	46	1303	99	21	0.9947	0.1027
1 + …+ 46	11643	2123	84	47	319	19	8	0.9925	0.1222
1 + …+ 47	11709	2142	84	48	604	50	13	0.9951	0.0994
1 + …+ 48	11866	2192	85	49	336	19	8	0.9919	0.1273
1 + …+ 49	11929	2211	85	50	255	23	7	0.9954	0.0955
1 + …+ 50	11965	2234	85	51	1323	116	20	0.9953	0.0975
1 + …+ 51	12387	2352	88	52	815	71	15	0.9949	0.1014
1 + …+ 52	12625	2423	90	53	864	75	14	0.9947	0.1034
1 + …+ 53	12860	2499	92	54	2417	245	30	0.9959	0.0911
1 + …+ 54	13927	2744	95	55	2680	321	33	0.9971	0.0755
1 + …+ 55	15015	3066	99	56	482	47	10	0.9945	0.1052

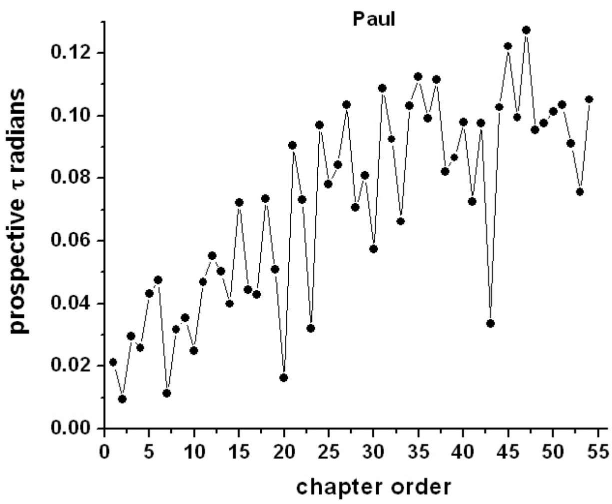


Figure 3.11. Prospective dissimilarity in Paul

Table 3.17
Prospective dissimilarity in Sealsfield

Chapters	***V***	***f*(1)**	***h***	**Chapter**	***V***	***f*(1)**	***h***	***cos τ***	***τ***
1	600	45	13	2	1825	142	27	1.0000	0.0074
1 + 2	2181	177	31	3	1197	114	21	0.9999	0.0144
1 + 2+ 3	2890	268	35	4	1399	161	24	0.9997	0.0227

1 + …+ 4	3655	380	41	5	1079	96	22	0.9998	0.0174
1 + …+ 5	4158	433	46	6	1010	67	20	0.9993	0.0385
1 + …+ 6	4638	497	51	7	1035	75	20	0.9994	0.0354
1 + …+ 7	5003	569	54	8	1333	138	27	0.9999	0.0138
1 + …+ 8	5516	707	62	9	2295	263	31	0.9999	0.0136
1 + …+ 9	6710	927	69	10	1620	138	26	0.9986	0.0526
1 + …+ 10	7395	1044	73	11	1265	98	26	0.9980	0.0638
1 + …+ 11	7785	1133	76	12	1191	95	20	0.9979	0.0653
1 + …+ 12	8149	1218	78	13	1071	89	17	0.9978	0.0658
1 + …+ 13	8552	1307	80	14	1198	82	19	0.9965	0.0836
1 + …+ 14	9019	1388	83	15	1545	164	27	0.9989	0.0477
1 + …+ 15	9603	1552	85	16	1602	137	26	0.9972	0.0753
1 + …+ 16	10116	1689	89	17	2273	192	30	0.9967	0.0813
1 + …+ 17	10899	1880	93	18	1252	285	24	0.9985	0.0540
1 + …+ 18	11366	1987	96	19	1653	171	29	0.9975	0.0706
1 + …+ 19	11937	2096	98	20	2735	273	35	0.9972	0.0745
1 + …+ 20	12994	2363	102	21	2040	220	29	0.9974	0.0727
1 + …+ 21	13605	2541	106	22	1655	157	29	0.9959	0.0906
1 + …+ 22	14031	2640	111	23	799	80	14	0.9962	0.0867
1 + …+ 23	14249	2696	111	24	753	68	14	0.9952	0.0975
1 + …+ 24	14451	2740	112	25	704	40	12	0.9914	0.1310
1 + …+ 25	14651	2771	112	26	679	44	15	0.9924	0.1231
1 + …+ 26	14775	2793	112	27	1516	179	24	0.9976	0.0698
1 + …+ 27	15194	2919	115	28	586	70	15	0.9973	0.0731

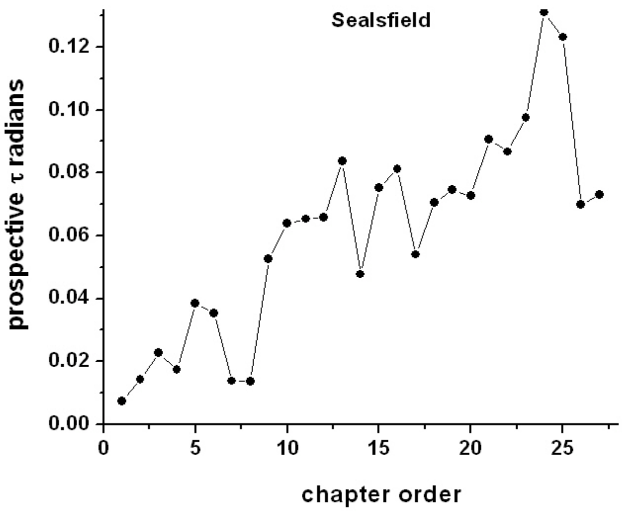


Figure 3.12. Prospective dissimilarity in Sealsfield

We omitted Tucholsky, Wedekind and Hoffmann because their texts were very short.

The prospective dissimilarity is in general increasing but the deviations from the smooth course are large hence no "smooth" curves could prove as adequate. Hence we must consider the deviations as part of structure. If we interpret local minima as chapters in which the writer recapitulates or evaluates the events contained in the previous chapters, which does not increase the dissimilarity of the new chapter, then such a fluctuating course is a characteristic feature of the work. In order to express this quality numerically, we use the fact that the number of local minima (m_L) cannot be greater than the half of the number of chapters ($C/2$). Hence the *recapitulative structure* (*RS*) can be expressed as

$$(3.6) \quad RS = \frac{2m_L}{n-1}.$$

Since this indicator is a simple proportion, there are no problems with its statistical treatment.

For example, in Table 3.17 or Figure 3.12 one finds 8 local minima and 27 comparisons. This is usually equal to the number of chapters minus 1, but in some cases one can omit a chapter for different reasons, e.g. chapter 34 in Paul containing only one sentence; in Kafka we considered only "Betrachtung" (see Appendix I). Hence *RS*(Sealsfield) = 2(8)/27 = 0.59. The indicator *RS* for all analyzed German authors is presented in increasing order in Table 3.18.

Table 3.18
The recapitulative structure (*RS*) of German authors

Author	m_L	n**-1**	***RS***
Meyer	2	10	0.40
Sealsfield	8	27	0.59
Chamisso	3	10	0.60
Eichendorff	3	9	0.67
Novalis	3	9	0.67
Paul	18	54	0.67
Kafka	6	17	0.71
Löns	5	12	0.83

As can be seen, neither the number of compared chapters nor the date of historical origin correlate with *RS*, hence *RS* can be considered a structural property of the text. Of course, an indicator equivalent to *RS* could be computed by direct vocabulary comparison, i.e. using method (A) indicated in Chapter 1.

However, the leaps between individual chapters may differ in spite of equal *RS*. In order to compare them, we compute again the prospective *AS* according to (3.5). In this case we obtain the results presented in Table 3.19.

Table 3.19
Absolute sequential difference of prospective τ-s

Author	***n*-1**	**Sum**	***AS***
Kafka	16	0.1958	0.0122
Sealsfield	26	0.3863	0.0149
Meyer	9	0.1622	0.0180
Eichendorff	8	0.1799	0.0225
Paul	53	1.1946	0.0225
Chamisso	9	0.2392	0.0266
Löns	11	0.3644	0.0331
Novalis	8	0.3794	0.0477

As can be seen again, the *AS*-values do not correlate with the number of comparisons (chapters taken into account). Kafka seems to have the most monotonous deployment – even if the respective figure is very fluctuating, a fact caused by the scaling of the *y*-axis. Novalis has the most oscillating deployment; he seems to be a writer with great thematic jumps.

4. Vectorial method of text comparison

4.1. Comparisons of texts

Needless to say, the vector can be used for intertextual comparisons, too. There are two possibilities:

(a) One compares each chapter of one text with each chapter of another text. The mean of all dissimilarities is considered the dissimilarity of the given texts.

(b) One takes each text as a whole and compares only the vectors resulting from the whole text.

The first kind of comparison is much more detailed and yields a deeper insight into the dynamics of the texts. The latter kind is rather categorical but sufficient for classification or the study of historical changes.

Since the number of texts in our investigation is not sufficient to study the history of German writing, we concentrate on the hypothesis that *internal dissimilarity* (when the compared parts belong to the same author) *is smaller than the external* (when the compared parts belong to different authors). Though the hypothesis is quite plausible, it need not hold in any case because we compare quite abstract formations detached from the text. Breaks within the text can be very great on reasons mentioned above. On the other hand, even a smaller intertextual dissimilarity may not be significantly smaller. In order to state it, one must perform a test (see below). In Table 4.1 the internal dissimilarities in Chamisso's *Peter Schlemihl* are presented.

Table 4.1
Chapter to chapter dissimilarities in Chamisso´s Peter Schlemihl

Ch.	1	2	3	4	5	6	7	8	9	10	11
1	0										
2	0.0111	0									
3	0.0075	0.0182	0								
4	0.0108	0.0039	0.0097	0							
5	0.0101	0.0183	0.0113	0.0008	0						
6	0.0010	0.0093	0.0173	0.0103	0.0012	0					
7	0.0278	0.0268	0.0359	0.0452	0.0380	0.0269	0				
8	0.0356	0.0079	0.0089	0.0037	0.0097	0.0025	0.0087	0			
9	0.0594	0.0950	0.0672	0.0682	0.0595	0.0498	0.0570	0.0681	0		
10	0.0638	0.0045	0.0311	0.0034	0.0044	0.0061	0.0140	0.0069	0.0042	0	
11	0.0036	0.0602	0.0014	0.0347	0.0069	0.0079	0.0049	0.0104	0.0037	0.0078	0
τ_{min} = 0.0008, τ_{max} = 0.0950, mean τ = 0.0222, stdev τ = 0.0233											

Table 4.2
Intertextual mean τ-angles

		Chamisso	Eichendorff	Hoffmann	Kafka	Löns	Meyer	Novalis	Paul	Sealsfield	Tucholsky	Wedekind
	Chapters	11	10	3	18	13	11	10	55	28	5	4
Chamisso	11	0.0222	0.0337	0.0315	0.0392	0.0677	0.0269	0.0333	0.0327	0.0266	0.0287	0.0268
Eichendorff	10	0.0337	0.0238	0.0478	0.0553	0.0444	0.0519	0.0265	0.0555	0.0424	0.0205	0.0368
Hoffmann	3	0.0315	0.0478	0.0408	0.0384	0.0850	0.0229	0.0461	0.0273	0.0301	0.0420	0.0338
Kafka	18	0.0392	0.0553	0.0384	0.0353	0.0893	0.0308	0.0557	0.0343	0.0400	0.0516	0.0418
Löns	13	0.0677	0.0444	0.0850	0.0893	0.0371	0.0905	0.0494	0.0933	0.0774	0.0481	0.0684
Meyer	11	0.0269	0.0519	0.0229	0.0308	0.0905	0.0091	0.0487	0.0175	0.0235	0.0460	0.0287
Novalis	10	0.0333	0.0265	0.0461	0.0557	0.0494	0.0487	0.0311	0.0528	0.0408	0.0239	0.0364
Paul	55	0.0327	0.0555	0.0273	0.0343	0.0933	0.0175	0.0528	0.0229	0.0300	0.0496	0.0344
Sealsfield	28	0.0266	0.0424	0.0301	0.0400	0.0774	0.0235	0.0408	0.0300	0.0289	0.0368	0.0306
Tucholsky	5	0.0287	0.0205	0.0420	0.0516	0.0481	0.0460	0.0239	0.0496	0.0368	0.0196	0.0317
Wedekind	4	0.0268	0.0368	0.0338	0.0418	0.0684	0.0287	0.0364	0.0344	0.0306	0.0317	0.0389
Total chapters	168											

If we compare the intertextual mean τ-angles as shown in Table 4.2, we see that Chamisso is internally more uniform than externally. His mean τ (marked grey) is smaller than the numbers in the same line representing other texts. However, looking at the column of Wedekind we see that the majority of authors have a smaller dissimilarity with him than Wedekind internally. This is a stimulus for considering the origin of his given work and the way of its writing, a task rather for literary scientists.

If we compare the sum of dissimilarities of individual authors, the year of writing and the number of chapters in the work, we do not find any correlation. From this fact we can conclude that at such an abstract level as we are working, each text is an original individual creation whose rather chaotic dissimilarities cannot be imitated. Nevertheless, comparing the mean internal dissimilarities with those of external ones, it can easily be seen (cf. Table 4.3) that – at least for the data at our disposal – except for two authors (Wedekind, Hoffmann) the external dissimilarities are greater, though the difference with these two authors is minimal. This testifies to the fact that there is some kind of internal structural unity in each work concealed behind a pattern in repeating words.

Table 4.3
Internal and external dissimilarities

Author	**Internal mean dissimilarities (increasing τ)**	**External mean dissimilarities**
Meyer	0.0091	0.0387
Tucholsky	0.0196	0.0377
Chamisso	0.0222	0.0347
Paul	0.0229	0.0427
Eichendorff	0.0238	0.0415
Sealsfield	0.0289	0.0324
Novalis	0.0311	0.0414
Kafka	0.0353	0.0476
Löns	0.0371	0.0714
Wedekind	0.0389	0.0330
Hoffmann	0.0408	0.0405

4.2. Cross-linguistic comparison

The most effective interlinguistic comparison of texts may be performed using the same text in all languages. In that case a number of undesired factors can be eliminated, e.g. individuality, style, genre, and one attains a kind of homogeneity at least in the object described by the text. Needless to say, different identical texts may yield different dissimilarities, hence operating with one single text is a good start but in no case a final result.

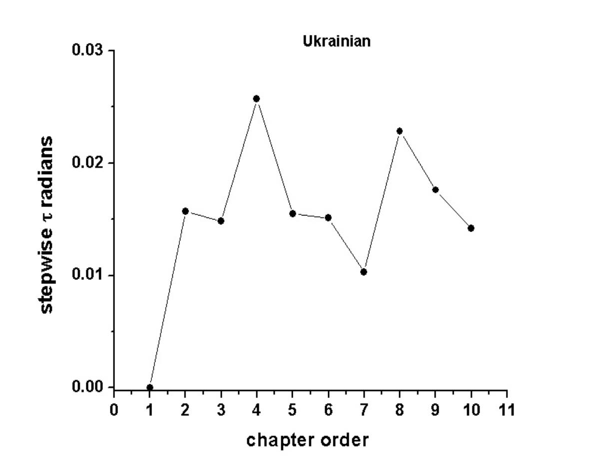
But even here several varieties of comparison are possible: (1) Comparing a chapter of the text in one language with the same chapter in another language and taking the mean τ, or (2) taking the given text in one language as a whole, computing its vector and comparing it with that of another language (taking the text also as a whole), or finally (3) performing in each language the same procedure as in German (comparing all chapters with the first) and finally compare the resulting sequences.

Here we shall use again E. Kelih's (2009, 2009a) Slavic parallel corpus made up of the first ten chapters of N. Ostrovskij's novel "How the steel was tempered" translated from Russian in 11 Slavic languages.

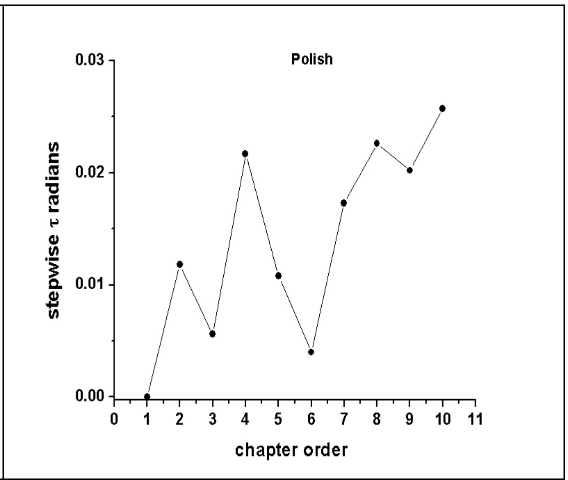
We begin with procedure (3) and perform the same analysis as above, i.e. each chapter of the individual translations will be compared with its first chapter. The same will be done with the original Russian text which serves as a comparative background. In this way we obtain the results presented in Table 4.4. The stepwise dissimilarities are graphically presented in Figure 4.1.

Table 4.4
Stepwise comparison of chapters in Ostrovskij's "How the steel was tempered" in 12 Slavic languages. The values represent τ radians

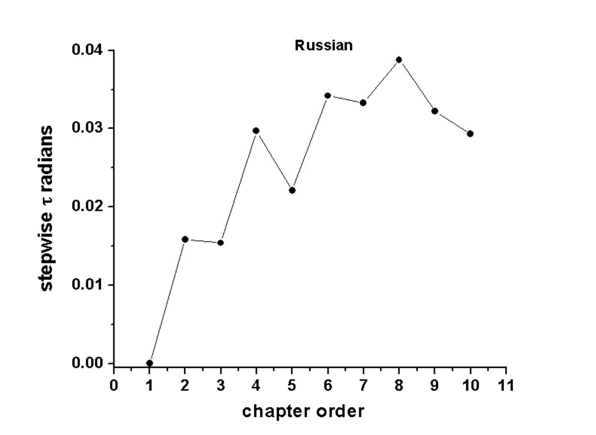
Chapter	**Russian**	**Belorussian**	**Ukrainian**	**Polish**	**Czech**	**Slovak**
1	0	0	0	0	0	0
2	0.0158	0.0177	0.0157	0.0118	0.0150	0.0092
3	0.0154	0.0148	0.0148	0.0056	0.0056	0.0037
4	0.0297	0.0302	0.0257	0.0217	0.0152	0.0201
5	0.0221	0.0238	0.0155	0.0108	0.0135	0.0100
6	0.0342	0.0356	0.0151	0.0040	0.0049	0.0037
7	0.0333	0.0377	0.0103	0.0173	0.0155	0.0190
8	0.0388	0.0387	0.0228	0.0226	0.0246	0.0176
9	0.0322	0.0382	0.0176	0.0202	0.0234	0.0156
10	0.0293	0.0390	0.0142	0.0257	0.0260	0.0246
Total	0.2509	0.2757	0.1519	0.1397	0.1437	0.1235
Chapter	**Sorbian**	**Bulgarian**	**Macedonian**	**Serbian**	**Croatian**	**Slovenian**
1	0	0	0	0	0	0
2	0.0226	0.0245	0.0177	0.0174	0.0179	0.0308
3	0.0139	0.0081	0.0033	0.0050	0.0041	0.0047
4	0.0207	0.0361	0.0367	0.0337	0.0316	0.0680
5	0.0304	0.0253	0.0247	0.0026	0.0038	0.0087
6	0.0194	0.0195	0.0075	0.0147	0.0126	0.0085
7	0.0281	0.0086	0.0051	0.0170	0.0176	0.0439
8	0.0405	0.0231	0.0062	0.0213	0.0207	0.0621
9	0.0480	0.0356	0.0270	0.0024	0.0009	0.0694
10	0.0329	0.0153	0.0111	0.0077	0.0114	0.0511
Total	0.2566	0.1963	0.1392	0.1218	0.1206	0.3473



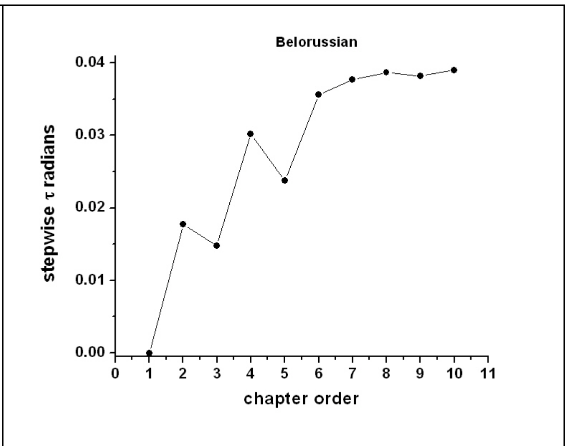
Ukrainian



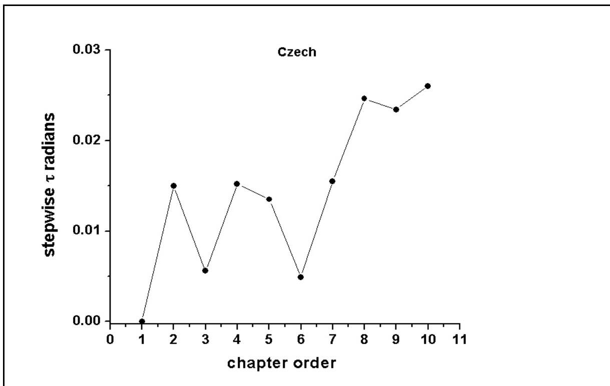
Polish



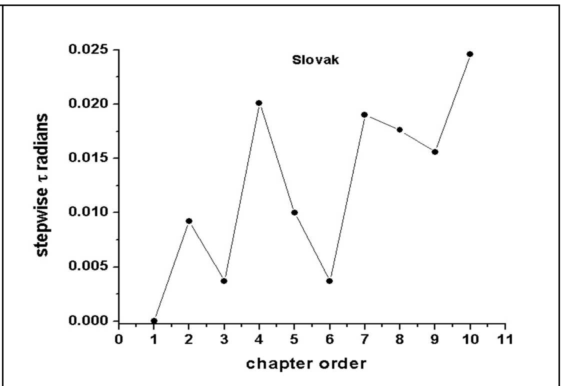
Russian



Belorussian



Czech



Slovak

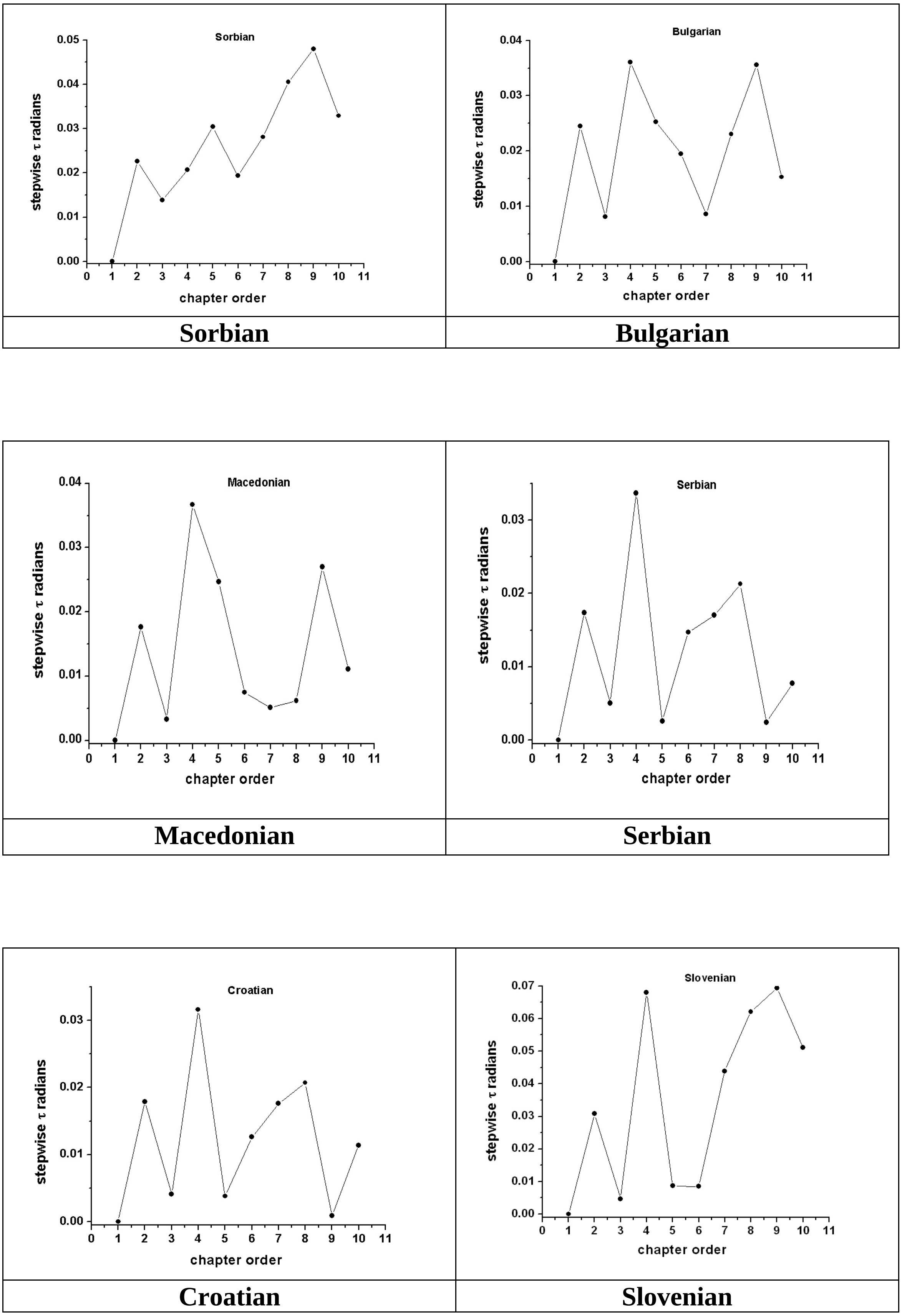


Figure 4.1. Stepwise dissimilarities of the same text in 12 Slavic languages

Now, if we take simply the *sum of these dissimilarities*, i.e. the sum of τ-angles, we obtain a very abstract picture of similarity between Russian and other Slavic languages. Looking at the "Total" row in Table 4.4 the following order can be discerned:

	Total τ
Slovenian	0.3473
Belorussian	0.2757
Sorbian	0.2566
Russian	0.2509
Bulgarian	0.1963
Ukrainian	0.1519
Czech	0.1437
Polish	0.1397
Macedonian	0.1392
Slovak	0.1235
Serbian	0.1218
Croatian	0.1206

However, in spite of the given sum the distance between individual steps may diverge differently. In order to measure it, we may use simply the standard deviation or some fractal measures. Computing the Euclidean distance would not be adequate because the τ-angles are too small as compared with chapter order – which is always 1. Let us first compute the standard deviation according to (3.4) of τ-angles and the mean of the absolute sequential differences. We obtain the following orders

	s		AS
Slovenian	0.0277	Macedonian	0.5393
Sorbian	0.0136	Slovenian	0.2589
Belorussian	0.0132	Bulgarian	0.1437
Macedonian	0.0121	Croatian	0.1341
Russian	0.0118	Serbian	0.1322
Bulgarian	0.0118	Sorbian	0.1023
Serbian	0.0106	Polish	0.0782
Croatian	0.0101	Slovak	0.0754
Czech	0.0088	Czech	0.0679
Polish	0.0088	Russian	0.0661
Slovak	0.0082	Ukrainian	0.0642
Ukrainian	0.0069	Belorussian	0.0587

The mean sequential difference shows the relationship of neighbouring chapters and may be a result of the breaks (stylistic deviations, explication, over-simplification, etc.) made by the translator between successive chapters. The elucidating of backgrounds of these facts must be left to future research.

Evidently, these results do not express exclusively the formal difference between languages but also the discrepancies which arose in the course of translating and editing the texts. In order to show whether the rankings are equal, we compute Kendall´s concordance coefficient for k = 3 rankings of n = 12 languages. If we transform the above results of *Total*, *s* and *AS* in ranks we obtain the rankings as given in Table 4.5. The measure expressing the agreement between the rankings is defined as (cf. Gibbons 1971: 252)

$$(4.1) \quad W = \frac{12S}{k^2 n(n^2 - 1)},$$

where

$$(4.2) \quad S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2 .$$

Taking the appropriate values from Table 4.5 we obtain

$$W = \frac{12(653)}{3^3(12)(12^2 - 1)} = 0.5074 .$$

Table 4.5
Ranking of τ-total, standard deviations and *AS* in 12 Slavic languages

Language	**τ-total**	***s***	***AS***	**Sum *R***	***S***
Belorussian	2	3	12	17	6.25
Bulgarian	5	6	3	14	30.25
Croatian	12	8	4	24	20.25
Czech	7	9	9	25	30.25
Macedonian	9	4	1	14	30.25
Polish	8	10	7	25	30.25
Russian	4	5	10	19	0.25
Serbian	11	7	5	23	12.25
Slovak	10	11	8	29	90.25
Slovenian	1	1	2	4	240.25
Sorbian	3	2	6	11	72.25
Ukrainian	6	12	11	29	90.25
				234	653

Since $W = 1$ designates perfect concordance and $W = 0$ no agreement, we see that $W = 0.5$ is exactly in the mid of the interval, designating rather randomness. Since $k(n\text{-}1)W$ is approximately distributed like a chi-square with 1 degree of freedom, we obtain $X^2 = 3(12\text{-}1)0.5074 = 16.74$ showing a significant disagreement of rankings, which indicates a personal factor in the course of translation. From the philological point of view, it means, that a translation – at least in this case – always contains a trace of translators individuality.

However, if the text is long and taken as a whole, breaks disappear and the result is statistically more stable. Pursuing variant (2) of the possibilities we compare the ten chapters of each language as a whole with Russian as a whole. In this way we obtain quite unequivocal result. The numerical values of τ radians can be seen in Table 4.6 and the graphical counterpart in Figure 4.2.

Table 4.6
Comparison of total texts with Russian

Language	**Text size *N***	$\boldsymbol{\tau}$
Russian	49663	0.0000
Belorussian	49874	0.0030
Polish	52736	0.0145
Ukrainian	49612	0.0241
Czech	52180	0.0488
Slovak	52093	0.0494
Bulgarian	57165	0.0592
Croatian	56415	0.0603
Serbian	56227	0.0629
Sorbian	58480	0.0649
Macedonian	58819	0.0935
Slovenian	62646	0.2088

This is a quite perfect result, which coincides with the analysis of the Type-Token relationship and in some way also with the geographical and areal affiliation of Slavic languages.

However, if we take into account also the text size, as shown in Table 4.6, we obtain a slightly different result presented in Figure 4.3 showing that the dissimilarity is linked with the difference in text size. This is a quite natural phenomenon: if there are more words in a Slavic language than in Russian, then the rank-frequency distributions (chapter or whole) display different Cartesian vector components and the dissimilarity increases. Thus both the difference in vocabulary size and in the vector is able to display the distance of a Slavic language from Russian (the basis of translation), which are mainly caused by morphological differences in the languages under examination.

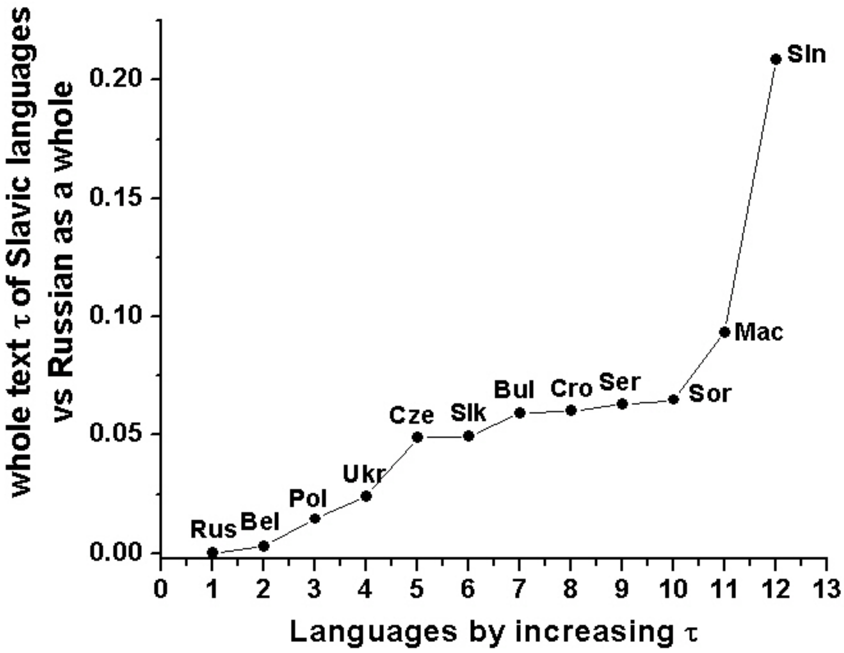


Figure 4.2. The dissimilarity of Slavic languages from Russian

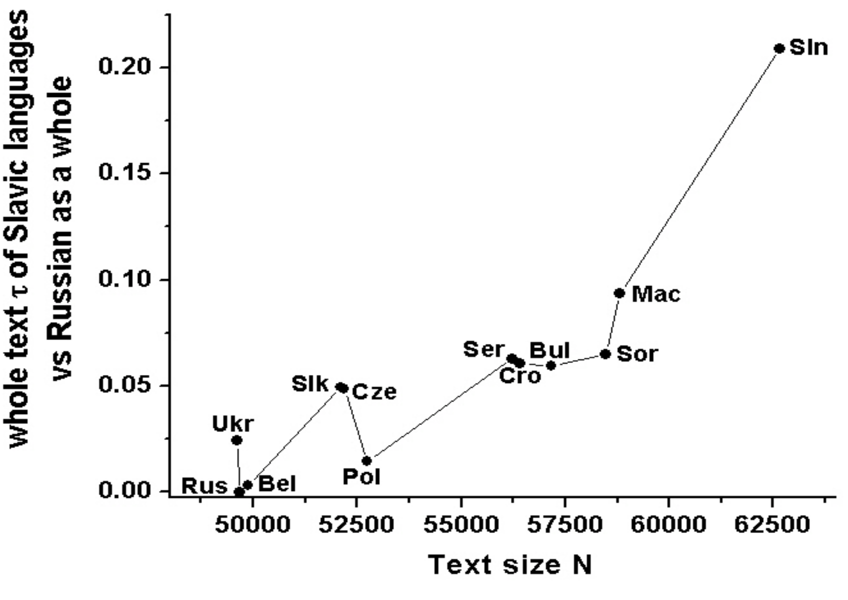


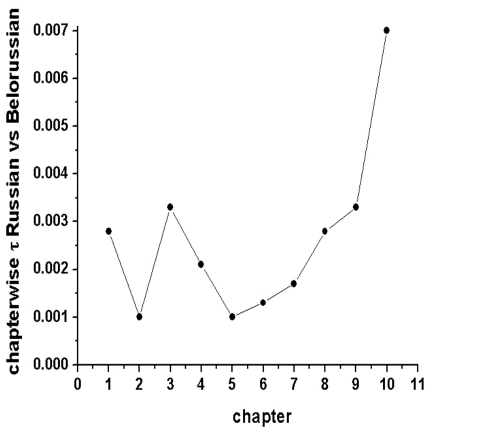
Figure 4.3. Dissimilarities of Slavic languages to Russian correlated with text size

Again, the relationship can be expressed by a power function with $R^2 = 0.90$. The parameters are not relevant because the independent variable is very large while the dependent one very small. A more lucid result could be attained if we performed some operation on both variables.

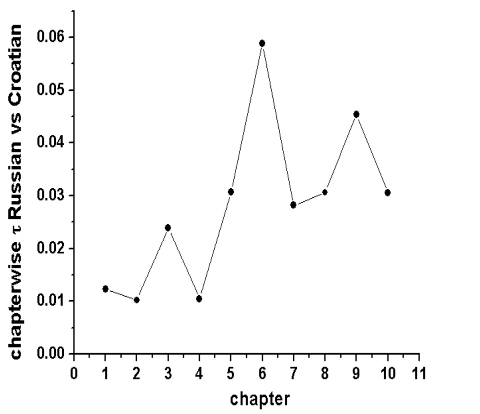
However, if we compare the Russian text chapter for chapter with parallel chapters in other languages, as can be seen in Table 4.7, the idea of a strong correlation of N with $\boldsymbol{\tau}$ must be abandoned. Ordering all τ-angles according to chapter size in any Slavic language we obtain a very strongly oscillating course which does not testify to a trend, as can be seen in Figure 4.4.

