Glottometrics 23
2012

RAM-Verlag

ISSN 2625-8226

Glottometrics

Glottometrics ist eine unregelmäßig erscheinende Zeitdchrift (2-3 Ausgaben pro Jahr) für die quantitative Erforschung von Sprache und Text.
Beiträge in Deutsch oder Englisch sollten an einen der Herausgeber in einem gängigen Textverarbeitungssystem (vorrangig WORD) geschickt werden.
Glottometrics kann aus dem **Internet** heruntergeladen werden (**Open Access**), auf **CD-ROM** (PDF-Format) oder als **Druckversion** bestellt werden.

Glottometrics is a scientific journal for the quantitative research on language and text published at irregular intervals (2-3 times a year).
Contributions in English or German written with a common text processing system (preferably WORD) should be sent to one of the editors.
Glottometrics can be downloaded from the **Internet (Open Access)**, obtained on **CD-ROM** (as PDF-file) or in form of **printed copies.**

Herausgeber – Editors

Bestellungen der CD-ROM oder der gedruckten Form sind zu richten an

Orders for CD-ROM or printed copies to RAM-Verlag RAM-Verlag@t-online.de

Herunterladen/ Downloading: https://www.ram-verlag.eu/journals-e-journals/glottometrics/

Die Deutsche Bibliothek – CIP-Einheitsaufnahme
Glottometrics. 23 (2012), Lüdenscheid: RAM-Verlag, 2012. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet (Open Access) unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar.
Bibliographische Deskription nach 23 (2012) **ISSN 2625-8226**

Contents

History of Quantitative Linguistics

Reviews

Glottometrics 23, 2012, 1-6

Längen von Komposita im Deutschen

Karl-Heinz Best, Göttingen

Abstract. The aim of this paper is to show that the lengths of compounds in German abide by a law. The findings lend support to the theory of word length distributions (Wimmer et alii 1994, Wimmer & Altmann 1996) once more.

Keywords: compounds, length, German

1. Verteilung von Einheiten unterschiedlicher Länge

Im Göttinger *Projekt Quantitative Linguistik* werden die Verteilungen sprachlicher Einheiten verschiedener Länge untersucht, um der Hypothese nachzugehen, dass diese in Texten und gegebenenfalls auch in Wörterbüchern gesetzmäßig verteilt sind. Eine allgemeine Annahme zur Häufigkeit von Komposita formuliert Altmann (1989, 104):

> „the number of compounds decreases with their increasing length.“

Untersuchungen zur Frage, ob, und wenn ja, welcher Verteilung Komposita unterschiedlicher Länge unterliegen, sind bisher Mangelware. Lediglich Poppe (2007) hat dem Thema eine eigenständige Studie gewidmet. Sie kam zu dem Ergebnis, dass zumindest bei den von ihr ausgewerteten 21 Pressetexten aus *GEOEpoche* und *Frankfurter Allgemeine Zeitung* eine Anpassung der 1-verschobenen Waring-Verteilung möglich war. Der Absicherung dieses Ergebnisses war die Staatsexamensarbeit von Klöpfel (2010) gewidmet, der insgesamt 80 Pressetexte aus *Frankfurter Allgemeine Zeitung* (4 Texte), *Süddeutsche Zeitung* (36 Texte) und *DIE ZEIT* (40 Texte) auswertete. Die Waring-Verteilung bewährte sich in 70 Fällen; bei 9 Texten konnte die Yule-Verteilung in 1-verschobener Form erfolgreich angewendet werden; bei einem Text ließ sich keine Verteilung anpassen.

Das Problem derartiger Untersuchungen besteht darin, dass in aller Regel nicht aufgrund theoretischer Überlegungen vorhergesagt werden kann, welche Verteilung für ein bestimmtes sprachliches Phänomen zu erwarten ist; stattdessen muss man sich damit begnügen, explorativ nachzuweisen, dass das Vorkommen von Einheiten verschiedener Länge nicht völlig chaotisch erfolgt.

2. Theoretische Grundlagen

Als theoretische Grundlage der Untersuchung wurden wie in allen Arbeiten des Göttinger Projekts die Beiträge von Wimmer et al. (1994) sowie von Wimmer & Altmann (1996) gewählt. Ihre generelle Hypothese besteht darin, dass die unterschiedlichen Längen sprachlicher Einheiten in Texten gemäß theoretisch begründeten Verteilungen vorkommen. Diese Annahme konnte in vielen Untersuchungen zu über 50 Sprachen oder Sprachentwicklungsstadien gestützt werden. Die unterschiedlichen Randbedingungen (Sprache, gewählte Einheit, Zeit, Autor, Textsorte...) führen dazu, dass nicht in allen Untersuchungen immer das gleiche Modell für die jeweiligen Daten gewählt werden kann.

3. Datenbasis und Verfahren

Datengrundlage für diese Untersuchung ist Klöpfel (2010), der für alle 80 Pressetexte einzeln gezeigt hat, dass sie der 1-verschobenen Waring-Verteilung (Wimmer & Altmann 1999, 643f.; 2005, 800)

$$(1) \qquad P_x = \frac{b}{b+n} \frac{n^{(x-1)}}{(b+n+1)^{(x-1)}}, \quad x = 1, 2, ...$$

bzw. in einigen Fällen der 1-verschobenen Yule-Verteilung, die ein Spezialfall der Waring-Verteilung ist (mit n = 1), folgen. Abschließend demonstriert Klöpfel (2010, 122), dass auch alle 80 Dateien zusammen diesem Modell folgen.

Die Auswahl der Texte erfolgte willkürlich, wenn gewährleistet war, dass die Texte mindestens drei verschiedene Kompositalängen aufwiesen. Sie stammen aus den Sparten Bildung, Feuilleton, Politik, Reisen und Sport. Die Länge der Komposita wurde danach bestimmt, aus wie vielen Lexemen sie bestehen; rein grammatische Morpheme (Derivateme, Flexeme, Fugenmorphe) spielen dabei keine Rolle.