Table 4.7
Chapterwise comparison with Russian

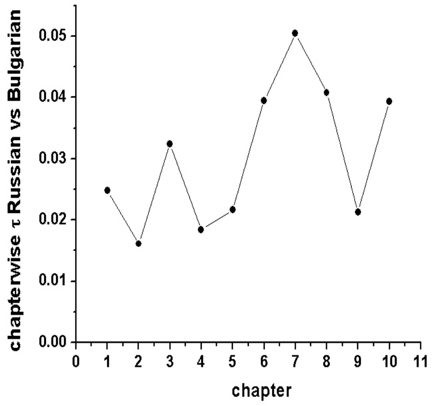
Chapter	Russian N	Belorussian N	τ
1	4107	4145	0.0028
2	4136	4177	0.0010
3	6323	6367	0.0033
4	3733	3791	0.0021
5	3769	3791	0.0010
6	7534	7547	0.0013
7	6019	6063	0.0017
8	5352	5362	0.0028
9	3291	3312	0.0033
10	5399	5319	0.0070



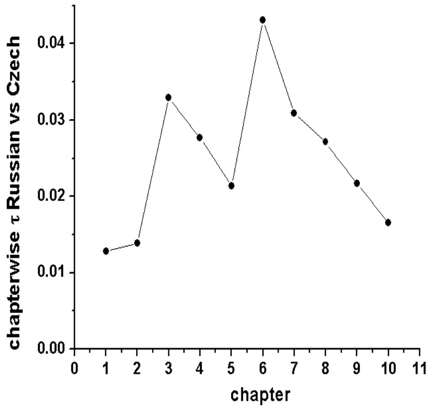
Chapter	Russian N	Croatian N	τ
1	4107	4582	0.0123
2	4136	4689	0.0102
3	6323	7160	0.0239
4	3733	4316	0.0105
5	3769	4255	0.0307
6	7534	8553	0.0589
7	6019	6841	0.0282
8	5352	6075	0.03061
9	3291	3760	0.0454
10	5399	6184	0.0306



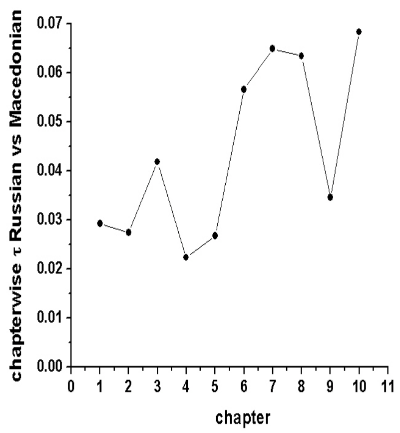
Chapter	Russian N	Bulgarian N	τ
1	4107	4653	0.0248
2	4136	4734	0.0161
3	6323	7224	0.0324
4	3733	4305	0.0184
5	3769	4277	0.0216
6	7534	8673	0.0395
7	6019	6992	0.0504
8	5352	6242	0.0408
9	3291	3787	0.0213
10	5399	6278	0.0393



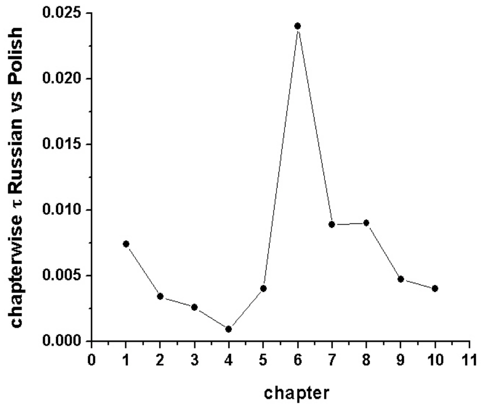
Chapter	Russian N	Czech N	τ
1	4107	3925	0.0128
2	4136	4381	0.0139
3	6323	6670	0.0329
4	3733	3920	0.0277
5	3769	3852	0.0214
6	7534	8117	0.0431
7	6019	6390	0.0309
8	5352	5738	0.0271
9	3291	3451	0.0217
10	5399	5736	0.0165



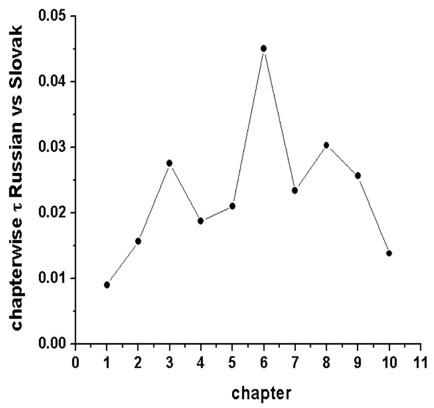
Chapter	Russian N	Macedonian N	τ
1	4107	4810	0.0293
2	4136	4898	0.0274
3	6323	7470	0.0418
4	3733	4424	0.0223
5	3769	4425	0.0267
6	7534	8914	0.0566
7	6019	7153	0.0648
8	5352	6414	0.0634
9	3291	3850	0.0345
10	5399	6461	0.0683



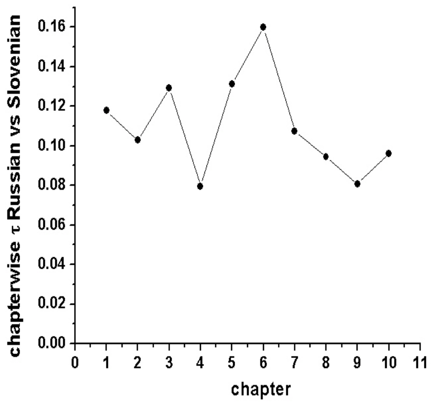
Chapter	Russian N	Polish N	τ
1	4107	4348	0.0074
2	4136	4368	0.0034
3	6323	6694	0.0026
4	3733	4003	0.0009
5	3769	3997	0.0040
6	7534	7937	0.0240
7	6019	6348	0.0089
8	5352	5753	0.0090
9	3291	3501	0.0047
10	5399	5786	0.0040



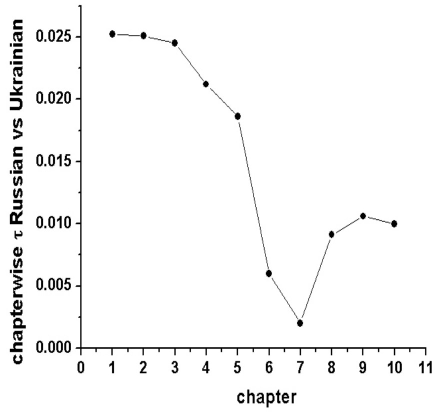
Chapter	Russian N	Slovak N	τ
1	4107	4275	0.0090
2	4136	4325	0.0156
3	6323	6496	0.0275
4	3733	3885	0.0187
5	3769	3862	0.0210
6	7534	8021	0.0450
7	6019	6337	0.0234
8	5352	5781	0.0303
9	3291	3412	0.0256
10	5399	5699	0.0138



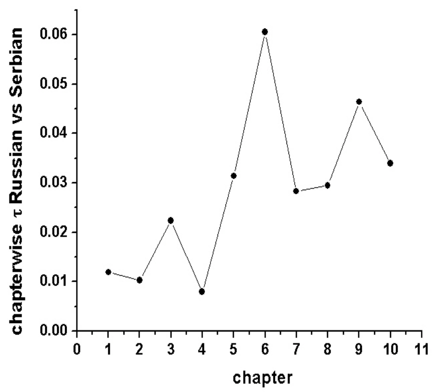
Chapter	Russian N	Slovenian N	τ
1	4107	5209	0.1178
2	4136	5199	0.1028
3	6323	7971	0.1293
4	3733	4787	0.0795
5	3769	4720	0.1312
6	7534	9546	0.1600
7	6019	7520	0.1074
8	5352	6822	0.0945
9	3291	4075	0.0807
10	5399	6797	0.0961



Chapter	Russian *N*	Ukrainian *N*	τ
1	4107	4119	0.0252
2	4136	4160	0.0251
3	6323	6282	0.0245
4	3733	3764	0.0212
5	3769	3755	0.0186
6	7534	7542	0.0060
7	6019	5999	0.0020
8	5352	5362	0.0091
9	3291	3278	0.0106
10	5399	5351	0.0100



Chapter	Russian *N*	Serbian *N*	τ
1	4107	4579	0.0119
2	4136	4656	0.0103
3	6323	7093	0.0224
4	3733	4290	0.0079
5	3769	4241	0.0314
6	7534	8566	0.0605
7	6019	6816	0.0283
8	5352	6029	0.0295
9	3291	3749	0.0464
10	5399	6208	0.0340



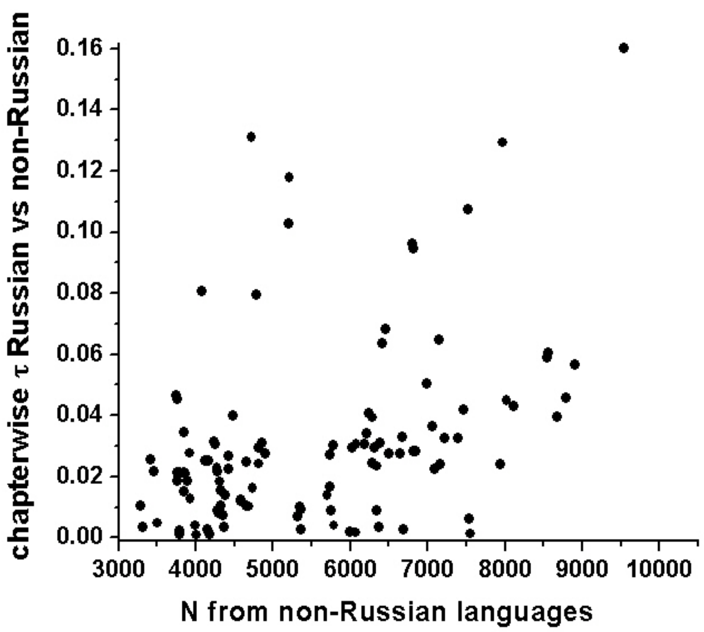


Figure 4.4. Chapterwise link between chapter size and τ for 11 languages

Using the results in Table 4.7 we compute again the *AS* indicator – which has a certain similarity with Lyapunov´s coefficient – for the individual chapter differences in τ, i.e. we compute the sum of the neighbouring τ-angles. For example, in Slovenian, we find the dissimilarities (τ radians) to parallel chapters of the Russian original as given in Table 4.7 and reproduced in Table 4.8. The absolute difference between the first two values is |0.1178 - 0.1028| = 0.0150; all of them are presented in the second column of Table 4.8 (all values rounded). It is not necessary to divide the sum by 9 because we have 10 chapters in all texts. Hence $AS = \sum|\tau_i - \tau_{i+1}|$, (i = 1,..9)

Table 4.8
τ-angles of parallel chapters in Russian and Slovenian

Chapter	$\boldsymbol{\tau}$	$\boldsymbol{D_{i,i+1}}$
1	0.1178	0.0150
2	0.1028	0.0264
3	0.1293	0.0498
4	0.0795	0.0517
5	0.1312	0.0289
6	0.1600	0.0527
7	0.1073	0.0129
8	0.0945	0.0138
9	0.0807	0.0154
10	0.0961	
	AS	0.2666

The sum of all $D_{i,i+1}$ in Slovenian is *AS* = 0.2666. If we perform the same computation for all languages as in Table 4.8, we obtain the results presented in Table 4.9 representing an almost perfect geographic positioning of Slavic languages in their relation to Russian

Table 4.9
Dissimilarity of Slavic languages with Russian

Order	**Language**	$\sum$ ***AS***
1	Belorussian	0.0122
2	Ukrainian	0.0325
3	Polish	0.0499
4	Czech	0.0799
5	Slovak	0.0988
6	Sorbian	0.1058
7	Bulgarian	0.1179

8	Croatian	0.1405
9	Macedonian	0.1424
10	Serbian	0.1433
11	Slovenian	0.2666

Since the geographic classification of Slavic languages (which of course show some kind of isomorphism with morphological characteristics of the languages analysed ─ a problem to be analysed in future) correlates with the kinship relations, the above method seems to be a powerful instrument of some disciplines of linguistics.

4.3. Vector distance

In the first four chapters we used the angle τ between the vectors of geometric properties. However, the same procedure can be performed also using the vector distance defined as

$$(4.3) \quad \delta_{12} = \left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \right]^{1/2},$$

where $x = V$, $y = f(1)$, $z = h$. The greater is δ, the greater is the dissimilarity of two texts. The geometric background is presented in Figure 4.5.

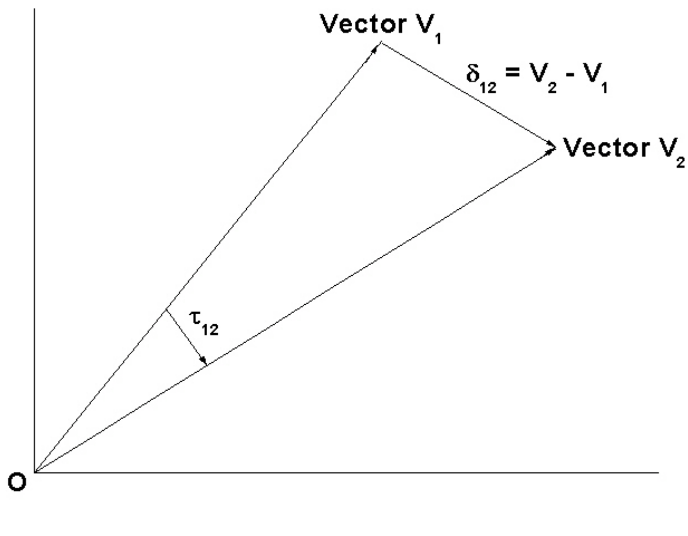


Figure 4.5. The geometric meaning of τ and δ

While τ captures the vectors so to say at their beginning, δ considers their end points. More specifically, τ represents the orientation difference of the considered vector pair whereas δ is the distance between their end points.[1] This presentation of dissimilarities is especially adequate for comparing the same text in different languages because here, there is no great difference in the points of V and $f(1)$ which may be decisive when we compare different chapters of a novel. Let us compare the first Chapter of Ostrovskij´s novel in Russian and in Belorussian in which we find

	V	$f(1)$	h
Russian	1907	169	20
Belorussian	1916	175	19

from which we obtain

$$\delta_{Russian,Belorussian}(\text{Chapter}\,1) = [(1907-1916)^2 + (169-175)^2 + (20-19)^2]^{1/2} = 10.86\,.$$

The coordinates of Russian are presented in Table 4.10, those of other languages together with the individual distances between parallel chapters in Table 4.11.

Table 4.10
The coordinates of 10 Chapters of Ostrovskij´s novel in Russian

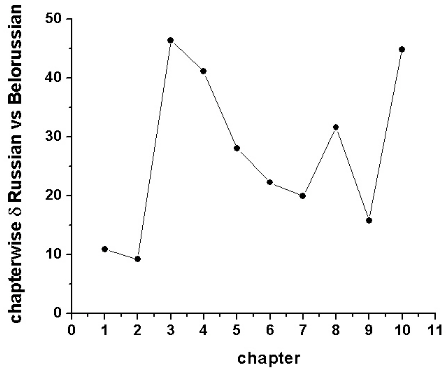
Chapter	***V***	***f*(1)**	***h***
1	1907	169	20.00
2	2088	152	18.50
3	2909	213	24.80
4	2157	127	17.00
5	1882	125	19.00
6	3369	183	26.33
7	2972	164	24.00
8	2814	140	20.75
9	1761	99	18.25
10	2853	169	23.50

[1] As pointed out before, $\tau = 0$ means that the considered vectors are collinear, hence their coordinates are in the same ratio $x_1/x_2 = y_1/y_2 = z_1/z_2$ = constant or, in other words, the corresponding rank-frequencies are fully similar. In particular, if the two end points coincide, then $\delta = 0$ and the considered vectors are identical, that is the above *constant* equals unity.

Table 4.11
The coordinates of Slavic languages and the distance to Russian chapters

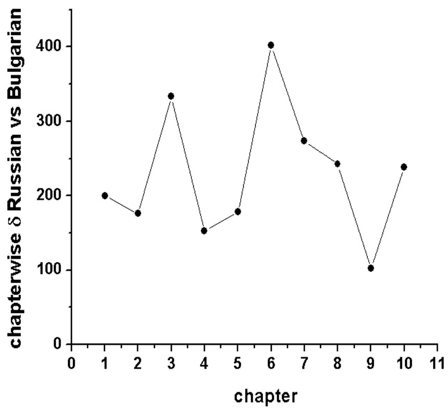
(a) Belorussian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1916	175	19.00	10.8628
2	2079	153	17.00	9.1788
3	2863	219	24.00	46.3966
4	2116	129	17.33	41.0501
5	1854	125	18.50	28.0045
6	3347	186	25.00	22.2434
7	2953	158	24.00	19.9249
8	2783	146	22.20	31.6086
9	1776	94	18.24	15.8114
10	2814	147	21.33	44.8298



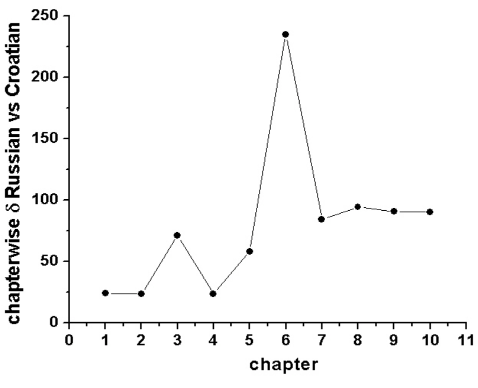
(b) Bulgarian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1709	194	23.00	199.5946
2	1913	170	21.62	175.9509
3	2581	273	28.00	333.4580
4	2007	155	20.00	152.6204
5	1706	150	23.00	177.8117
6	2979	280	31.00	401.9090
7	2729	289	25.00	273.2673
8	2591	235	22.00	242.3955
9	1663	129	20.25	102.5085
10	2633	260	23.33	238.0778



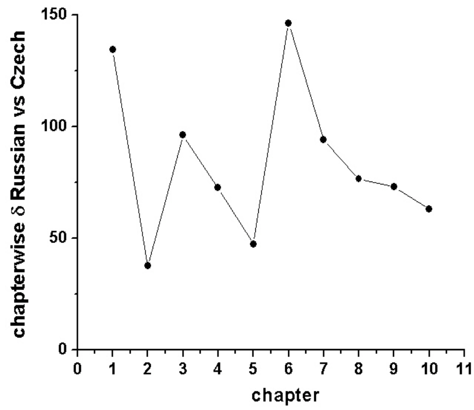
(c) Croatian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1900	192	21.00	24.0624
2	2096	174	20.00	23.4574
3	2888	281	27.00	71.2028
4	2149	149	19.21	23.5135
5	1881	183	19.50	58.0108
6	3222	366	28.50	234.7397
7	2958	247	24.00	84.1724
8	2845	229	22.00	94.2527
9	1795	183	19.33	90.6265
10	2823	254	22.66	90.1427



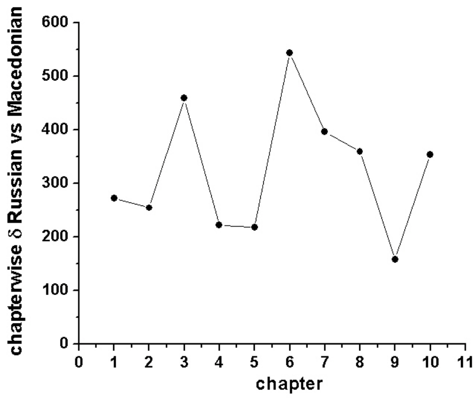
(d) Czech

Chapter	*V*	*f*(1)	*h*	*δ*
1	1773	180	20.33	134.4511
2	2109	183	17.00	37.4733
3	2904	309	25.00	96.1303
4	2111	183	16.00	72.4776
5	1854	163	19.00	47.2017
6	3369	329	28.50	146.0161
7	2945	254	24.00	93.9628
8	2805	216	21.33	76.5332
9	1820	142	17.00	73.0175
10	2891	219	20.00	62.8987



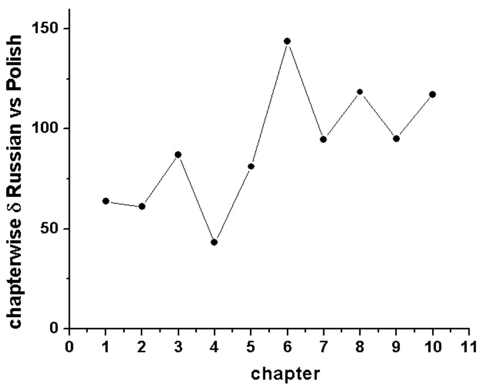
(e) Macedonian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1636	193	23.62	272.0847
2	1836	184	22.25	254.0513
3	2456	283	30.73	458.4148
4	1937	157	21.50	222.0816
5	1667	155	23.66	217.1329
6	2842	316	31.50	543.5483
7	2606	314	26.00	395.5502
8	2484	282	25.00	359.2799
9	1610	146	23.50	158.2326
10	2536	325	24.60	353.3075



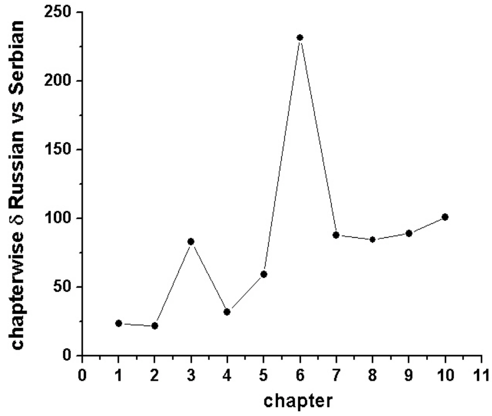
(f) Polish

Chapter	*V*	*f*(1)	*h*	*δ*
1	1970	160	18.80	63.6509
2	2149	149	19.00	61.0758
3	2995	227	24.24	87.1339
4	2200	131	16.00	43.1972
5	1962	138	18.33	81.0521
6	3481	273	23.00	143.7188
7	3061	196	21.00	94.6256
8	2928	172	19.00	118.4190
9	1855	113	17.50	95.0398
10	2970	165	20.00	117.1207



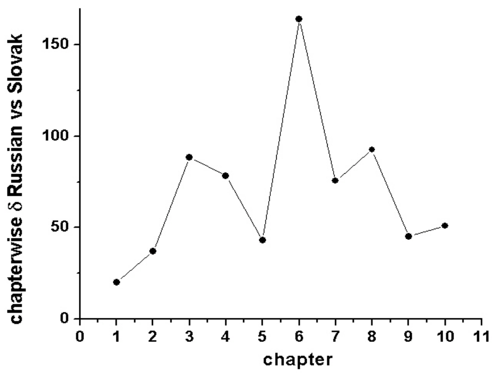
(g) Serbian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1899	191	20.73	23.4208
2	2082	173	20.00	21.8918
3	2852	273	27.00	82.7879
4	2129	142	20.00	31.9061
5	1877	184	19.00	59.2115
6	3237	373	28.00	231.3586
7	2941	246	25.00	87.6698
8	2823	224	21.50	84.4841
9	1787	184	18.50	88.8879
10	2816	263	22.50	101.0247



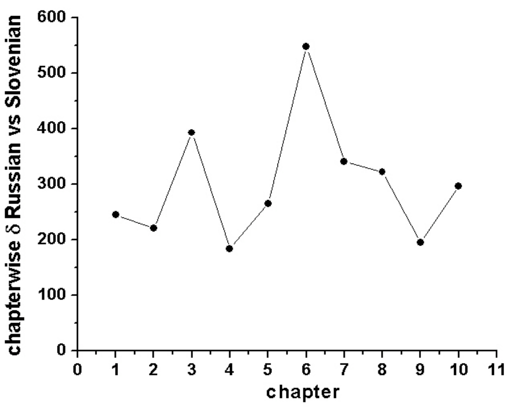
(h) Slovak

Chapter	*V*	*f*(1)	*h*	*δ*
1	1895	185	21.50	20.0562
2	2068	183	20.00	36.9222
3	2864	289	27.50	88.3645
4	2087	162	16.50	78.2640
5	1862	163	20.25	42.9600
6	3292	328	27.00	164.1781
7	2937	231	25.25	75.6013
8	2771	222	24.00	92.6475
9	1757	144	18.50	45.1781
10	2818	206	22.00	50.9534



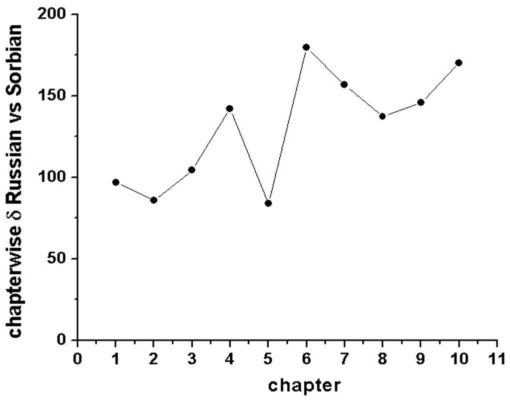
(i) Slovenian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1955	409	24.00	244.7856
2	2098	372	22.00	220.2550
3	2944	604	28.00	392.5764
4	2199	306	21.00	183.9049
5	1929	386	24.00	265.2452
6	3354	730	32.00	547.2350
7	3038	498	26.00	340.4644
8	2955	429	27.00	321.6225
9	1874	258	21.00	195.0835
10	2920	457	26.00	295.7013



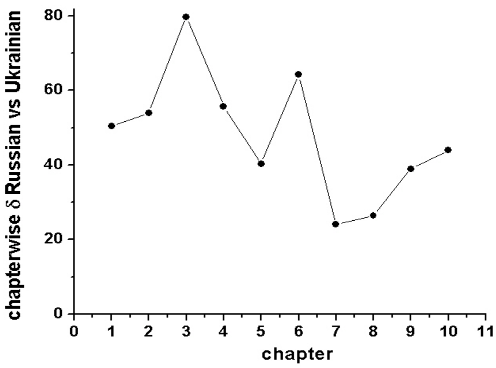
(j) Sorbian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1976	237	22.00	96.8969
2	2152	209	21.00	85.7394
3	2942	312	26.00	104.3621
4	2261	224	20.00	142.2463
5	1950	174	20.33	83.8258
6	3444	346	28.67	179.4421
7	3075	282	23.50	156.6309
8	2917	231	21.50	137.4429
9	1902	136	18.50	145.7740
10	2997	260	25.00	170.3504



(k) Ukrainian

Chapter	*V*	*f*(1)	*h*	*δ*
1	1895	120	19.00	50.4579
2	2078	99	18.00	53.9375
3	2877	140	22.67	79.7342
4	2127	80	16.75	55.7590
5	1864	89	17.33	40.2839
6	3309	160	25.00	64.2711
7	2949	157	23.00	24.0624
8	2809	114	20.33	26.4797
9	1796	82	16.40	38.9541
10	2821	139	21.00	43.9346



In order to evaluate the results in elementary way we proceed in the same way as above. We see, that the course of δ differs from that of τ and that its course is different for each Slavic language. We suppose that here not only the difference between Russian and other Slavic language but also the personal style of translators play probably a certain role.

Again, we can characterize the given language in its difference to Russian (1) by the mean distance of chapters

$$(4.4) \quad \bar{\delta} = \frac{1}{10}\sum_{i=1}^{10}\delta_i ,$$

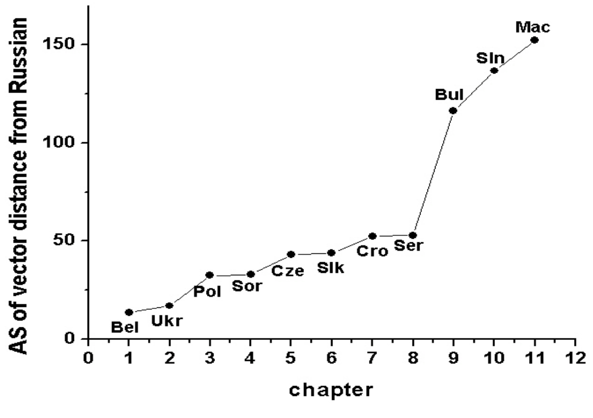
the values of which are presented in increasing order in Table 4.12, or (2) by using the sequential distances and computing *AS* (Formula 3.5) whose values can be found in Table 4.13. We did not present them together in one table because they yield quite different pictures. While mean delta destroys the usual distance between Slavic languages, *AS* yields what we are used to see.

Table 4.12
Mean of the vector distance from Russian

Language	$\bar{\delta}$
Belorussian	26.9911
Ukrainian	47.7874
Slovak	69.5125
Croatian	79.4181
Serbian	81.2643
Czech	84.0162
Polish	90.5034
Sorbian	130.2711
Bulgarian	229.7590
Slovenian	300.6874
Macedonian	323.3684

Table 4.13
AS of vector distance from Russian

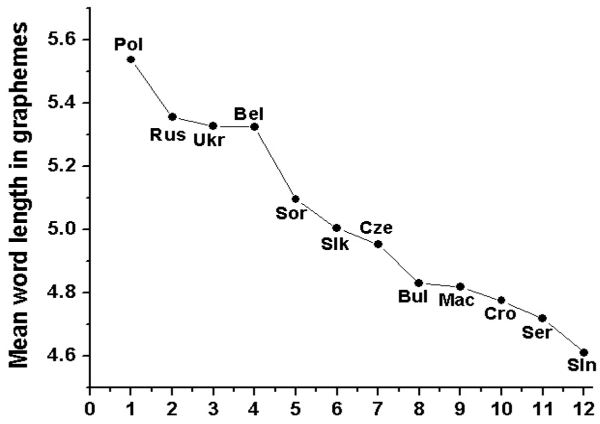
Language	***AS***
Belorussian	13.5414
Ukrainian	16.9772
Polish	32.3820
Sorbian	32.9564
Czech	42.9439
Slovak	43.7552
Croatian	52.4473
Serbian	52.9083
Bulgarian	116.2496
Slovenian	136.7359
Macedonian	152.2760



The classification into East-, South- and Westslavic languages is here expressed in the same way as in E. Kelih´s (2009) comparison of mean word lengths ascertained in the same texts except for Polish, and this is easy to explain: in comparison to other Slavic languages Polish has probably a very low phoneme-grapheme correspondence, thus the word length (measured in terms of grapheme numbers) is slightly higher. In his table (2009: 119) we find the results presented in Table 4.14.

Table 4.14
Mean word length in Slavic languages measured in graphemes (from Kelih 2009)

Language	Mean word length
Polish	5.5365
Russian	5.3559
Ukrainian	5.3270
Belorussian	5.3237
Sorbian	5.0953
Slovak	5.0041
Czech	4.9512
Bulgarian	4.8297
Macedonian	4.8186
Croatian	4.7743
Serbian	4.7189
Slovenian	4.6106



5. The ternary plot

The *T*-vector could be plotted in three-dimensional Cartesian coordinates, but its two-dimensional presentation yields very obscure images. A much more lucid image can be achieved by the ternary plot introduced for this aim in a previous publication (cf. Popescu, Mačutek, Altmann 2009: 40ff). If the three vector components are normalized, one obtains the image presented in Figure 5.1.

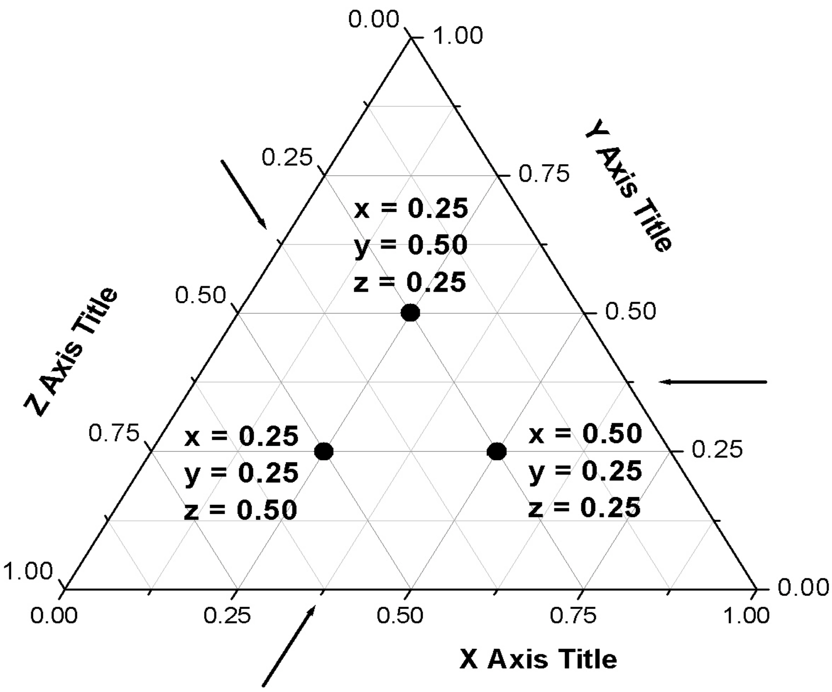


Figure 5.1. Ternary plot

In the ternary plot, there is no common origin (0,0,0); the vector is represented only by the corresponding point (*x*, *y*, *z*) of the ternary plot. Here we shall use a third vector $U(V,f_1,L)$ whose elements are normalized forms of the vocabulary *V*, the highest frequency f_1, and the arc length *L* defined as

(5.1) $$L = \sum_{r=1}^{V-1} D_r = \sum_{r=1}^{V-1} [(f(r) - f(r+1))^2 + 1]^{1/2},$$

where D_r are the Euclidian distances between the adjacent individual frequencies. We define the vector

(5.2) $$U(x,y,z),$$

where

$$x = \frac{X}{X+Y+Z}, \quad y = \frac{Y}{X+Y+Z}, \quad z = \frac{Z}{X+Y+Z}$$

and

(5.3) $$X = \frac{V - V_{min}}{V_{max} - V_{min}}, \quad Y = \frac{f_1 - f_{1,min}}{f_{1,max} - f_{1,min}}, \quad Z = \frac{L - L_{min}}{L_{max} - L_{min}}.$$

Since the maximum and the minimum change according to the dataset, after adding a new text or language to the original dataset the minima and the maxima may change, hence the evaluation of *X*, *Y* and *Z* must be computed anew if necessary. Further, if one characterizes a writer or a language by several texts, then it is more appropriate to use the average *x,y,z* in order to obtain a unique point.

Let us begin with the characterization of 20 languages taking the data from Popescu et al. (2009). The raw data and their transformation is presented in Table 5.1. Since in this data V_{min} = 116, V_{max} = 6073, $f_{1,min}$ = 10, $f_{1,max}$ = 901, L_{min} = 125.56, L_{max} = 6722.04, we obtain the *X*, *Y* and *Z* values as

$$X = \frac{V - 116}{6073 - 116}, \qquad Y = \frac{f_1 - 10}{901 - 10}, \qquad Z = \frac{L - 125.56}{6722.04 - 125.56}.$$

The *x*,*y*,*z* values are given in the pertinent columns of Table 5.1. It is to be noted that $x + y + z = 1$.

Table 5.1
Computation of the components *x,y,z* of the vector *U* of 100 texts in 20 languages
(B – Bulgarian, Cz – Czech, E – English, G – German, H – Hungarian, Hw – Hawaiian, I – Italian, Kn – Kannada, Lk – Lakota, Lt – Latin, M – Maori, Mq – Marquesan, Mr – Marathi, R – Romanian, Rt – Rarotongan, Ru – Russian, Sl – Slovenian, Sm – Samoan, T – Tagalog)

ID	*V*	*f*(1)	*L*	*X*	*Y*	*Z*	Sum	*x*	*y*	*z*
B 01	400	40	428.5	0.0477	0.0337	0.0459	0.1273	0.3746	0.2646	0.3608
B 02	201	13	205.4	0.0143	0.0034	0.0121	0.0297	0.4798	0.1132	0.4069
B 03	285	15	289.8	0.0284	0.0056	0.0249	0.0589	0.4818	0.0953	0.4229
B 04	286	21	297	0.0285	0.0123	0.0260	0.0669	0.4267	0.1846	0.3887
B 05	238	19	247.3	0.0205	0.0101	0.0185	0.0490	0.4177	0.2060	0.3764
Cz 01	638	58	684.2	0.0876	0.0539	0.0847	0.2262	0.3874	0.2382	0.3744
Cz 02	543	56	586.2	0.0717	0.0516	0.0698	0.1931	0.3711	0.2673	0.3616
Cz 03	1274	182	1432	0.1944	0.1930	0.1981	0.5855	0.3320	0.3297	0.3383
Cz 04	323	27	342	0.0347	0.0191	0.0328	0.0866	0.4011	0.2202	0.3787
Cz 05	556	84	627	0.0739	0.0831	0.0760	0.2329	0.3171	0.3566	0.3263

E 01	939	126	1043	0.1382	0.1302	0.1391	0.4074	0.3391	0.3196	0.3413
E 02	1017	168	1157	0.1513	0.1773	0.1564	0.4850	0.3119	0.3656	0.3225
E 03	1001	229	1205	0.1486	0.2458	0.1636	0.5580	0.2663	0.4405	0.2932
E 04	1232	366	1567	0.1873	0.3996	0.2186	0.8055	0.2326	0.4961	0.2714
E 05	1495	297	1761	0.2315	0.3221	0.2479	0.8015	0.2888	0.4019	0.3093
E 07	1597	237	1801	0.2486	0.2548	0.2539	0.7573	0.3283	0.3364	0.3353
E 13	1659	780	2388	0.2590	0.8642	0.3430	1.4663	0.1767	0.5894	0.2340
G 05	332	30	351.4	0.0363	0.0224	0.0342	0.0929	0.3901	0.2415	0.3684
G 09	379	30	398.4	0.0441	0.0224	0.0414	0.1080	0.4089	0.2079	0.3832
G 10	301	18	309.8	0.0311	0.0090	0.0279	0.0680	0.4569	0.1321	0.4110
G 11	297	18	306.8	0.0304	0.0090	0.0275	0.0668	0.4546	0.1343	0.4111
G 12	169	14	175.4	0.0089	0.0045	0.0076	0.0209	0.4247	0.2143	0.3610
G 14	129	10	132.5	0.0022	0.0000	0.0011	0.0032	0.6735	0.0000	0.3265
G 17	124	11	128	0.0013	0.0011	0.0004	0.0028	0.4747	0.3967	0.1286
H 01	1079	225	1289	0.1617	0.2413	0.1763	0.5793	0.2791	0.4165	0.3044
H 02	789	130	907.2	0.1130	0.1347	0.1185	0.3661	0.3086	0.3678	0.3236
H 03	291	48	332.4	0.0294	0.0426	0.0314	0.1034	0.2841	0.4125	0.3033
H 04	609	76	674.1	0.0828	0.0741	0.0832	0.2400	0.3449	0.3087	0.3465
H 05	290	32	314.4	0.0292	0.0247	0.0286	0.0825	0.3539	0.2992	0.3469
Hw 03	521	277	764.3	0.0680	0.2997	0.0968	0.4645	0.1464	0.6452	0.2085
Hw 04	744	535	1229	0.1054	0.5892	0.1673	0.8620	0.1223	0.6836	0.1941
Hw 05	680	416	1047	0.0947	0.4557	0.1398	0.6901	0.1372	0.6603	0.2025
Hw 06	1039	901	1877	0.1549	1.0000	0.2655	1.4204	0.1091	0.7040	0.1869
I 01	3667	388	4007	0.5961	0.4242	0.5884	1.6088	0.3705	0.2637	0.3658
I 02	2203	257	2426	0.3503	0.2772	0.3488	0.9764	0.3588	0.2839	0.3572
I 03	483	64	534.3	0.0616	0.0606	0.0620	0.1842	0.3345	0.3291	0.3364
I 04	1237	118	1330	0.1882	0.1212	0.1825	0.4919	0.3825	0.2464	0.3711
I 05	512	42	537.5	0.0665	0.0359	0.0624	0.1648	0.4033	0.2179	0.3788
In 01	221	16	228.5	0.0176	0.0067	0.0156	0.0400	0.4411	0.1685	0.3904
In 02	209	18	218.6	0.0156	0.0090	0.0141	0.0387	0.4034	0.2320	0.3646
In 03	194	14	199.9	0.0131	0.0045	0.0113	0.0288	0.4539	0.1556	0.3904
In 04	213	11	217.4	0.0163	0.0011	0.0139	0.0313	0.5198	0.0358	0.4443
In 05	188	16	195.7	0.0121	0.0067	0.0106	0.0294	0.4105	0.2287	0.3608
Kn 003	1833	74	1891	0.2882	0.0718	0.2677	0.6277	0.4592	0.1144	0.4264

Kn 004	720	23	733.3	0.1014	0.0146	0.0921	0.2081	0.4872	0.0701	0.4427
Kn 005	2477	101	2558	0.3963	0.1021	0.3688	0.8673	0.4570	0.1178	0.4252
Kn 006	2433	74	2481	0.3890	0.0718	0.3571	0.8179	0.4755	0.0878	0.4366
Kn 011	2516	63	2558	0.4029	0.0595	0.3687	0.8311	0.4848	0.0716	0.4436
Lk 01	174	20	184.8	0.0097	0.0112	0.0090	0.0299	0.3252	0.3749	0.2998
Lk 02	479	124	580	0.0609	0.1279	0.0689	0.2578	0.2364	0.4964	0.2672
Lk 03	272	62	317.6	0.0262	0.0584	0.0291	0.1137	0.2304	0.5134	0.2562
Lk 04	116	18	125.6	0.0000	0.0090	0.0000	0.0090	0.0000	1.0000	0.0000
Lt 01	2211	133	2328	0.3517	0.1380	0.3339	0.8236	0.4270	0.1676	0.4054
Lt 02	2334	190	2502	0.3723	0.2020	0.3603	0.9346	0.3984	0.2162	0.3855
Lt 03	2703	103	2783	0.4343	0.1044	0.4029	0.9415	0.4613	0.1109	0.4279
Lt 04	1910	99	1983	0.3012	0.0999	0.2816	0.6826	0.4412	0.1463	0.4125
Lt 05	909	33	930	0.1331	0.0258	0.1219	0.2809	0.4739	0.0919	0.4342
Lt 06	609	19	621	0.0828	0.0101	0.0751	0.1680	0.4927	0.0601	0.4472
M 01	398	152	526.9	0.0473	0.1594	0.0608	0.2676	0.1769	0.5957	0.2274
M 02	277	127	386	0.0270	0.1313	0.0395	0.1978	0.1366	0.6638	0.1996
M 03	277	128	384.6	0.0270	0.1324	0.0393	0.1987	0.1360	0.6664	0.1976
M 04	326	137	444.3	0.0353	0.1425	0.0483	0.2261	0.1559	0.6304	0.2137
M 05	514	234	715.2	0.0668	0.2514	0.0894	0.4076	0.1639	0.6168	0.2193
Mq 01	289	247	507	0.0290	0.2660	0.0578	0.3529	0.0823	0.7538	0.1639
Mq 02	150	42	178.6	0.0057	0.0359	0.0080	0.0497	0.1149	0.7232	0.1619
Mq 03	301	218	500.4	0.0311	0.2334	0.0568	0.3213	0.0967	0.7265	0.1768
Mr 001	1555	75	1612	0.2416	0.0730	0.2254	0.5399	0.4474	0.1351	0.4175
Mr 018	1788	126	1890	0.2807	0.1302	0.2675	0.6784	0.4137	0.1919	0.3944
Mr 026	2038	84	2099	0.3226	0.0831	0.2992	0.7049	0.4577	0.1178	0.4244
Mr 027	1400	92	1468	0.2155	0.0920	0.2035	0.5110	0.4218	0.1801	0.3981
Mr 288	2079	84	2141	0.3295	0.0831	0.3055	0.7181	0.4589	0.1157	0.4255
R 01	843	62	886.4	0.1220	0.0584	0.1153	0.2957	0.4127	0.1973	0.3900
R 02	1179	110	1269	0.1784	0.1122	0.1734	0.4640	0.3846	0.2419	0.3736
R 03	719	65	770.2	0.1012	0.0617	0.0977	0.2607	0.3883	0.2368	0.3749
R 04	729	49	764.4	0.1029	0.0438	0.0968	0.2435	0.4226	0.1797	0.3977
R 05	567	46	599.2	0.0757	0.0404	0.0718	0.1879	0.4029	0.2150	0.3821
R 06	432	30	451.8	0.0530	0.0224	0.0494	0.1249	0.4246	0.1797	0.3958
Rt 01	223	111	315.9	0.0180	0.1134	0.0289	0.1602	0.1121	0.7077	0.1802

Rt 02	214	69	264.8	0.0165	0.0662	0.0211	0.1038	0.1585	0.6381	0.2033
Rt 03	207	66	255.9	0.0153	0.0629	0.0198	0.0979	0.1561	0.6421	0.2018
Rt 04	181	49	215.6	0.0109	0.0438	0.0136	0.0683	0.1597	0.6406	0.1997
Rt 05	197	74	250.7	0.0136	0.0718	0.0190	0.1044	0.1302	0.6880	0.1817
Ru 01	422	31	441	0.0514	0.0236	0.0478	0.1228	0.4184	0.1920	0.3896
Ru 02	1240	138	1357	0.1887	0.1437	0.1866	0.5190	0.3636	0.2768	0.3596
Ru 03	1792	144	1909	0.2813	0.1504	0.2704	0.7021	0.4007	0.2142	0.3851
Ru 04	2536	228	2732	0.4062	0.2447	0.3951	1.0460	0.3884	0.2339	0.3777
Ru 05	6073	701	6722	1.0000	0.7755	1.0000	2.7755	0.3603	0.2794	0.3603
Sl 01	457	47	493.7	0.0572	0.0415	0.0558	0.1546	0.3703	0.2686	0.3610
Sl 02	603	66	651.1	0.0818	0.0629	0.0797	0.2243	0.3645	0.2802	0.3552
Sl 03	907	102	990.9	0.1328	0.1033	0.1312	0.3672	0.3616	0.2812	0.3572
Sl 04	1102	328	1404	0.1655	0.3569	0.1938	0.7162	0.2311	0.4983	0.2706
Sl 05	2223	193	2385	0.3537	0.2054	0.3426	0.9017	0.3923	0.2278	0.3799
Sm 01	267	159	403.2	0.0253	0.1672	0.0421	0.2347	0.1080	0.7126	0.1793
Sm 02	222	103	303.9	0.0178	0.1044	0.0270	0.1492	0.1193	0.6995	0.1812
Sm 03	140	45	168.4	0.0040	0.0393	0.0065	0.0498	0.0809	0.7887	0.1304
Sm 04	153	78	214.2	0.0062	0.0763	0.0134	0.0960	0.0647	0.7953	0.1400
Sm 05	124	39	149.5	0.0013	0.0325	0.0036	0.0375	0.0358	0.8675	0.0967
T 01	611	89	681	0.0831	0.0887	0.0842	0.2560	0.3246	0.3464	0.3290
T 02	720	107	807.5	0.1014	0.1089	0.1034	0.3136	0.3233	0.3471	0.3296
T 03	645	128	748.5	0.0888	0.1324	0.0944	0.3157	0.2813	0.4195	0.2992

The averages of x, y, and z in individual languages are presented in Table 5.2. For example the mean x in Tagalog is

$$\bar{x}(Tagalog) = (0.3246 + 0.3233 + 0.2813)/3 = 0.3097.$$

The ranking in Table 5.2 is performed according to the component $\bar{x}$. As can be seen, the languages are ordered approximately according to their degree of analyticity. However, many other texts must be analyzed in order to obtain a stable ranking. The ternary plot of all texts is presented in Figure 5.2.