Im Unterschied zu Klöpfel (2010) werden hier die Daten der drei untersuchten Zeitungen je für sich zusammengefasst und in dieser Form ebenfalls der Anpassung der Waring-Verteilung unterzogen. Die Gesamtdatei wird von Klöpfel (2010, 122) übernommen und durch Anpassung des gleichen Modells an eine Datei zu den Komposita in Werbetexten ergänzt, die auf einer Auswertung von Sowinski (1979, 110; 1998, 67) beruht und für die gezeigt werden konnte, dass sie der 1-verschobenen Hyperpoisson-Verteilung entspricht (Best 2006, 47); auch die Waring-Verteilung lässt sich daran anpassen, wenn auch nicht mit ganz so gutem Ergebnis. Die Anpassungen erfolgten mit Hilfe des *Altmann-Fitters* (1997).

Offenbar handelt es sich hier um eine Familie von Verteilungen, die Spezialfälle oder Grenzfälle der Hyperpascal-Verteilung sind: Die Waring- und die Yule-Verteilung sind Spezialfälle, die Hyperpoisson-Verteilung ist ein Grenzfall. Die Hyperpascal-Verteilung stellt die Lösung der Differenzengleichung

$$(2) \qquad P_x = \frac{k+x-1}{m+x-1} q P_{x-1}, \quad \text{für } x = 1, 2, ...$$

dar, die zeigt, wie die Häufigkeiten der Nachbarklassen reguliert werden. Im Zähler steht die Diversifikationskraft des Sprechers (k-1), im Nenner die Unifikationskraft des Hörers (m-1), der extreme Längenveränderungen abbremst; die Konstante q stellt die normierende Regulierung der Sprache dar. Ist die Lage in der Sprache einigermaßen im Gleichgewicht, d.h. q = 1, dann kann sich die Beziehung vereinfachen. Setzt man in (2) $k = n$ und $m = b+n+1$, dann bekommt man die Waring-Verteilung. Wenn auch k = 1 und $m = b+2$, dann erhält man die Yule-Verteilung. Zu bemerken ist, dass mit m = 1 Formel (2) zu der negativen Binomialverteilung führt, und mit $k = m$ zu der geometrischen Verteilung. Wenn jedoch k und q extreme Werte annehmen, d.h. $k \to \infty$ und $q \to 0$ und ihr Produkt $kq \to a$, dann resultiert die Hyperpoisson-Verteilung. Die gleiche Konvergenz, jedoch mit m = 1, führt zu der Poisson-

Verteilung. Wie man sieht, ist die Regulierung der Kompositalänge in einer Sprache durch ein einfaches Modell erfassbar. Das Testen an mehreren Sprachen würde zeigen, welche Technik der Kompositabildung sich durchsetzt bzw. ob das Modell hinreichend ist.

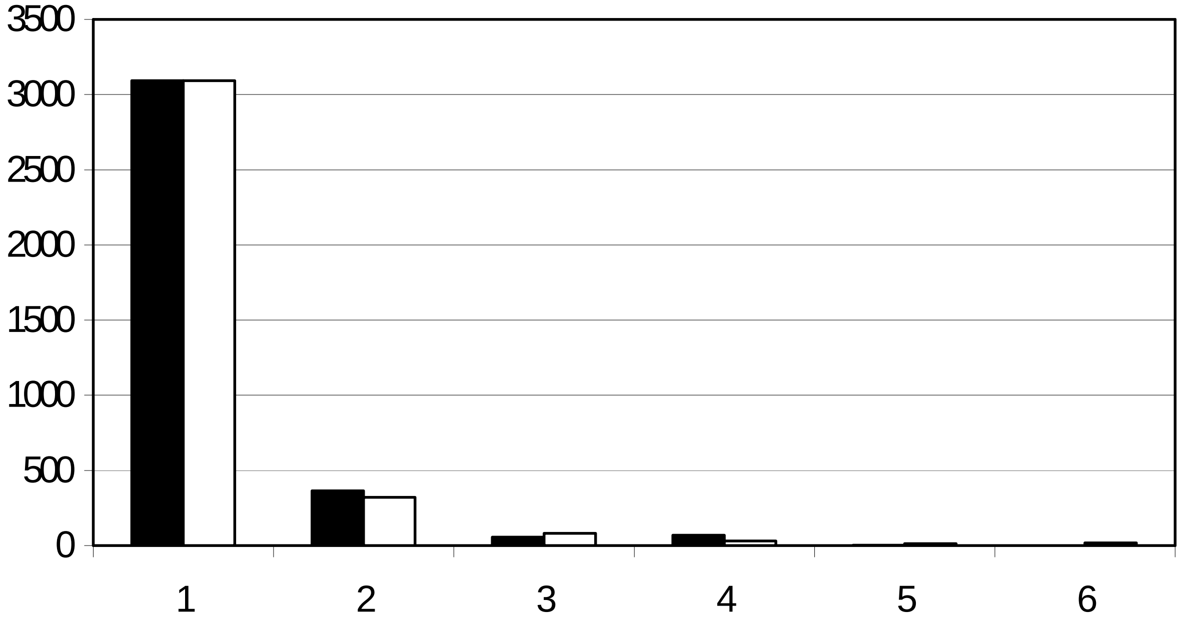
4. Ergebnisse der Anpassung der 1-verschobenen Waring-Verteilung an die Pressetexte

Die Anpasssung der 1-verschobenen Waring-Verteilung erbrachte die in Tabelle 1 dargestellten Ergebnisse:

Tabelle 1
Anpassung der 1-verschobenen Waring-Verteilung
an die Kompositalängen in Texten der Presseorgane

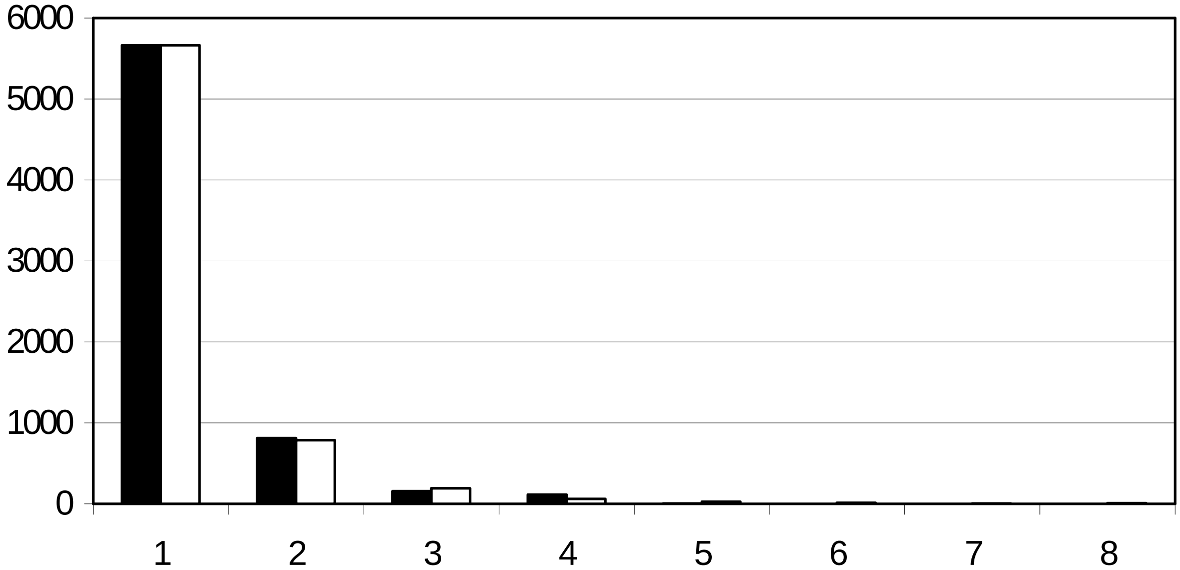
	Frankfurter Allgemeine Zeitung		*Süddeutsche Zeitung*		*DIE ZEIT*		Gesamtdatei aller drei Zeitungen (Klöpfel 2010, 122)	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1	291	291.00	2280	2280.00	3093	3121.31	5664	5664.00
2	42	41.98	407	409.89	364	321.92	813	787.08
3	4	9.50|	97	93.07	57	82.44	157	192.05
4	10	4.52|	31	25.05|	70	30.05|	112	62.65|
5			3	7.69|	2	13.38|	5	24.58|
6			1	2.62|	1	17.91|	2	10.97|
7			0	0.98|			0	5.40|
8			1	0.70|			1	7.27|
Σ =	347		2820		3587		6754	
b =	7.0614		12.4085		3.3621		5.1970	
n =	1.3589		2.9389		0.5016		1.0001	
C =	0.0000		0.0001		0.0044		0.0012	

Legende zur Tabelle (Weitere Erläuterungen in Best 2006: 29-33.): Σ ist die Zahl der Komposita der jeweiligen Datei; x: Länge der Komposita (in Lexemen), wobei x = 1 für Komposita steht, die aus zwei Lexemen bestehen, x = 2 für Komposita aus drei Lexemen, usw.; n_x: beobachtete Zahl der Komposita mit der Lexemzahl x; NP_x: durch Anpassung der 1-verschobenen Waring-Verteilung errechnete Zahl der Komposita mit der Lexemzahl x; b, n: Parameter der Verteilung; C ist der Diskrepanzkoeffizient, der eingesetzt wird, wenn P mangels Freiheitsgraden nicht bestimmt werden kann oder eine sehr umfangreiche Datei bearbeitet wird; er soll die Bedingung $C \leq 0.01$ erfüllen; dies ist in allen Fällen gegeben. Die senkrechten Linien in den Tabellen zeigen eine Zusammenfassung der so markierten Längenklassen an. Die Graphik gibt als Beispiel das gute Ergebnis zu den Texten von *DIE ZEIT* wieder:



Graphik 1: Beobachtete (schwarz) und berechnete Werte (weiß) zu *DIE ZEIT*

Das Ergebnis für die Gesamtdatei lässt sich wie folgt darstellen:



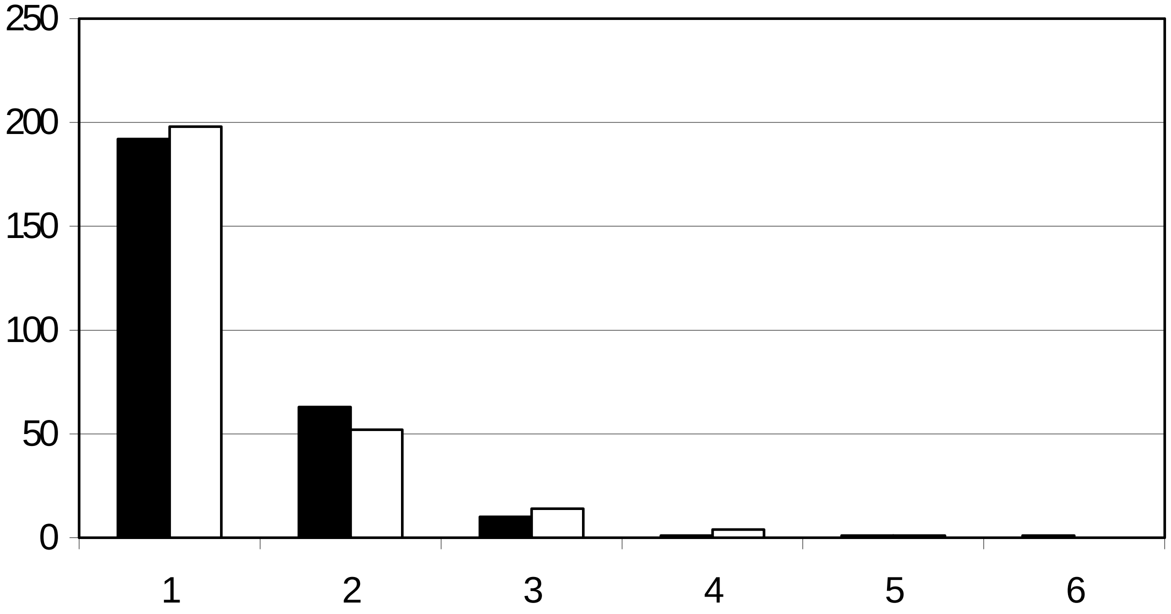
Graphik 2: Beobachtete (schwarz) und berechnete Werte (weiß) zu allen ausgewerteten Pressetexten