Table 5.2
Mean values of the vector $U(x,y,z)$ in 20 languages
(ordered by increasing $\bar{x}$)

Language	$\bar{x}$	$\bar{y}$	$\bar{z}$
Sm	0.0817	0.7727	0.1455
Mq	0.0980	0.7345	0.1675
Hw	0.1287	0.6733	0.1980
Rt	0.1433	0.6633	0.1933
M	0.1539	0.6346	0.2115
Lk	0.1980	0.5962	0.2058
E	0.2777	0.4213	0.3010
T	0.3097	0.3710	0.3192
Hu	0.3141	0.3609	0.3249
Sl	0.3440	0.3112	0.3448
Cz	0.3617	0.2824	0.3559
It	0.3699	0.2682	0.3619
Ru	0.3863	0.2393	0.3745
R	0.4059	0.2084	0.3857
Bu	0.4361	0.1727	0.3911
Mr	0.4399	0.1481	0.4120
In	0.4457	0.1641	0.3901
Lt	0.4491	0.1322	0.4188
G	0.4691	0.1896	0.3414
Kn	0.4727	0.0923	0.4349

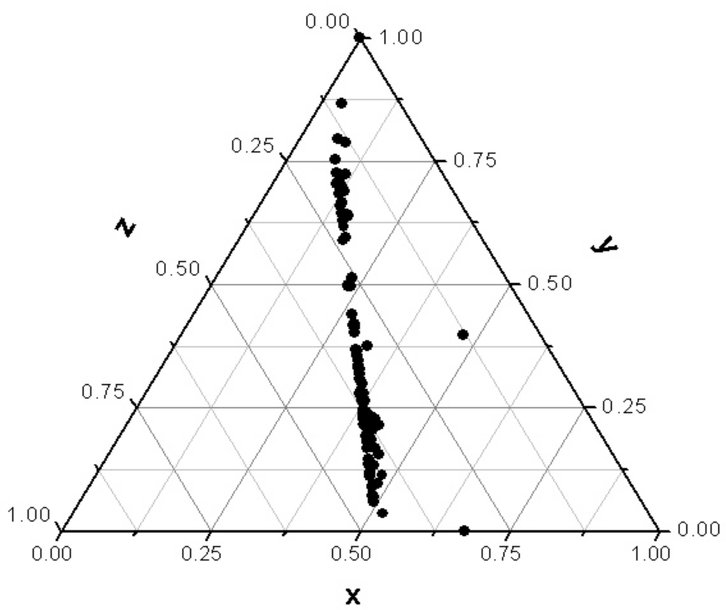


Figure 5.2. The ternary plot of normalized vectors U
of 100 texts in 20 languages

The ternary plot of means is presented in Figure 5.3. As can be seen, the two outliers in Figure 5.2 disappear if the means are taken.

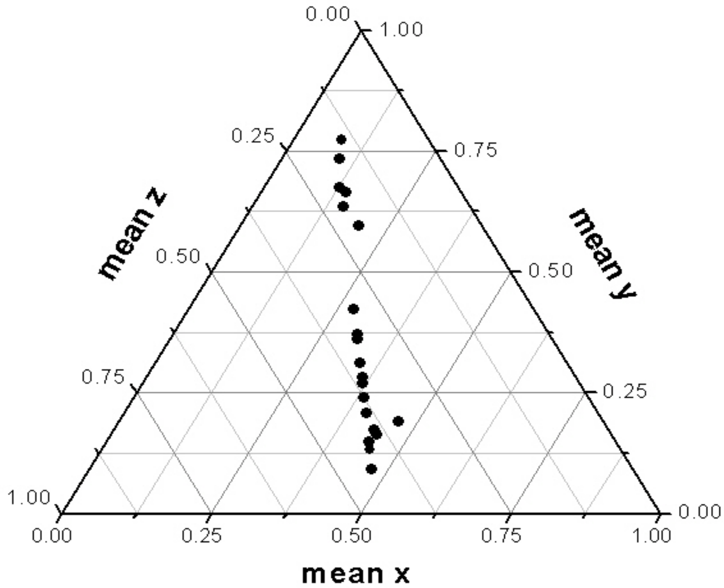


Figure 5.3. Ternary plot of mean normalized vectors U of 20 languages

The results are relatively stable. If we consider the identical texts in 12 Slavic languages each consisting of 10 chapters using the corpus of E. Kelih (2009, 2009a), we obtain the results presented in Table 5.3. Here we have

$$X = \frac{V - 1610}{3481 - 1610}, \quad Y = \frac{f_1 - 80}{730 - 80}, \quad Z = \frac{L - 1728.7413}{4044.1405 - 1728.7413}.$$

Table 5.3
The vector U in 12 Slavic languages (from Kelih 2009, 2009a)

Chapter	***N***	***V***	***f*(1)**	***L***	***X***	***Y***	***Z***	**Sum**	***x***	***y***	***z***
Bel_01	4145	1916	175	2067	0.1635	0.1462	0.1459	0.4556	0.3590	0.3208	0.3203
Bel_02	4177	2079	153	2208	0.2507	0.1123	0.2070	0.5699	0.4398	0.1971	0.3631
Bel_03	6367	2863	219	3050	0.6697	0.2138	0.5705	1.4541	0.4606	0.1471	0.3924
Bel_04	3791	2116	129	2224	0.2704	0.0754	0.2139	0.5597	0.4832	0.1347	0.3821
Bel_05	3791	1854	125	1955	0.1304	0.0692	0.0976	0.2972	0.4388	0.2329	0.3283
Bel_06	7547	3347	186	3501	0.9284	0.1631	0.7653	1.8568	0.5000	0.0878	0.4122
Bel_07	6063	2953	158	3083	0.7178	0.1200	0.5847	1.4225	0.5046	0.0844	0.4110
Bel_08	5362	2783	146	2902	0.6269	0.1015	0.5069	1.2354	0.5075	0.0822	0.4103
Bel_09	3312	1776	94	1850	0.0887	0.0215	0.0522	0.1624	0.5462	0.1326	0.3212

Bel_10	5319	2814	147	2936	0.6435	0.1031	0.5215	1.2680	0.5075	0.0813	0.4112
Bul_01	4653	1709	194	1872	0.0529	0.1754	0.0620	0.2903	0.1823	0.6042	0.2135
Bul_02	4734	1913	170	2055	0.1619	0.1385	0.1409	0.4414	0.3669	0.3137	0.3194
Bul_03	7224	2581	273	2819	0.5190	0.2969	0.4709	1.2868	0.4033	0.2307	0.3660
Bul_04	4305	2007	155	2135	0.2122	0.1154	0.1756	0.5032	0.4217	0.2293	0.3490
Bul_05	4277	1706	150	1827	0.0513	0.1077	0.0425	0.2015	0.2546	0.5343	0.2111
Bul_06	8673	2979	280	3220	0.7317	0.3077	0.6439	1.6833	0.4347	0.1828	0.3825
Bul_07	6992	2729	289	2984	0.5981	0.3215	0.5423	1.4619	0.4091	0.2199	0.3710
Bul_08	6242	2591	235	2796	0.5243	0.2385	0.4611	1.2239	0.4284	0.1948	0.3768
Bul_09	3787	1663	129	1766	0.0283	0.0754	0.0160	0.1197	0.2366	0.6296	0.1339
Bul_10	6278	2633	260	2861	0.5468	0.2769	0.4890	1.3126	0.4165	0.2110	0.3725
Cro_01	4582	1900	192	2065	0.1550	0.1723	0.1451	0.4724	0.3281	0.3647	0.3072
Cro_02	4689	2096	174	2244	0.2598	0.1446	0.2226	0.6269	0.4143	0.2307	0.3550
Cro_03	7160	2888	281	3136	0.6831	0.3092	0.6078	1.6001	0.4269	0.1933	0.3799
Cro_04	4316	2149	149	2275	0.2881	0.1062	0.2359	0.6301	0.4572	0.1685	0.3743
Cro_05	4255	1881	183	2038	0.1448	0.1585	0.1337	0.4370	0.3315	0.3626	0.3059
Cro_06	8553	3222	366	3551	0.8616	0.4400	0.7870	2.0885	0.4125	0.2107	0.3768
Cro_07	6841	2958	247	3173	0.7205	0.2569	0.6237	1.6011	0.4500	0.1605	0.3895
Cro_08	6075	2845	229	3046	0.6601	0.2292	0.5691	1.4584	0.4526	0.1572	0.3902
Cro_09	3760	1795	183	1956	0.0989	0.1585	0.0980	0.3553	0.2783	0.4460	0.2758
Cro_10	6184	2823	254	3048	0.6483	0.2677	0.5698	1.4858	0.4363	0.1802	0.3835
Cze_01	3925	1773	180	1929	0.0871	0.1538	0.0864	0.3274	0.2661	0.4699	0.2639
Cze_02	4381	2109	183	2268	0.2667	0.1585	0.2329	0.6580	0.4053	0.2408	0.3539
Cze_03	6670	2904	309	3182	0.6916	0.3523	0.6277	1.6716	0.4137	0.2108	0.3755
Cze_04	3920	2111	183	2272	0.2678	0.1585	0.2348	0.6610	0.4051	0.2397	0.3552
Cze_05	3852	1854	163	1993	0.1304	0.1277	0.1140	0.3721	0.3505	0.3432	0.3063
Cze_06	8117	3369	329	3664	0.9401	0.3831	0.8359	2.1592	0.4354	0.1774	0.3872
Cze_07	6390	2945	254	3170	0.7135	0.2677	0.6227	1.6039	0.4449	0.1669	0.3882
Cze_08	5738	2805	216	2995	0.6387	0.2092	0.5468	1.3947	0.4580	0.1500	0.3920
Cze_09	3451	1820	142	1940	0.1122	0.0954	0.0914	0.2990	0.3754	0.3190	0.3056
Cze_10	5736	2891	219	3083	0.6847	0.2138	0.5850	1.4835	0.4615	0.1441	0.3943
Mac_01	4810	1636	193	1799	0.0139	0.1738	0.0302	0.2180	0.0638	0.7976	0.1387
Mac_02	4898	1836	184	1989	0.1208	0.1600	0.1123	0.3930	0.3073	0.4071	0.2856
Mac_03	7470	2456	283	2698	0.4522	0.3123	0.4188	1.1833	0.3821	0.2639	0.3539
Mac_04	4424	1937	157	2066	0.1748	0.1185	0.1457	0.4389	0.3982	0.2699	0.3320
Mac_05	4425	1667	155	1793	0.0305	0.1154	0.0276	0.1735	0.1756	0.6651	0.1593
Mac_06	8914	2842	316	3118	0.6585	0.3631	0.5998	1.6213	0.4061	0.2239	0.3699

Mac_07	7153	2606	314	2884	0.5323	0.3600	0.4989	1.3913	0.3826	0.2588	0.3586
Mac_08	6414	2484	282	2731	0.4671	0.3108	0.4330	1.2109	0.3858	0.2566	0.3576
Mac_09	3850	1610	146	1729	0.0000	0.1015	0.0000	0.1015	0.0000	1.0000	0.0000
Mac_10	6461	2536	325	2830	0.4949	0.3769	0.4754	1.3473	0.3674	0.2798	0.3529
Pol_01	4348	1970	160	2107	0.1924	0.1231	0.1632	0.4787	0.4020	0.2571	0.3409
Pol_02	4368	2149	149	2274	0.2881	0.1062	0.2353	0.6295	0.4576	0.1686	0.3738
Pol_03	6694	2995	227	3191	0.7402	0.2262	0.6314	1.5978	0.4633	0.1415	0.3952
Pol_04	4003	2200	131	2311	0.3153	0.0785	0.2513	0.6451	0.4888	0.1216	0.3895
Pol_05	3997	1962	138	2077	0.1881	0.0892	0.1504	0.4278	0.4398	0.2086	0.3517
Pol_06	7937	3481	273	3724	1.0000	0.2969	0.8616	2.1585	0.4633	0.1376	0.3992
Pol_07	6348	3061	196	3229	0.7755	0.1785	0.6478	1.6018	0.4842	0.1114	0.4044
Pol_08	5753	2928	172	3074	0.7044	0.1415	0.5810	1.4270	0.4936	0.0992	0.4072
Pol_09	3501	1855	113	1945	0.1309	0.0508	0.0935	0.2752	0.4758	0.1845	0.3397
Pol_10	5786	2970	165	3109	0.7269	0.1308	0.5959	1.4536	0.5001	0.0900	0.4100
Rus_01	4107	1907	169	2051	0.1587	0.1369	0.1390	0.4347	0.3652	0.3150	0.3198
Rus_02	4136	2088	152	2217	0.2555	0.1108	0.2107	0.5770	0.4428	0.1920	0.3653
Rus_03	6323	2909	213	3091	0.6943	0.2046	0.5885	1.4874	0.4668	0.1376	0.3957
Rus_04	3733	2157	127	2264	0.2924	0.0723	0.2310	0.5957	0.4908	0.1214	0.3878
Rus_05	3769	1882	125	1982	0.1454	0.0692	0.1095	0.3241	0.4485	0.2136	0.3379
Rus_06	7534	3369	183	3519	0.9401	0.1585	0.7731	1.8717	0.5023	0.0847	0.4131
Rus_07	6019	2972	164	3106	0.7280	0.1292	0.5949	1.4521	0.5013	0.0890	0.4097
Rus_08	5352	2814	140	2927	0.6435	0.0923	0.5175	1.2534	0.5134	0.0736	0.4129
Rus_09	3291	1761	99	1839	0.0807	0.0292	0.0477	0.1577	0.5119	0.1854	0.3027
Rus_10	5399	2853	169	2995	0.6644	0.1369	0.5471	1.3484	0.4927	0.1015	0.4057
Ser_01	4579	1899	191	2063	0.1545	0.1708	0.1445	0.4697	0.3288	0.3636	0.3076
Ser_02	4656	2082	173	2230	0.2523	0.1431	0.2165	0.6118	0.4123	0.2339	0.3538
Ser_03	7093	2852	273	3092	0.6638	0.2969	0.5888	1.5496	0.4284	0.1916	0.3800
Ser_04	4290	2129	142	2248	0.2774	0.0954	0.2243	0.5971	0.4646	0.1598	0.3757
Ser_05	4241	1877	184	2035	0.1427	0.1600	0.1322	0.4349	0.3281	0.3679	0.3040
Ser_06	8566	3237	373	3574	0.8696	0.4508	0.7971	2.1174	0.4107	0.2129	0.3764
Ser_07	6816	2941	246	3156	0.7114	0.2554	0.6166	1.5833	0.4493	0.1613	0.3894
Ser_08	6029	2823	224	3019	0.6483	0.2215	0.5573	1.4272	0.4543	0.1552	0.3905
Ser_09	3749	1787	184	1949	0.0946	0.1600	0.0950	0.3496	0.2706	0.4576	0.2718
Ser_10	6208	2816	263	3052	0.6446	0.2815	0.5716	1.4978	0.4304	0.1880	0.3817
Slk_01	4275	1895	185	2053	0.1523	0.1615	0.1401	0.4539	0.3356	0.3559	0.3086
Slk_02	4325	2068	183	2228	0.2448	0.1585	0.2155	0.6188	0.3956	0.2561	0.3483
Slk_03	6496	2864	289	3120	0.6702	0.3215	0.6009	1.5926	0.4208	0.2019	0.3773

Slk_04	3885	2087	162	2226	0.2549	0.1262	0.2149	0.5960	0.4277	0.2117	0.3606
Slk_05	3862	1862	163	2002	0.1347	0.1277	0.1180	0.3804	0.3541	0.3357	0.3102
Slk_06	8021	3292	328	3586	0.8990	0.3815	0.8020	2.0825	0.4317	0.1832	0.3851
Slk_07	6337	2937	231	3137	0.7092	0.2323	0.6083	1.5498	0.4576	0.1499	0.3925
Slk_08	5781	2771	222	2965	0.6205	0.2185	0.5338	1.3728	0.4520	0.1591	0.3888
Slk_09	3412	1757	144	1879	0.0786	0.0985	0.0649	0.2419	0.3248	0.4070	0.2682
Slk_10	5699	2818	206	2997	0.6456	0.1938	0.5478	1.3873	0.4654	0.1397	0.3949
Sln_01	5209	1955	409	2332	0.1844	0.5062	0.2607	0.9512	0.1938	0.5321	0.2740
Sln_02	5199	2098	372	2440	0.2608	0.4492	0.3074	1.0174	0.2564	0.4415	0.3021
Sln_03	7971	2944	604	3513	0.7130	0.8062	0.7704	2.2895	0.3114	0.3521	0.3365
Sln_04	4787	2199	306	2477	0.3148	0.3477	0.3234	0.9859	0.3193	0.3527	0.3280
Sln_05	4720	1929	386	2286	0.1705	0.4708	0.2407	0.8820	0.1933	0.5338	0.2729
Sln_06	9546	3354	730	4044	0.9321	1.0000	1.0000	2.9321	0.3179	0.3410	0.3410
Sln_07	7520	3038	498	3501	0.7632	0.6431	0.7653	2.1716	0.3515	0.2961	0.3524
Sln_08	6822	2955	429	3350	0.7189	0.5369	0.7003	1.9561	0.3675	0.2745	0.3580
Sln_09	4075	1874	258	2105	0.1411	0.2738	0.1627	0.5776	0.2443	0.4741	0.2816
Sln_10	6797	2920	457	3345	0.7002	0.5800	0.6982	1.9784	0.3539	0.2932	0.3529
Sor_01	4851	1976	237	2184	0.1956	0.2415	0.1967	0.6339	0.3086	0.3810	0.3104
Sor_02	4812	2152	209	2334	0.2897	0.1985	0.2612	0.7493	0.3866	0.2648	0.3486
Sor_03	7395	2942	312	3219	0.7119	0.3569	0.6438	1.7126	0.4157	0.2084	0.3759
Sor_04	4483	2261	224	2461	0.3479	0.2215	0.3162	0.8857	0.3929	0.2501	0.3570
Sor_05	4272	1950	174	2099	0.1817	0.1446	0.1599	0.4862	0.3737	0.2974	0.3289
Sor_06	8795	3444	346	3752	0.9802	0.4092	0.8740	2.2634	0.4331	0.1808	0.3861
Sor_07	7058	3075	282	3326	0.7830	0.3108	0.6898	1.7835	0.4390	0.1742	0.3867
Sor_08	6316	2917	231	3120	0.6986	0.2323	0.6007	1.5315	0.4561	0.1517	0.3922
Sor_09	3850	1902	136	2015	0.1561	0.0862	0.1235	0.3658	0.4267	0.2356	0.3377
Sor_10	6648	2997	260	3228	0.7413	0.2769	0.6476	1.6658	0.4450	0.1662	0.3888
Ukr_01	4119	1895	120	1991	0.1523	0.0615	0.1134	0.3272	0.4655	0.1881	0.3465
Ukr_02	4160	2078	99	2152	0.2501	0.0292	0.1827	0.4621	0.5413	0.0633	0.3954
Ukr_03	6282	2877	140	2986	0.6772	0.0923	0.5432	1.3127	0.5159	0.0703	0.4138
Ukr_04	3764	2127	80	2184	0.2763	0.0000	0.1967	0.4730	0.5842	0.0000	0.4158
Ukr_05	3755	1864	89	1929	0.1358	0.0138	0.0864	0.2360	0.5752	0.0587	0.3662
Ukr_06	7542	3309	160	3435	0.9081	0.1231	0.7371	1.7682	0.5136	0.0696	0.4168
Ukr_07	5999	2949	157	3077	0.7157	0.1185	0.5821	1.4162	0.5053	0.0836	0.4110
Ukr_08	5362	2809	114	2897	0.6408	0.0523	0.5047	1.1979	0.5350	0.0437	0.4214
Ukr_09	3278	1796	82	1856	0.0994	0.0031	0.0551	0.1576	0.6309	0.0195	0.3496
Ukr_10	5351	2821	139	2933	0.6472	0.0908	0.5200	1.2580	0.5145	0.0722	0.4133

The corresponding ternary plot is presented in Figure 5.4.

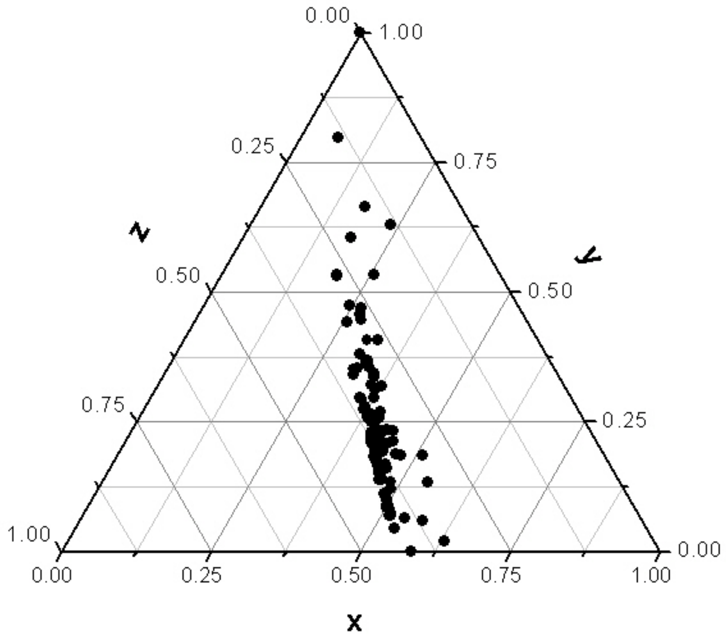


Figure 5.4. Ternary plot of identical texts in 12 Slavic languages

Taking only the mean values we obtain the results in Table 5.4 and Figure 5.5. The Slavic languages seem to keep in all dimensions a position < 0.5

Table 5.4
Mean vector *U* in 12 Slavic languages
(ordered according to $\bar{x}$)

Language	$\bar{x}$	$\bar{y}$	$\bar{z}$
Macedonian	0.2869	0.4423	0.2709
Slovenian	0.2909	0.3891	0.3200
Bulgarian	0.3554	0.3350	0.3096
Serbian	0.3977	0.2492	0.3531
Croatian	0.3988	0.2474	0.3538
Czech	0.4016	0.2462	0.3522
Slovak	0.4065	0.2400	0.3534
Sorbian	0.4077	0.2310	0.3612
Polish	0.4668	0.1520	0.3811
Russian	0.4736	0.1514	0.3751
Belorussian	0.4747	0.1501	0.3752
Ukrainian	0.5381	0.0669	0.3950

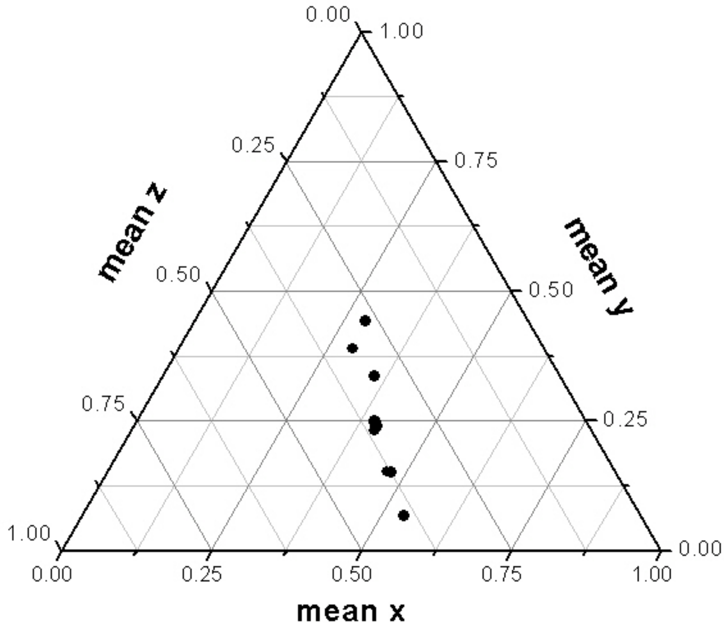


Figure 5.5. Ternary plot of the mean vector *U* for 12 Slavic languages

All figures presented above show that the most "mobile" component is *y*, and the location of languages or texts in the ternary plot occupies a very narrow corridor. This corridor seems to represent a law-like connection of the three components of the vector *U*. Even if we add further texts in which some minima and maxima are more extreme, the corridor will be preserved. Some texts or languages may slightly change their position but do not leave the corridor. If we collect all 533 texts in 30 languages that are at our disposal (cf. Popescu et al. 2009; Popescu, Mačutek, Altmann 2009) and re-normalize the vector *U* for word forms, we obtain an image presented in Figure 5.6. Here the extreme values were: V_{min} = 33, V_{max} = 6397, $f_{1,min}$ = 2, $f_{1,max}$ = 1399, L_{min} = 33, L_{max} = 7427 found with German writers. That means, a text may move within the given corridor according to the six extreme values.

However, a text or group of texts (language) may exchange places with another group of texts if the boundary conditions (external conditions) change. In order to show this possibility, we present in Table 5.5 the data from 20 languages as presented in Table 5.2 under internal conditioning (left part) and under external conditioning (right part).

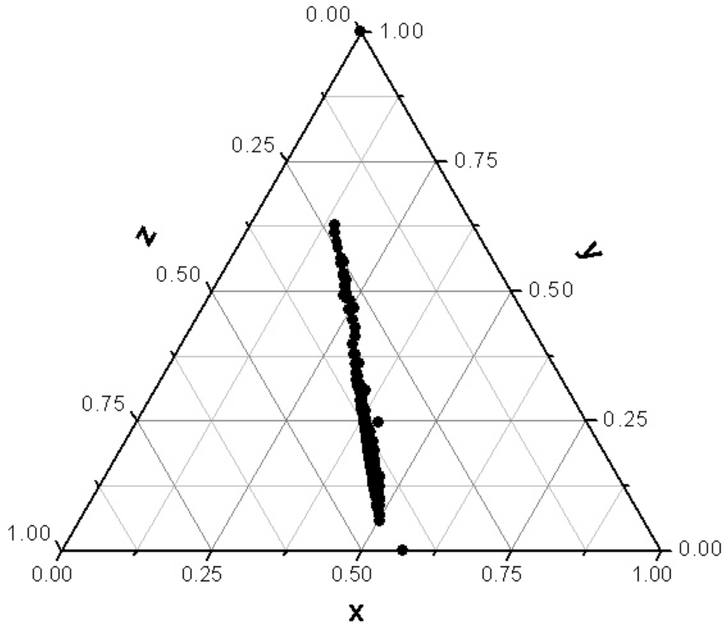


Figure 5.6. Ternary plot of the *U* vector for word forms of 533 texts in 30 languages

Table 5.5
The *U* vector of means in 20 languages with internal and external conditioning (ordered according to increasing $\bar{x}$)

	Internal conditioning					External conditioning		
	$\bar{x}$	$\bar{y}$	$\bar{z}$			$\bar{x}$	$\bar{y}$	$\bar{z}$
Sm	0.0775	0.7764	0.1461		Hw	0.1792	0.5693	0.2515
Mq	0.0954	0.7366	0.1680		Mq	0.1938	0.5504	0.2558
Hw	0.1284	0.6735	0.1981		Sm	0.2199	0.5155	0.2645
Rt	0.1405	0.6655	0.1940		M	0.2285	0.4941	0.2775
M	0.1527	0.6355	0.2118		Rt	0.2519	0.4635	0.2846
Lk	0.1856	0.6075	0.2069		E	0.3288	0.3320	0.3392
E	0.2774	0.4215	0.3011		Lk	0.3424	0.3199	0.3377
T	0.3090	0.3714	0.3196		T	0.3640	0.2795	0.3566
H	0.3128	0.3616	0.3256		H	0.3695	0.2672	0.3633
Sl	0.3433	0.3115	0.3452		Sl	0.3890	0.2379	0.3731
Cz	0.3607	0.2828	0.3565		Cz	0.4049	0.2131	0.3820
I	0.3694	0.2684	0.3622		I	0.4123	0.2019	0.3858

Ru	0.3858	0.2394	0.3747		Ru	0.4242	0.1819	0.3939
R	0.4051	0.2087	0.3862		R	0.4392	0.1597	0.4011
B	0.4327	0.1737	0.3936		G	0.4414	0.1615	0.3971
Mr	0.4397	0.1482	0.4121		In	0.4465	0.1536	0.3999
In	0.4401	0.1658	0.3941		B	0.4468	0.1526	0.4005
Lt	0.4487	0.1322	0.4190		Mr	0.4682	0.1124	0.4194
G	0.4509	0.1981	0.3510		Lt	0.4726	0.1046	0.4228
Kn	0.4725	0.0924	0.4351		Kn	0.4921	0.0740	0.4338

Here, for example, the Polynesian languages exchanged their places, and German moved some places upwards. The method is, nevertheless, adequate for showing some internal states of a language or text, e.g. analytism. For searching for laws one should increase the set of texts but for classification one should use only the internal conditioning of the given group of texts.

Another peculiarity is the mutual relationship of the individual normalized components of the vector *U*. Consider first the relationship $y = f(x)$ obtained for 533 texts in 30 languages presented in Figure 5.7. As can easily be seen, except for two outliers one obtains a perfect straight line whose slope is b = -1.60722 and we suppose that after adding more texts the slope will converge to the golden ratio Φ = 1.618 0339 887… with negative sign.

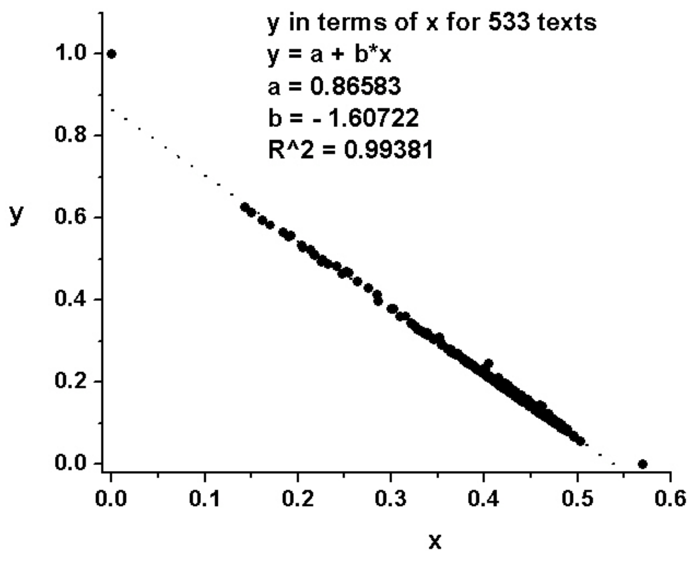


Figure 5.7. The relationship of normalized components *x* and *y* in 533 texts from 30 languages

Taking the relationship $z = f(x)$ we obtain the result presented in Figure 5.8. Again, there are two outliers, one of which is anomalous, and the slope of the straight line is b = 0.6072, which seems to converge to a function of Φ, namely $1/\Phi = \Phi - 1 = 0.6180$.

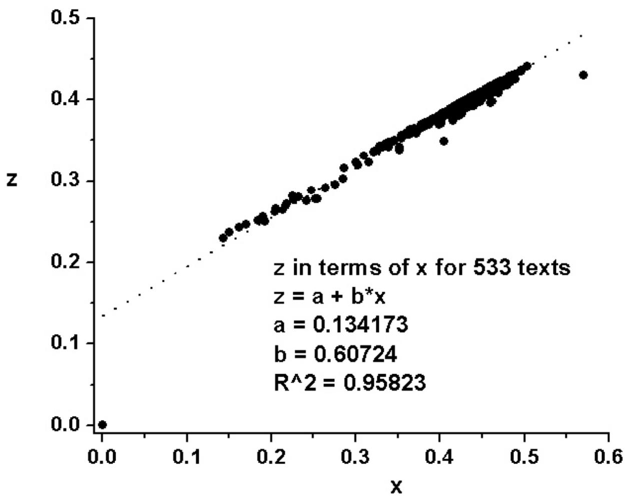


Figure 5.8. The relationship of normalized components x and z in 533 texts from 30 languages

Hence the third relationship, namely that between z and y, $z = f(y)$, is also a straight line whose slope seems to converge to $-(1 - 1/\Phi) = -0.381$.

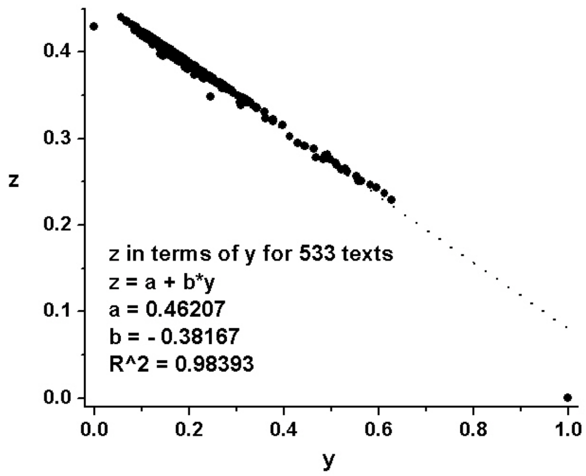


Figure 5.9. The relationship of normalized components y and z in 533 texts from 30 languages

Summarizing, we obtain the general rule according to which with increasing vocabulary V the length L increases and the maximal frequency f_1 decreases, and this notwithstanding the language, genre, and author. Moreover, for sufficiently large data sets, the normalized values x, y, z, of the word frequency quantities V, f_1, L tend to vary linearly with each other, the absolute values of the slopes being in a golden relationship. For 533 texts in 30 languages we have the result given in Table 5.6.

Table 5.6
The golden relationship between the components of vector *U* for word forms

Slope	**Absolute value for 533 texts**	**Close to**
y, *x*	1.60722	Φ = 1.618 …
z, *x*	0.60724	1/ Φ = 0.618 …
z, *y*	0.38167	1 – 1/ Φ = 0.381 ...

It is not easy to find the linguistic background of this peculiar agreement with the golden section. The text authors cannot be aware of it at all and even if they knew what the golden section in texts is, they would not be able to create consciously a text in accordance with it. Since this regularity controls the writing process like an invisible hand, we may consider it a law-like phenomenon. It cannot be captured by simple inspection; it appears only after different transformations as has been shown above.

Needless to say, in small and specific texts sets, deviations can appear, and it is just this deviation that shows us the specificity of text or language.

Automatically the question arises whether (a) other word-like units behave in the same way, and (b) whether other units, for example morphs, syllables, phonemes etc. display the same tendency. Here we shall touch only one of the Köhlerian motifs (cf. Köhler, Naumann 2009; Mačutek 2009), namely the word frequency motif. To this end we used 53 stories written in Russian. Let us compute the frequencies of individual words in each text separately and replace the words by the respective frequency. In that case we obtain a sequence of numbers. A frequency motif is a non-decreasing sequence of numbers, e.g. 1,1,4,52 or 5,17,23,… Motifs have the status of very abstract linguistic units. Since they originate in words, we suppose that the ternary plot will be similar to the general trend but the individual relationships between the variables *x,y,z* may differ. As can be seen in Figure 5.10, the basic trend remains, even if the overall direction is slightly rotated in clockwise direction and the slopes of the individual functions for word frequency motifs differ from the ones for word forms. A test of deviation could be performed but at this stage of research we may dispense with it. A comparison of the *b*-values collected in Table 5.7 with those in Table 5.6 is sufficient.