Für die Kompositalängen in Werbetexten wurde anhand eines relativ kleinen Korpus bereits gezeigt, dass sie der 1-verschobenen Hyperpoisson-Verteilung ($P(X^2) = 0.34$) entsprechen. Da diese Verteilung sich bei Kompositalängen aber mehrfach nicht bewährt hat, soll hier demonstriert werden, dass sie auch der 1-verschobenen Waring-Verteilung folgen, wenn auch mit weniger gutem Ergebnis und auffallend hohen Parameterwerten:

Tabelle 2
Anpassung der 1-verschobenen Waring-Verteilung
an die Kompositalängen in Werbetexten

x	n_x	NP_x	x	n_x	NP_x
1	192	197.86	4	1	3.55
2	63	51.78	5	1	0.93|
3	10	13.55	6	1	0.33|
$\sum = 268$ $b = 139920.8810$ $n = 49598.8198$ $X^2 = 5.802$ $FG = 2$ $P = 0.06$					

Legende zu Tabelle 2: *X²*: Chiquadrat; *FG*: Freiheitsgrade. Da es sich um eine recht kleine Datei handelt, wird ein anderes Prüfkriterium angewendet: *P* ist die Überschreitungswahrscheinlichkeit des *X²* und zeigt mit $P \geq 0.05$ ein gutes Ergebnis an; diese Bedingung ist hier erfüllt.



Graphik 3: Beobachtete (schwarz) und berechnete Werte (weiß) zu Werbetexten

5. Ergebnisse der Anpassung der 1-verschobenen Waring-Verteilung an die Verteilung der Längen von Komposita

Die Untersuchung hat gezeigt, dass Komposita verschiedener Länge in deutschen Pressetexten der 1-verschobenen Waring-Verteilung folgen. Das Gleiche kann auch für ein kleines Korpus aus Werbetexten festgestellt werden. Die Ergebnisse von Poppe (2007) finden hier weitere Bestätigung; die Datenbasis ist mit rund 7000 Komposita deutlich verbessert.

Komposita verschiedener Länge sind ein Spezialfall von Wortlängen; zumindest bei deutschen Texten scheint es jedoch so zu sein, dass Wortlängen (in Silben oder Morphen gemessen) einer anderen Verteilung, der 1-verschobenen Hyperpoisson-Verteilung, folgen als die Komposita.

Vergleichbare Untersuchungen unter anderen Bedingungen (andere Textsorte, Zeit, Sprache,...) können zu dem Ergebnis kommen, dass auch andere Verteilungen verwendet werden müssen. Die Hypothese, dass sprachliche Erscheinungen bestimmten Gesetzen folgen, wurde jedoch auch dieses Mal unterstützt.

Literatur

Altmann, Gabriel (1989). Hypotheses about compounds. In: Hammerl, Rolf (ed..), *Glottometrika 10, 100-107*. Bochum: Brockmeyer.

Best, Karl-Heinz (2006). *Quantitative Linguistik: Eine Annäherung.* 3., stark überarbeitete und ergänzte Auflage. Göttingen: Peust & Gutschmidt.

Klöpfel, Henning (2010). *Quantitative Untersuchungen zu Komposita im Deutschen. Zur Länge von Komposita in deutschen Pressetexten.* Staatsexamensarbeit, Göttingen.

Poppe, Stefanie (2007). Die Verteilung von Kompositalängen in deutschen journalistischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 15, 79-85.*

Sowinski, Bernhard (1979). *Werbeanzeigen und Werbesendungen.* München: Oldenbourg.

Sowinski, Bernhard (1998). *Werbung*. Tübingen: Niemeyer.

Wimmer, Gejza, & Altmann, Gabriel (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (ed..), *Glottometrika 15, 112-133.* Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, & Altmann, Gabriel (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, Gejza, & Altmann, Gabriel (2005). Unified Derivation of Some Linguistic Laws. In: Köhler, Reinhard, Altmann, Gabriel, & Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik - Quantitative Linguistics. Ein internationales Handbuch: 791-807.* Berlin/ N.Y.: de Gruyter.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Informationen zum Göttinger *Projekt Quantitative Linguistik*: Homepage: http://wwwuser.gwdg.de/~kbest/

Glottometrics 23, 2012, 7-22

Distinct word length frequencies: distributions and symbol entropies

Reginald Smith, Rochester, NY[1]

Abstract: The distribution of frequency counts of distinct words by length in a language's vocabulary will be analyzed using two methods. The first, will look at the empirical distributions of several languages and derive a distribution that reasonably explains the number of distinct words as a function of length. We will be able to derive the frequency count, mean word length, and variance of word length based on the marginal probability of letters and spaces. The second, based on information theory, will demonstrate that the conditional entropies can also be used to estimate the frequency of distinct words of a given length in a language. In addition, it will be shown how these techniques can also be applied to estimate higher order entropies using vocabulary word length.

Keywords: entropy, word frequency, n-gram, vocabulary

1. Introduction

The literature on word-length frequency distributions is one of the most vast amongst quantitative linguistics (cf. Best 1997, 2001; Grzybek 2006; Schmidt 1996). Word length frequencies typically investigate the frequency of words of different lengths in syllables. These distributions are common amongst texts and are typically interpreted as a type of negative binomial distribution (Altmann 1988; Wimmer & Altmann 1996) or Hyper-Poisson (Best 1998). Despite these different distributions though, they do have variations that can be used for applications such as authorship analysis (Williams 1970).