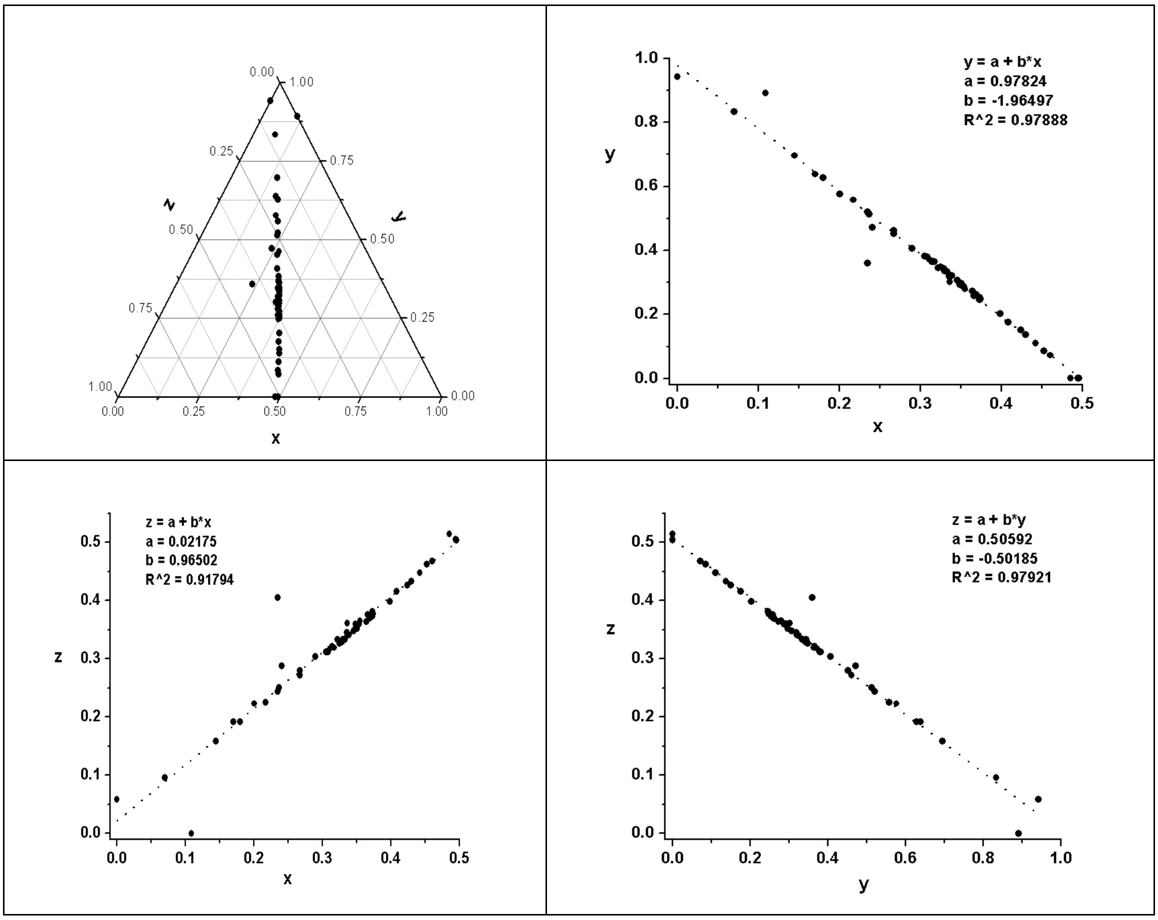


Figure 5.10. Word frequency motifs from 53 Russian texts

Table 5.7
The parameters *b* of the frequency motifs

Slope	**Absolute value of *b* for 53 Russian frequency motifs**	**Golden ratio**
y, *x*	1.9650	Φ = 1.618…
z, *x*	0.9650	$1/\Phi$ = 0.618…
z, *y*	0.5019	$1 - 1/\Phi$ = 0.381...

The dependencies deviate from the functions of the golden ratio. We suppose that units of different levels have their own domains in the ternary plot and the slope *b* of the straight line dependencies will take on specific values. The examination of this phenomenon will be postponed. Here we merely show that words have their specific status when compared with other phenomena. In Figure 5.11 the normalized *U*-vector of random numbers is presented. One can see that there is no “reserved” place for the components.

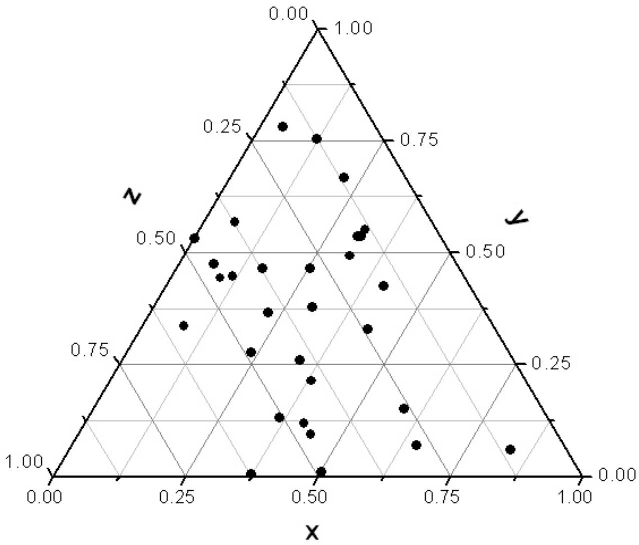


Figure 5.11. Ternary plot of the normalized vector *U* of random numbers

Figure 5.12 presents the normalized *U*-vector of French word associations (cf. Nemcová, Popescu, Altmann 2010) showing that this phenomenon has a very broad corridor in the plot and a number of outliers signalizing great freedom in association. On the other hand, the meaning diversification of English words as presented in Figure 5.13 (Fan, Popescu, Altmann 2008) has a very narrow corridor showing that diversification of meaning underlies a more rigorous control than free association of words.

Thus the ternary plot of the above vectors is a method of locating linguistic phenomena in a specific domain. For outliers a linguistic explication, i.e. the boundary conditions of text generation must be given, a task rather for literary exegetists and specialists in individual languages. Many examinations of different phenomena in various languages will be necessary in order to be able to capture the mechanisms controlling the shape of the ternary plot and propose the first hypotheses.

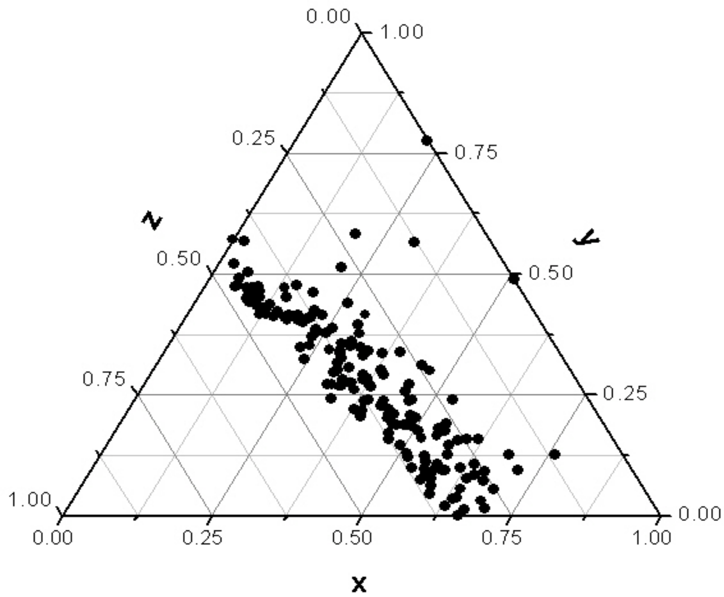


Figure 5.12. Ternary plot of French word associations
(cf. Nemcová, Popescu, Altmann 2010)

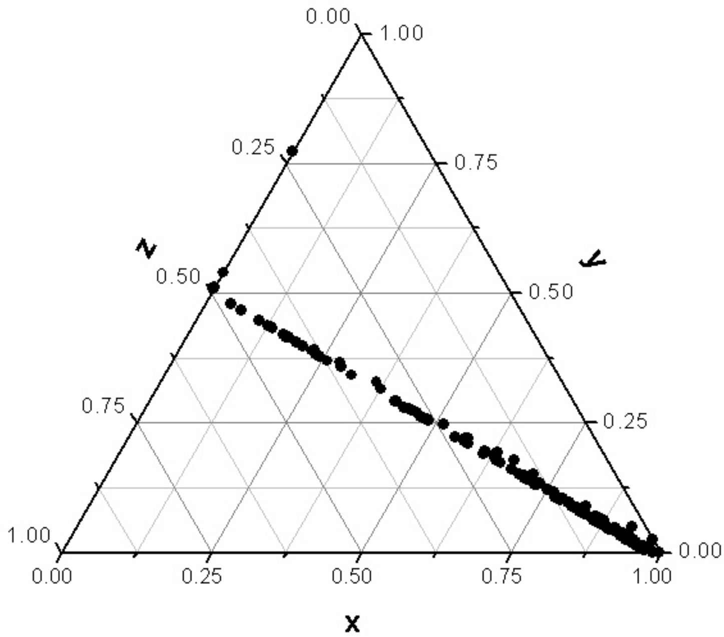


Figure 5.13. Ternary plot of meaning diversification of English words
(from Fan, Popescu, Altmann 2008)

6. Further simple methods for measuring text dynamics

In textual time series, time x is always represented by integer steps x = 1,2,3,… because before we measure, we partition the text in discrete units, e.g. chapters, sentences, clauses, words, morphemes, syllables, etc. and perform some measurements on these units. Needless to say, this partitioning is not always simple or unequivocal because many units may be discontinuous and we determine their identity and position by definition, i.e. using some conventional criterion. But since in language neither units nor the criteria of their identification are something "natural" but in all cases our conceptual constructs, we can conceive them *ad libitum*, however, not losing sight of the fact that we do it in order to test a hypothesis – or, in qualitative linguistics, try to establish a grammatical rule. The hypotheses in quantitative linguistics concern some mechanism, dependences, development or simply the behaviour of a property in text or language.

Since text is a linear formation, something moves along it. This "something" is either the degrees of measured properties or the differences between them. Their sequences form either monotonous (straight line, simple curve), regular repetitive (rhythm) or irregular oscillating (chaotic, fractal) movements whose properties are object of special sciences.

Here we present two simple indicators capturing aspects of the dynamics. The first one takes into account only the changes in up and down direction but not their amplitude. Changes of this kind are signalized by the extreme points in the sequence of values. Consider for example the sequence in Figure 3.1. Here there are n = 11 points out of which m = 8 are extremes (minima or maxima), as can easily be seen. However, two points, namely the first and the last, are necessarily extremes and must be subtracted both from n and from m. In this way we obtain a simple proportion of extremes among all point and define the *non-smoothness indicator*

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

It does not take into account the value of the extreme, hence it is a non-weighted indicator. The indicator has the property of proportion and can easily be statistically processed. If the sequence is monotonous, then NS = 0, if all points are extremes, then NS = 1. The greater the frequency of the oscillation, the greater is NS. The indicator does not say anything either about the regularity of the oscillation or its amplitude; it merely characterizes its presence.

For the data in Figure 3.1, we obtain NS = (8 – 2)/(11 – 2) = 0.6667. In the following we suppose that extremes can occur in every point with the same probability (except for the first and the last one, which are not considered). It is a huge simplification, but at this stage of research the behaviour of extremes cannot be determined more specifically. Under the assumption, the number of extremes has the binomial distribution with the parameters *p* and *n*-2. The indicator *NS* is in fact an estimation of the unknown parameter *p*. Hence, the variance of *NS* is

$$Var(NS) = \frac{NS(1-NS)}{n-2}, \tag{6.2}$$

so that on a very abstract level the dynamics of two texts can be compared. For our German data we obtain the results presented in Table 6.1

Table 6.1
Non-smoothness indicator for German authors

	Retrospective dissimilarity			Prospective dissimilarity		
Writer	***n***	***m***	***NS***	***n***	***m***	***NS***
Meyer	11	7	0.5556	10	4	0.2500
Chamisso	11	8	0.6667	10	6	0.5000
Kafka	18	12	0.6250	17	11	0.6000
Sealsfield	28	16	0.5385	27	16	0.5600
Eichendorff	10	6	0.5000	9	6	0.5714
Novalis	10	7	0.6250	9	6	0.5714
Paul	55	35	0.6226	54	39	0.7115
Löns	13	11	0.8182	12	10	0.8000
Hoffmann	3	2	0.0000			
Wedekind	4	3	0.5000			
Tucholsky	5	5	1.0000			

The authors were arranged according to prospective dissimilarity. It can easily be shown that e.g. Meyer´s prospective and retrospective dissimilarity do not differ significantly (u = 1.36) even if we compute the t-test. But the prospective dissimilarities of Meyer (0.2500) and Löns (0.8000) differ significantly (u = 2.64).

Considering the non-smoothness of Slavic languages in terms of τ radians (Table 4.4) we obtain a relatively uniform picture as shown in Table 6.2.

The differences are probably due rather to stylistic than to linguistic causes.

The non-smoothness indicator *NS* does not express the amplitude of oscillation, it merely shows the proportion of extremes. Hence two different sequences may yield the same value of *NS*. One can consider it the first step in evaluating the non-smoothness.

Table 6.2
The non-smoothness of the translation from Russian
in Slavic languages in terms of τ radians
(based on texts from the Kelih corpus)

Language	***n***	***m***	***NS* (decreasing)**
Russian	10	9	0.875
Belorussian	10	8	0.750
Polish	10	8	0.750
Czech	10	8	0.750
Slovak	10	8	0.750
Serbian	10	8	0.750
Croatian	10	8	0.750
Ukrainian	10	7	0.625
Sorbian	10	7	0.625
Bulgarian	10	7	0.625
Slovenian	10	7	0.625
Macedonian	10	7	0.625

In order to express the amplitude of the oscillation, a sequence is normalized (i.e., all its terms are divided by their maximum) and the indicator *NS* will be multiplied by the arc length *L* and divided by the maximum arc length which is given as

$$L_{\max} = \sum_{i=1}^{n-1}\left[(0-1)^2 + 1^2\right]^{1/2} = (n-1)\sqrt{2},$$

hence we obtain a weighted indicator of *roughness* as

$$R = \frac{(m-2)L}{(n-2)(n-1)\sqrt{2}} \tag{6.3}$$

The indicator *R* attains values from the interval $[0,1]$. The value $R=1$ characterizes exclusively sequences which oscillate regularly between 0 and 1. The lower bound is attained only if a sequence is strictly monotonous.

We present Tucholsky's stepwise dissimilarity (cf. Table 3.4) as an example in Table 6.3. All terms in the original sequence are divided by the maximum value, i.e., by 0.0360.

For the normalized sequence we have *L* = 4.9683, *n* = 5, *m* = 5 and *R* = 0.8783.

Table 6.3
Roughness for the text by Tucholsky
(*S* – stepwise dissimilarity, *N* – normalized stepwise dissimilarity)

Part	***S***	***N***
1	0	0
2	0.0360	1
3	0.0089	0.2472
4	0.0342	0.9500
5	0.0195	0.5417

The values of *R* for German authors are presented in Table 6.4 and 6.5, those of Slavic translation from Russian in Table 6.6

Table 6.4
Retrospective roughness of German writers

Writer	***n***	***m***	***L***	***R*** **(increasing)**
Hoffmann	3	2	2.2844	0.0000
Eichendorff	10	6	9.2915	0.3650
Sealsfield	28	16	27.7628	0.3915
Wedekind	4	3	3.4241	0.4035
Meyer	11	7	10.5594	0.4148
Paul	55	35	55.7773	0.4548
Kafka	18	12	17.8802	0.4648
Novalis	10	7	10.0324	0.4926
Chamisso	11	8	10.8391	0.5110
Löns	13	11	12.6266	0.6088
Tucholsky	5	5	4.9691	0.8784

Table 6.5
Prospective roughness of German writers

Writer	***n***	***m***	***L***	***R*** **(incresing)_**
Meyer	11	4	9.1189	0.1433
Chamisso	11	6	9.5994	0.3017
Eichendorff	10	6	8.1723	0.3210
Novalis	10	6	8.5616	0.3363
Sealsfield	28	16	26.3041	0.3709
Kafka	18	11	16.7383	0.3916
Löns	13	10	11.1181	0.4765
Paul	55	39	54.2704	0.4961

As can be seen, the prospective and the retrospective roughness are not equal and the order of writers is different.

Table 6.6
Roughness of translations from Russian

Language	***n***	***m***	***L***	***R* (decreasing)**
Russian	10	9	9.2395	0.6352
Belorussian	10	8	9.2363	0.5443
Croatian	10	8	10.1872	0.6003
Serbian	10	8	10.0970	0.5950
Slovak	10	8	9.6548	0.5689
Polish	10	8	9.6037	0.5659
Czech	10	8	9.4847	0.5589
Slovenian	10	7	10.0659	0.4943
Bulgarian	10	7	9.9568	0.4889
Macedonian	10	7	9.9349	0.4878
Ukrainian	10	7	9.4930	0.4662
Sorbian	10	7	9.2873	0.4560

The weighted indicator of dynamics can be used for classification and for testing. As the variance of the roughness indicator R is not known, a simulation study could help to estimate it. However, we face another problem here – we need random numbers generated from a rank-frequency distribution (i.e., not only a probability mass function, but also the generated frequencies must be non-increasing). The algorithm for generating random numbers with the mentioned property is being developed and it will be addressed in a separate paper; now we give only a very general recommendation how to obtain an estimate for the variance of R.

Let us have a text containing V words.

1) Rank the word frequencies (i.e., construct a rank-frequency distribution).
2) Generate a rank-frequency distribution of the same type and with the same sample size as is obtained in the first step (we emphasize again that the result must be a non-increasing sequence). Evaluate the arc length L and the roughness indicator R (cf. formula 5.1 and 6.3).
3) Repeat Step 2) until one has a reasonable number of roughness indicators (what is the reasonable number depends on several factors – hardware and software of a computer, homogeinity of obtained results, etc.). In general, the more repetitions, the more reliable estimate. One must find a compromise between time costs and desired exactness.
4) Evaluate the variance of the obtained roughness indicators and use it as an estimation / approximation in a test.

7. The binary code of sentence

7.1. Goedelization

Since there are more than 200 definitions of sentence, we do not want to add a new one. In general, one can consider it as a linear realisation of a nonlinear thought, which is, of course, no definition. The linearization differs in different languages. Syntax offers models of parts of utterances which we call sentences. There are a number of schools that present different models, such as the classical Latin grammar, phrase structure grammar, dependence grammar, stratification grammar, functional grammar, etc. They are all alternative representations containing an aspect of truth (but not the whole), and use mostly some kind of graph. But graphs have their own properties which can be evaluated numerically. Here we shall present only one possibility, namely the binary code of sentence given by a number from which the structure of sentence can be reconstructed, if one knows the number of words in the sentence. The proposal has been made by V.Altmann and G.Altmann (2008: 175ff) whose example will be presented here because the poem can easily be downloaded from the Internet. The procedure is a kind of Goedelization allowing us to associate any sentence structure with a unique number.

One can use any type of grammar that shows the relations of words in the sentence. The results may be different but if performed consequently, one obtains both a characteristic measure and the dynamic behaviour of the texts. Consider for example the first line of Goethe´s famous poem “Erlkönig”, which can be alternatively analyzed as presented in Figure 7.1.

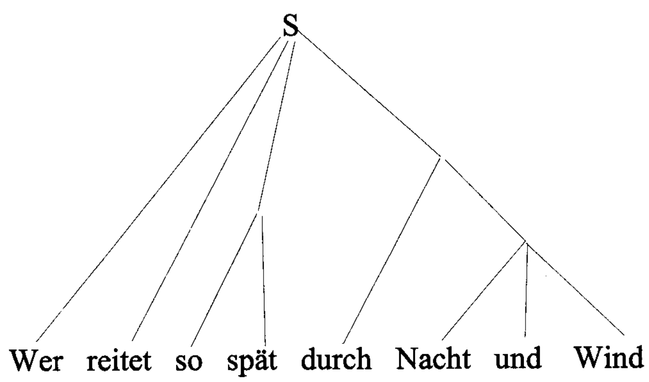


Figure 7.1. One of the possibilities

Now, the vertices will be numerated from top to bottom and from left to right in order to obtain Figure 7.2.

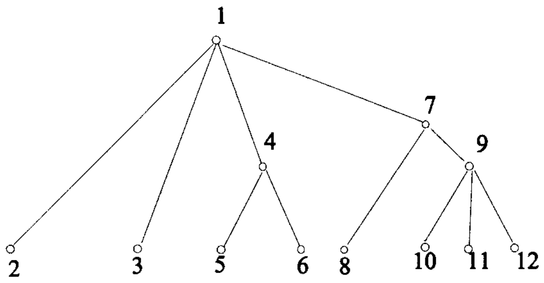


Figure 7.2. Sequence and adjacency of vertices

This graph can be presented in form of an adjacency matrix in which an existing adjacency obtains the value of 1, a non-existing one the value of 0, i.e.

$$(7.1) \quad a_{ij} = \begin{cases} 0, & \text{the vertices } i \text{ and } j \text{ are not adjacent} \\ 1, & \text{the vertices } i \text{ and } j \text{ are adjacent (joined with an edge),} \end{cases}$$

We restrict ourselves to the upper triangular matrix (because of symmetry), the diagonal and the lower triangular matrix will be ignored. Thus we obtain the adjacency matrix in Table 7.1.

Tabelle 7.1
Upper triangular adjacency matrix of the graph in Figure 7.2

v	1	2	3	4	5	6	7	8	9	10	11	12
1	-	1	1	1	0	0	1	0	0	0	0	0
2		-	0	0	0	0	0	0	0	0	0	0
3			-	0	0	0	0	0	0	0	0	0
4				-	1	1	0	0	0	0	0	0
5					-	0	0	0	0	0	0	0
6						-	0	0	0	0	0	0
7							-	1	1	0	0	0
8								-	0	0	0	0
9									-	1	1	1
10										-	0	0
11											-	0
12												-

The binary code (*BC*) will be computed in form of a sum (c.f. Balakrishnan 1997):

$$(7.2) \quad BC = a_{12}2^0 + a_{13}2^1 + \ldots + a_{1n}2^{n-2} + a_{23}2^{n-1} + \ldots + a_{2n}2^{2n-3} + \ldots + a_{n-1,n}2^{k-1},$$

where a_{ij} are the weights given by formula (4.1), n ist he number of vertices, $k = n(n-1)/2$ and the summing begins with the cell (1,2). For the given matrix we obtain

$$BC = 1(2^0) + 1(2^1) + 1(2^2) + 1(2^5) + 1(2^{30}) + 1(2^{31}) + 1(2^{51}) + 1(2^{52}) + \\ +1(2^{60}) + 1(2^{61}) + 1(2^{62}) = 8{,}077{,}205{,}934{,}910{,}210{,}087.$$

In order to normalize this number, one divides it by the maximum which would be attained if all pairs of vertices would be adjacent, i.e.

$$(7.3) \quad BC_{\max} = \sum_{i=0}^{\frac{n(n-1)}{2}-1} 2^i = 2^{\frac{n(n-1)}{2}} - 1.$$

For n = 12 we obtain BC_{max} = 73,786,976,294,838,206,463. Hence the relative binary code is given as

$$(7.4) \quad BC_{rel} = \frac{BC}{BC_{\max}}.$$

In the example it is 8,077,205,934,910,210,087 / 73,786,976,294,838,206,463 = 0.1095. For the complete Goethe´s poem analyzed in this way we obtain the sequence

0.1095; 0.3779; 0.3779; 0.0147; 0.0147; 0.4286; 0.3751; 0.0469; 1.0000; 0.1095; 0.4286; 0.3752; 0.3783; 1.0000; 0.3783; 0.3750; 0.3799; 0.3779; 0.4286; 0.4286; 0.0009; 0.4286; 0.4286; 0.4286; 0.3750; 0.0469; 0.4286; 0.0000001; 0.4286; 0.4286; 0.3779; 0.4286; 0.0147; 0.3751; 0.1111; 0.0146; 0.4286; 0.4286; 0.0009; 0.0469; 0.09557; 0.1111; 0.1095; 0.0029.

This is a quite usual time series representing very abstractly the syntactic structure of the given poem. The sequence is presented in Figure 7.3.

The probability distribution in the upper triangular matrix is very simple, because we do not assume any a priori restrictions or conditions. These can be taken into account only if we are concerned with a specific type of sentences. But since we treat the syntactic structure in general, the probability of a_{ij} = 1 in the upper triangular matrix is 0.5, hence also $p(a_{ij} = 0)$ = 0.5. Thus, each adjacency abides by the zero-one (Bernoulli) distribution given as

$$(7.5) \quad P_x = p^x q^{1-x} = 0.5^x (0.5^{1-x}), \quad x = 0,1.$$

The mean of (7.5) is p = 0.5 and the variance is pq = 0.5(0.5) = 0.25. In the following, the binary code (BC) will be simply denoted by B. In order to compute $Var(B)$ we write

$$(7.6) \quad Var(B) = Var\left(a_{1,2}2^0 + a_{1,3}2^1 + \ldots + a_{n-1,n}2^{\frac{n(n-1)}{2}-1}\right),$$

Since $a(i,j)$ are assumed to be independent, we obtain

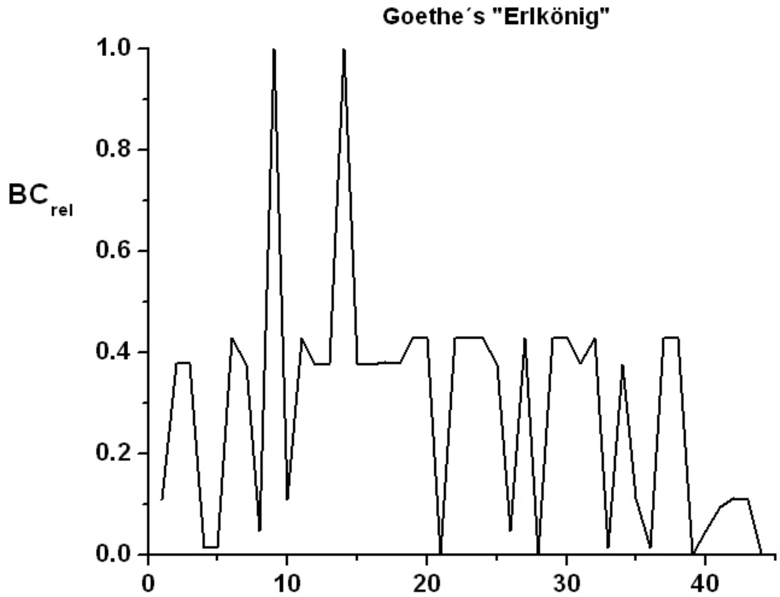


Figure 7.3. The relative binary code of "Erlkönig" (from Altmann, Altmann 2008: 178)

(7.7) $$Var(B) = (2^0)^2 Var(a_{1,2}) + (2^1)^2 Var(a_{1,3}) + \ldots + (2^{\frac{n(n-1)}{2}} - 1)^2 Var(a_{n-1,n}).$$

From (7.5) we know that $Var(a_{i,j}) = 0.25$ and B_{max} is a constant, so that we obtain .

(7.8) $$\frac{Var(B)}{B_{\max}^2} = 0.25 \frac{2^{0^2} + 2^{1^2} + \ldots + \left(2^{\frac{n(n-1)}{2}} - 1\right)^2}{B_{\max}^2}$$

yielding finally

(7.9) $$Var(B / B_{\max}) = \frac{4^{\frac{n(n-1)}{2}} - 1}{12\left(2^{\frac{n(n-1)}{2}} - 1\right)^2}$$

For example, if $n = 12$, we obtain $Var(B/B_{max}) = 0.08333$. Using the variance we can set up asymptotic tests for the comparison of texts and if using the same text, even for the comparison of languages. But perhaps even without parallel texts some characteristic properties of languages can be shown.

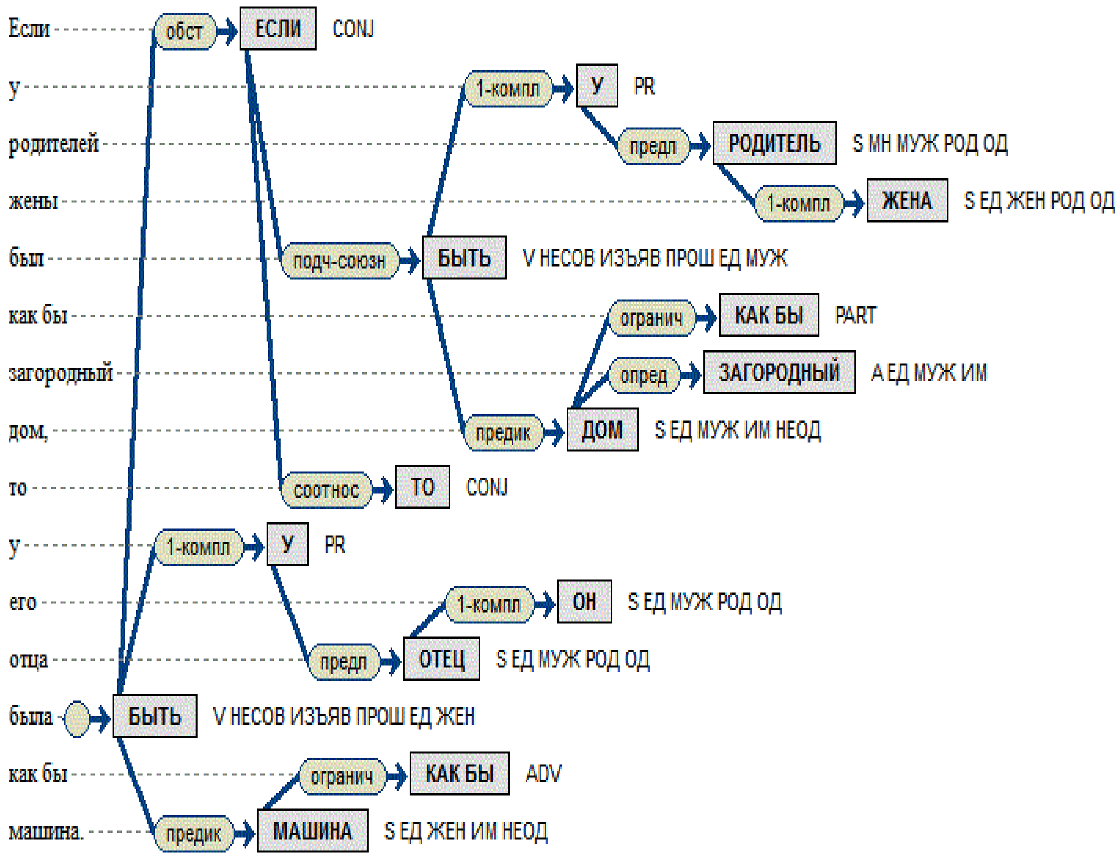
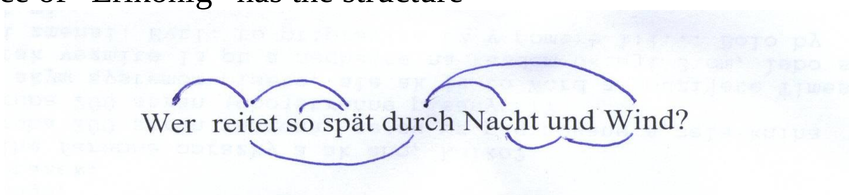


Figure 7.4. A sentence from the Russian Upsala corpus as analyzed by the Computational Linguistic Laboratory of the Institute for problems of information transmission of the Academy of Sciences in Moscow.

The analysis of this kind, if performed with pencil and paper, may be very tiresome. But using a well prepared corpus it can be performed mechanically. We present a Russian example in Figure 7.4 in order to show the possible complexity

A simpler way of analysing the sentence without drawing trees is the marking of the relation between governor and governed, with or without the orientation of the edges (because we consider only the upper triangular matrix). Instead of numerating the nodes we numerate the words. For example the first sentence of “Erlkönig” has the structure



1 2 3 4 5 6 7 8

from which we easily obtain the sequence

12, 24, 25, 34, 56, 58, 67, 78

Since there are n = 8 words, our binary sequence will have the form

1,0,0,0,0,0,0,0,1,1,0,0,0,1,0,0,0,0,0,0,0,1,0,1,1,0,1

whose B = 188,752,641 and B_{max} = 268,435,455 hence

BC_{rel} = 188,752,641 / 268,435,455 = 0.7032.

Every method of describing the sentence structure leads to different results which are not commensurable. However, the variance can be computed according to (7.9).

In the sequel we will present tabular data of the binary code from 20 Russian and 20 Czech texts and will study the properties of the text in the next section.

Table 7.2
Binary codes of sentence structures in 20 Russian texts
[First row: number of words in sentence. Second row: Binary code]
"Upsala"-Korpus: http://www.slaviska.uu.se/ryska/index.html

Russian text 1
40,10,9,3,25,15,4,6,4,4,14,38,13,9,16,13,17,20,6,10,3,45,26,37,6,39,14,7,16,8,14,8,30, 15,16,25,38,37,21,20,13,16,31,45,34,22,23,18,24,15,15,72,3,6,11,29,4,8,100,20,35,67, 48,10,60,11,38,50,16,28,5,38,35,35,28,26,9,21,7,7,5,25,25,12,46,10,11,5,17,15,33,22, 46,24,20,10,12,20,28,10,21,26,35,11,10,23,17,35,29,18,18,24,13,23,44,18,15,49,42,4, 24,7,10,14,44,13,37,70,7,12,45,31,63,22,12,8,21,11,5,12,11,23,9,16,2,6,18,22,18,8,24, 46,53,21,12,17,14,5,6,3,1,18,27,14,15,20,33,22,20,5,3,4,3,10,7,7,8,8,4,6,24,2,6,21,31,6, 30,17,10,34,40,6,27,10,5,7,21,25,7,14,29,6,3,29,6,4,7,10,1,13,25,11,15,30,18,30,30,24, 4,39,27,7,1,3,2,18,9,3,9,9,5,3,2,6,4,12,5,12,7,21,16,16,31,10,22,19,15,6,2,18,4,17,36,35 (n = 254)
0.6719,0.5011,0.7500,0.7143,0.7500,0.7500,0.7778,0.7512,0.5873,0.3333,0.5625, 0.5313,0.7500,0.7500,0.7188,0.5002,0.5000,0.1250,0.5636,0.5000,0.8571,0.7500, 0.3750,0.9102,0.5054,0.6250,0.7500,0.5469,0.4688,0.7520,0.9375,0.5049,0.5000, 0.5000,0.5625,0.7500,0.7500,0.5000,0.5000,0.7500,0.5001,0.7500,0.5000,0.5000, 0.5469,0.5000,0.5000,0.8125,0.5000,0.5000,0.5000,0.5088,0.7143,0.7507,0.8750, 0.2500,0.8889,0.5059,0.5000,0.2500,0.6250,0.9375,0.8750,0.8750,0.8672,0.5004, 0.5000,0.7500,0.7500,0.5000,0.8768,0.5000,0.5000,0.7500,0.5000,0.7500,0.5020, 0.5000,0.7190,0.7503,0.2229,0.0625,0.8750,0.4075,0.5000,0.5005,0.5002,0.5376, 0.5000,0.5000,0.5000,0.7500,0.7500,0.7500,0.5000,0.7500,0.5002,0.5000,0.7500, 0.5010,0.5000,0.7500,0.5000,0.5005,0.7559,0.7715,0.5000,0.6055,0.4375,0.5000, 0.5000,0.8750,0.5001,0.5000,0.3750,0.5000,0.5000,0.5000,0.7500,0.8889,0.5000, 0.2582,0.5020,0.5001,0.5000,0.6641,0.4375,0.7812,0.5079,0.5002,0.5000,0.5000, 0.5000,0.5000,0.0627,0.5059,0.7500,0.5005,0.7566,0.8750,0.5004,0.5000,0.8750, 0.3750,1.0,0.9376,0.5000,0.7500,0.5000,0.2540,0.3750,0.7500,0.5000,0.5000,0.1251,

0.5000,0.8750,0.5367,0.5051,0.7143,0.0,0.5000,0.5000,0.5001,0.5000,0.3750,0.5000, 0.5000,0.7500,0.7556,0.7143,0.5556,0.7143,0.5010,0.5081,0.5081,0.8750,0.5054, 0.5556,0.0793,0.5000,1.0,0.9688,0.5000,0.1250,0.8127,0.4063,0.5313,0.5003,0.5000, 0.5000,0.9376,0.5625,0.5010,0.8143,0.5117,0.5000,0.3750,0.7503,0.9376,0.9375, 0.5255,0.8571,0.7500,0.5090,0.8889,0.5082,0.1885,0.0,0.5001,0.5000,0.5001,0.7500, 0.5000,0.5000,0.7500,0.5000,0.5000,0.6032,0.2500,0.6250,0.5098,0.0,0.7143,1.0, 0.5000,0.6270,0.7143,0.5020,0.5020,0.5406,0.7143,1.0,0.6261,0.7778,0.1565,0.5484, 0.8750,0.1992,.5000,0.5000,0.7500,0.5000,0.5001,0.7500,0.9219,0.7500,0.5236,1.0, 0.8750,0.5873,0.5000,0.5000,0.7505.
Russian text 2
1,16,14,21,13,28,22,3,30,49,52,30,31,12,5,8,47,22,31,47,13,15,8,12,27,27,6,15,17,37, 53,17,82,12,45,21,7,32,58,108,15,7,30,14,35,27,10,15,31,53,55,49,21,7,23,7,30,13,17, 11,39,22,31,45,4,6,28,12,27,19,68,9,21,9,19,14,15,12,17,26,12,10,7,32,40,23,11,5,14, 48, 48,35,35,34,14,10,15,29,15,31,5,6,73,6,43,8,24,26,11,4,7,3,11,5,12,15,7,17,34,13, 20,25,6,9,3,39,28,8,7,13,22,17,28,16,24,5,2,46,19,3,16,18,8,66,11,13,9,10,3,20,3,16,18, 7, 62,18,28,7,22,6,19,13,6,12,4,14,7,15,11,33,9,5,23,29,15,20,15,7,17,22,26,7,24,44,22, 8,20,41,25, 2,33,31,36,25,13,26,23,10,17,9,16,24,79,21,6,13,16,24,17,4,40,2,4,12,51,8, 8,109,1,6,6,24,58,19,18,12,16,42,45 (n = 229)
0.0,0.3906,0.5001,0.8125,0.5013,0.5000,0.5000,0.7143,0.7500,0.7500,0.7734,0.5000, 0.8750,0.7500,0.5347,0.5216,0.8125,0.5449,0.5000,0.5000,0.5001,0.3750,0.5313, 0.5313,0.5000,0.7656,0.5167,0.5000,0.5000,0.2500,0.5000,0.7500,0.6875,0.5002, 0.5000,0.7500,0.5081,0.8750,0.5000,0.7500,0.5000,0.5098,0.7500,0.7500,0.6250, 0.6250,0.7500,0.9375,0.5000,0.5000,0.8750,0.2500,0.5000,0.5081,0.7813,0.5081, 0.5000,0.5001,0.5000,0.5005,0.7500,0.5000,0.5000,0.6543,0.5873,0.5073,0.5000, 0.9375,0.0081,0.5000,0.6250,0.5020,0.2500,0.7500,0.2500,0.7813,0.5000,0.8750, 0.5000,0.3770,0.1254,0.5010,0.7504,0.5000,0.5000,0.6250,0.5005,0.7566,0.5012, 0.5000,0.5000,0.5000,0.5000,0.5625,0.7500,0.7500,0.5938,0.7500,0.9375,0.8750, 0.9384,0.9688,0.5000,0.5163,0.5000,0.5042,0.6250,0.5000,0.5005,0.6032,0.5042, 0.4286,0.5001,0.7556,0.7500,0.5000,0.8906,0.5000,0.1875,0.7500,0.5000,0.5000, 0.5160,0.5020,0.8571,0.7500,0.5000,0.5079,0.7501,0.7500,0.5000,0.2539,0.5000, 0.0005,0.0938,0.5210,1.0,0.5000,0.5000,0.7143,0.5000,0.0156,0.5040,0.5000,0.9531, 0.6251,0.5020,0.5015,0.7143,0.7500,0.7143,0.7500,0.7500,0.8284,0.4375,0.5000, 0.5000,0.5316,0.7500,0.5168,0.7500,0.5001,0.6408,0.9375,0.6032,0.7500,0.4141, 0.5000,0.7500,0.5000,0.5029,0.5484,0.7500,0.7500,0.5000,0.7500,0.7109,0.5079, 0.5000,0.5000,0.6250,0.1332,0.5000,0.8750,0.5000,0.9844,0.7500,0.7500,0.5000, 0.1250,0.3750,0.5000,0.5000,0.8750,0.5001,0.5000,0.5000,0.5013,0.5313,0.5020, 0.7500,0.5000,0.6875,0.2500,0.5070,0.5002,0.7500,0.5000,0.5000,0.5873,0.7500,1.0, 0.5556,0.7500,0.5078,0.2540,0.7501,0.7500,0.0,0.5168,0.5035,0.8750,0.5000,0.5000, 0.5000,0.5000,0.8750,0.1484,0.7500,
Russian text 3
7,11,6,6,39,6,14,4,7,22,9,6,3,44,41,12,13,14,21,7,20,16,9,2,12,5,10,5,5,32,13,8, 4,5,2,18,12,6,5,11,11,5,13,20,16,12,10,3,4,5,2,5,5,5,8,23,7,5,8,9,4,8,9,2,18,11, 11,28,22,5,5,16,26,39,14,8,31,10,12,17,12,7,8,6,9,4,7,4,20,49,65,1,14,9,10,7,6,3, 3,10,12,8,15,14,19,19,1,22,6,27,19,16,13,13,11,15,7,15,14,8,24,29,25,13,7,26, 14,7,19,15,10,49,10,15,12,17,22,16,6,6,37,11,11,18,7,25,13,10,7,5,4,3,3,4,1,16, 9,5,6,5,2,13,3,2,6,15,10,7,13,7,9,13,7,7,26,5,9,15,18,6,5,7,7,3,4,16,8,4,7,4,6,3,5, 1,2,8,3,3,2,5,2,4,2,3,3,2,4,2,2,4,5,19,13,15,7,2,9,17,11,4,15,18,3,8,5,13,4,9,4,10,

15,13,13,6,11,3,7,5,20,11,12,4,5,10,3,7,3,14,15,7,19,12,1,9,4,7,10,11,2,5,12,16,6,7,19,4,6,12,10,11,15,6,17,12,9,2,7,4,2,6,3,20,2,3,4,2,4,15,14,2,2,2,1,10,2,10,4,16,24,8,5,8,7,15,4,8,14,5,10,3,10,3,4,7,5,17,8,18,8,21,13,5,7,7,4,10,3,4,24,6,12,16,2,5,9,4,12,8,12,14,11,14,29,9,4,6,16,12,11,10,4,5,16,15,6,14,6,6,6,4,5,6,30,11,8,2,5,11,8,14,9,18,10,6,18,3,4,10,12,9,11,21,4,12,7,3,6,3,4,6,1,3,8,3,3,10,7,4,10,15,25,10,4,5,10,11,6,5,14,2,12,10,4,6,12,9,2,10,21,8,10,1,12,3,10,6,1,5,2,5,5,3,4,4,8,17,7,5,12,8,8,10,15,9,6,1,3,2,7,3,5,2,20,16,17,12,22,13,6,11,9,3,6,9,10,6,5,4,4,6,12,4,9,4,13,4,2,13,6,12,5,2,3,3,7,6,9,18,12,23,8,3 (n = 492)
0.5081,0.7656,0.5168,0.7507,0.3750,0.6261,0.7500,0.3333,0.8750,0.5000,0.7500,0.8752,0.8571,0.3750,0.7500,0.4065,0.7500,0.5000,0.5000,0.5117,0.5000,0.8750,0.7500,1.0,0.6956,0.8768,0.7539,0.7195,0.2991,0.7500,0.5001,0.5040,0.3492,0.5210,1.0,0.5313,0.5039,0.5035,0.8768,0.7500,0.5005,0.5484,0.7500,0.7500,0.5000,0.7500,0.5010,0.7143,0.4444,0.5279,1.0,0.6921,0.6921,0.6921,0.5039,0.5000,0.5023,0.6921,0.5040,0.5003,0.6032,0.8906,0.5006,1.0,0.5000,0.5005,0.5005,0.7500,0.7500,0.5376,0.5367,0.2500,0.0625,0.6250,0.5001,0.5059,0.5000,0.5625,0.5002,0.5000,0.7813,0.6253,0.8750,0.9688,0.3770,0.3333,0.5023,0.6032,0.7813,0.7500,0.7500,0.0,0.5001,0.6270,0.5000,0.6211,0.2663,0.8571,0.8571,0.7500,0.7500,0.8750,0.5000,0.9375,0.5000,0.5000,0.0,0.5000,0.7515,0.5000,0.7500,0.8750,0.7500,0.7500,0.5005,0.7500,0.7346,0.7500,0.5001,0.5040,0.7500,0.5000,0.7500,0.7656,0.5088,0.5000,0.6250,0.1367,0.5000,0.8125,0.5010,0.0317,0.5013,0.5000,0.5002,0.5000,0.5000,0.7500,0.5197,0.5236,0.7500,0.7520,0.0096,0.5000,0.6254,0.5000,0.5001,0.5010,0.5083,0.7566,0.5873,0.7143,0.8571,0.6032,0.0,0.5000,0.7500,0.7556,0.7512,0.1457,1.0,0.6250,0.7143,1.0,0.7512,0.5000,0.5015,0.6565,0.5001,0.5022,0.7500,0.7500,0.6097,0.5079,0.5000,0.5279,0.5020,0.2500,0.5000,0.5070,0.5347,0.5079,0.5061,0.8571,0.2063,0.7500,0.5040,0.5556,0.5098,0.5873,0.5046,0.8571,0.5376,0.0,1.0,0.5049,0.8571,0.8571,1.0,0.6305,1.0,0.5873,1.0,0.8571,0.8571,1.0,0.6984,1.0,1.0,0.3333,0.4076,0.9375,0.5001,0.5470,0.5083,1.0,0.5005,0.5000,0.9844,0.8889,0.5078,0.5000,0.7143,0.5040,0.4076,0.5001,0.6032,0.5020,0.8889,0.4385,0.7500,0.0627,0.3439,0.7513,0.5005,0.7143,0.9844,0.9384,0.5000,0.5005,0.6252,0.8889,0.5376,0.7500,0.4286,0.5117,0.4286,0.8750,0.6094,0.5044,0.6250,0.6252,0.0,0.5034,0.5873,0.5082,0.9375,0.8789,1.0,0.1457,0.5002,0.7188,0.5236,0.5081,0.5000,0.8889,0.7512,0.5001,0.1890,0.5005,0.6250,0.2660,0.7500,0.5002,0.2598,1.0,0.7502,0.4444,1.0,0.8751,0.7143,0.5352,1.0,0.8571,0.8889,1.0,0.8889,0.5000,0.5001,1.0,1.0,1.0,0.0,0.5010,1.0,0.5625,0.7778,0.5000,0.2500,0.8125,0.3451,0.5006,0.5146,0.5000,0.7778,0.7501,0.3751,0.5689,0.9375,0.8571,0.5002,0.7143,0.8889,0.5082,0.5210,0.5000,0.5313,0.5000,0.5003,0.1797,0.0007,0.1730,0.5628,0.5042,0.5873,0.7500,0.7143,0.5556,0.0938,0.2668,0.5332,0.7500,1.0,0.4457,0.5003,0.6508,0.6252,0.1290,0.7500,0.5001,0.6250,0.7500,0.5000,0.9375,0.6032,0.5479,0.0313,0.9375,0.5005,0.5640,0.6508,0.5376,0.5000,0.8750,0.8752,0.6875,0.9688,0.5293,0.3595,0.7778,0.7537,0.8752,0.9375,0.5005,0.9375,1.0,0.5406,0.7500,0.5073,0.5001,0.5029,0.0156,0.5015,0.7515,0.5625,0.7143,0.6508,0.5006,0.6250,0.8750,0.5005,0.5000,0.8889,0.7500,

0.5081,0.7143,0.5168,0.8571,0.7778,0.5236,0.0,0.8571,0.5020,0.8571,0.8571, 0.7500,0.5117,0.5556,0.5015,0.5000,0.6250,0.5015,0.5873,0.7537,0.5625, 0.5005,0.8752,0.7556,0.7500,1.0,0.6250,0.5005,0.4127,0.5274,0.5002,0.5001, 1.0,0.5010,0.5000,0.5059,0.1338,0.0,0.5003,0.7143,0.5010,0.5177,0.0,0.5142, 1.0,0.4076,0.2845,0.8571,0.8889,0.6032,0.5002,0.7500,0.5023,0.2845,0.5002, 0.5040,0.2539,0.5007,0.1250,0.5029,0.2663,0.0,0.8571,1.0,0.5082,0.8571, 0.8172,1.0,0.7500,0.7500,0.7500,0.7500,0.7500,0.5234,0.9376,0.7813,0.2598, 0.8571,0.6719,0.5020,0.0642,0.7827,0.5552,0.5873,0.6032,0.9688,0.7500, 0.5873,0.5020,0.8889,0.5001,0.5873,1.0,0.8750,0.5168,0.0941,0.5367,1.0, 0.7143,0.8571,0.5088,0.8751,0.5011,0.5000,0.5001,0.7500,0.7500,0.8571
Russian Text 4
2,13,9,7,19,10,16,18,18,5,8,6,6,13,4,8,16,8,8,10,18,8,11,10,10,14,10,17,17,12, 16,7,18,14,8,16,14,30,11,7,9,8,12,30,6,10,16,27,17,13,29,12,11,10,7,13,10,13,6, 7,8,5,6,2,5,14,5,10,6,17,2,3,10,9,10,8,22,16,10,15,16,24,16,15,5,9,5,8,8,4,15,14, 2,13,25,13,3,13,10,7,14,10,4,6,6,5,9,14,7,16,22,9,3,8,15,9,17,12,6,6,19,7,11,9,7, 13,9,19,13,17,3,15,9,4,3,5,18,7,11,8,4,8,12,9,15,10,5,20,14,13,12,16,9,6,6,8,10, 14,7,9,12,20,11,14,9,18,10,6,13,19,13,15,10,4,8,7,10,2,22,4,4,7,15,27,7,6,12,24, 16,12,9,13,6,8,5,13,5,7,6,3,6,6,13,13,8,10,5,8,13,14,6,11,10,2,19,12,19,20,22,14, 11,3,6,9,6,10,8,6,7,4,7,6,11,9,8,10,2,4,10,4,9,5,7,26,4,2,18,3,3,6,7,11,4,6,5,2,2,2, 2,7,3,1,2,2,3,18,3,7,3,11,7,12,10,14,14,9,12,14,14,25,5,16,5,25,14,10,17,10,10, 28,12,13,11,4,15,4,4,6,35,18,16,6,3,9,10,11,18,8,8,17,9,12,24,10,21,1,3,9,4,10,7, 4,3,10,6,6,4,12,10,7,4,6,8,4,1,9,4,3,7,7,12,13,3,4,4,3,19,5,2,3,3,3,9,5,3,7,20,13,2, 2,12,10,28,15,11,26,30,13,9,9,11,7,12,21,13,9,8,9,5,6,17,14,7,6,7,11,17,7,14,18, 19,20,5,28,15,11,21,11,9,5,7,8,22,8,5,8,8,9,10,19,2,1,33,18,11,15,8,10,15,9,11, 14,14,7,11,4,11,5,6,4,10,14,9,6,8,13,41,15,14,4,5,8,17,11,2,17,4,14,25,12,14,20, 5,6,3,3,2,4,7,3,1,1,3,8,15,8,8,30,14,4,13,8,10,11,11,11,13,6,4,5 (n = 480)
1.0,0.8750,0.7500,0.7656,0.7500,0.7500,0.5000,0.5000,0.5000,0.5142,0.7501, 0.5323,0.7512,0.1954,0.5873,0.5002,0.5000,0.7501,0.5040,0.5010,0.5000, 0.5313,0.8750,0.5010,0.7500,0.2501,0.5005,0.5000,0.5000,0.8750,0.5000, 0.7503,0.7500,0.5625,0.5040,0.5000,0.9688,0.6885,0.6250,0.2642,0.8750, 0.7500,0.5004,0.5000,0.5178,0.8750,0.8750,0.5781,0.5000,0.8750,0.5000, 0.5001,0.5005,0.5010,0.5081,0.5001,0.5010,0.6250,0.5160,0.5628,0.8984, 0.8768,0.7515,1.0,0.5953,0.6250,0.5347,0.7500,0.5169,0.5000,1.0,0.7143, 0.5010,0.5020,0.5015,0.9531,0.7500,0.5313,0.5010,0.8750,0.5313,0.5000, 0.7500,0.5001,0.6647,0.8828,0.8768,0.5039,0.7501,0.6508,0.5000,0.8750,1.0, 0.8750,0.7500,0.7500,0.8571,0.9375,0.6875,0.5081,0.6250,0.5078,0.6508, 0.5160,0.5171,0.5367,0.2520,0.7188,0.5629,0.1641,0.5000,0.5020,0.7143, 0.6876,0.5000,0.5020,0.7500,0.8750,0.5165,0.5168,0.5000,0.7503,0.5003, 0.5029,0.5042,0.5000,0.5020,0.5000,0.5001,0.5156,0.4286,0.7500,0.5010, 0.2222,0.7143,0.7537,0.5000,0.5122,0.5005,0.7735,0.1746,0.5020,0.7500, 0.5029,0.7500,0.5015,0.8172,0.7500,0.7820,0.6250,0.5002,0.5000,0.7500, 0.5070,0.7512,0.7501,0.5010,0.7500,0.7503,0.5010,0.5314,0.5000,0.5313, 0.3126,0.5001,0.5000,0.5000,0.5236,0.5001,0.6250,0.7500,0.5000,0.5000, 0.5873,0.5040,0.5081,0.5010,1.0,0.5469,0.4444,0.7778,0.5117,0.5000,0.5000,

0.7501,0.5216,0.5003,0.3750,0.5000,0.5002,0.5010,0.6895,0.5173,0.5039, 0.8143,0.2501,0.5367,0.2073,0.5054,0.8571,0.9063,0.5236,0.5001,0.5001, 0.8750,0.7500,0.5210,0.7501,0.5001,0.8750,0.5171,0.5005,0.7500,1.0,0.7500, 0.7500,0.7500,0.5322,0.5000,0.5001,0.5002,0.8571,0.5163,0.5020,0.4844, 0.3154,0.5040,0.6573,0.7503,0.6508,0.5081,0.8752,0.5009,0.7500,0.5011, 0.9375,1.0,0.5556,0.8770,0.3016,0.5020,0.5376,0.7504,0.7500,0.5873,1.0, 0.6875,0.8571,0.7143,0.2663,0.5046,0.7500,0.8889,0.2660,0.9384,1.0,1.0,1.0, 1.0,0.7503,0.7143,0.0,1.0,1.0,0.7143,0.7500,0.4286,0.6409,0.8571,0.5005, 0.5785,0.8125,0.5019,0.8750,0.5000,0.7500,0.7500,0.5000,0.7500,0.6250, 0.5279,0.3750,0.5376,0.3750,0.5001,0.5010,0.3750,0.5010,0.5016,0.6250, 0.5002,0.6094,0.2349,0.5873,0.1875,0.5556,0.4444,0.5163,0.2813,0.5782, 0.5000,0.7507,0.8571,0.2520,0.5010,0.5005,0.5000,0.7500,0.7501,0.0039, 0.5020,0.5002,0.5078,0.7520,0.5000,0.0,0.7143,0.5010,0.7778,0.7500,0.8125, 0.5873,0.4286,0.5001,0.7512,0.7514,0.5556,0.7500,0.5012,0.5079,0.5873, 0.7512,0.5040,0.4444,0.0,0.1895,0.6032,0.8571,0.7109,0.5027,0.7346,0.5002, 0.8571,0.8889,0.5873,0.8571,0.7500,0.5484,1.0,0.7143,0.7143,0.7143,0.5020, 0.9384,0.8571,0.5081,0.8281,0.7500,1.0,1.0,0.4690,0.6250,0.5000,0.7500, 0.7500,0.5000,0.6875,0.5001,0.7500,0.5020,0.5000,0.5081,0.5002,0.5000, 0.7344,0.8750,0.6563,0.7500,0.5484,0.5167,0.2500,0.7500,0.5042,0.5168, 0.7503,0.5001,0.9375,0.5273,0.5001,0.5000,0.5000,0.7500,0.5210,0.5000, 0.5000,0.5005,0.7500,0.5005,0.8125,0.5367,0.5083,0.5023,0.5488,0.7500, 0.7556,0.7501,0.5039,0.0645,0.7500,0.6250,1.0,0.0,0.1250,0.0627,0.5005, 0.5000,0.5039,0.7500,0.5000,0.5020,0.7500,0.1251,0.2501,0.7501,0.7500, 0.4444,0.2505,0.5367,0.5165,0.7778,0.5010,0.5001,0.2676,0.5254,0.7969, 0.8130,0.5000,0.5000,0.5001,0.5556,0.7537,0.2540,0.5000,0.7500,1.0,0.2500, 0.6032,0.7500,0.1250,0.5004,0.5001,0.7500,0.5435,0.5323,0.8571,0.7143,1.0, 0.6984,0.7503,0.4286,0.0,0.0,0.8571,0.5040,0.5000,0.7500,0.6251,0.5000, 0.5001,0.8889,0.5010,0.0996,0.5018,0.7500,0.5005,0.6250,0.5001,0.2671, 0.5873,0.5376
Russian Text 5
9,18,18,5,7,7,23,30,17,31,8,12,12,11,14,10,7,9,8,17,13,11,4,4,11,7,13,5,12,11, 18,7,18,6,18,24,18,23,8,15,19,8,13,8,12,10,7,19,12,19,9,14,2,17,9,13,5,7,4,5,19, 22,12,10,9,3,5,11,8,6,7,17,28,4,4,5,8,8,17,11,11,9,19,17,18,6,8,18,20,7,11,10,24, 9,7,17,10,7,8,11,13,20,17,9,3,12,20,4,5,10,5,10,9,25,14,21,6,14,4,10,11,28,7,15, 13,11,4,10,16,3,12,8,12,7,6,21,11,10,13,10,6,4,22,7,22,4,5,18,10,5,7,5,3,8,7,6, 10,3,5,8,3,6,4,3,8,7,7,6,4,7,5,8,6,9,10,4,2,9,10,9,11,11,15,20,20,11,18,9,11,16,5, 11,14,7,11,7,7,10,6,11,8,3,10,10,16,10,15,8,7,18,12,8,12,5,10,16,8,17,23,10,9, 11,13,4,7,14,14,8,13,17,7,6,4,25,12,8,10,9,6,5,8,11,10,16,7,11,19,15,10,2,18,18, 7,3,25,10,4,12,11,5,23,14,16,22,6,14,10,2,20,17,5,6,16,1,2,4,2,3,5,2,3,2,4,9,9,8, 10,6,7,3,6,11,8,6,7,4,6,2,6,6,6,11,6,3,7,11,5,6,6,9,2,9,10,4,4,1,2,11,6,3,9,3,4,1,1, 10,14,5,9,9,17,18,4,3,4,6,12,4,8,5,12,2,9,5,17,7,10,8,4,5,12,5,5,6,3,13,9,5,5,4,9, 8,6,8,8,14,2,18,21,11,11,9,5,10,4,8,17,23,6,14,19,9,19,2,22,11,8,4,3,4,8,6,7,13, 27,16,9,6,13,19,23,20,3,15,16,12,11,8,6,12,17,8,8,9,11,13,24,5,19,7,13,13,13,24, 8,10,17,7,13,6,8,13,13,6,8,17,10,14,11,6,8,8,17,7,9,9,5,7,11,6,8,10,11,12,3,12,

21,26,12,8,18,23,15,19,5,18,17,9,4,7,9,16,11,11,8,3,16,8,10,13,11,10,12,15,3,6, 7,7,14 (n = 489)
0.7500,0.8750,0.5000,0.5484,0.0706,0.8750,0.7500,0.5000,0.7500,0.0625, 0.3867,0.5625,0.7500,0.7500,0.5001,0.5003,0.5022,0.5020,0.5020,0.5000, 0.8750,0.8750,0.6508,0.7778,0.7500,0.7502,0.5002,0.5367,0.7813,0.5007, 0.5000,0.7504,0.5000,0.5169,0.1250,0.5000,0.7500,0.5000,0.5040,0.5000, 0.8750,0.7578,0.5001,0.5020,0.6094,0.5015,0.9531,0.7500,0.5004,0.8750, 0.5020,0.7500,1.0,0.5000,0.5020,0.2501,0.8768,0.5012,0.5556,0.5376,0.7500, 0.5000,0.0626,0.7500,0.5020,0.8571,0.7556,0.3755,0.7501,0.7517,0.5005, 0.7500,0.7500,0.8889,0.8889,0.5376,0.5039,0.5040,0.5000,0.5005,0.7500, 0.3145,0.5000,0.7188,0.8750,0.7344,0.2520,0.5000,0.4688,0.7503,0.5005, 0.7500,0.5000,0.5020,0.7520,0.5000,0.5010,0.6328,0.6251,0.4380,0.8438, 0.7500,0.5000,0.8750,0.7143,0.8750,0.5000,0.7778,0.5210,0.5015,0.5679, 0.5012,0.5010,0.5078,0.5001,0.5000,0.5160,0.3751,0.8889,0.5002,0.5004, 0.8750,0.5081,0.5000,0.5001,0.5156,0.5873,0.5010,0.5000,0.7143,0.7500, 0.6367,0.5002,0.6328,0.8751,0.5000,0.6875,0.5010,0.5000,0.8438,0.7512, 0.5556,0.5000,0.5079,0.1250,0.5873,0.7556,0.5000,0.5015,0.5347,0.5093, 0.5132,0.8571,0.5020,0.6409,0.7197,0.5010,0.8571,0.5347,0.5044,0.7143, 0.4219,0.5556,0.8571,0.6289,0.2582,0.8750,0.5173,0.7778,0.5137,0.0968, 0.5001,0.8127,0.3613,0.5010,0.7778,1.0,0.7500,0.5002,0.7500,0.5009,0.5005, 0.7500,0.5000,0.6875,0.5313,0.7500,0.6250,0.7500,0.2500,0.9384,0.5005, 0.5001,0.9219,0.7188,0.5080,0.5626,0.7500,0.5173,0.8125,0.5039,0.7143, 0.5391,0.3760,0.6250,0.5017,0.5000,0.5025,0.5022,0.5000,0.5002,0.7032, 0.7500,0.5367,0.5010,0.5000,0.8750,0.8750,0.6875,0.5010,0.5020,0.5005, 0.6875,0.7778,0.5051,0.7188,0.2501,0.6251,0.7500,0.1484,0.8750,0.8127, 0.8889,0.8125,0.5002,0.5020,0.8750,0.5020,0.5167,0.5660,0.5059,0.5007, 0.5010,0.5000,0.9375,0.5005,0.8750,0.7500,0.5001,1.0,0.7500,0.5000,0.6328, 0.8571,0.7500,0.3760,0.6984,0.7500,0.5009,0.5376,0.6250,0.5001,0.2500, 0.5000,0.8127,0.7500,0.6563,1.0,0.5000,0.5000,0.5367,0.5169,0.7500,0.0,1.0, 0.5556,1.0,0.7143,0.2346,1.0,0.4286,1.0,0.8889,0.5020,0.5020,0.5039,0.5010, 0.5163,0.5042,0.8571,0.2970,0.8750,0.7656,0.8751,0.2582,0.6032,0.7511,1.0, 0.5024,0.3908,0.5037,0.2505,0.5178,0.8571,0.2581,0.7813,0.8768,0.5168, 0.5169,0.5318,1.0,0.5079,0.2510,0.8889,0.7778,0.0,1.0,0.5006,0.3908,0.8571, 0.8750,0.7143,0.5873,0.0,0.0,0.5010,0.5001,0.7537,0.5020,0.7500,0.2500, 0.2188,0.5556,0.7143,0.5873,0.6281,0.5001,0.6032,0.5587,0.0762,0.7500,1.0, 0.7500,0.1623,0.6094,0.2305,0.5001,0.7501,0.6984,0.5220,0.5002,0.8768, 0.7537,0.7505,0.8571,0.7500,0.7500,0.7556,0.5347,0.3333,0.5020,0.7501, 0.7507,0.5010,0.7501,0.2501,1.0,0.5000,0.5000,0.8750,0.7500,0.5938,0.8172, 0.7500,0.5873,0.7500,0.5000,0.5000,0.5168,0.5001,0.7500,0.7500,0.5625,1.0, 0.5000,0.5005,0.4414,0.8889,0.7143,0.7778,0.5010,0.2354,0.5013,0.7500, 0.7500,0.7500,0.7500,0.5168,0.5001,0.7500,0.7500,0.2969,0.7143,0.7500, 0.9375,0.5002,0.5005,0.7501,0.7814,0.1254,0.5000,0.3789,0.7578,0.9375, 0.8750,0.5002,0.7500,0.6921,0.8750,0.5081,0.5002,0.6251,0.5001,0.7500, 0.7501,0.8750,0.5000,0.5083,0.7500,0.5163,0.5059,0.5001,0.7500,0.7512,

0.8750,0.5000,0.7500,0.5001,0.5001,0.6261,0.7500,0.7501,0.2500,0.5042, 0.5034,0.5010,0.5347,0.8594,0.5005,0.5168,0.5059,0.6885,0.5005,0.5001, 0.7143,0.8750,0.7500,0.5000,0.4377,0.5059,0.5000,0.5000,0.5001,0.6875, 0.5279,0.5000,0.5000,0.5020,0.6032,0.5082,0.5029,0.5000,0.5006,0.5005, 0.7501,0.7143,0.5000,0.5020,0.7500,0.7500,0.9395,0.5015,0.5003,0.5000, 0.8571,0.5235,0.5117,0.7500,0.7500
Russian Text 6
10,4,15,5,12,24,2,3,10,2,11,7,19,21,17,12,29,5,14,13,12,14,10,11,9,20,8,31,17,9, 22,9,19,22,9,4,8,8,9,3,5,7,19,7,4,2,14,9,7,13,1,1,2,9,5,13,12,12,12,6,3,15,8,9,8,1, 1,4,3,35,31,7,10,24,8,1,1,5,11,15,13,16,24,9,10,11,14,6,1,1,9,14,20,6,3,6,6,18, 16,13,12,11,14,26,7,12,8,13,6,6,8,4,15,12,14,18,6,11,15,2,8,19,8,6,28,9,24,6,6,5, 6,4,5,5,19,14,4,7,18,5,5,14,12,26,12,17,13,24,13,19,7,7,16,5,6,10,9,8,16,13,22, 14,6,15,15,9,7,20,12,10,13,14,15,8,23,12,7,11,8,14,9,5,27,12,6,2,4,4,4,3,2,1,6,1, 2,12,7,30,27,18,4,10,8,5,15,4,18,13,7,2,2,2,10,6,11,10,9,27,11,10,13,4,2,5,4,7,5, 1,8,2,2,6,2,2,2,5,4,3,14,9,2,17,3,27,20,24,13,31,11,12,25,14,19,10,18,19,11,7,4, 9,6,6,5,5,13,10,9,6,12,12,6,2,12,7,5,3,2,3,6,4,4,13,15,9,9,2,7,4,8,6,3,3,5,4,2,5,9, 4,6,2,6,2,10,5,5,5,8,5,6,6,9,2,5,4,3,2,7,10,6,4,1,16,16,15,4,6,12,8,17,20,18,13,8, 5,11,8,5,6,2,4,8,26,12,4,5,2,18,4,3,6,2,8,6,6,8,9,6,15,8,10,8,4,22,5,9,19,15,7,21, 13,20,13,9,11,13,14,2,10,10,24,7,7,4,12,10,12,4,4,24,7,6,15,7,11,12,1,17,24,17, 31,30,14,19,15,14,23,19,28,10,7,7,17,4,3,10,17,7,30,23,16,6,8,8,18,3,16,17,33, 15,7,10,18,4,18,20,3,2,1,8,3,5,5,5,7,2,6,4,28,15,14,13,13,16,8,5,19,29,14,16,3,6, 25,18,18,7,16,8,6,19,4,1,13,10,11,21,3,15,24,7,13,7 (n = 481)
0.5015,0.5873,0.6250,0.2864,0.7500,0.8750,1.0,0.8571,0.8125,1.0,0.7500, 0.7501,0.5000,0.5000,0.7500,0.8125,0.2500,0.6921,0.7500,0.5001,0.7500, 0.5000,0.5010,0.5005,0.3770,0.5000,0.7501,0.8438,0.7500,0.5023,0.7500, 0.5010,0.5469,0.7813,0.5007,0.7778,0.7588,0.5055,0.5020,0.8571,0.9384, 0.5042,0.5000,0.7503,0.6508,1.0,0.6563,0.7500,0.5082,0.0079,0.0,0.0,1.0, 0.5020,0.7556,0.7500,0.9375,0.5002,0.5002,0.9063,0.7143,0.7500,0.5059, 0.7500,0.5020,0.0,0.0,0.8889,0.7143,0.8750,0.7500,0.5079,0.5010,0.8750, 0.5059,0.0,0.0,0.9384,0.5002,0.7500,0.5625,0.5000,0.8750,0.5024,0.5010, 0.7500,0.5000,0.5163,0.0,0.0,0.5020,0.7500,0.7500,0.5173,0.7143,0.5090, 0.5168,0.2500,0.4375,0.3751,0.5002,0.7500,0.5001,0.5000,0.5042,0.7500, 0.8750,0.7500,0.5090,0.5173,0.7501,0.7778,0.7500,0.7500,0.2501,0.8750, 0.5235,0.7500,0.5000,1.0,0.5040,0.6328,0.5040,0.7517,0.7500,0.5020,0.5000, 0.5171,0.5173,0.5367,0.5173,0.5873,0.9384,0.8143,0.5000,0.5001,0.8889, 0.3985,0.5000,0.1799,0.8172,0.5001,0.7500,0.5000,0.5000,0.5000,0.6876, 0.5039,0.5002,0.5000,0.5088,0.7501,0.5000,0.5376,0.5147,0.5007,0.5010, 0.1290,0.7500,0.8750,0.7500,0.7500,0.7512,0.5000,0.5000,0.7500,0.9375, 0.8125,0.5002,0.8750,0.5000,0.7500,0.7500,0.7501,0.5000,0.5002,0.5081, 0.5005,0.4414,0.7500,0.5020,0.5367,0.5000,0.8750,0.7815,1.0,0.4444,0.5556, 0.3333,0.7143,1.0,0.0,0.5108,0.0,1.0,0.8750,0.5081,0.2500,0.5000,0.7500, 0.7778,0.5017,0.5040,0.5210,0.7500,0.5556,0.5000,0.7500,0.5046,1.0,1.0,1.0, 0.5010,0.7517,0.5005,0.5002,0.5020,0.5000,0.5005,0.7500,0.2501,0.7778,1.0, 0.5347,0.7778,0.7501,0.7556,0.0,0.2540,1.0,1.0,0.5051,1.0,1.0,1.0,0.5367,

0.7778,0.4286,0.8438,0.5020,1.0,0.7500,0.4286,0.7656,0.4688,0.5000,0.5002, 0.8750,0.3755,0.5002,0.5000,0.5001,0.6260,0.7500,0.5000,0.7500,0.5005, 0.7501,0.5873,0.7500,0.7515,0.3907,0.5367,0.5484,0.7500,0.5010,0.8750, 0.8752,0.7500,0.7500,0.5031,1.0,0.5002,0.5315,0.5210,0.7143,1.0,0.8571, 0.5632,0.5556,0.5556,0.5002,0.5000,0.7656,0.7500,1.0,0.5081,0.7778,0.5059, 0.6720,0.7143,0.8571,0.5406,0.5873,1.0,0.5142,0.5016,0.1746,0.5171,1.0, 0.5471,1.0,0.5012,0.4076,0.5376,0.6921,0.7822,0.2864,0.0548,0.7515,0.5020, 1.0,0.7674,0.5873,0.7143,1.0,0.7503,0.7500,0.7515,0.8889,0.0,0.5000,0.5000, 0.8750,0.5873,0.5171,0.5002,0.3789,0.4609,0.7813,0.5000,0.5002,0.6876, 0.2874,0.5005,0.8828,0.5484,0.5051,1.0,0.8889,0.5059,0.2500,0.7188,0.5556, 0.5279,1.0,0.1250,0.8889,0.7143,0.5147,1.0,0.5020,0.4219,0.9376,0.5040, 0.5010,0.7517,0.5000,0.5010,0.5016,0.5049,0.6032,0.5000,0.5367,0.5020, 0.5000,0.5000,0.7500,0.5000,0.1876,0.8750,0.5001,0.3145,0.7500,0.7500, 0.2501,1.0,0.5625,0.5017,0.7500,0.3984,0.5022,0.3333,0.7500,0.5010,0.6250, 0.2222,0.8889,0.7500,0.6253,0.5168,0.5000,0.6333,0.7500,0.5002,0.0,0.2500, 0.5000,0.5000,0.5000,0.7500,0.5001,0.8911,0.5000,0.8750,0.8750,0.8750, 0.8750,0.5625,0.7503,0.7503,0.7500,0.6032,0.7143,0.5010,0.5000,0.5235, 0.5000,0.7500,0.7656,0.7515,0.5020,0.7501,0.9375,0.7143,0.7500,0.5000, 0.9375,0.5000,0.5080,0.8125,0.7500,0.6508,0.8750,0.6250,0.7143,1.0,0.0, 0.5005,0.7143,0.6921,0.5142,0.7566,0.5080,1.0,0.5161,0.8889,0.7500,0.7500, 0.7500,0.8750,0.7500,0.8750,0.7500,0.5376,0.5000,0.8438,0.5001,0.2500, 0.8571,0.7512,0.5000,0.3750,0.7500,0.3828,0.7344,0.5040,0.6408,0.7500, 0.5873,0.0,0.5001,0.7813,0.5002,0.7500,0.8571,0.1250,0.5000,0.5007,0.2033, 0.7501
Russian text 7
12,21,13,27,10,3,6,17,23,20,18,4,29,15,9,12,22,26,33,2,7,19,5,8,11,21,22,7,21, 19,5,12,2,15,11,12,31,5,31,6,12,26,24,20,28,8,14,17,17,15 (n = 50)
0.5002,0.5000,0.5002,0.5000,0.5016,0.8571,0.5046,0.5000,0.5000,0.8750, 0.7500,0.7778,0.5000,0.5000,0.6250,0.7500,0.1406,0.6250,0.5625,1.0,0.7500, 0.5000,0.7537,0.5040,0.8750,0.1875,0.5000,0.5081,0.5938,0.5000,0.9384, 0.7500,1.0,0.5000,0.5007,0.6406,0.5000,0.7537,0.5000,0.5170,0.8750,0.8750, 0.5000,0.5000,0.5000,0.8750,0.7500,0.7500,0.6250,0.5000
Russian text 8
5,13,19,6,18,17,5,11,7,33,24,9,11,25,24,30,9,37,49,4,23,5,18,7,23,16,9,6,18,4, 6,5,14,33,13,14,13,6,31,43,18,8,15,22,20,10,34,27,21,11,13,15,35,3,3,3,5,13,8, 2,9,16,7,18,4,4,15,1,26,21,22,18,4,16,14,13,20,14,17,9,4,16,8,23,8,17 (n = 86)
0.8768,0.8750,0.7500,0.5637,0.8750,0.7505,0.1105,0.5938,0.5100,0.5000, 0.5000,0.7852,0.5006,0.7500,0.8750,0.7500,0.5029,0.8750,0.6250,0.8889, 0.2524,0.7566,0.0664,0.7502,0.5000,0.7500,0.5029,0.5206,0.7500,0.7778, 0.5171,0.5484,0.7500,0.5000,0.5002,0.8750,0.5001,0.5235,0.7500,0.5000, 0.5004,0.5045,0.5000,0.7500,0.7500,0.8750,0.5000,0.7500,0.0469,0.5005, 0.8750,0.7656,0.5000,0.8571,0.4286,0.7143,0.6921,0.7500,0.5039,1.0,0.5029, 0.7500,0.5122,0.7500,0.6984,0.6984,0.5000,0.0,0.5000,0.7500,0.5000,0.5000, 0.5873,0.7500,0.8750,0.7500,0.7500,0.7500,0.5000,0.5024,0.8889,0.5000,

0.7501,0.5000,0.1309,0.7500
Russian text 9
6,16,35,10,19,14,26,16,12,12,14,7,9,10,32,5,16,35,30,28,42,12,30,15,35,11,30,5, 41,34,11,17,22,14,16,9,15,22,17,9,25,7,17,4,30,22,15,5,8,25,17,12,25,20,16,31,8 (n = 57)
0.5167,0.5000,0.5000,0.6885,0.8750,0.7856,0.5000,0.6523,0.1877,0.5938, 0.5001,0.0862,0.5020,0.0068,0.5000,0.7537,0.5000,0.5000,0.5000,0.5000, 0.5000,0.7500,0.8184,0.7500,0.5000,0.7500,0.5000,0.5552,0.8750,0.5000, 0.1567,0.2188,0.8750,0.8751,0.5000,0.5001,0.5000,0.0781,0.5000,0.5022, 0.5000,0.7504,0.7500,0.7778,0.5000,0.5000,0.7500,0.8778,0.5007,0.7500, 0.7500,0.7500,0.5000,0.5000,0.7500,0.5000,0.9375
Russian text 10
1,3,12,40,17,10,7,1,6,23,19,42,18,2,16,30,22,17,11,2,3,32,15,7,9,10,19,5,8,5, 14,5,17,16,18,12,2,10,10,3,8,25,10,16,6,11,8 (n = 47)
0.0,0.7143,0.7500,0.7500,0.5000,0.5640,0.5081,0.0,0.6026,0.5000,0.7500, 0.7500,0.8750,1.0,0.6250,0.7500,0.5000,0.5000,0.5313,1.0,0.8571,0.6250, 0.1250,0.7503,0.5020,0.7520,0.7500,0.7566,0.7734,0.0753,0.5001,0.5435, 0.1250,0.7500,0.7500,0.2502,1.0,0.5010,0.6172,0.8571,0.5046,0.8750,0.5010, 0.6250,0.5163,0.5005,0.5235
Russian text 11
4,9,25,17,12,17,13,17,16,5,10,6,5,7,22,15,13,11,14,6,12,17,14,23,27,5,9,4,6,13, 11,8,3,10,8,10,11,28,28,18,6,4,9,7,11,11,13,11,8,18,3,9,9,6,5,18,12,12 (n = 58)
0.5873,0.7500,0.5000,0.1563,0.5002,0.7500,0.5002,0.5000,0.5000,0.3451, 0.5015,0.5236,0.5132,0.7503,0.5000,0.7500,0.5001,0.8760,0.7500,0.5216, 0.5002,0.7500,0.5001,0.5000,0.6250,0.6305,0.6270,0.7778,0.5109,0.7500, 0.5007,0.5313,0.4286,0.5010,0.7501,0.5010,0.8750,0.7500,0.1250,0.5000, 0.5254,0.5873,0.5005,0.5081,0.5005,0.8750,0.5001,0.5005,0.5313,0.7500, 0.8571,0.5020,0.6250,0.5160,0.7566,0.8750,0.7500,0.7500
Russian text 12
4,13,22,27,10,8,2,10,32,29,10,13,13,19,15,18,4,12,24,13,16,23,9,15,24,5,17,21, 10,24,3,17,24,34,7,12,10,43,21,16,31,22,15,14,13,12,17,14 (n = 48)
0.5873,0.8125,0.7500,0.5000,0.5010,0.7501,1.0,0.6260,0.5000,0.5000,0.5010, 0.7500,0.7500,0.5000,0.7500,0.5000,0.6508,0.5004,0.5000,0.5001,0.5000, 0.5000,0.7500,0.7500,0.7500,0.7556,0.8750,0.7500,0.7500,0.7500,0.8571, 0.5000,0.5000,0.7500,0.5117,0.7500,0.5215,0.5000,0.5000,0.5000,0.5000, 0.6875,0.7500,0.7500,0.7500,0.7500,0.7500,0.5001
Russian text 13
2,10,17,10,15,17,13,9,18,9,12,9,9,17,14,9,24,8,9,23,19,18,15,7,17,15,2,20,6,26, 6,16,24,7,30,13,7,4,14,15,6,13,19,17,18,30,31,11,14 (n = 49)
1.0,0.7559,0.7500,0.7500,0.5000,0.5313,0.7500,0.7500,0.5000,0.5025,0.7500, 0.5020,0.5645,0.5000,0.5000,0.5005,0.5625,0.5003,0.7500,0.7500,0.5000, 0.5000,0.7500,0.5079,0.5000,0.5000,1.0,0.5938,0.7512,0.8750,0.7513,0.5000, 0.5000,0.0742,0.7500,0.6251,0.7501,0.7778,0.5001,0.5000,0.5178,0.7500, 0.5625, 0.5000,0.8438,0.2031,0.5313,0.8125,0.5490