Often, the studies are conducted on the running text of a document or a corpus in order to determine the word length distribution of words within these textual sources. There is also a second tradition of word length and text analysis wherein the letter or grapheme, instead of the syllable, is used as the basic unit. This tradition has existed in parallel and has typically been used in mathematical studies in the tradition of (Shannon 1951) who analyzed the entropy of letters in texts. In this paper, we will take a limited view of word length distributions using this second tradition. In particular, we will be interested in the word length distribution only amongst distinct words in a language's vocabulary. These studies have been done in the past on the distribution of dictionary word lengths such as that in English (Rothschild 1986).

In this paper, we will investigate distinct word length distributions from two aspects. In the first part (section 2), we will investigate the typical distinct word length distribution and describe it using a derived distribution and goodness of fit tests. We will also discuss some connections between this distribution and the partition function. In the second part (sections 3-6), we will look at a different perspective based off of combinatorics to estimate the number of distinct words. Instead of the typical estimation of n-gram combinations being based only on the

[1] Address correspondence to: Reginald Smith E-mail: rsmith@bouchet-franklin.org

zero or first order entropy, using higher order conditional n-gram entropies can provide a reasonable approximation to the actual distribution.

2. Distinct word length distributions

The distinct words are calculated using files from WinEdt iSpell spell check dictionaries for a given language. Spell check dictionaries are not comprehensive by design since making them too large will include many rare or archaic words that should not be passed as correct in most writing. However, they give a good sample of commonly used terms in a language and the population of distinct words found in most texts excluding some proper nouns. The exception is the extinct language of Meroitic which used a corpus adopted from a previous paper (Smith, 2007). Below in Figure 1 is a visual comparison of the distinct word distributions by language (excluding Latin).

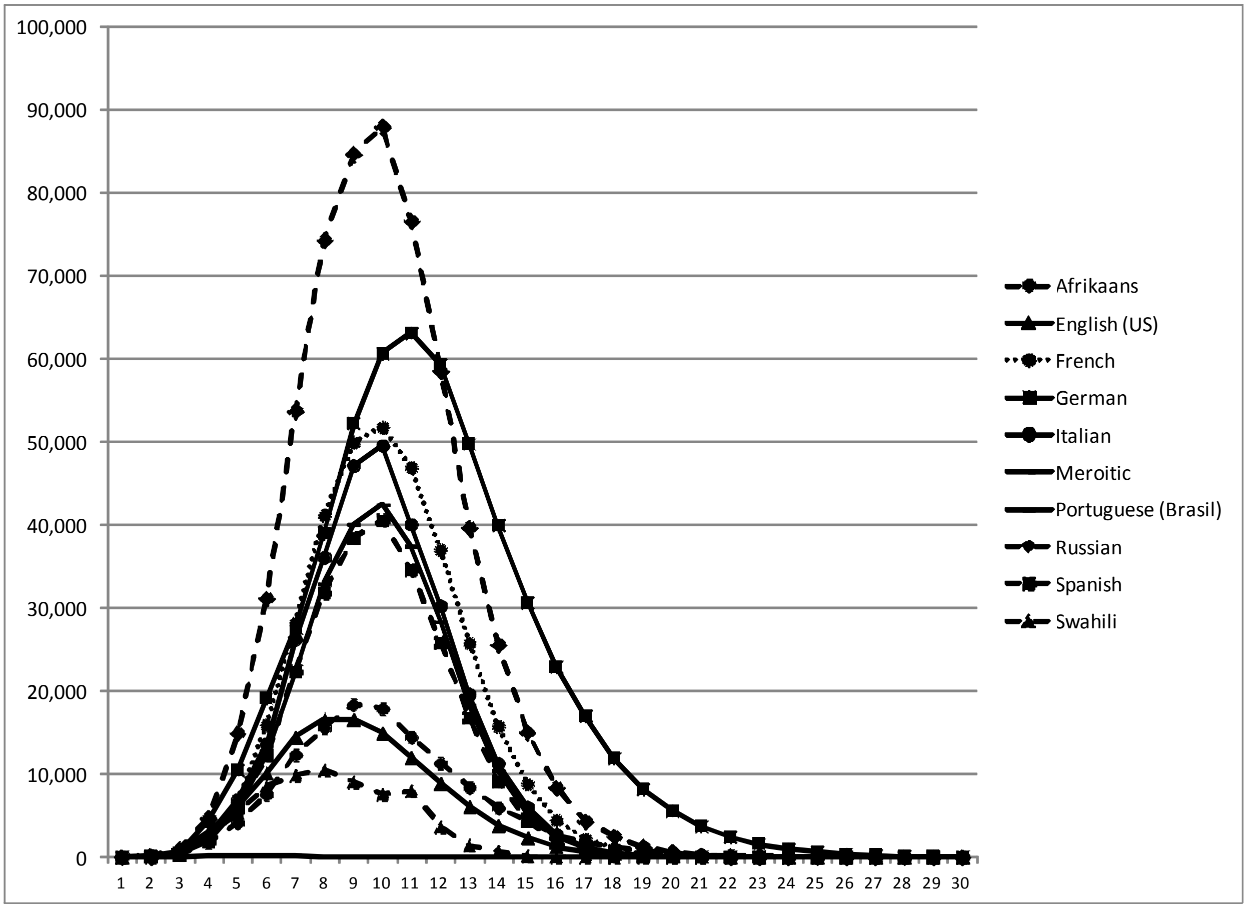


Figure 1: Graph of the frequency of distinct words by word length (Latin not shown due to a large vocabulary size that distorts the graph scale)

A key question is that given the similar nature of the distributions, is there a general distribution that can be used to describe the frequency count of distinct words in a language? If we assume the symbol set of a written language is restricted to its letters and the space/word separator character (of total symbol count *L*), a simple model of word generation can be described as follows. Assume we have a bag containing each symbol, including the space, where we can draw letters to form words and replace them after each drawing. A word is considered “finished” if you draw a space symbol from the bag. There is only one boundary condition, namely the first character must be a letter, not a space since there can be no zero-length words. Therefore, we can see that drawing from the bag can have two marginal probabilities: the probability of drawing a non-space letter (*p*) and the probability of drawing a space (“ “) (*1-p*). Therefore, for a given word of length *N*, the probability of drawing a word is $p^N(1-p)$ and the probability of pulling *N* straight letters is p^N. Granted, this simple model ignores the fact that letters have higher order conditional probabilities for digrams, trigrams, etc. that alter the probability of letters or even spaces being subsequently drawn.