Russian text 14
3,5,4,3,6,10,9,20,19,10,13,5,3,14,20,34,5,14,31,23,29,13,32,19,12,7,2,21,7,22, 18, 27,10,17,15,12,25,14,25,10,17,13 (n = 42)
0.7143,0.7556,0.7778,0.7143,0.5051,0.5011,0.2524,0.3750,0.5000,0.5012, 0.5001,0.6921,0.8571,0.8125,0.5625,0.5000,0.7556,0.3790,0.7500,0.5000, 0.5313,0.1876,0.2500,0.5000,0.7500,0.7969,1.0,0.7670,0.5042,0.7500,0.7500, 0.5000,0.8750,0.5000,0.4375,0.2893,0.7500,0.7500,0.5625,0.7500,0.5000, 0.7500
Russian text 15
5,3,4,7,5,12,18,22,15,16,18,25,8,7,22,18,10,19,11,9,12,10,7,10,9,12 (n = 26)
0.5279,0.7143,0.8889,0.7034,0.7566,0.2546,0.7500,0.5000,0.5000,0.7500, 0.0742,0.5039,0.7500,0.8750,0.8750,0.5000,0.5010,0.7500,0.3755,0.7500, 0.5000,0.5010,0.6484,0.1104,0.5029,0.5002
Russian text 16
2,7,10,22,7,6,13,6,16,13,7,24,11,12,15,12,16,15,2,17,5,12,14,14,19,21,8,17,9,4, 6,9,14,18,18,7,5,18,5,5,24,19,9,8,14,8,13,3,11,15,7,7,4,8,13,19,6,13,24,16,17,12, 20,7 (n = 64)
1.0,0.5110,0.0106,0.8750,0.5081,0.5470,0.7657,0.5949,0.5000,0.7500,0.7501, 0.3760,0.7500,0.1877,0.5000,0.5004,0.5000,0.7500,1.0,0.5000,0.5376,0.5003, 0.5000,0.7500,0.5000,0.2813,0.7501,0.5000,0.1270,0.8889,0.6408,0.2051, 0.7500,0.0938,0.5000,0.7503,0.5660,0.5000,0.5376,0.2346,0.5000,0.5000, 0.5020,0.7501,0.5001,0.5002,0.9375,0.8571,0.5001,0.9375,0.0862,0.7503, 0.8889,0.7501,0.5002,0.3750,0.7513,0.0704,0.6250,0.5000,0.5000,0.9375, 0.5000,0.8127
Russian text 17
3,13,6,17,13,11,14,7,18,10,18,24,17,21,15,17,11,6,27,16,11,10,5,13,32,8,27,8, 19,14,9,12,10,23,7,26,9,13,12,29,17,12,13,16,14,5,11,8,21 (n = 49)
0.7143,0.8750,0.5177,0.5000,0.5001,0.5005,0.5000,0.0203,0.7500,0.7500, 0.7500,0.5000,0.5000,0.5000,0.1719,0.8750,0.8750,0.7515,0.3750,0.7500,0.750 0,0.7500,0.0968,0.5001,0.5000,0.8750,0.8750,0.7501,0.5000,0.5000,0.7500,0.54 69,0.7500,0.5000,0.7503,0.8750,0.7500,0.5001,0.5001,0.8750,0.3750,0.8130,0.7 500,0.5000,0.5000,0.5406,0.5938,0.5040,0.5000
Russian Text 18
3,6,15,22,22,15,13,12,10,9,5,6,13,6,8,9,14,1,13,26,2,22,6,31,11,16,6,4,11,22,11, 4,28,12,6,13,11,13 (n = 38)
0.8571,0.7512,0.5000,0.6250,0.7500,0.7500,0.5002,0.5004,0.5010,0.5020,0.537 6,0.5031,0.5000,0.6262,0.9375,0.2520,0.7504,0.0,0.7500,0.5000,1.0,0.7500,0.65 73,0.9043,0.6250,0.7500,0.5168,0.4444,0.5005,0.7500,0.5625,0.7778,0.5000,0.7 500,0.5236,0.7500,0.6250,0.7500
Russian text 19
3,2,6,12,10,1,10,17,3,22,10,17,7,5,4,16,6,6,17,12,14,7,17,14,3,6,9,9,5,3,7,11,15, 6,8,20,11,18,6,11,9,3,4,7,5,13,3,21,4,3,4,8,13,8,13,7,2,8,15,3,12,5,10,8,8,8,7,13, 15,15,7,7,13,15,7,2,15,12,9,15,8,15,3,10,12,24,12,13,19,6,17,5,5,7,7,4,14,11,30, 2 (n = 100)

0.7143, 1.0, 0.5171, 0.6252, 0.5005, 0.0, 0.0947, 0.6250, 0.7143, 0.5625, 0.0328, 0.5000, 0.5042, 0.5484, 0.5873, 0.5000, 0.6720, 0.0343, 0.7500, 0.7500, 0.5001, 0.5785, 0.7500, 0.9375, 0.7143, 0.7512, 0.5024, 0.5020, 0.5552, 0.7143, 0.5081, 0.3130, 0.2504, 0.5178, 0.1524, 0.7500, 0.8750, 0.7500, 0.5324, 0.0474, 0.2520, 0.7143, 0.7778, 0.7503, 0.6305, 0.7500, 0.8571, 0.5000, 0.5873, 0.7143, 0.7778, 0.5626, 0.5001, 0.8750, 0.5001, 0.7503, 1.0, 0.5021, 0.5000, 0.7143, 0.7500, 0.0762, 0.7500, 0.7578, 0.6289, 0.7501, 0.5082, 0.5000, 0.5000, 0.0635, 0.1992, 0.7501, 0.5001, 0.7500, 0.5082, 1.0, 0.7500, 0.7500, 0.8750, 0.5000, 0.7501, 0.7500, 0.8571, 0.8750, 0.7190, 0.5000, 0.5001, 0.5000, 0.7500, 0.5165, 0.7500, 0.5435, 0.7556, 0.5547, 0.5081, 0.7778, 0.7500, 0.7500, 0.7500, 1.0
Russian text 20
3,8,17,33,21,22,10,25,16,40,10,29,12,2,38,4,9,22,36,12,16,12,15,13,32,27,8,1, 16,20,19,11,50,8,25,14 (n = 36)
0.7143, 0.5313, 0.8750, 0.9375, 0.5625, 0.7500, 0.6260, 0.6250, 0.7500, 0.7500, 0.2197, 0.7500, 0.5002, 1.0, 0.5000, 0.8889, 0.7500, 0.9844, 0.8750, 0.5004, 0.7500, 0.8125, 0.8750, 0.7500, 0.9375, 0.9082, 0.7501, 0.0, 0.7500, 0.5000, 0.7500, 0.1880, 0.5000, 0.5045, 0.8750, 0.7500

Table 7.3
Binary codes of sentence structures in 20 Czech texts
Prague dependency Treebank 2.0 (Hajič et al. 2006)[1]
[First line: number of words in sentence. Second line: Binary code]

1. M.Slezák, Hra o Tengovo dědictví v Číne začíná. Lidové noviny, 200/1994
7,2,24,16,10,14,8,8,17,16,5,9,24,7,25,8,13,21,18,37,6,15,13,18,4,8,14,8,9,12,14, 11,18,8,11,9,16,4,9,4,9,21,7,4,7,15,7,8,4 (n = 49)
0.1417, 0.5000, 0.6602, 0.7512, 0.7530, 0.7530, 0.7666, 0.7511, 0.5205, 0.7032, 0.7529, 0.6407, 0.5157, 0.7666, 0.8135, 0.6651, 0.7667, 0.1094, 0.7666, 0.0743, 0.7666, 0.8955, 0.7666, 0.5510, 0.5156, 0.6299, 0.7510, 0.7666, 0.6416, 0.7667, 0.8145, 0.7666, 0.6573, 0.7658, 0.6953, 0.8203, 0.8135, 0.7656, 0.5220, 0.6406, 0.8126, 0.7530, 0.6416, 0.6406, 0.6260, 0.6407, 0.5217, 0.5157, 0.6406
2. M. Slezák, Dračí emisar u kremelského orla. Lidové noviny, 211/1994
5,9,5,16,18,41,11,42,27,16,28,9,26,18,28,24,19,16,25,22,4,12,14,19,19,7,10,5, 11,14,14 (n = 31)
0.7510, 0.8135, 0.6260, 0.7672, 0.8125, 0.8135, 0.1533, 0.8145, 0.5469, 0.5782, 0.7676, 0.5254, 0.8135, 0.7657, 0.6427, 0.5791, 0.5217, 0.2266, 0.7677, 0.7041, 0.5781, 0.6875, 0.7529, 0.6416, 0.7676, 0.8135, 0.7668, 0.5479, 0.7744, 0.7042, 0.1685,(n = 31)
3. I. Krčálová, Podnikatelská banka nabírá dech. Lidové noviny, 202/1994.
4,9,2,12,15,32,11,18,11,5,20,32,14,10,6,28,21,22,9,16,9,25,6,16,8,9,22,10,16,15, 8 (n = 31)

[1] [1] We thank Petr Pajas for help with analysis of Prague Dependency Treebank

0.6406, 0.7666, 0.5000, 0.5176, 0.6457, 0.8135, 0.8125, 0.1573, 0.6416, 0.7529, 0.8135, 0.7666, 0.6406, 0.5791, 0.7667, 0.7510, 0.7510, 0.7666, 0.3760, 0.7676, 0.6427, 0.8135, 0.7666, 0.6563, 0.7657, 0.7510, 0.7657, 0.8135, 0.6584, 0.7667, 0.6426
4. I. Krčálová, Poločas bankovní nerovnováhy. Lidové noviny, 209/1994.
3,2,13,11,25,19,15,6,10,5,20,25,11,20,5,15,30,10,9,27,12,40,19,19,17,13,14,12 (n = 28)
0.7500, 0.5000, 0.7667, 0.7657, 0.6416, 0.6885, 0.8135, 0.6408, 0.6582, 0.7666, 0.7735, 0.5469, 0.8125, 0.7667, 0.5254, 0.6953, 0.5791, 0.1955, 0.7510, 0.8135, 0.8135, 0.8130, 0.7657, 0.2198, 0.0176, 0.7657, 0.2266, 0.8135
5. J. Stuchlíkova, Ulster: tajný trumf katolíků. Lidové noviny, 209/1994.
4, 7, 17, 11, 14, 10, 7, 5, 15, 11, 17, 39, 24, 10, 24, 8, 12, 15, 8, 18, 30, 30, 8, 8, 9, 13, 25, 32, 8, 18, 17, 12, 12, 20, 6, 14, 17, 5, 7, 18, 8, 16, 27, 5, 11, 10, 10, 13, 7, 13, 6, 30 (n = 52)
0.6250, 0.5162, 0.7657, 0.7666, 0.7734, 0.5469, 0.7530, 0.5107, 0.6563, 0.8135, 0.5259, 0.7510, 0.5255, 0.7511, 0.7657, 0.6301, 0.6646, 0.6263, 0.7666, 0.6602, 0.3756, 0.4453, 0.8208, 0.6457, 0.8135, 0.7667, 0.7666, 0.7063, 0.3765, 0.7676, 0.6573, 0.7677, 0.6407, 0.3843, 0.5470, 0.7678, 0.7667, 0.7656, 0.6573, 0.6426, 0.7666, 0.7667, 0.6563, 0.8203, 0.6416, 0.7510, 0.5172, 0.5864, 0.7588, 0.6261, 0.6885, 0.6408
6. J. Stuchlíková, Ulster: protestanti v úzkých. Lidove noviny, 211/1994
4,12,3,33,15,5,15,9,8,17,7,21,5,18,22,29,17,9,23,32,20,17,10,9,17,9,21,13,18 (n = 29)
0.6563,0.6407,0.1250,0.7744,0.7666,0.8135,0.5158,0.3916,0.5181,0.7657, 0.6885,0.2042,0.8135,0.7667,0.5791,0.6407,0.6407,0.5014,0.7511,0.6281, 0.7666,0.7750,0.2198,0.6563,0.7668,0.5010,0.5157,0.7510,0.7510
7. V. Klaus, Životní prostředí a společenský systém. Lidové noviny, 02/1994.
5, 5, 25, 27, 37, 30, 8, 33, 40, 45, 10, 5, 24, 49, 10, 69, 20, 39, 31, 25, 16, 48, 24, 23, 37, 53, 26, 23, 17 (n = 29)
0.7666,0.7510,0.7657,0.7667,0.6504,0.8125,0.7662,0.7672,0.0168,0.4385, 0.4453,0.8135,0.7666,0.7676,0.7667,0.7510,0.8135,0.6575,0.6494,0.8125, 0.6426,0.7657,0.6262,0.7667,0.7038,0.1563,0.6426,0.6573,0.7666,
8. V. Klaus, Rozpočet a věčné levicové vábení. Lidové noviny, 208/1994.
5, 4, 5, 11, 12, 6, 8, 7, 20, 17, 6, 9, 41, 10, 14, 23, 4, 26, 6, 15, 25, 33, 10, 34, 45, 9, 18, 17, 41, 48, 24, 13, 41, 45, 39, 6, 27, 28, 35, 7, 16, (n = 41)
0.8135, 0.5156, 0.7666, 0.7667, 0.6563, 0.7510, 0.7510, 0.7510, 0.7657, 0.8135, 0.6407, 0.8208, 0.6890, 0.5070, 0.6417, 0.7666, 0.3438, 0.7081, 0.1095, 0.5859, 0.5244, 0.5157, 0.6604, 0.7032, 0.7041, 0.7657, 0.7657, 0.6426, 0.8125, 0.6602, 0.8135, 0.7511, 0.8167, 0.7657, 0.6494, 0.8135, 0.7041, 0.5176, 0.6417, 0.7676, 0.8208
9. M. Achremenko, Sběrné suroviny se chovají podle poptávky. Lidové noviny, 200/1994.
6, 17, 2, 9, 13, 16, 13, 6, 20, 17, 19, 13, 14, 24, 3, 25, 28, 16, 10, 8, 6, 19, 8, 10, 19, 14, 20, 28, 22, 10, 18, 16, 11, 7, 13, 12, 10, 5, 16, 13 (n = 40)

0.6426, 0.7667, 0.5000, 0.1358, 0.5479, 0.6416, 0.7511, 0.6418, 0.1358, 0.7823, 0.7511, 0.5864, 0.6445, 0.7666, 0.6250, 0.6426, 0.5479, 0.8135, 0.7667, 0.6407, 0.4385, 0.7657, 0.7657, 0.6338, 0.6416, 0.6572, 0.6445, 0.6885, 0.7677, 0.6573, 0.6582, 0.7657, 0.8135, 0.6260, 0.7666, 0.8140, 0.3763, 0.0010, 0.6575, 0.6573
10. M. Achremenko, SIF a Milo Olomouc mezi elitou. Lidové noviny, 204/1994.
6, 10, 2, 17, 16, 25, 10, 9, 17, 10, 18, 31, 11, 18, 9, 26, 9, 27, 10, 44, 16, 20, 14, 7 (n = 24)
0.5166, 0.7657, 0.5000, 0.6260, 0.6885, 0.6641, 0.6416, 0.6281, 0.6563, 0.7530, 0.8976, 0.7666, 0.7510, 0.5166, 0.7815, 0.7676, 0.6573, 0.5469, 0.6260, 0.0293, 0.5010, 0.3916, 0.8135, 0.6953
11. V. Cílek, Milankovičův cyklus zpochybněn. Vesmír 72, 1993/3.
3, 2, 20, 15, 17, 30, 9, 28, 19, 15, 4, 14, 16, 19, 17, 27, 14, 15, 7, 19, 19, 24, 14, 17, 23, 14, 40, 4, 25, 19, 13, 10, 10, 14, 24, 14, 26, 14, 24, 15, 24, 10, 36, 6, 22, 23, 29, 21 (n = 48)
0.1250, 0.5000, 0.6416, 0.6573, 0.6426, 0.8135, 0.8135, 0.6426, 0.5244, 0.6418, 0.2188, 0.6602, 0.7588, 0.5215, 0.7658, 0.5783, 0.6416, 0.6426, 0.7667, 0.6445, 0.6416, 0.7512, 0.6563, 0.7129, 0.8203, 0.8135, 0.6421, 0.6406, 0.7666, 0.6426, 0.4385, 0.7657, 0.8135, 0.6407, 0.7042, 0.2188, 0.6250, 0.6953, 0.7530, 0.6417, 0.5169, 0.5166, 0.1104, 0.6280, 0.7676, 0.7667, 0.6416, 0.2266
12. V. Cílek, Apokalypsa, nebo eden? Vesmír 72 1993/1.
3, 8, 2, 43, 26, 21, 24, 12, 2, 25, 21, 25, 21, 23, 10, 14, 15, 5, 9, 11, 23, 14, 17, 7, 7, 26, 12, 6, 10, 18, 26, 3, 18, 17, 6, 18, 12, 2, 8, 31, 15, 35, 17, 14, 16, 13, 16, 36, 20, 15, 15, 34, 16, 8, 14, 3, 13, 16, 22, 12, 21, 16, 6, 12, 11, 7, 17, 4, 20, 4, 8, 14, 31, 14, 9, 22, 27, 27, 36, 6, 22, 11, 16, 20, 61, 23 (n = 86)
0.6250, 0.7746, 0.5000, 0.7676, 0.7510, 0.6416, 0.7666, 0.7511, 0.5000, 0.8125, 0.7515, 0.6426, 0.7676, 0.6407, 0.3917, 0.8135, 0.1573, 0.6416, 0.6417, 0.7676, 0.8135, 0.6281, 0.7666, 0.7666, 0.6426, 0.6260, 0.7588, 0.6426, 0.7515, 0.8126, 0.7666, 0.3750, 0.6260, 0.5479, 0.5479, 0.7666, 0.5178, 0.5000, 0.7657, 0.7749, 0.7510, 0.6426, 0.7110, 0.7744, 0.8125, 0.7657, 0.6407, 0.7041, 0.5781, 0.7114, 0.6301, 0.7510, 0.7735, 0.6260, 0.7744, 0.6250, 0.3760, 0.7666, 0.4453, 0.7666, 0.6263, 0.7667, 0.6426, 0.7676, 0.7110, 0.0012, 0.7515, 0.7656, 0.5782, 0.5156, 0.5254, 0.5791, 0.7676, 0.3916, 0.6885, 0.6407, 0.6426, 0.6572, 0.6262, 0.7676, 0.6572, 0.7676, 0.7657, 0.7512, 0.8135, 0.5479
13. J J. Chalupský, Larvy nesou smrt. Vesmír 72, 1993/1.
3, 4, 2, 28, 9, 20, 20, 23, 5, 11, 27, 40, 6, 6, 17, 10, 7, 15, 22, 13, 16, 5, 13, 3, 14, 10, 12, 26, 20, 4, 26, 21, 26, 17, 11, 18, 8, 10, 38, 20, 19, 12, 14, 28, 22, 11, 12, 22, 8, 35 (n = 51)
0.6250, 0.6406, 0.5000, 0.6504, 0.5479, 0.7735, 0.6565, 0.8126, 0.6426, 0.5068, 0.6417, 0.6280, 0.5245, 0.7676, 0.7530, 0.5782, 0.6338, 0.6573, 0.7041, 0.3916, 0.8135, 0.7588, 0.6279, 0.6250, 0.5859, 0.7511, 0.6416,

0.8204, 0.7657, 0.6406, 0.6416, 0.7667, 0.8145, 0.8145, 0.6563, 0.5791, 0.8203, 0.6416, 0.6494, 0.7658, 0.7080, 0.7657, 0.5031, 0.6417, 0.6422, 0.5479, 0.7666, 0.6260, 0.5167, 0.1651
14. J. Chalupský, Vášně kolem "DNA otisků prstů". Vesmír 71, 1992/11.
5, 2, 24, 16, 4, 4, 36, 23, 24, 17, 15, 17, 22, 27, 13, 14, 8, 29, 27, 7, 28, 14, 25, 6, 11, 4, 30, 15, 28, 12, 18, 10, 12, 23, 17, 17, 4, 11, 9, 17, 9, 38, 4, 6, 17, 12, 23, 19, 14, 26, 12, 16, 15, 5, 20, 39, 23, 21, 18, 15, 12, 10, 17, 44, 22, 6, 17, 29, 7, 30, 18, 15, 11, 32, 23, 13, 20, 29 (n = 78)
0.6416, 0.5000, 0.1104, 0.1260, 0.3906, 0.1094, 0.3760, 0.6416, 0.7510, 0.6417, 0.6646, 0.7510, 0.7114, 0.8135, 0.3838, 0.1104, 0.8028, 0.2266, 0.1250, 0.4375, 0.7677, 0.2041, 0.7061, 0.8208, 0.7667, 0.6563, 0.6604, 0.6563, 0.7744, 0.4385, 0.7666, 0.6416, 0.5781, 0.1094, 0.7658, 0.7041, 0.5781, 0.6417, 0.8203, 0.4458, 0.1095, 0.0000, 0.5781, 0.6426, 0.3926, 0.6583, 0.6417, 0.8125, 0.7666, 0.2114, 0.6416, 0.8135, 0.4385, 0.6426, 0.7667, 0.1414, 0.7676, 0.8135, 0.6885, 0.0000, 0.2032, 0.7657, 0.7666, 0.7046, 0.8125, 0.6417, 0.6573, 0.1573, 0.7667, 0.8135, 0.7510, 0.7657, 0.7666, 0.1094, 0.6416, 0.7657, 0.6573, 0.8916
15. S. Komárek, K historickým a psychologickým kořenům pojmu mimikry. Vesmír 72, 1993/3.
7, 2, 27, 17, 9, 25, 30, 23, 14, 25, 32, 18, 24, 15, 30, 29, 37, 38, 14, 10, 38, 27 (n = 22)
0.6418, 0.5000, 0.6416, 0.5479, 0.6602, 0.8125, 0.6260, 0.1260, 0.6563, 0.7677, 0.5791, 0.7734, 0.6426, 0.3918, 0.7667, 0.6416, 0.7510, 0.7667, 0.5176, 0.5860, 0.5782, 0.6250
16. S. Komárek, Jarmark a chrám. Vesmír 71, 1992/11.
3, 2, 14, 16, 29, 26, 24, 19, 28, 32, 10, 25, 20, 22, 16, 18, 26, 22, 12, 42, 38, 19, 17, 21, 16, 5, 15, 24 (n = 28)
0.6250, 0.5000, 0.7657, 0.5254, 0.7744, 0.6426, 0.6407, 0.7667, 0.5479, 0.6279, 0.7677, 0.7682, 0.8919, 0.5217, 0.8204, 0.6504, 0.1573, 0.7676, 0.5176, 0.6281, 0.7510, 0.6407, 0.6573, 0.7510, 0.6563, 0.6953, 0.5169, 0.3926
17. F. Koukolík, O bytí a porozumění. Vesmir 71, 1992/11.
4, 2, 38, 27, 21, 6, 39, 10, 13, 49, 12, 23, 21, 26, 5, 9, 22, 26, 15, 20, 6, 25, 25, 10, 22, 25, 29, 36, 20, 26, 33, 17, 2, 13, 28, 18, 13, 9, 20, 27, 12, 27, 23, 20, 16, 24, 20, 12, 12, 12, 9, 15, 3, 30, 16, 9, 19, 16, 21, 37, 21, 39, 7, 25, 20, 24, 47, 26, 12, 12, 8, 4, 20, 11, 15, 31, 18, 12, 5, 14, 5, 6, 9, 14, 25, 9, 10, 9, 20, 6, 4, 4, 2, 8, 7, 6, 4, 10 (n = 100)
0.7031, 0.5000, 0.3467, 0.8135, 0.6604, 0.7657, 0.7657, 0.6417, 0.7500, 0.1104, 0.6563, 0.6446, 0.6282, 0.7666, 0.5244, 0.5782, 0.6262, 0.7813, 0.8203, 0.7686, 0.6426, 0.7666, 0.3907, 0.5254, 0.6426, 0.5480, 0.6875, 0.5864, 0.7041, 0.8140, 0.6427, 0.7676, 0.5000, 0.1455, 0.6953, 0.6573, 0.7666, 0.7042, 0.6416, 0.6602, 0.5791, 0.7657, 0.6260, 0.7739, 0.6260, 0.8135, 0.7041, 0.5177, 0.6261, 0.6260, 0.6407, 0.6407, 0.6250, 0.7666, 0.7657, 0.7510, 0.6573, 0.8135, 0.1432, 0.6427, 0.5479, 0.7530, 0.6446,

0.8208, 0.6319, 0.7676, 0.5864, 0.1538, 0.7744, 0.1416, 0.5782, 0.0469, 0.6279, 0.7667, 0.5010, 0.0010, 0.5469, 0.7512, 0.7529, 0.6426, 0.8135, 0.5236, 0.7667, 0.7510, 0.7676, 0.7502, 0.6260, 0.7658, 0.8146, 0.1413, 0.6719, 0.7656, 0.5000, 0.7588, 0.2032, 0.7036, 0.5781, 0.6408
18. F. Koukolík, Spatříme - a poznáme. Vesmir 72, 1993/1.
3, 6, 2, 28, 17, 8, 5, 5, 17, 10, 14, 26, 32, 20, 13, 22, 34, 34, 27, 13, 24, 3, 3, 6, 18, 23, 23, 42, 25, 32, 20, 17, 18, 14, 16, 37, 40, 35, 13, 43, 44, 16, 24, 8, 3, 37, 16, 17, 14, 14, 11, 12, 21, 24, 22, 27, 33, 17, 9, 17, 5, 21, 34, 36, 17, 33, 42, 27, 11, 23, 11, 24, 26, 17, 35, 22, 2, 23, 5, 25, 18, 12, 6, 4, 9, 16, 26, 17, 11, 20, 21, 9, 22, 11, 13, 12, 26, 21, 8, 8, 31, 17, 19, 10, 17, 20, 23, 18, 18, 24, 16, 3, 26, 27, 30, 5, 48, 16, 31, 28, 12, 35, 17, 21, 25, 29, 6, 18, 45, 8, 39, 8 (n = 134)
0.6250, 0.7512, 0.5000, 0.7510, 0.4458, 0.5177, 0.6426, 0.8135, 0.8135, 0.7510, 0.7110, 0.5166, 0.6451, 0.6279, 0.7667, 0.7510, 0.5244, 0.6729, 0.6417, 0.6458, 0.5196, 0.3750, 0.7500, 0.7510, 0.7676, 0.5177, 0.7657, 0.6255, 0.7676, 0.7676, 0.7667, 0.7109, 0.5176, 0.7676, 0.8135, 0.5169, 0.8208, 0.7512, 0.6421, 0.8203, 0.7666, 0.8203, 0.7676, 0.6417, 0.6250, 0.7109, 0.7676, 0.6426, 0.6416, 0.5032, 0.6953, 0.8135, 0.7657, 0.6573, 0.7041, 0.7032, 0.6426, 0.6418, 0.6429, 0.6563, 0.6279, 0.2188, 0.8203, 0.4375, 0.6407, 0.6958, 0.5010, 0.5797, 0.7110, 0.1262, 0.5179, 0.7530, 0.8203, 0.8135, 0.7666, 0.5479, 0.5000, 0.6416, 0.1650, 0.5168, 0.6416, 0.7676, 0.7659, 0.6406, 0.5215, 0.6260, 0.6422, 0.5205, 0.7038, 0.6407, 0.8135, 0.2032, 0.7676, 0.8125, 0.8140, 0.7510, 0.7110, 0.6263, 0.5205, 0.8126, 0.5169, 0.6563, 0.7042, 0.6446, 0.7672, 0.6573, 0.8135, 0.7667, 0.7667, 0.5177, 0.6670, 0.1250, 0.3907, 0.6260, 0.6417, 0.6426, 0.6416, 0.6260, 0.7676, 0.4453, 0.6427, 0.5254, 0.7666, 0.3916, 0.7110, 0.7676, 0.7666, 0.7676, 0.5254, 0.8135, 0.6890, 0.7677
19. M. Mareš, Uprostřed Evropy? Vesmír 71, 1992/12.
2, 2, 29, 17, 6, 23, 12, 12, 13, 23, 12, 30, 17, 16, 44, 17, 8, 9, 4, 20, 13, 11, 5, 18, 20, 11, 35, 11, 17, 16, 30, 26, 19, 15 (n = 34)
0.5000, 0.5000, 0.3843, 0.8203, 0.6263, 0.6279, 0.0183, 0.4688, 0.6563, 0.6602, 0.8208, 0.6446, 0.8135, 0.6260, 0.7110, 0.6887, 0.2041, 0.7667, 0.1094, 0.5479, 0.7677, 0.7080, 0.7588, 0.7041, 0.6299, 0.6250, 0.7135, 0.8135, 0.8204, 0.6563, 0.7666, 0.6953, 0.7041, 0.0176
20. M. Mareš, Jak se dozvídáme, co si myslíme. Vesmír 71, 1992/12.
6, 2, 8, 16, 13, 17, 22, 22, 6, 3, 33, 6, 24, 18, 21, 33, 14, 14, 6, 10, 7, 15, 19, 3, 36, 26, 20, 10, 16, 24, 5, 8, 46, 17, 24, 4, 18, 23, 5, 20, 15, 12, 37, 14, 6, 12, 13, 6, 23, 12, 13, 22, 20, 21, 13, 16, 33, 13, 17, 13, 14, 12, 30, 5, 9, 15, 15, 15, 30, 20, 19, 27, 10, 17, 22, 15, 13, 30, 21, 17, 26, 7, 20, 10, 53, 25, 16, 35, 12, 12, 14, 11, 27, 10, 51, 28, 2, 25, 8, 20, 29, 21, 3, 23, 12, 17, 3, 21, 12, 11, 15, 16, 19, 15, 23, 28, 23, 7, 33, 19, 9, 20, 15, 10, 21 (n = 125)
0.8135, 0.5000, 0.7666, 0.7041, 0.6563, 0.7042, 0.5259, 0.2271, 0.6255, 0.7500, 0.8135, 0.5864, 0.8135, 0.5479, 0.7510, 0.6299, 0.6260, 0.7667, 0.6573, 0.6570, 0.6260, 0.3984, 0.7513, 0.6250, 0.7506, 0.6953, 0.6602, 0.5860, 0.7666, 0.1281, 0.8203, 0.5469, 0.5859, 0.8135, 0.8125, 0.6250,

0.5215,	0.5016,	0.7529,	0.5480,	0.6416,	0.8126,	0.7666,	0.7667,	0.5160,
0.8135,	0.7658,	0.5208,	0.5166,	0.7512,	0.6250,	0.8135,	0.1266,	0.7735,
0.7677,	0.5157,	0.6408,	0.8203,	0.6426,	0.4385,	0.7593,	0.6563,	0.7356,
0.4082,	0.7667,	0.7658,	0.6885,	0.7667,	0.1172,	0.7529,	0.6953,	0.7530,
0.7510,	0.8135,	0.7511,	0.6416,	0.8126,	0.6426,	0.7666,	0.6261,	0.7512,
0.8125,	0.5166,	0.7667,	0.5259,	0.7667,	0.8135,	0.3907,	0.1577,	0.6260,
0.6573,	0.7593,	0.7658,	0.6573,	0.6426,	0.7667,	0.5000,	0.8135,	0.7735,
0.8125,	0.8135,	0.6602,	0.6250,	0.2198,	0.6409,	0.7666,	0.7500,	0.3760,
0.6280,	0.7656,	0.7666,	0.5217,	0.1281,	0.4453,	0.6407,	0.6575,	0.6885,
0.3760,	0.1416,	0.8208,	0.6954,	0.7657,	0.7129,	0.7666,	0.5157	

7.2. Breaks in the sequence

The binary code is only one of the many possibilities of characterizing an aspect of the sentence structure. Since it is expressed quantitatively, the sequence of B_{rel}-values can be examined further. In this chapter we show one of the methods of finding breaks, i.e. the places where a significant jump in the sequence occurs. In order to find such a place, one must compare all neighbouring B_{rel}-values. Knowing the variance of B_{rel} we set up the testing criterion (normal distribution) in the following form

$$(7.10) \qquad u = \frac{B_{rel,1} - B_{rel,2}}{\sqrt{Var(B_{rel,1}) + Var(B_{rel,2})}}$$

$$= \frac{B_{rel,1} - B_{rel,2}}{\sqrt{\dfrac{4^{\frac{n_1(n_1-1)}{2}} - 1}{12\left(2^{\frac{n_1(n_1-1)}{2}} - 1\right)^2} + \dfrac{4^{\frac{n_2(n_2-1)}{2}} - 1}{12\left(2^{\frac{n_2(n_2-1)}{2}} - 1\right)^2}}}$$

where n_1 and n_2 are the respective sentence lengths measured in terms of number of words. In order to illustrate the computation let us compute the difference between the first and the second sentence in the Russian text 1. Here we have

$$n_1 = 40, \qquad n_2 = 10, \qquad B_{rel,1} = 0.6719, \qquad B_{rel,2} = 0.5011.$$

Inserting these values in formula (7.10), we obtain a complex expression which can be simplified if we take limits. Here we shall compute exactly, insert the given numbers in (7.10) and obtain

$$u = \frac{0.6719 - 0.5011}{\sqrt{\dfrac{4^{\frac{40(40-1)}{2}} - 1}{12\left(2^{\frac{40(40-1)}{2}} - 1\right)^2} + \dfrac{4^{\frac{10(10-1)}{2}} - 1}{12\left(2^{\frac{10(10-1)}{2}} - 1\right)^2}}} = 0.4184$$

which is not significant at the one-sided 95% level. Hence there is no significant difference between the binary codes of the first two sentences. It is true that as a matter of fact we compute the *t*-test and the quantile of *t* should be determined according to the number of degrees of freedom ($n_1 + n_2 - 2$), but one can rest content with the simpler case and decide that if $u > 1.64$ we have a significant jump downwards, i.e. to a simpler sentence structure; if $u < -1.64$ we have a significant jump to a more complex sentence structure.

In order to show the ***progressive syntactic dependence fragmentation*** of a text we show the computation using the Russian text 15 as presented in Table 7.4. As can be seen, there is only one significant jump downwards as shown in the forth column of Table 7.4; the rest of the differences are not significant (according to our criterion). Hence the text is not strongly syntactically fragmented. In order to express the fragmentation quantitatively, we simply establish the indicator

(7.11) $$F = \frac{D + U}{N - 1}$$

where *D* is the number of downward jumps, *U* the number of upward jumps and *N* is the number of sentences in text (*N*-1 is the number of subsequent differences). In the given text (Russian 15) we have $N = 26$, $D = 1$, $U = 0$, hence

F(Russian 15) = (1 + 0)/25 = 0.04.

Table 7.4
Progressive syntactic dependence fragmentation in Russian text 15

***n* words**	B_{rel}	***u***	**DOWN**	**UP**
5	0.5279	-0.4420	0	0
3	0.7143	-0.4126	0	0
4	0.8889	0.4526	0	0
7	0.7034	-0.1303	0	0
5	0.7566	1.2293	0	0
12	0.2546	-1.2135	0	0
18	0.7500	0.6124	0	0

22	0.5000	0.0000	0	0
15	0.5000	-0.6124	0	0
16	0.7500	1.6554	1	0
18	0.0742	-1.0525	0	0
25	0.5039	-0.6028	0	0
8	0.7500	-0.3062	0	0
7	0.875	0.0000	0	0
22	0.875	0.9186	0	0
18	0.5	-0.0024	0	0
10	0.501	-0.6099	0	0
19	0.75	0.9173	0	0
11	0.3755	-0.9173	0	0
9	0.75	0.6124	0	0
12	0.5	-0.0024	0	0
10	0.501	-0.3611	0	0
7	0.6484	1.3178	0	0
10	0.1104	-0.9614	0	0
9	0.5029	0.0066	0	0
12	0.5002	0.8664	0	0

We have chosen this simple expression of syntactic fragmentation because it represents a proportion (if we allow ourselves an idealised assumption that any two jumps are mutually independent and that all of them can occur with the same probability) lying in the <0,1> range and warranting an easy comparability of texts. This task will be performed in the next chapter.

The results of computations concerning Russian and Czech texts are summarized in Tables 7.5 and 7.6.

Table 7.5
Progressive syntactic dependence fragmentation of 20 Russian texts

Text	***D***	***U***	***N***	$F = \frac{D+U}{N-1}$
Russian 01	1	2	254	0.0119
Russian 02	2	0	229	0.0088
Russian 03	3	3	492	0.0122
Russian 04	1	2	480	0.0063
Russian 05	1	5	489	0.0123
Russian 06	2	3	481	0.0104
Russian 07	1	0	50	0.0204
Russian 08	2	1	86	0.0353

Russian 09	0	0	57	0.0000
Russian 10	1	0	47	0.0217
Russian 11	0	0	58	0.0000
Russian 12	0	0	48	0.0000
Russian 13	0	1	49	0.0208
Russian 14	0	0	42	0.0000
Russian 15	1	0	26	0.0400
Russian 16	2	2	64	0.0635
Russian 17	0	2	49	0.0417
Russian 18	1	0	38	0.0270
Russian 19	1	2	100	0.0303
Russian 20	0	0	36	0.0000

Table 7.6
Progressive syntactic dependence fragmentation of 20 Czech texts

Text	***D***	***U***	***N***	$F = \dfrac{D+U}{N-1}$
Czech 01	1	1	49	0.0417
Czech 02	0	0	31	0.0000
Czech 03	0	0	31	0.0000
Czech 04	0	1	28	0.0370
Czech 05	0	0	52	0.0000
Czech 06	0	0	29	0.0000
Czech 07	1	0	29	0.0357
Czech 08	0	0	41	0.0000
Czech 09	0	0	40	0.0000
Czech 10	0	0	24	0.0000
Czech 11	0	0	48	0.0000
Czech 12	1	1	86	0.0235
Czech 13	0	0	50	0.0000
Czech 14	1	1	78	0.0260
Czech 15	0	0	22	0.0000
Czech 16	0	0	28	0.0000
Czech 17	2	0	98	0.0206
Czech 18	0	0	132	0.0000
Czech 19	1	0	34	0.0303
Czech 20	1	2	125	0.0242

8. The binary code of text

8.1. The classical method

One can join the sentences of a text in the same way as one joins the individual words of a sentence on the basis of their grammatical or semantic associations. The sentences are joined on the basis of the occurrence of the same word or its synonym and by reference. The direction of association is irrelevant because in texts the reference or association is directed always backwards. However, there is a great discrepancy between the opinions about the existence of a reference.

A special direction in textology initiated by L. Hřebíček (1997, 2000) operates with supra-sentence units which have been called to his honour "hrebs". Hreb is an entity containing all sentences of the texts in which the same sign occurs or which contain some reference to one another. The concept of hreb can be extended to different subunits (morphemes, words, phrases, clauses,…). A sentence can belong simultaneously to several hrebs.