For distinct words of length *N* the virtual word length (in probabilistic terms) is Np^N. We term this the virtual word length since the words are obviously of length *N* but the word formation process dictates an expected value that is much lower due to the possibility of pulling a space in any one of the 1 to *N-1* selections. The virtual word length is useful since it can be used with the number of written symbols to calculate the total number of words of distinct length , W_N, as

$$(1) \qquad W_N = L^{Np^N} - 1$$

The final term subtracting one is to reflect the approach to zero as *N* gets large. This, using a fixed value of *p*, is similar to the geometric distribution with the last term *(1-p)* absent. However, by fitting the distribution with the correct value of *p* we can accurately model the distinct word-length distributions. In Table 1 the results are shown for calculating the *p* value that minimizes the chi-square error between the data and the distribution. The graphical results are shown in Figure 2.

In addition, there is a further observation that allows us to estimate the average word length. In particular, the average word length can be given by

$$(2) \qquad \bar{W} = \sum_{k=1}^{M} k \frac{L^{kp^k} - 1}{\sum_{j=1}^{M} L^{jp^j} - 1}$$

where *M* is the number of letters in the word of longest length. One can recognize the term in the denominator as the partition function and the average word length distribution as structurally similar to the average energy calculation in statistical mechanics. It is also the total number of distinct words, in other words, the vocabulary size.

The types of distributions represented in both equations 1 and 2 are double exponentials. Typically, closed form solutions are difficult to achieve except in the most simple of cases. Therefore, the author used numerical approximations, regressions, and simulations over various values of *p* and *L* to determine the approximate equations. From these methods, the expressions for the total vocabulary size, mean word length, and the variance of word length were found and are shown in equations 3-5:

$$(3) \qquad W_{total} = \sum_{j=1}^{M} L^{jp^j} - 1 \approx AL^{b \ln L \frac{p}{1-p}}$$

where both *A* and *b* are constants, likely the ratios of fundamental quantities.

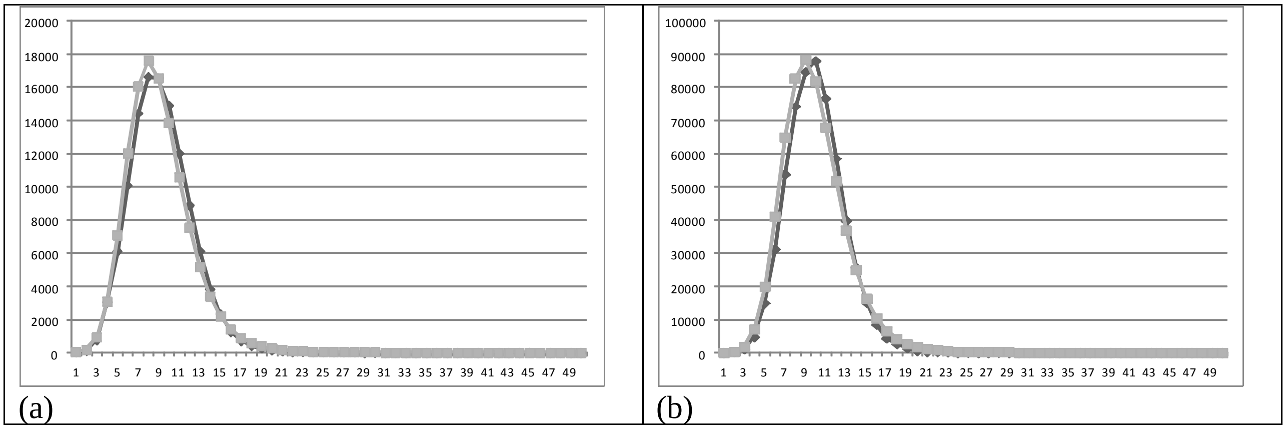
(4) $$\overline{W} = -\frac{1}{p\ln p}$$

(5) $$\sigma_W^2 = \frac{1}{[p(1-p)]^2}$$

These expressions, with the constants *A* and *b* determined by minimizing chi-square error, are compared against real data in Tables 1, 2, and 3.

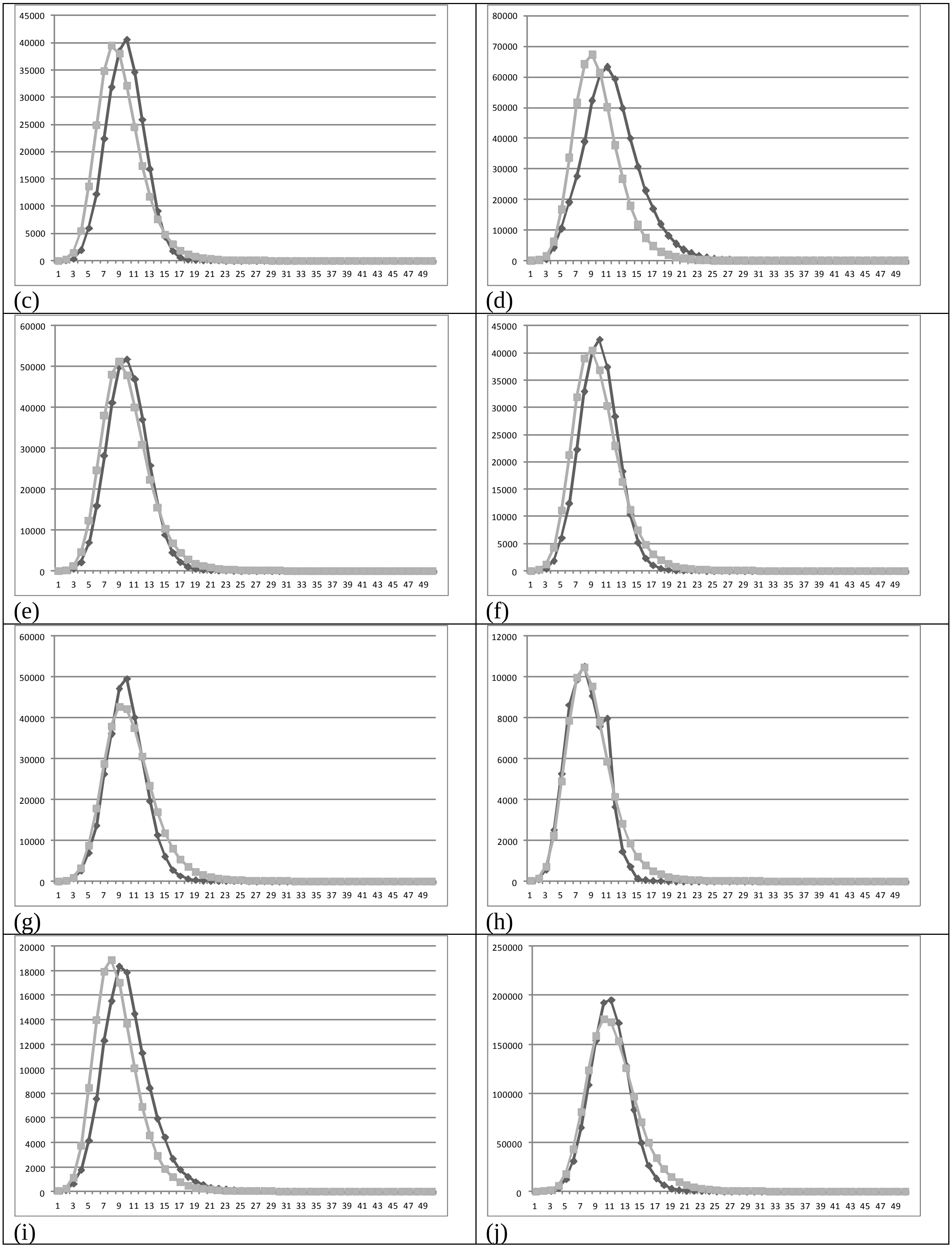
Table 1
The solved for *p* values for the theoretical distinct word distributions (*M* = 50) along with the average word lengths (observed and based on -1/*p* ln *p*), their percent difference, and the chi-square statistic (df=48) and probability for the given *p* value.

Language	**Symbols**	**p**	**avg. word length (obs.)**	**avg. word length (exp.)**	**% difference**	**χ2**	**P(χ2)**
English	27	0.883	9.2	9.1	1%	2000	0
Russian	32	0.894	10.0	10.0	0%	16176	0
Spanish	33	0.886	9.8	9.3	5%	37215	0
German	31	0.893	11.7	9.9	18%	284122	0
French	27	0.894	10.1	10.0	1%	21204	0
Portuguese	27	0.892	9.9	9.8	1%	23786	0
Italian	22	0.899	9.9	10.4	5%	24482	0
Swahili	25	0.880	8.3	8.9	7%	5561	0
Afrikaans	31	0.880	10.1	8.9	14%	31508	0
Latin	24	0.908	10.9	11.4	5%	96491	0
Meroitic	24	0.809	6.4	5.8	10%	73	0.01



(a)

(b)


45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
80000
70000
60000
50000
40000
30000
20000
10000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
60000
50000
40000
30000
20000
10000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
60000
50000
40000
30000
20000
10000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
12000
10000
8000
6000
4000
2000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
250000
200000
150000
100000
50000
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49

(c) (d)

(e) (f)

(g) (h)

(i) (j)

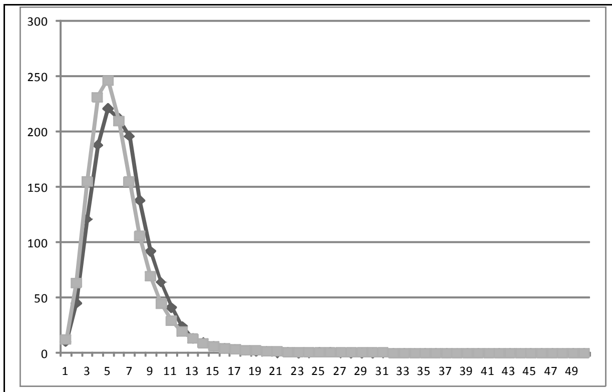
 (k)	

Figure 2: The distribution of distinct words by length from the spell check dictionaries (diamonds) and the theoretical distribution based on the *p* value in Table 1 (squares). The languages shown (a-k) are English, Russian, Spanish, German, French, Portuguese, Italian, Swahili, Afrikaans, Classical Latin, and Meroitic.

Table 2
The calculated value of the standard deviation of word length using equation 5 compared to actual measured values along with the percent difference.

Language	**Symbols**	**p**	**word length σ_w (obs.)**	**word length σ_w (exp.)**	**% difference**
English	27	0.883	9.7	9.7	0%
Russian	32	0.894	10.3	10.6	2%
Spanish	33	0.886	10.1	9.9	2%
German	31	0.893	12.2	10.4	17%
French	27	0.894	10.4	10.6	1%
Portuguese	27	0.892	10.2	10.4	2%
Italian	22	0.899	10.2	11.0	8%
Swahili	25	0.880	8.6	9.5	9%
Afrikaans	31	0.880	10.6	9.4	13%
Latin	24	0.908	11.2	11.9	7%
Meroitic	24	0.809	7.0	6.5	9%

Table 3
The calculated total vocabulary size along with the fitted parameters for equation 3 with a fitted $A = 7.45$.