Here we dispense with the direct construction of herbs, which can be used for various characterisations of texts (cf. Ziegler, Altmann 2002) and restrict ourselves to the existence or non-existence of a referential, associative, repetitive etc. relation between two sentences. In order to exemplify the procedure we analyse the text "The vertical fields" by Fielding Dawson (1930-2002) http://www.classicshorts.com/stories/vrtclfld.html (accessed Dec. 20, 2009). Here we partition it in sentence-like sections considering the dot, the colon and sometimes also the semicolon as boundary signals. Since texts can be partitioned in different ways, we consider it as one of many possibilities. The individual sections are put in separate lines and the lines are numerated.

1. *On Christmas Eve around 1942, when I was a boy, after having the traditional punch and cookies and after having sung 'round the fire (my Aunty Mary at the piano), I, with my sister, my mother and my aunts, and Emma Jackman and her son, got into Emma Jackman's car and drove down Taylor Avenue to church for the midnight service:*
2. *I looked out the rear window at passing houses, doors adorned with holly wreaths, I looked into windows--catching glimpses of tinseled trees and men and women and children moving through rooms into my mind and memory forever;*
3. *the car slowed to the corner stop at Jefferson and the action seemed like a greater action, of Christmas in a cold damp Missouri night;*
4. *patches of snow lay on the ground and in the car the dark figures of my mother and sister and aunts talked around me and the car began to move along in an air of sky--at bottom dark and cold, seeming to transform the car, my face, and hands, pressed close to the glass as I saw my friends with their parents in*

their cars take the left turn onto Argonne Drive and look for a parking place near the church;

5. Emma Jackman followed, and I watched heavily coated figures make their exists, and move down the winter walk toward the jewel-like glittering church--up the steps into the full light of the doorway--fathers and sons and mothers and daughters I knew and understood them all, I gazed at them with blazing eyes:

6. light poured from open doors;

7. high arched stained glass windows cast downward slanting shafts of color across the cold churchyard, and the organ boomed inside while we parked and got out and walked along the sidewalk, I holding my mother's right arm, my sister held mother's left arm (mother letting us a little support her)--down the sidewalk to join others at the warmly good noisy familiar threshold:

8. spirits swirled up the steps into the church and Billy Berthold handed out the Christmas leaflets, I gripped mine.

9. I looked at the dominant blue illustration of Birth in white and yellow rays moving outward to form a circle around the Christ child's skull as Mary downward gazed; Joseph;

10. kneeling wisemen downward gazed;

11. I gazed down the long center aisle at the rising altar's dazzling cross and we moved down the aisle, slipped in front of Mr. and Mrs. Sloan and my buddy Lorry, Mr. and Mrs. Dart and my buddy Charles, Mr. and Mrs. Reid and my buddy Gene and his brother Ed--we then knelt away the conscious realization of our selves among music in the House of the Lord, I conscious of a voice that, slowly, coarsely, wandered--the I (eye) in see, hear me (I), we were on our feet singing, and the choir swept down the aisle, their familiar faces moving side to side as collective voices raised in anthem I held the hymnbook open and my mother and sister and I sang in celebration of God the crowded and brightly decorated--pine boughs and holly wreaths hung around the walls with candles high on each pew, I glanced at the gleaming cross--my spine arched, and far beyond the church, beyond the front door, beyond the land of the last sentence in James Joyce's _Dubliners_ a distant door seemed to open away beyond pungent green of pine gathered around rich red hollyberry clusters, red velvet, white-yellow center of candle flame, white of silk, gold of tassle, and gleaming glittering eternally cubistic gold cross and darkness of wooden beams powerfully sweeping upward--apex for the strange smoky penuma that so exhilarated me, I who smiled and reeled in a vast cold cold gaze down at myself listening to Charles Kean's Christian existentialist sermon in time before the plate was passed and the choir had singing, gone, and we were outside, I standing by my sister;

12. my mother and aunts were shaking Charles's hand, I shook that solid hand warmly, and I walked down the steps, my mother and sister and aunts again, again, once again it rushed through me taking my breath, my spine arched *toward trees and streets walking slowly breathing deep I moved down the sidewalk, eyes crystallizing streets yards houses and all lives within;*

13. my perception forked upward through treetops into the vertical fields of space, and a moment later, in the crowded back seat of the car, as Emma Jackman started the engine, I breathed vapor on the rear window, and with my finger, I signed my name.

The analysis can be programmed but not all repetitions are references or associations (e.g. prepositions, articles, conjunctions) and some words must be lemmatized in order to identify them (e.g. I, my, mine, me, we, our, us,…) or all of its synonyms must be stated (e.g. car, automobile, Cadillac,…). If one analyzes long or many texts, then both the partitioning of the text and the way of determining the associations of sentences must always be the same in order to be comparable. In modern texts this is usually no great problem, but in medieval German texts having no punctuation or in folklore in which the punctuation was made by a researcher *a posteriori*, it may turn out to be a problem. In poems one can choose the sentence or the verse as the frame of reference, even clauses in coordinated sentences. However, we must realize that any partitioning of a text into whatever segments is our conceptual construct. We do not discover but construct linguistic entities. Thus hrebs or Köhler´s motifs (Köhler 2006; Köhler, Naumann 2007, 2009; Mačutek 2009) are as well justified as any other traditional linguistic construct, e.g. word or phoneme. Our conceptual constructs are not true or false but prolific or not prolific. The most prolific concepts are those which allow us to formulate hypotheses or at least to devise means for expressing some text qualities and testing the differences. The situation is not different in other sciences. Even in physics we work with entities which were constructed conceptually and we try to find their correlates in the reality.

Table 8.1
Sentence associations in F. Dawson´s text

	2	3	4	5	6	7	8	9	10	11	12	13
1	1	1	1	1	0	1	1	1	0	1	1	1
2		0	1	1	0	1	1	1	0	1	1	1
3			1	0	0	1	1	0	0	1	0	1
4				1	0	1	1	1	0	1	1	1
5					1	1	1	1	1	1	1	1
6						0	0	0	0	1	0	0
7							1	1	0	1	1	1
8								1	0	1	1	1
9									1	1	1	1
10										1	0	0
11											1	1
12												1

Using pencil and paper and processing a text manually, the simplest way is first to prepare a frequency dictionary of words and check for each whether it associates two sentences in any possible way. Even programmed results must be thoroughly checked by hand, so this examination is very time consuming. We hope that in the future it will be possible to write satisfying programs.

For the above text we obtain the upper triangle matrix as presented in Table 8.1. The resulting text vector is

[1,1,1,1,0,1,1,1,0,1,1,1,0,1,1,0,1,1,1,0,1,1,1,1,0,0,1,1,0,0,1,0,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,0,0,0,1,0,0,1,1,0,1,1,1,1,0,1,1,1,1,1,1,1,1,0,0,1,1,1]

and the binary code of this text yields

$$BC = 273878107826428688842878 / 30223145490365729367654 = 0.9062,$$

saying that the text is very concentrated, because $BC \in \langle 0, 1\rangle$. It does not matter whether we call this property text concentration or cohesion – thereby we simply coin a concept which can be measured in different ways.

Considering the result we may ask whether short texts are always highly coherent and the deployment of the text reduces the concentration, or is it the property of the given concrete text. In order to solve this problem, many texts in many languages must be examined. It is a work with potential problems, intuitive decisions, trial and error, and can be performed only automatically.

However, it could help to give partial answers to questions like: Is there a fixed hierarchy of cohesion/concentration in genres? Is a fairy tale more concentrated in one language than in another, i.e. are there differences between languages in this respect? What is the status of scientific texts among which mathematical texts should have the strongest concentration? Etc.

8.2. Other methods

Since in very long texts, 2^n is beyond the capacity of many computers, the above method can be modified in different ways. The simplest way is to add the ones in the upper triangle matrix and divide the sum by the number of cases in order to obtain a simple proportion which can be processed statistically. Since there are $n(n-1)/2$ cases, we obtain a simple cohesion measure as

$$(8.1) \quad C_1 = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i<j}}^{n} a_{ij} \,,$$

where a_{ij} = 0,1. Here, instead of 2^n we simply placed 1. For example, in Table 8.1 there are n = 13 sentences; if two sentences form a pair, there are 13(12)/2 pairs. Hence 2/[13(12)] = 0.0128205. Since the number of ones is 56, we obtain

$$C_1 = 0.0128205(56) = 0.7179.$$

This cohesion indicator has the advantage of not emphasizing the place of greater or smaller cohesion. – as it is done by *BC* – and its use for testing is straightforward.

Needless to say, instead of 2^n other kinds of weighting could be used, say 1.1^n or any other number yielding moderate powers.

Since one can present the text vector in form of a binary sequence as shown above, different aspects of its properties can be captured using nonparametric statistical methods.

Looking at Table 7.5 we can easily establish a simple measure of *dependence fragmentation* of the text. The *F*-column presents a sequence interrupted by significant jumps up and down signalizing the change in dependence structure. A simple measure of *F* is given by the proportion of significant jumps ($J = D + U$) in any direction. Since between *N* neighbouring sentences there are *N*-1 possible jumps, in Chapter 7 we obtained the proportion

$$(8.2) \quad F = \frac{J}{N-1}.$$

The indicator $F \in <0,1>$ where 0 means a total smoothness or syntactic monotonousness of the text while 1 means a very agitated text. This fact can perhaps be used for distinguishing text sorts, for characterization of persons in a drama and for other literary purposes, but also, using the same text, for comparing the dependence structure in languages.

Since *F* is a simple proportion, the dependence structure of texts can easily be compared using the methods presented in previous chapters. The structure of the next sentence does not directly depend on that of the preceding one, hence – unless there is a special cause – the expectation of *F* is 0.5, because the probability of a jump or a smooth transition is equal, even if in the Russian and Czech texts we never obtained an *F* greater than 0.5. Hence $E(F) = 0.5$. The variance of *F* is $Var(F) = 0.5(0.5)/(N-1) = 0.25/(N-1)$ and the normal two-sided test for the difference of two texts can be approximated by computing

$$(8.3) \quad u = \frac{|F_1 - F_2|}{\sqrt{\frac{0.25}{N_1 - 1} + \frac{0.25}{N_2 - 1}}} = \frac{2|F_1 - F_2|}{\sqrt{\frac{1}{N_1 - 1} + \frac{1}{N_2 - 1}}}.$$

For example, the difference in dependence fragmentation of the Russian Text 1 and Text 2 yields (using Table 7.5)

$$u = \frac{2\,|\,0.0119 - 0.0088\,|}{\sqrt{\dfrac{1}{253} + \dfrac{1}{228}}} = 0.07\,,$$

which is not significant. As can easily be stated, neither Russian nor Czech texts respectively differ significantly from one another because all *F*-values are very low.

However, if we proceed using the empirical mean *F* for a given language (and not 0.5), then we must compute it anew after adding a further text. Hence other text sorts must be examined in order to state whether the given low values of *F* are characteristic just for the given texts.

Nevertheless, a comparison of text groups is possible using the same method as above. Let us have K_i (i = 1,2) texts in two groups, e.g. K_1 = 20 for Russian and K_2 = 20 for Czech. Let further J_1 be the number of significant jumps in group 1 and J_2 in group 2. For Russian texts in Table 7.5 we obtain the sum of the second and third column as J_1 = 19 + 23 = 42, for Czech texts in Table 7.6, J_2 = 8 + 6 = 14. The number of sentences is given as S_i = $\sum_{j=1}^{K_i}(N_j - 1) = \sum_{j=1}^{K_i} N_j - K_i$, yielding for Russian 3175-20 = 3155 and for Czech 1055-20 = 1035. For comparing the two groups we compute first the weighted means $\bar{p}_i$ yielding $\bar{p}_1$ = 42/3155 = 0.0133 for Russian and $\bar{p}_2$ = 14/1035 = 0.0135 for Czech. As a matter of fact, the means are almost identical, i.e. no test would be necessary, but for the sake of completeness we show at least the method. Adding the two groups, we compute the common expectation in form of a weighted mean of both groups, namely as

$$\text{(8.4)} \quad \hat{p} = \frac{J_1 + J_2}{S_1 + S_2}\,,$$

yielding $\hat{p}$ = (42 + 14)/(3155 + 1035) = 56/4190 = 0.0134. The test criterion yields

$$\text{(8.5)} \quad u = \frac{|\,\bar{p}_1 - \bar{p}_2\,|}{\sqrt{\hat{p}\hat{q}\left(\dfrac{1}{S_1} + \dfrac{1}{S_2}\right)}}\,.$$

Inserting the above numbers in (8.5) we obtain for the difference of Russian and Czech texts

$$u = \frac{|0.0133 - 0.0135|}{\sqrt{0.0134(0.9866)\left(\frac{1}{3155} + \frac{1}{1035}\right)}} = 0.0243,$$

signalizing that there is no significant difference between the dependence fragmentation in these two text groups. In order to examine the given indicators and tests in more detail one needs texts of different sorts. It can be expected that stage play texts and poetry would change this monotonous picture.

8.3. Using the binary code

Having performed the evaluation of stepwise retrospective dissimilarity, i.e. the comparison of individual chapters with chapter 1 (cf. e.g. Table 3.4), we can reorder the table according to τ. In this way each chapter (besides those with the lowest and highest τ) obtains two neighbours, as can be seen for Chamisso in Table 8.2. A transformation in similarities would not change the order.

Table 8.2
Stepwise dissimilarity in Chamisso

Chapter	τ
1	0.0000
5	0.0008
6	0.0012
10	0.0042
11	0.0078
8	0.0087
4	0.0097
2	0.0111
3	0.0182
7	0.0269
9	0.0681

Marking the direct neighbours of a chapter with 1, we can again obtain the upper triangular matrix from which the binary code of the text can be computed. The matrix corresponding to Table 8.2 is presented in Table 8.3.

Table 8.3
The similarity matrix of Chamisso's *Peter Schlemihl*

	1	2	3	4	5	6	7	8	9	10	11
1					1						
2			1	1							
3							1				
4								1			
5						1					
6										1	
7									1		
8											1
9											
10											1
11											

The computation of the *binary code of chapter (dis)similarities* will be performed in the usual way. Since the maximum BC_{max} would be attained if all chapters had the same τ, i.e. yielding (7.3), we would obtain for Chamisso

$$BC_{rel} = 20345381418175500/(2^{55} - 1) = 0.5647.$$

Analyzing all German texts, we obtain the results shown in Table 8.4. It is to be noted that the association of "end" chapters with "beginning" chapters can yield a zero BC_{rel}. (cf. Paul)

Table 8.4
The relative binary code of chapter (dis)similarities
in German texts (s. p. 29)

Text	**# of chapters (*n*)**	$\boldsymbol{BC_{rel}}$
Novalis	10	0.1487
Paul	55	0.0000
Chamisso	11	0.5647
Hoffmann	3	0.7143
Eichendorff	10	0.2900
Sealsfield	28	0.5010
Meyer	11	0.1251
Wedekind	4	0.7778
Löns	13	0.4375
Kafka	18	0.5625
Tucholsky	5	0.7840

The binary code used for this purpose is essentially a stylistic indicator showing the fragmentation of the text and jumps in word variability. The greater the binary code using retrospective dissimilarity, the smoother the transition from one chapter to the next. If the chapters have the same lexical structure, then all chapters are equal and each of them has n-1 equal neighbours, yielding $BC_{rel} = 1$. This is, of course, a rather improbable event.

Let us consider the binary code of the same text in all Slavic languages. The τ radians are presented in Table 3.23. Here we work with $n = 10$ chapters in each language. The result does not correspond exactly with the classification of Slavic languages: evidently the translators left their stylistic traces in the texts or the reduction to *BC* is not sensitive enough because we did not perform direct comparisons. The binary codes are presented in Table 8.5.

Table 8.5
The relative binary code of chapters of Ostrovskiy's "How The Steel Was Tempered" in 12 Slavic languages using
stepwise retrospective dissimilarity

Language	**# of chapters**	BC_{rel} **increasing**
Serbian	10	0.0090
Croatian	10	0.0090
Macedonian	10	0.0256
Russian	10	0.0347
Bulgarian	10	0.0725
Slovak	10	0.1409
Ukrainian	10	0.1875
Polish	10	0.2813
Slovenian	10	0.3125
Sorbian	10	0.3756
Czech	10	0.4063
Belorussian	10	0.4072

In order to obtain more sensitive results, we consider the *τ* radians of the stepwise retrospective dissimilarity as Cartesian components of a vector in a 10-dimensional space. All these vectors are presented in Table 3.23. Again, we compute the cosines of the angles between two vectors using formula (3.2). For example, we have the vectors

T(Russian) =
(0, 0.0158, 0.0154, 0.0297, 0.0221, 0.0342, 0.0333, 0.0388, 0.0322, 0.0293),

T(Macedonian) =
(0, 0.0177, 0.0033, 0.0367, 0.0247, 0.0075, 0.0051, 0.0062, 0.0270, 0.0111),

then the cosine of their angle is

$$\cos\tau(Russ, Mac) = \frac{[0(0) + 0.0158(0.0177) + ... + 0.0293(0.0111)}{\sqrt{0^2 + 0.0158^2 + ... + 0.0293^2}\sqrt{0^2 + 0.0177^2 + ... + 0.0111^2}}$$

$$= 0.7730$$

from which we obtain 0.6872 radians. In this way each of the 11 Slavic languages was compared with Russian; the results are presented in Table 8.6

Table 8.6
Tau radians for the comparison of retrospective dissimilarities of 11 Slavic languages with Russian based on Ostrovskij´s novel

Slavic language vs Russian	**cos τ decreasing**	**τ rad increasing**
Russian	1.0000	0.0000
Belorussian	0.9953	0.0970
Sorbian	0.9560	0.2977
Ukrainian	0.9496	0.3188
Polish	0.9279	0.3821
Czech	0.9263	0.3863
Slovak	0.9187	0.4060
Bulgarian	0.9014	0.4478
Slovenian	0.8923	0.4684
Croatian	0.8389	0.5755
Serbian	0.8323	0.5876
Macedonian	0.7730	0.6872

As can easily be seen, this kind of comparison places the Slavic languages at a more adequate distance from Russian, and clearly separates the South Slavic languages from the others.

The method can be used even if the two compared vectors do not have the same number of dimensions. The missing ones can be filled with zeroes. Looking, for example, at Table 3.4 and comparing Hoffmann with Tucholsky, we obtain the vectors

T(Hoffmann) = <0.0, 0.0264, 0.0348, 0.0, 0.0>,
T(Tucholsky) = <0.0, 0.0369, 0.0089, 0.0342, 0.0195>,

where we added two zeroes in T(Hoffmann). The tau angle between these vectors and that between the vectors (where we present an adapted “Tucholsky 2”)

T(Hoffmann) = <0.0, 0.0264, 0.0348, 0.0, 0.0>,
T(Tucholsky 2) = <0.0, 0.0369, 0.0089, 0.0, 0.0>,

is, however, the same. Consequently, according to this definition, if the two compared vectors do not have the same number of dimensions, the τ angle is determined by the space with fewer dimensions. Or, in other words, if we compare two texts each with a different number of chapters, the τ angle is fixed by the text with fewer chapters, while the additional chapters of the other text are disregarded. This drawback could be avoided by taking *the average tau angle* of all possible combinations such as

T(Hoffmann 1) = <0.0, 0.0264, 0.0348, 0.0, 0.0>,
T(Tucholsky) = <0.0, 0.0369, 0.0089, 0.0342, 0.0195>,

T(Hoffmann 2) = <0.0, 0.0, 0.0264, 0.0348, 0.0>,
T(Tucholsky) = <0.0, 0.0369, 0.0089, 0.0342, 0.0195>,

T(Hoffmann 3) = <0.0, 0.0, 0.0, 0.0264, 0.0348 >,
T(Tucholsky) = <0.0, 0.0369, 0.0089, 0.0342, 0.0195>.

The primary τ computed for the parts of the same text yields an image of dissimilarities in the deployment of the texts, while the secondary τ comparing different texts shows the dissimilarity of this process in two different texts. Since the dissimilarity is fully quantified here, the procedure can be used for long texts and their comparison in the same or different languages.

Mutatis mutandis, any set of properties of texts can be processed in this way.

9. Belza – Skorochod'ko chaining

Text cohesion is one of the rarely studied properties, as far as quantitative methods are concerned. The concept of cohesion can be defined and measured in various ways (cf. Köhler, Altmann 2009: 57). Co-reference is a basic concept in this context, mainly thought of in terms of anaphora. In this chapter, we will concentrate on the ideas proposed by Belza (1971) and propagated by Skorochod'ko (1981). Belza measures text cohesion using a simple measure, which he called the *chaining coefficient*. Co-reference should, in the most general sense, take into account every linguistic entity which is able to refer to an object or relation, i.e. all kinds of phrases including nominal, adverbial and verb phrases. To simplify measuring, Belza restricted his method to counting the number of adjacent sentences with co-referring elements, in fact to sentences which contain identical words or synonyms or matching pronouns. Although this is a brutal simplification, even a linguistically dubious one, we will use it here for our illustrative purposes because of its simplicity.

The applied method of obtaining references is therefore almost the same as in the previous chapter but here one counts only the number of sentences which are adjacent. A chain is an uninterrupted sequence of sentences joined by repetition of words, (quasi-)synonyms or pronouns; anaphoras referring to the complete sentence were omitted. In spite of these criteria different authors may define the chaining differently, and some decisions are always ad-hoc because it depends also on the interpretation. Skorochoďko (1981: 31) uses only repetitions of identical words as chaining elements, but in some literary texts it is rather a sign of stylistic incompetence.

Regardless of how coherence is measured, the sentences of a text must be identified and segmented. Determining the sentence boundaries is not always simple; there are many cases of ambiguities and different ways of interpretation. Punctuation does not always help because the marks are ambiguous: A dot can indicate the end of a sentence (full stop), show that a number is an ordinal one or that a character string is an abbreviation, or stand for a number of omitted characters, numbers, or words. Similar problems are found with other punctuation marks. The researcher will have to determine a procedure for sentence segmentation, taking into account that several punctuation marks such as parentheses and dashes can embrace complete sentences embedded in other ones etc. All these and other problems make automated sentence segmentation rather unreliable, although methods of computational linguistics are constantly improved, mainly with the help of statistical learning algorithms.

There are two ways of stating a chain: (1) The longest chains are taken into account and no other chain can begin within these chains, i.e. changing referents is not allowed; (2) all sequences of sentences with different referents are taken into account. We shall adhere to the latter method.

Chains of length 1 are possible. Let the length of a chain be k_i and the number of chains in the text c, then Belza´s chaining coefficient is defined as

$$C = \frac{1}{c}\sum_{i=1}^{c} k_i \,, \tag{9.1}$$

i.e. it is the mean length of chains in a text. Belza (1971) states that in Russian, the chaining coefficient of technical texts is C = 7.4, of popular scientific texts C = 6.6, and of fiction texts C = 5.3. That means, there is a hierarchy of texts which can be determined empirically. Since there is surely great dispersion in every text sort, the indicator can be used also for stylistic purposes.

For the sake of illustration we present a detailed analysis of a Czech text “O punkevním vodníku Jaroslavovi” by Pavel Bubla (http://palmknihy.cz/www/download.php?ID=7072), under the following conditions: the signs “.?!” are sentence boundaries; a complex sentence of whatever kind is one sentence; direct speech not an independent unit; chaining units are the same lemma, synonyms (but not hypernyms and hyponyms), referring pronouns; two sentences can form two chains if there are two different joining words. The chain forming words are underlined.

1. *V každé pořádné řece či rybníku žijí ryby, raci a všelijaká jiná žoužel.*
2. *V pohádkách pak ještě navíc <u>vodníci</u>.*
3. *V ponorné říčce Punkvě má revír i <u>vodník Jaroslav</u>.*
4. *Tak jako Filípek s Ondřejem hlídají jeskynní průvan a jeskyňky krápníky, tak <u>vodník Jaroslav</u> hlídá vodu a hospodaří s ní v jeskyních i na povrchu.*
5. *<u>Vodník Jaroslav</u> má však dvě velké slabosti.*
6. *Rád si pospí a ještě raději <u>hraje</u> karty se svým přítelem, lesním mužem Otou.*
7. *Obyčejně <u>hrají</u> lízaný mariáš o rybí šupinky.*
8. *Jak ti dva braši zasednou ke kartám, zapomenou na celý svět.*
9. *Tak se jednou přihodilo, že Jaroslav po dobrém obědě <u>usnul</u> na svém oblíbeném kamenném sedátku nad vodní hladinou Pohádkového jezírka v Punkevních jeskyních.*
10. *"To jsem si pěkně <u>zdříml</u>," liboval si, když se probudil.*
11. *Jenže vzápětí zanaříkal: "Achich, achich, já mám ale žízeň!"*
12. *Pískl na kropenatého pstruha, který právě plul kolem, a přikázal mu, aby přinesl vodu z Vilémovického potoka.*
13. *"Mám strašnou žízeň.*
14. *Šel bych si pro ni sám, ale <u>nemám</u> <u>čas</u>," říkal <u>Jaroslav</u> důležitě.*
15. *"<u>Nemáš</u> <u>čas</u>, protože <u>si</u> musíš přepočítat rybí <u>šupinky</u>, abys věděl, kolik <u>ti</u> <u>jich</u> zůstalo ze včerejška.*
16. *Však jsi <u>jich</u> s Otou prohrál celý kopec," ozvalo se zpod klenby jeskyně.*
17. *<u>Vodník</u> zamžoural do šera a podle hlasů poznal <u>Filípka</u> a Ondřeje, skřítky*

z Jezerní jeskyně.

18. "Jak víš, Filípku, že jsem včera prohrál vrchovatý kopec rybích šupinek?" ptal se Jaroslav.
19. "To je jednoduché.
20. Jak zasednete s Otou ke kartám, oba zapomenete na celý svět, nevidíte a neslyšíte, i kdyby se jeskyně bořily," odpověděl mu Filípek.
21. "Zapomeneme, zapomeneme," huhlal vodník, "mám mnoho starostí s vodou.
22. Je velké sucho, vody je málo a musím ji rozdělit tak, aby jí bylo všude dost," vymlouval se.
23. "To bys měl udělat ale hodně rychle, protože všichni v krasu mají žízeň, ne jenom ty," přidal se k Filípkovi Ondřej.
24. "Poslali jsme za tebou jeskynní průvan se vzkazem, ale ty jsi ho vůbec nevzal na vědomí.
25. Měl ti vyřídit, abys přišel k nám na poradu do Jezerní jeskyně, kterou svolává čarodějnice Dobromila, abychom se poradili, co uděláme proti suchu, které tolik sužuje kras.
26. Jak ale s Otou zasednete ke kartám, není s vámi řeč, nevidíte a neslyšíte," znovu vyčítal vodníkovi Filípek.
27. "Opravdu jste pro mne poslali jeskynní průvan se vzkazem?
28. Chtěl jsem oplatit Otovi prohru z minulé neděle, a ne a ne přijít karta, a tak se stalo, že jsem na všechno zapomněl," omlouval se zkroušeně skřítkům Jaroslav.
29. "Když se jeskynní průvan vrátil s nepořízenou, přišli jsme tedy za tebou sami.
30. Musíme se dohodnout, kam pošleme vodu.
31. Vody v našich jezírkách ubývá a na jeskyně je žalostný pohled, a nejenom na jeskyně," vysvětlovali skřítci vodníkovi.
32. "Co mám s vámi dělat?
33. A každou chvíli přijde Ota," povzdechl si vodník a chystal se se skřítky na cestu.
34. Jen to dořekl, ozvalo se za ohybem Punkvy: "Jaroslave, vstávej a připrav karty, hned budeme hrát a můžeš mi oplatit svou prohru ze včerejška."
35. Lesní muž Ota si již v duchu představoval další hromádku rybích šupinek, kterou vyhraje nad Jaroslavem.
36. Ale jakmile spatřil skřítky, poznal, že z karet nic nebude.
37. "Dobrý den, Oto," pozdravili skřítci.
38. Filípek, který vodníku Jaroslavovi dělal kázání, se obrátil k lesnímu muži: "Právě jsme mluvili o tom, že jste včera hráli s vodníkem karty a hádali jste se o každou rybí šupinu, až se jeskyně téměř otřásaly.
39. Měli jste oči jen pro karty a žádný jste neodpověděli jeskynnímu průvanu, který jsme poslali za Jaroslavem, aby poslal do jeskyní vodu.
40. Jeskyňky si stěžovaly, že krápníky nemají vodu, a Dobromila se zlobí

také."

41. *Filípkovo kázání se snažil Ota přerušit poznámkou: "Tak to byl jeskynní průvan, co mi rozházel karty," rozpomínal se.*
42. *Ondřej se vmísil do rozmluvy: "Letos je v krasu velké sucho, že to určitě pociťují i tvé stromy v lesích, viď, Oto?*
43. *Byl jsem u holuba Karla a viděl jsem, že je ve žlebech sucho, i tam, kde bývají skály vždy vlhké a mokré, jak je rok dlouhý."*
44. *"To víš, že mají žízeň!*
45. *Ale horší je to s malými stromky," souhlasil s Ondřejem Ota.*
46. *"Oto, Oto, místo abys požádal Jaroslava o pomoc, tak hrajete karty a ještě se u toho hádáte," plísnil jej znovu Filípek.*
47. *Ještě chtěl něco říci, když jej přerušil Vincek:*
48. *"Jaroslave, pověz nám, kam máme letět a co máme zařídit.*
49. *Je nejvyšší čas něco pro kras udělat!"*
50. *Vodník byl Vinckovi vděčný, že svým dotazem ukončil Filípkovo kázání.*
51. *Vyslal netopýry za macarátem Ričmundem se vzkazem, aby poslal vodu do Sloupských jeskyní a Pustého žlebu.*
52. *Rovněž má pustit vodu z Holštejna a z Ostrova do Suchého žlebu, který byl nyní suchý nejen svým pojmenováním.*
53. *Potom ať vyhledají v Holštejně nebo v Ostrově hejkala Janka a řeknou mu, aby shromáždil bludičky, hejkaly a víly u Horního můstku Macochy a tam počkali na Otu.*
54. *A až to všechno vyřídí, mají se vrátit do Punkevních jeskyní.*
55. *Vodník Jaroslav pak poslal své podřízené kropenaté pstruhy proti proudu Punkvy připravit vodě cestu.*
56. *Také Ota se rozloučil a pospíchal za hejkaly, bludičkami a vílami, aby napojili stromy ve žlebech a dali napít i modrým zvonkům a jahodám rostoucím v jejich stínu.*
57. *Voda přišla i do jeskyní, svlažila vzduch a jeskynní pára pomalu vyplňovala jejich prostory.*
58. *Když se netopýři vrátili do Punkevních jeskyní, poděkovali skřítci vodníku Jaroslavovi za pomoc žíznivému krasu.*
59. *Pak Vincek s Franckem vzali skřítky na záda, vypískli vodníkovi pozdrav na rozloučenou a ztratili se Jaroslavovi ve tmě a jeskynní páře.*

The sentences forming chains are presented in Table 9.1. There are 59 sentences and the sum of lengths in Table 9.1 is 83. There are some chains of length 1, namely sentences 1,8,11,12,13,19,44,48,49,50, hence we obtain $C = (10 + 83)/59 = 1.5763$.

Table 9.1
Chains of length > 1 in the Czech text
P. Bubla: "O punkevním vodníku Jaroslavovi "

sentences	**subject**	**length**
2-5	vodníci, vodník Jaroslav	4
6-7	hraje, hrají	2
9-10	usnul, zdříml	2
14-15	nemám, nemáš	2
14-15	čas	2
15-16	šupinky, jich	2
14-18	Jaroslav, si, ti, vodník	5
17-18	Filípka, Filípku	2
20-21	zapomenete, zapomeneme	2
21-22	vodou, vody, ji, jí	2
23-29	ty, ti, vodníkovi, mne, Jaroslav, tebou	7
28-29	skřítkům, sami	2
28-29	přijít, přišli	2
30-31	vodu, vody	2
31-33	skřítci, vámi, skřítky	3
33-35	vodník, Jaroslave, Jaroslavem	3
33-35	Ota, mi	3
34-35	hrát, vyhraje	2
36-37	skřítky, skřítci	2
37-38	Oto, lesnímu muži	2
38-39	karty	2
38-39	vodníku Jaroslavovi,vodníkem, Jaroslavem	2
38-39	jeskyně, jeskyní	2
39-40	vodu	2
41-42	Ota, Oto	2
42-43	sucho	2
45-46	Ota, Oto, jej	2
46-47	Filípek, jej	2
51-52	vodu	2
51-52	žlebu	2
52-53	Holštejna, Holštějně	2
52-53	Ostrova, Ostrově	2
57-58	jeskyní	2
58-59	skřítci, skřítky	2
58-59	vodníku, Jaroslavovi, vodníkovi	2

The results of analysis of some further Czech texts are presented in Table 9.2.

Table 9.2
Belza-coefficient of 6 Czech fairy tales
(all texts accessed March 25, 2010)

No	Text	k_i	C
1	.P. Bubla: O punkevním vodníku Jaroslavovi, published 22.2.2007, (http://palmknihy.cz/www/download.php?ID=7072)	1,4,2,1,2,1,1,1,2,2,2, 5,2, 1,2,2,7,2,2, 2,3,3,3,2,2,2, 2,2,2, 2,2,2,1,2,2,1,1,1,2, 2,2,2,2,2,2, (c = 45) (n = 59)	1.5763
2	P. Bubla: O pohádkových jeskyních, published 31.7.2007, (http://palmknihy.cz/www/download.php?ID=7492)	4,2,2,2,1,3,5,2,2,2,2,2,2, 8,2, 2,2,2,1, 1,1,1,4,2,2,1, 2,2,1,2,2,2,1,1,1,8,5,3, 1, 1,1,1,2,4,2,2,2,2,2, 2,2,1, 2,3,2,1.2,2,2 (c = 59) (n = 68)	1.8971
3	P. Bubla: O Alenčiných jmeninách, published 7.2.2007, (http://palmknihy.cz/www/download.php?ID=7028)	2,2,1,5,3,6,2,2,2,2,1,2,1, 2,1, 1,1,2,1,1,4,4,2,2,2,2, 1,1,3,3,1,1,1,2,2,2,2,2,4,1, 2,2,1,1,2,1,1,2,2,1,1,2,1,2 (c = 54) (n = 71)	1.4507
4	P. Kováč: O hoře, published 7.11.2009, (http://www.firesnake.eu/pohadky/o_hore.htm)	1,1,1,2,1,2,1,1,1,1,1,2,1, 2,1, 2,1,1,2,1,1,2,1,1,1,3, 1,2,2, 2, 1,1,1,1,1,2,2,1,1, 2,1,1,1,1, 1, 2,2,1,1,2,3,2, 1,1,1,1,1,2, 2, 2,1,2,1,2,1,2,1,1,1,1,1, 3,3,3 (c = 74) (n = 94)	1.1383
5	. P. Kováč: O myčce Bošce, published 23.11.2009, (http://www.firesnake.eu/pohadky/o_bosce.htm)	1,1,2,2,1,1,2,2,2,1,1,1,1,1,1, 1,1,1,1,1,1,1,1,1,2,2,1,2,2, 1, 2,1,1,1,1,1,2,1,1,1,1,1,1, 1,2, 2,1,2,2,1,2,1,1,1 (c = 54) (n = 61)	1.1475
6	P. Kováč: O závorách, published 7.11.2009, (http://www.firesnake.eu/pohadky/o_zavorach.htm)	1,1,1,1,1,1,1,1,1,2,2,2,1,1,2, 1,1,1,1,1,1,1,1,2,1,2,1,1,1,1, 1,1,1,1,1,1,1 (c = 37) (n = 39)	1.1026

The results of chaining analysis of German press texts is presented in Table 9.3.

Table 9.3
Chaining in 10 German press texts

No	Text	k_i	*C*
1	Jens Heitmann: Messe will ILA nach Hannover holen. ET, 13.2.10, p. 5.	2,4,3,4,2,1,1,2,2,2,1,1,1,1 (c =14, n = 27)	1.93
2	Florian Oel: Kommt das Ende der Telefondose? ET, 13.2.10, p. 5	3,2,1,1,1,2,3,2,1,2,2 (c = 11,n = 20)	1.82
3	Politiker kochen eigenes Süpp-chen. ET, 13.2.10, p. 7	1,2,3,1,2,2 (c = 6, n = 11)	1.83
4	Hanne-Dore Schumacher: Bio-technologie legt weiter zu. ET, 13.2.10, p. 7	2,2,3,2,2,3,2,2,2,1,1 (c = 11, n = 22)	2.00
5	Hanne-Dore Schumacher: Gellert setzt auf Göttinger Einzelhändler. ET, 13.2.10, p. 7	6,3,1,1,1,2,2,1,2,2,2,4,3,1 (c = 14, n = 31)	2.21
6	Die Welt im Griff des Chaos-wetters. ET, 13.2.10, p. 8	1,1,1,3,2,2,2,2,1,3,2 (c = 11, n = 20)	1.82
7	Köhler heißt jetzt Schröder. ET, 13.2.10, p. 8	3,3,1,3,2,2,3,5,2,2,2 (c = 11, n = 28)	2.55
8	Klaus Wallbaum: Land verliert Steuereinnahmen. ET, 13.2.10, p. 1	2,1,1,2,2,1,1,5,2,1,1,2,2 (c = 13, n = 23)	1.77
9	Klaus von der Brelie: Afghanistan will mehr Hilfe bei Polizeiausbil-dung. ET, 13.2.10, p. 1.	2,3,2,3,3,1,2,2,3,1 (c = 10, n = 22)	2.20
10	Thomas Borchert: Der Weg ist frei für russischen Gas. ET, 13.2.10, p. 2	1,3,1,1,3,2,2,1,2,2,1 (c = 11, n = 19)	1.73

For the sake of comparison, in Table 9.4 we present results of nine Slovak press texts from the online journals “SME” and “Plus 7 dní” analyzed on the basis of the same criteria.

A purely visual comparison shows that German texts do not have such a dispersion as Slovak ones, and the Czech fairy tales have a very low coefficient, but the volume of data is preliminarily very small to make more conclusive judgments. Here we shall rather develop a testing procedure.

Table 9.4
Belza-coefficient of 9 Slovak press texts

No	Text	k_i	C
1	http://www.sme.sk/c/5167031/sviatocny-pocit-nestoji-na-tom-ze-mame-trojmetrovu-kopu-darcekov.html (accessed December 24, 2009)	4,2,3,4,2,1,2,1,1,4,4,2,2,2 (n = 30, c = 14)	2.4286
2	http://www.sme.sk/c/5166986/medvedev-nase-vztahy-s-putinom-sa-nezmenili-a-ani-nezmenia.html (accessed December 24, 2009)	2,12,2,3 (n = 14, c = 4)	4.7500
3	http://www.sme.sk/c/5166799/brazilski-lekari-odstranili-chlapcovi-z-tela-14-ihiel.html (accessed December 24, 2009)	2,2,3,3,1,4,1,1,4,1,1,2,1 (n = 26, c = 13)	2.0000
4	http://komentare.sme.sk/c/5162547/odrocena-katarzia.html (accessed December 27, 2009)	2,2,2,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,1,1 (n = 24, c = 21)	1.2381
5	http://komentare.sme.sk/c/5156029/utrpenie-profesionalnych-wertherov.html (accessed December 27, 2009)	2,1,1,1,1,1,1,1,1,3,2,2,1,1,1,2 (n = 20, c = 16)	1.3750
6	http://komentare.sme.sk/c/5162593/kodanska-zmena-klimy-zlocin-a-trest.html (accessed December. 27, 2009)	1,2,1,2,2,1,2,2,2,2,1,1,2,1 (n = 22, c = 14)	1.5714
7	http://komentare.sme.sk/c/5159383/jazykovy-zakon-je-nielen-zbytocny-ale-aj-skodlivy.html (accessed December 28, 2009)	5,7,3,5,8,2,3,1,4 (n = 38, c = 9)	4.2222
8	http://www.plus7dni.sk/plus7dni/historia/akcia-david.html (accessed December 27, 2009)	2,4,3,3,2,2,2,3,2,2,1,2,1,5, 2,1,1,1,4,3 (n = 39, c = 20)	4.2300
9	http://trencin.sme.sk/c/5169766/babkoherec-tvu-sa-v-meste-venuju-devatdesiat-rokov.html (accessed December 28, 2009)	1,1,1,1,1,1,1,3,2,2,2,2,2,1,1,2,2,3,1,1,1,2,1 (n = 31, c = 23)	1.5200

Since one does not know the behaviour of C, we shall normalize it and develop a test for differences. Here, we must distinguish two cases as shown above: either the chains are not intersecting/embedded or they may intersect. In both cases we simply normalize the indicator C as follows

(9.2) $$C_{rel} = \frac{C - C_{\min}}{C_{\max} - C_{\min}}.$$

The minimal chaining is in both cases equal to 1, i.e. there is no chaining and all sentences form a separate chain, which means $C_{min} = 1$. However, C_{max} is different for the two cases. If no intersection or embedding is possible, then the longest chain is n, the number of sentences in text, i.e., all sentences form one single chain. If a chain can begin within another chain, then the longest chain contains n sentences, the second longest chain n-1 sentences, etc., hence the maximal number of chains is $n + (n-1) + (n-2) + \ldots + 3+2+1 = n(n+1)/2$.

Thus in case of *no intersections,* we obtain the normalized indicator

(9.3) $$C_{1,rel} = \frac{C-1}{n-1},$$

and in case of *possible embedding*

(9.4) $$C_{2,rel} = \frac{C-1}{\frac{n(n+1)}{2} - 1} = \frac{2(C-1)}{n(n+1)-2}.$$

Both indicators lie in the interval <0,1> but usually they are very small. One could multiply them with a constant but this would be only a visual adaptation.

For Table 9.4 where we allow embedding we obtain the results presented in Table 9.5

Table 9.5
The indicator $C_{2,rel}$ for Slovak texts (in Table 9.4)

Text No	***n***	***c***	$\boldsymbol{C_{2,rel}}$
1	30	14	0.1020
2	14	4	0.6250
3	26	13	0.0833
4	24	21	0.0216
5	20	16	0.0417
6	22	14	0.0571
7	38	9	0.1790
8	39	20	0.1746
9	31	23	0.0359

The variance of the individual values of a text is given as s^2 as usual. The variance of the mean (i.e. of C) is s^2/c where c is the number of chains and the variance of $C_{1,rel}$ is

(9.5) $$Var\left(C_{1,rel}\right) = \frac{s_x^2}{c(n-1)^2}$$

while the variance of $C_{2,rel}$ is given as

(9.6) $$Var\left(C_{2,rel}\right) = \frac{4s_x^2}{c[n(n+1)-2]^2}.$$

An asymptotic test can be set up as usually in form

(9.7) $$u = \frac{C_{1,rel}(Text 1) - C_{1,rel}(Text 2)}{\sqrt{Var(C_{1,rel}(Text 1)) + Var(C_{1,rel}(Text 2))}}$$

and analogously for $C_{2,rel}$.

For the sake of illustration let us compare the difference in $C_{2,rel}$ of the first two Slovak texts. First we obtain the simple variances s^2(Text 1) = 1.3407, s^2(Text 2) = 23.5833. Collecting all numbers and inserting them in (9.7) we obtain

$$u = \frac{|0.1020 - 0.6250|}{\sqrt{\frac{4(1.3407)}{14[30(31)-2]^2} + \frac{4(23.5833)}{4[14(15)-2]^2}}} = 22.39$$

signalizing a highly significant difference. The chaining in the two texts is very different.

In German press texts we obtain $C_{2,rel}$ as presented in Table 9.6.

Table 9.6
$C_{2,rel}$ in German press texts

Text No	***n***	***c***	$\boldsymbol{C_{2,rel}}$
1	27	14	0.24
2	20	11	0.11
3	11	6	0.12
4	22	11	0.33
5	31	14	0.54
6	20	11	0.11
7	28	11	1.00
8	23	13	0.05
9	22	10	0.57
10	19	11	0.00

The results in Czech fairy tales are shown in Table 9.7.

Table 9.7
$C_{2,rel}$ in Czech fairy tales

Text No	***n***	***c***	$C_{2,rel}$
1	59	45	1.38
2	68	59	1.00
3	71	54	0.44
4	94	74	0.05
5	61	54	0.06
6	39	37	0.00

Besides the means signalizing the strength of chaining, the sequence of chains can be considered a time series and at the same time it can be partitioned in Köhlerian motifs. For example, the first text in Table 9.1 contains the following motifs: 4, 2-3-4, 2, 1-2, 1-1-4-4, 2-2-2. The Belza-motifs signalize increasing concentration of content. Motifs are abstractions of second order. A third order abstraction can be set up using the sequence of motif lengths, etc. The relevance of these abstractions must be studied on very extensive set of texts up to the order which is identical with a sequence of random numbers.

However, chaining is not restricted to identical words, synonyms or references; it can be generalized to any kind of entities. One can study the chaining of phoneme groups, syllables, morphs, assonances, alliterations, speech acts, phrases, semantic word groups, etc. One automatically arrives at the problem of perseveration, Skinner´s formal strengthening, inertia of neuron firing, extinction of verbal stimulus and other aspects of speech or text which represent some latent psychological or neuronal mechanisms (cf. Möller, Laux, Deister 2009).

10. Conclusions

The study of texts is an infinite enterprise. The number of aspects increases almost yearly; one discovers continuously new vistas. The situation is complicated by the fact that all aspects are associated with one another and build up a labyrinth for which we do not have a guide. The beginning has been accomplished but not much more than first, tentative steps have been taken so far; textology develops in so many directions that one can only examine a small number of issues. Either one presents new methods or uses a specific method and examines a limited set of texts in some languages in order to obtain more reliable results.

In the present book we restricted ourselves to the evaluation of some data from the surface of texts. We concentrated on vectors, codes and chains, tried to characterize texts and showed some methods of evaluation and testing. Each of the aspects chosen can be further developed; we leave it to colleagues who can choose one of the directions, extend its scope and integrate it in a more complex theory.

However, the future of this research will perhaps be as complex as that of natural sciences. On the one hand, specialization will be ever more rigorous, and completely new textological disciplines will arise; on the other, the trend for unification will be ever more urgently requested. The time will perhaps come in which an evolutionary superstructure will furnish explanatory theories in three senses: First, the phylogenetic evolution of language will give us arguments for the finding and foundation of causes and mechanisms which are not text-inherent any more; second, the ontogenesis in language acquisition may help us to understand the subconscious rise of mechanisms which are not innate but learned by repetition, and third, the dynamics of text growth or deployment unveiling some mysteries of phylogeny and ontogeny. Since texts represent facts traded through millennia in different languages, genres and forms, they represent the surface through which we shall try to enter deeper levels not only of language itself but also those of the communicating persons. Our aim is to study both the subconscious regularities (mechanisms) arising at certain stages of phylogeny, ontogeny or text and the conscious/learnable ones which are necessary for mastering a language. Needless to say, the latter domain is the traditional linguistics encompassing a lot of classifications, descriptions and standardizations. Dynamic text analysis pursued in this book is rather a battle with probabilities, processes, dependencies and functions, which is not possible without at least elementary mathematics. One uses it not as an end in itself but as a means for measurement, characterization, testing and inference.

Appendix I. Texts used

Bulgarian (private letters) (in Table 3.3.4)

B 01 Boris 2 (Letter)
B 02 Ceneva1 (Letter)
B 03 Ceneva 2 (Letter)
B 04 Janko1 (Letter)
B 05 Janko 3 (Letter)

Czech (short stories by Bohumil Hrabal) (in Table 3.34)

Cz 01 Hrabal 310: Expozé panu ministru informací (Jarmilka, 44–47)
Cz 02 Hrabal 315: Lednová povídka (Jarmilka, 58–61)
Cz 03 Hrabal 316: Únorová povídka (Jarmilka, 62–69)
Cz 04 Hrabal 319: Blitzkrieg (Jarmilka, 86–87)
Cz 05 Hrabal 323: Protokol (Jarmilka, 129–131)

English (taken from http://nobelprize.org/nobel_prizes/lists/all/) **(**in Table 2.6, 3.34)

E 01: Jimmy Carter,	Nobel lecture (Peace 2002)
E 02: Toni Morrison,	Nobel lecture (Literature 1993)
E 03: George C. Marshall,	Nobel lecture (Peace 1953)
E 04: James M. Buchanan Jr.,	Nobel lecture (Economy 1986)
E 05: Saul Bellow,	Nobel lecture (Literature 1976)
E 07: Sinclair Lewis,	Nobel lecture (Literature 1930)
E 08: Ernest Rutherford,	Nobel lecture (Chemistry 1908)
E 13: Richard P. Feynman**,**	Nobel lecture (Physics 1965)

German (in Table 2.1, 3.34)

Arnim 01	Der tolle Invalide auf dem Fort Ratonneau
Arnim 02	Des ersten Bergmanns ewige Jugend
Arnim 03	Frau von Saverne
Busch 01	Eduards Traum
Chamisso 01-11	Peter Schlemihls wundersame Geschichte I-XI
Droste 01	Die Judenbuche
Droste 02	Der Tod des Erzbischofs Engelbert
Droste 03	Das Fegefeuer

Droste 04	Der Fundator
Droste 05	Die Schwestern
Droste 08	Der Geierpfiff
Eichendorff 01-10	Aus dem Leben eines Taugenichts 1-10
Goethe 01	Die neue Melusine
Goethe 05	Der Gott und die Bajadere
Goethe 09	Elegie 19
Goethe 10	Elegie 13
Goethe 11	Elegie 15
Goethe 12	Elegie 2
Goethe 14	Elegie 5
Goethe 17	Der Erlkönig
Heine 01	Die Harzreise
Heine 02	Die Heimkehr - Götterdämmerung
Heine 03	Die Heimkehr - Die Wallfahrt nach Kevlaar
Heine 04	Ideen. Das Buch Le Grand
Heine 07	Belsazar
Hoffmann 01-03	Der Sandmann
Immermann 01	Der Karneval und die Somnambule
Kafka 01	In der Strafkolonie
Kafka 02	Ein Bericht für eine Akademie
Kafka 03	Betrachtung - Kinder auf der Landstraße
Kafka 04	Betrachtung - Entlarvung eines Bauernfängers
Kafka 05	Betrachtung - Der plötzliche Spaziergang
Kafka 06	Betrachtung - Entschlüsse
Kafka 07	Betrachtung - Der Ausflug ins Gebirge
Kafka 08	Betrachtung - Das Unglück des Junggesellen
Kafka 09	Betrachtung - Der Kaufmann
Kafka 10	Betrachtung - Zerstreutes Hinausschaun
Kafka 11	Betrachtung - Der Nachhauseweg
Kafka 12	Betrachtung - Die Vorüberlaufenden
Kafka 13	Betrachtung - Der Fahrgast
Kafka 14	Betrachtung - Kleider
Kafka 15	Betrachtung - Die Abweisung
Kafka 16	Betrachtung - Zum Nachdenken für Herrenreiter
Kafka 17	Betrachtung - Das Gassenfenster
Kafka 18	Betrachtung - Wunsch, Indianer zu werden
Kafka 19	Betrachtung - Die Bäume
Kafka 20	Betrachtung - Unglücklichsein
Kafka 21	Ein Brudermord
Kafka 22	Ein Landarzt
Kafka 23	Der Geier
Kafka 24	Vor dem Gesetz
Kafka 25	Ein Hungerkünstler

Kafka 26	Nachts
Kafka 27	Das Schweigen der Sirenen
Kafka 28	Die Sorge des Hausvaters
Keller 01	Romeo und Julia auf dem Dorfe
Keller 02	Vom Fichtenbaum
Keller 03	Spiegel, das Kätzchen
Keller 04	Das Tanzlegendchen
Lessing 01	Der Besitzer des Bogens
Lessing 02	Die Erscheinung
Lessing 03	Der Esel mit dem Löwen
Lessing 04	Der Fuchs
Lessing 05	Die Furien
Lessing 06	Jupiter und das Schaf
Lessing 07	Der Knabe und die Schlange
Lessing 08	Minerva
Lessing 09	Der Rangstreit der Tiere
Lessing 10	Zeus und das Pferd
Löns 01-13	Der Werwolf 1-13
Meyer 01-11	Der Schuss von der Kanzel 1-11
Novalis 01-09	Heinrich von Ofterdingen - Die Erwartung 1-9
Novalis 11	Hyazinth und Rosenblütchen
Novalis 12	Neue Fragmente - Sophie
Novalis 13	Neue Fragmente - Traktat vom Licht
Paul 01	Dr. Katzenbergers Badereise 1.
Paul 02	Dr. Katzenbergers Badereise 2. Reisezwecke
Paul 03	Dr. Katzenbergers Badereise 3. Ein Reisegefaehrte
Paul 04	Dr. Katzenbergers Badereise 4. Bona
Paul 05	Dr. Katzenbergers Badereise 5. Herr von Niess
Paul 06	Dr. Katzenbergers Badereise 6. Fortsetzung der Abreise
Paul 07	Dr. Katzenbergers Badereise 7. Fortgesetzte Fortsetzung der Abreise
Paul 08	Dr. Katzenbergers Badereise 8. Beschluss der Abreise
Paul 09	Dr. Katzenbergers Badereise 9. Halbtagfahrt nach St. Wolfgang
Paul 10	Dr. Katzenbergers Badereise 10. Mittags-Abenteuer
Paul 11	Dr. Katzenbergers Badereise 11. Wagen-Sieste
Paul 12	Dr. Katzenbergers Badereise 12. die Avantuere
Paul 13	Dr. Katzenbergers Badereise 13. Theodas ersten Tages Buch
Paul 14	Dr. Katzenbergers Badereise 14. Missgeburten-Adel
Paul 15	Dr. Katzenbergers Badereise 15. Hasenkrieg
Paul 16	Dr. Katzenbergers Badereise 16. Ankunft-Sitzung
Paul 17	Dr. Katzenbergers Badereise I. Huldigungpredigt
Paul 18	Dr. Katzenbergers Badereise II. Ueber Hebels alemannische Gedichte

Paul 19	Dr. Katzenbergers Badereise III. Rat zu urdeutschen Taufnamen
Paul 20	Dr. Katzenbergers Badereise IIII. Dr. Fenks Leichenrede
Paul 21	Dr. Katzenbergers Badereise V. Ueber den Tod nach dem Tode
Paul 22	Dr. Katzenbergers Badereise 17. Blosse Station
Paul 23	Dr. Katzenbergers Badereise 18. Maennikes Seegefecht
Paul 24	Dr. Katzenbergers Badereise 19. Mondbelustigungen
Paul 25	Dr. Katzenbergers Badereise 20. Zweiten Tages Buch
Paul 26	Dr. Katzenbergers Badereise 21. Hemmrad der Ankunft im Badeorte
Paul 27	Dr. Katzenbergers Badereise 22. Niessiana
Paul 28	Dr. Katzenbergers Badereise 23. Ein Brief
Paul 29	Dr. Katzenbergers Badereise 24. Mittagtischreden
Paul 30	Dr. Katzenbergers Badereise 25. Musikalisches Deklamatorium
Paul 31	Dr. Katzenbergers Badereise 26. Neuer Gastrollenspieler
Paul 32	Dr. Katzenbergers Badereise 27. Nachtrag
Paul 33	Dr. Katzenbergers Badereise 28. Darum
Paul 35	Dr. Katzenbergers Badereise 30. Tischgebet und Suppe
Paul 36	Dr. Katzenbergers Badereise 31. Aufdeckung und Sternbedeckung
Paul 37	Dr. Katzenbergers Badereise 32. Erkennszene
Paul 38	Dr. Katzenbergers Badereise 33. Abendtisch-Reden über Schauspiele
Paul 39	Dr. Katzenbergers Badereise 34. Brunnen-Beaengstigungen
Paul 40	Dr. Katzenbergers Badereise 35. Theodas Brief an Bona
Paul 41	Dr. Katzenbergers Badereise 36. Herzens-Interim
Paul 42	Dr. Katzenbergers Badereise 37. Neue Mitarbeiter an allem
Paul 43	Dr. Katzenbergers Badereise I. Die Kunst, einzuschlafen
Paul 44	Dr. Katzenbergers Badereise II. Das Glueck
Paul 45	Dr. Katzenbergers Badereise III. Die Vernichtung
Paul 46	Dr. Katzenbergers Badereise 38. Wie Katzenberger …
Paul 47	Dr. Katzenbergers Badereise 39. Doktors Hoehlen-Besuch
Paul 48	Dr. Katzenbergers Badereise 40. Theodas Hoehlen-Besuch
Paul 49	Dr. Katzenbergers Badereise 41. Drei Abreisen
Paul 50	Dr. Katzenbergers Badereise 42. Theodas kuerzeste Nacht der Reise
Paul 51	Dr. Katzenbergers Badereise 43. Praeliminar-Frieden ...
Paul 52	Dr. Katzenbergers Badereise 44. Die Stuben-Treffen
Paul 53	Dr. Katzenbergers Badereise 45. Ende der Reisen und Noeten
Paul 54	Dr. Katzenbergers Badereise I. Wuensche fuer Luthers Denkmal
Paul 55	Dr. Katzenbergers Badereise II. Ueber Charlotte Corday

Paul 56	Dr. Katzenbergers Badereise III. Polymeter
Pseudonym 01	Eine kleine Geschichte mit der Zeit
Pseudonym 02	Taumelnde Realitaet
Raabe 01	Im Siegeskranze
Raabe 02	Eine Silvester-Stimmung
Raabe 03	Ein Besuch
Raabe 04	Deutscher Mondschein
Raabe 05	Theklas Erbschaft
Rieder 01	Liebe Mutter
Rieder 02	Brief an einen Toten
Rückert 01	Barbarossa
Rückert 02	Amor ein Besenbinder
Rückert 03	Der Frost
Rückert 04	Die goldne Hochzeit
Rückert 05	Erscheinung der Schnitterengel
Schnitzler 01	Der Sohn
Schnitzler 02	Albine
Schnitzler 03	Amerika
Schnitzler 04	Der Andere
Schnitzler 05	Die Braut
Schnitzler 06	Erbschaft
Schnitzler 07	Die Frau des Weisen
Schnitzler 08	Der Fürst ist im Hause
Schnitzler 09	Das Schicksal
Schnitzler 10	Welch eine Melodie
Schnitzler 11	Frühlingsnacht im Seziersaal
Schnitzler 12	Die Toten schweigen
Schnitzler 13	Er wartet auf den vazierenden Gott
Schnitzler 14	Mein Freund Ypsilon
Sealsfield 01	Das Cajuetenbuch - Die Praerie am Jacinto
Sealsfield 02-17	Das Cajuetenbuch 1-16
Sealsfield 18	Das Cajuetenbuch - Der Fluch Kishogues
Sealsfield 19	Das Cajuetenbuch - Der Kapitaen
Sealsfield 20	Das Cajuetenbuch - Callao 1825
Sealsfield 21	Das Cajuetenbuch - Havanna 1816
Sealsfield 22	Das Cajuetenbuch - Sehr Seltsam!
Sealsfield 23	Das Cajuetenbuch - Ein Morgen im Paradiese
Sealsfield 24	Das Cajuetenbuch - Selige Stunden
Sealsfield 25	Das Cajuetenbuch - Das Diner
Sealsfield 26	Das Cajuetenbuch - Der Abend
Sealsfield 27	Das Cajuetenbuch - Die Fahrt und die Kajuete
Sealsfield 28	Das Cajuetenbuch - Das Paradies der Liebe
Storm 01	Der Schimmelreiter
Sudermann 01	Die Reise nach Tilsit

Tucholsky 01-05	Schloss Gripsholm 1-5
Wedekind 01-04	Mine-Haha I-IV
Wedekind 05	Rabbi Esra
Wedekind 06	Frühlingsstürme
Wedekind 07	Silvester
Wedekind 08	Der Verführer

Hawaiian (in Table 2.6, 3.34)

Hw 03: Moolelo, Kawelo, Mokuna I - KE KUAUHAU O KAWELO, http://www2.hawaii.edu/~kroddy/moolelo/kawelo/mokuna1.htm
Hw 04: Moolelo, Kawelo, Mokuna II - KA HANAU ANA O KAWELO http://www2.hawaii.edu/~kroddy/moolelo/kawelo/mokuna2.htm
Hw 05: Moolelo, Kawelo, Mokuna III - KA HOOLELE LUPE ANA O KAUA HOA ME KAWELO, http://www2.hawaii.edu/~kroddy/moolelo/kawelo/mokuna3.htm
Hw 06: Moolelo, Kawelo, Mokuna IV - KA IKE ANA O KO KAWELO UHANE IA UHUMAKAIKAI, http://www2.hawaii.edu/~kroddy/moolelo/kawelo/mokuna4.htm

Hungarian (online newspaper texts) (in Table 2.6, 3.34)

H 01: Orbán Viktor beszéde az Astoriánál
H 02: A nominalizmus forradalma
H 03: Népszavazás
H 04: Egyre több
H 05: Kunczekolbász

Indonesian (online newspaper texts) (in Table 2.6, 3.34)

In 01: Assagaf-Ali Baba Jadi Asisten
In 02: BRI Siap Cetak Miliarder Dalam Dua Bulan
In 03: Pengurus PSM Terbelah
In 04: Pemerintah Andalkan Hujan
In 05: Pelni Jamin Tiket Tidak Habis

Italian (in Table 2.2)
End-of-year speeches of Italian presidents 1949-2008

(In Table 3.34)
I 01: Silvio Pellico Le mie prigioni

I 02: Alessandro Manzoni	I promessi sposi
I 03: Giacomo Leopardi	Canti
I 04: Grazia Deledda	Canne al vento
I 05: Edmondo de Amicis	Il cuore

Kannada (in Table 2.6, 3.34)

Kn 003: Pradhana Gurudhat: Aadalitha Bashe Kelavu Vicharagalu(1984), 71-92
Kn 004: Pradhana Gurudhat: Aadalitha Bashe Kelavu Vicharagalu(1984), 93-103
Kn 005: T.R.Nagappa: Vayskara Shikshana mathu swayam seve (1988), 1-15
Kn 006: T.R.Nagappa: Vayskara Shikshana mathu swayam seve (1988), 16-42
Kn 011: D.N.S.Murthy:Shreshta arthashasthagnayaru (1990), 3-53

Lakota (in Table 2.6, 3.34)

Lk 01: The fly on the window. Neva Standing Bear tape-recorded 11/16/1994 in Denver, Colorado, USA
Lk 02: Iktomi meets the prairie chicken and Blood Clot Boy. Neva Standing Bear tape-recorded 9/12/1994 in Denver, Colorado, USA
Lk 03: Iktomi meets two women and Iya. Neva Standing Bear tape-recorded 9/19/1994 in Denver, Colorado, USA
Lk 04: Bean, grass, and fire. Florine Red Ear Horse tape-recorded 9/19/1995 in Denver, Colorado, USA

Latin (in Table 3.34)

Lt 01: Vergil	Georgicon liber primus
Lt 02: Apuleius	Fables, Book 1
Lt 03: Ovidius	Ars amatoria, liber primus
Lt 04: Cicero	Post reditum in senatu oratio
Lt 05: Martialis	Epigrammata
Lt 06: Horatius	Sermones.Liber 1, Sermo 1

Maori (in Table 2.6, 3.34)

M 01 Maori Nga Mahi a Nga Tupuna, ed. George Grey. Wellington, L. T. 3rd edition 1953
M 02 KO TE PAAMU TUATAHI WHAKATIPUTIPU KAU A TE MAORI . TE AO HOU The New World [electronic resource] No. 5 (Spring 1953)

M 03 A TAWHAKI,TE TOHUNGA RAPU TUNA. TE AO HOU The New World [electronic resource] S, No. 10 (April 1955)
M 04 KA PU TE RUHA KA HAO TE RANGATAHI. Accessible in NGA KORERO A REWETI KOHERE MA, in New Zealand Electronic Texts (NZETC, Auckland, Internet)
M 05 KA KIMI A MAUI I ONA MATUA, In TE AO HOU, No. 8, Winter 1954

Marathi (in Table 2.6, 3.34)

Mr 001: B.P.Joshi: Nisar Sheti (1991), pp 77-97
Mr 018: V.L.Pandy: Thumcha chehara thumche yaktimatv, (1990), pp 9-89
Mr 026: Kanchan Ganekar: Nath ha majha (1989), pp 1-17
Mr 027: Sarangar: Rashtriy Uthpann (1985), pp 1-104
Mr 288: Madhav Gadkari :Chaupher (1988), pp 1-14

Marquesan (in Table 2.6, 3.34)

Mq 01 Story Kopuhoroto'e II from the collection Henri Lavondès: Récits marquisiens dits par Kehuenui avec ls collaboration de S. Teikihuupoko. Publication provisoire. Papeete, Centre ORSTOM 1964, pp. 25-37
Mq 02 Ka´akai o Te Henua ´Enana. A Story of the Country of People recorded by Sam H. Elbert.
Mq 03 TE HAKAMANU. LA DANSE DE L´OISEAU. Légende marquisienne. Texte marquisien: Lucien Teikikeuhina Kimitete Papeete, Haere Po no Tahiti 1990

Rarotongan (all texts in: Legends from the Atolls. Editor Kauraka Kauraka, Suva 1983) (in Table 2.6)

Rt 01 Akamaramaanga, by Kauraka Kauraka himself, 1983
Rt 02 Ko Paraka e te Kehe, by Tepania Puroku, 1977
Rt 03 Ko Tamaro e ana uhi, by Herekaiura Atama, 1977
Rt 04 Te toa ko Teikapongi, by Temu Piniata, 1982
Rt 05 Te toa ko Herehuaroa e Araitetonga, by Kaimaria Nikoro, 1982

Romanian (http://www.romanianvoice.com/poezii/poeti/eminescu.php) (in Table 2.6, 3.34)

R 01: Eminescu, M.: Luceafarul - Lucifer
R 02: Eminescu, M.: Scrisoarea III - Satire III
R 03: Eminescu, M.: Scrisoarea IV - Satire IV

R 04: Eminescu, M.: Scrisoarea I - Satire I
R 05: Eminescu, M.: Scrisoarea V - Satire V
R 06: Eminescu, M.: Scrisoarea II - Satire II

Russian (in Table 3.34)

Ru 01 Fedor M. Dostoevskij: Prestuplenie i nakazanie (p. I, ch. 1)
Ru 02 Nikolaj G. Gogol': Portret
Ru 03 Viktor Pelevin: Buben verchnego mira
Ru 04 Lev N. Tolstoy: Metel'
Ru 05 Ivan S. Turgenev: Bežin lug

Samoan (texts in: Tala o le Vavau. The Myths, Legends and Customs of Old Samoa. Polynesian Press Samoa House, Auckland 1987 (in Table 2.6, 3.34)

Sm 01 O le mea na maua ai le ava, pp. 15 – 16
Sm 02 O le tala ia Sina ma lana tuna, pp. 17 – 19
Sm 03 O le tala ia Tamafaiga, pp. 49 – 52
Sm 04 O le faalemigao, pp. 91 – 92
Sm 05 O upu faifai ma le gaoi, p. 95

Slavic languages (in Table 2.3, 3.36)

Translations of 10 chapters of N. Ostrovskij´s "How the steel was tempered" from Russian.

Slovenian (in Table 3.34)

Sl 01 Ivan Cankar: V temi
Sl 02 Slavko Grum: Vrata
Sl 03 Josip Jurčič: Sosedov sin (ch. I)
Sl 04 Ferdo Kočevar: Grof in menih
Sl 05 Fran Levstik: Zveženj

Tagalog/Pilipino (in Table 2.6, 3.34)
(from http://www.seasite.niu.edu/Tagalog/tagalog_short_stories_fs.htm)

T 01 A.V. Hernandez: Magpinsan
T 02 A.V. Hernandez: Limang Alas, Tatlong Santo
T 03: A.B.L. Rosales: Kristal Na Tubig

References

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen.* Lüdenscheid: RAM [= Studies in Quantittative Linguistics 2]

Balakrishnan, V.K. (1997). *Graph theory*. New York: McGraw-Hill.

Bazan, J., Szczuka, M., Wojna, A., Wojnarski, M. (2004). On the evolution of rough set exploration system. *Proceedings of the RSCTC 2004: 592–601*.

Belza, M.I. (1971). K voprosu o nekotorych osobennostjach semantičeskoj struktury svjaznych tekstov. In: *Semantičeskie problemy avtomatizacii informacionnogo potoka: 58-73*. Kiev.

Brunet, E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. Le nombre et le texte. *Revue informatique et statistique dans les sciences humaines 24(1-4), 81-116.*

Dubois, D., Prade, H. (1990). Rough fuzzy sets and fuzzy rough sets. *International Journal of General Systems 17, 191–209.*

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics 17, 2008, 79-86.*

Gibbons, J.D. (1971). *Nonparametric statistical inference*. New York: McGraw-Hill.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štepánek, J., Havelka,, J., Mikulová, M. (2006). *Prague Dependency Treebank 2.0.* Philadelphia Linguistic Data Consortium.

Hřebíček, L. (1997). *Lectures on text theory*. Prague: Oriental Institute

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

Kelih, E. (2009). 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. (2009a): 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. (2006). The frequency distribution of the lengths of length sequences. In: Genzor, J., Bucková, M. (eds.), *Favete linguis. Studies in honour of Victor Krupa: 145-152*. Bratislava: Slovak Academy of Sciences.

Köhler, R., **Naumann, S.** (2007). Quantitative text analysis using L, F- and T-segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data Analysis, Machine Learning and Applications: 637-646.* Berlin-Heidelberg: Springer.

Köhler, R., Naumann, S. (2009). A contribution to quantitative studies on the sentence level. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 34-45*. Lüdenscheid: RAM-Verlag.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics 14, 33-80.*

Labbé, C., Labbé, D. (2001). Inter-textual distance and authorship attribution Corneille and Moliére. *Journal of Quantitative Linguistics 8, 213-231.*

Labbé, C., Labbé, D. (2003). La distance intertextuelle. *Corpus 2, 95-118.*

Labbé, C., Labbé, D. (2006). A tool for literary studies. Inter-textual distance and tree-classification. *Literary and Linguistic Computing 21, 311 - 326.*

Mačutek, J. (2009). Motif richness. In: Köhler, R. (ed.), *Issues in Quantitative Linguistics: 51-60.* Lüdenscheid: RAM-Verlag [= Studies in Quantitative Linguistics 5]

Mačutek, J., Popescu, I.-I., Altmann, G. (2007). Confidence intervals and tests fort he h-point and related text characteristics. *Glottometrics 15, 42-52.*

Mandelbrot, B.B. (1982). *The fractal geometry of nature*. New York: Freeman.

Merriam, T. (2002). Intertextual distances between Shakespeare plays, with special reference to *Henry V* (verse). *Journal of Quantitative Linguistics 9(3), 261-273.*

Möller, H.J., Laux, G., Deister, A. (2009): *Psychiatrie und Psychotherapie.* 4. Auflage. Stuttgart: Thieme.

Muller, C. (1992). *Principes et méthodes de statistique lexicale.* Paris: Champion.

Nemcová, E., Popescu, I.-I., Altmann, G. (2010). Word associations in French. In: Berndt, A., Böcker, J. (eds.), *Sprachlehrforschung: Theorie und Empirie: 223-237*. Frankfurt: Lang.

Ord, J.K. (1972). *Families of frequency distributions*. London: Griffin.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston.

Paivio, A., Yuille, J., Madigan, S. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology, 76 (Supp.l), 1-25.*

Pawlak, Z. (1991). *Rough Sets: Theoretical Aspects of Reasoning About Data*. Dordrecht: Kluwer Academic Publishing

Popescu, I.-I. et al. (2009). *Word frequency studies*. Berlin-New York: Mouton de Gruyter.

Popescu, I.-I., Mačutek, J., Altmann, G. (2009). *Aspects of word frequencies.* Lüdenscheid: RAM. [= Studies in Quantitative Linguistics 3]

Rudman, J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities 351-365.*

Skorochod'ko, E.F. (1981). *Semantische Relationen in der Lexik und in Texten.* Bochum: Brockmeyer.

Tuldava, J. (1971). Statističeskij metod sravnenija leksičeskogo sostava dvuch tekstov. *Linguistica 4, 199-220.*

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie.* Trier: Wissenschaftlicher Verlag.

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009a). Zipf´s law in Italian texts. *Journal of Quantitative Linguistics 16(4), 354-367.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2009b). Parts-of-speech diversification in Italian tesxts. *Glottometrics 19, 42-48.*

Tuzzi, A., Popescu, I.-I., Altmann, G. (2010). *Quantitative analysis of Italian texts*. Lüdenscheid: RAM [= Studies in Quantitative Linguistics 6]

Viprey, J.-M., Ledoux, C.N. (2006). About Labbe´s "Intertextual Distance". *Journal of Quantitative Linguistics 13(2-3), 265-283.*

Ziegler, A., Altmann, G. (2002). *Denotative Textanalyse*. Wien: Praesens.

Author index

Altmann, G. 4,5,12,26,76,87,93,94, 100,103,124,135
Altmann, V. 100,103
Balakrishnan, V.K. 101
Bazan, J. 21
Belza, M.I. 135
Brunet, E. 1
Deister, A. 145
Dubois, D. 21
Fan, F. 93,94
Gibbons, J.D. 60
Hajič, J. 115
Hřebíček, L. 124
Kelih, E. 15,56,64,74,82,97
Kendall, M.G. 60
Köhler, R. 91,126,135
Labbé, C. 1
Labbé, D. 1
Laux, G., 145
Ledoux, C.N. 1
Mačutek, J. 4,5,26,76,87,91,126
Madigan, S. 2
Merriam, T. 1
Möller, H.J. 145
Muller, C. 1
Naumann, S..91,126
Nemcová, E. 93,94
Ord, J.K. 1
Paivio, A. 2
Pawlak, Z. 21
Popescu, I.-I. 4,5,12,22,24,26,76,77, 87,93,94
Prade, H. 21
Rudman, J. 1
Skorochoďko, E.F. 135
Szczuka, M. 21
Tuldava, J. 1
Tuzzi, A. 12,26
Viprey, J.-M. 1
Wojna, A. 21
Wojnarski, M. 21
Yuille, J. 2
Ziegler, A. 124

Subject index

absolute sequential difference 39
adjacency matrix 101
analytism 20,24,80,89
anaphora 135
arc length 76,90,97-99
association 124, 126
attractor 2
Belza-motif 145
Belza´s chaining coefficient 136
Bernoulli distribution 102
binary code (*BC*) 3,100-134
- of chapter 131
- of sentence 100-123
- of text 2,3,124-134

binomial distribution 96
break 29,53,59,120-123
chaining 135-145
chaining coefficient 135-145
chaos 40
cohesion 127,128,135
comparison
- crosslinguistic 55-68
- of texts 2,53-75
- of vocabularies 1

concentration 127,145
concordance coefficient 60
co-reference 135
cos τ 26
dependence fragmentation 121,122, 128-130
dependence structure 128
development 2
- historical 2
- ontogenetic 2

dispersion 38-40
dissimilarity 26-28,30,38,53-59,61, 62,67,68,129-131,134
- cumulative 26,29-37,43
- external 53-55
- internal 53-55
- measure of- 26
- prospective 42-52,96
- retrospective 26-38,96,130,
- 132
- stepwise 26-40,97,98,130,132

distance 59,68,69,76
- mean- 73
- vector- 68-74

distribution 2
diversification 93,94
drama 25,28,128
dynamics 95-99
frequency motif 91,92
frequency spectrum 1
general textology 2
genre 2
goedelization 100-120
golden ratio 91,92
h-point 4,5
hreb 124,126
indicator
- *A* 4-25
- *AS* 39,40,52,59,60,67,73,74
- *B, BC* 101-103,105-120,127,
- 128,131,132
- *C* 136,140-145
- C_1 127,128
- C_{rel} 143-145
- *δ* 68-74
- *F* 121-123,128,129
- *NS* (non-smoothness) 95-97
- *R* (of roughness) 97-99
- *RS* 51,52
- τ 25-29,33-36,39-50,52,54-56,59,62-68,96,130,132-134

Italian presidents 12-15,25
jump 120-122,128,132
Köhlerian motif 91,126,145
language
- Belorussian 15,16,19-21,24, 56,57,59-63,67,69,70,74, 75,82,83,86,97,99,132,133

- Bulgarian 15,16,19-21,24,56, 58-62,64,67,70,74,77,81, 83,86,89,97,99,132,133
- Croatian 15,16,19-21,24,56, 58-63,68,70,74,75,83,86, 97,99,132,133
- Czech 15,16,19-21,24,56,57, 59-62,64,67,71,74,75,77, 81,83,86,88,97,99,115-120,122,123,128,129, 132,133,136,139-141,145
- East Slavic 21,74
- English 22,24,78,81,88,93,94
- French 93,94
- German 5-12,14,24,25,27-37, 51,56,78,81,87,89,96,98, 126,131,140,141,144
- Hawaiian 22,24,78,81,88
- Hungarian 22,24,78,81,88
- Indonesian 22,24,78,81,89
- Italian 12-15,24,25,78,81,88
- Kannada 22,24,78,79,81,89
- Lakota 22-24,79,81,88
- Latin 23-24,79,81,89
- Macedonian 15-17,19-21,24, 56,58-62,64,68,71,74,75, 83,84,86,97,99,132,133
- Maori 23,24,79,81,88
- Marathi 24,79,81,89
- Marquesan 23,24,79,81,88
- Polish 15,17,19-21,24,56,57, 59-62,65,67,71,74.75,84, 86,97,99,132,133
- Polynesian 89
- Rarotongan 23,24,79-81,88
- Romanian 23-25,79,81,89
- Russian 15,17,20,21,24,56, 57,59-67,69,73,74,75,80, 81,84,86,89,92,97-99, 104-115,120-123,128,129, 132,133,136
- Samoan 23,24,80,81,88
- Serbian 15,17,21,24,56,58-62, 66,68,72,74,75,84,86,97, 99,132,133
- Slavic 15-21,56,59,61,62,67, 68,70,73,82,86,87,96-98, 132, 133
- Slovak 15,17-18.21,24,56, 57,59-62,65,67,72,74,75, 84-86,97,99,132,133,141-144
- Slovenian 15,18-21,24,56,58-62,65,67,68,72,74,75,80, 81,85,86,88,97,99,132,133
- Sorbian 15,18-21,24,56,58-62,67,73-75,85,86,97,99, 132,133
- South Slavic 21,74,133
- Tagalog 23,24,80,81,88
- Ukrainian 18-21,24,56,57,59-62,66,67,73-75,85,86,97, 99,132,133
- West Slavic 21,74

Lyapunov coefficient 67
modulus 4-25
non- smoothness 95-97
normal test 40,120,128
order 40
Ord’s-scheme 1
oscillation 95-97
power function 37,63
progressive syntactic dependence fragmentation 121,122
randomness 40
random walk 37
rank-frequency distribution 1,26,61, 99
rank-frequency sequence 4,43
recapitulative structure (RS) 51,52
reference 124,126
rough set 21
roughness 97-99
- prospective 98,99
- retrospective 98,99

self-regulation 2
sentence association 125
sentence boundary 135

sentence length 120
similarity 1,26,29
Skinner effect 29
smoothness 128
spontaneity 38
stability 2
stage play 2,130
standard deviation 38-40,59
structure 40
style 2,14,39
synthetism 21,24
ternary plot 76-94
text size 4,61,62
text theory 2
time series 102,145
TTR 61
unfolding 2
uniform distribution 40
variability 38
vector distance 68
vector P 4
vector T 3,26-52
vector U 76-94
vocabulary size 61,90
word association 93,94
word length 74
zero-one distribution 101