Language	Symbols	p	b	Vocab. size (obs.)	Vocab. size (exp.)	% difference
English	27	0.883	0.118	118,619	118,619	0%
Russian	32	0.894	0.111	584,929	584,929	0%
Spanish	33	0.886	0.110	247,049	247,048	0%
German	31	0.893	0.114	532,276	532,274	0%
French	27	0.894	0.117	338,989	338,989	0%
Portuguese	27	0.892	0.117	261,798	261,798	0%
Italian	22	0.899	0.125	294,977	294,975	0%
Swahili	25	0.880	0.120	67,988	67,988	0%
Afrikaans	31	0.880	0.113	130,564	130,564	0%
Latin	24	0.908	0.121	1,243,950	1,243,949	0%
Meroitic	24	0.809	0.122	1,396	1,396	0%

The distribution from equation 1 tightly fits the measured counts of distinct words for all languages. In addition, the formulas of average word length, standard deviation, and total vocabulary work reasonably well for a parameter based on such a high level measure. Finally, it is interesting to note the relatively narrow range of p from 0.88 to 0.90 for all languages except Meroitic. The variable b also shows a relatively narrow range of values across languages. The shorter value for Meroitic may be partially due to the corpus separating certain bound morpheme suffixes into separate words for analysis (Smith 2007) as well as a small sample size. This likely demonstrates that language content, despite the numbers of letters or sounds, is quite similar in how words are transliterated into written language.

It must be mentioned there is an inaccuracy in the calculated distinct word length distribution from equation 1 as well as the estimate in equation 3. In particular, the equation overestimates the frequency count of large words. As an example, to estimate the largest word in a language, you can use equation 1 and set $W_n = 1$. Using the parameters for English, this estimates that the longest word in English is around 43 letters. In fact, the longest non-coined word in English, according to Oxford Dictionary, is *floccinaucinihilipilification*, at 29 letters. There is a technical term *pneumonoultramicroscopicsilicovolcanoconiosis* at 45 letters, however, this is an extreme exception. Equation 1 estimates English should have 13 different words of 29 letters each. In the WinEdt spellcheck dictionary used for the empirical distinct word counts, the longest single word is 29 letters but you must drop to 24 letter words to find more than 10 distinct words. This pattern is repeated in other languages.

So equation 1 should be used with caution when calculating the frequency count of distinct words significantly larger than the mean word length. In particular, though accuracy varies by language, typically past one standard deviation greater than the mean, the frequency counts become very inaccurate and overestimate the number of distinct words. However, because words of this length are relatively rare compared to the size of the vocabulary, the overall fit for the distribution can still be strong despite these inaccuracies. For example in English words of length greater than the mean plus one standard deviation (19 letters or greater), account for about 0.5% of all distinct words.

In the next section, we will look at estimates of the distinct word count in a very different manner borrowing from the tools of information theory.

3. Information Theory and Linguistics

Claude Shannon, not content to sit on his laurels after his 1948 magnum opus *A Mathematical Theory of Communication*, turned his attention to the applications of information theory to human language representations, particularly written English. In (Shannon 1951), he used both statistical analysis of text and best guesses by volunteers of the next letters in fragmented texts to estimate the entropies of different orders for letters, including spaces. In (Shannon 1951) the entropy of a given order is defined as the conditional entropy of a given letter coming after an n-gram sequence.

$$(6) \qquad H_N = -\sum\nolimits_{i,j} p(b_i\,, j) log_2 p_{b_i}(j)$$

In the above, $p(b_i, j)$ is the probability of the n-gram b_i, j and $p_{bi}(j)$ is the conditional probability of the letter j after the sequence b_i. Note, it is a common source of confusion that this is the conditional entropy for an n-gram based on n-1 symbols, not the joint entropy of n symbols.

His effort was rapidly replicated amongst many other written languages from all parts of the world. These include German (Küpfmüller 1954; Söder 1999), Russian (Lebedev & Garmaš 1959), French (Petrova et. al. 1964), Italian (Manfrino 1960; Capocelli & Ricciardi 1980), Arabic (Wanas et. al. 1976), Brazilian Portuguese (Manfrino 1970; Gomes 2007), Farsi (Darrudi, Hejazi, Oroumchian 2004) and Spanish (Guerrero & Perez 2008; Guerrero 2009). A great overview of some of these studies and the research in general is given by (Yaglom, Yaglom 1983). In addition, the symbol entropy has been used to analyze some other non-human communication repertoires such as black capped chickadees (Hailman et. al. 1985), dolphin whistles (McCowan et. al. 1999; McCowan et. al. 2002; Ferrer-i-Cancho & McCowan 2009), and humpbacked whale sounds (Doyle et. al. 2008). The additional languages of Afrikaans, Swahili, Classical Latin, and Meroitic have also been computed by the author for this paper. The Afrikaans and Swahili Corpora are due to (Scannell 2007; Roos 2009), the Latin Corpora use a selection of classical Latin texts from the Latin Library (http://www.thelatinlibrary.com) and the Meroitic corpus was constructed from transliterated texts from the *Répertoire d'épigraphie méroïtique* (REM) (Leclant, 2000). The entropy of Afrikaans was calculated using a 26 letter Latin alphabet plus a space and the accented characters ô, ê, ë, á for a total of 31 symbols. Swahili used 23 characters plus the character 'ch' and space for a total of 25 symbols. Latin used the 23 symbols of the classical Latin script plus spaces (24 characters) and Meroitic used the 24 characters of the alphabet plus the word separator character (25 characters).

In each case, the entropies, summarized in Table 5, are both a measure of the order and redundancy for a given n-gram. A lower n-gram entropy indicates increased redundancy and the n-gram entropy at any order n >1 must be less than or equal to the n-gram entropy of a lower order.

4. Word length combinatorics

In this section, we will look at the distinct word distribution from another perspective, that of the entropy of the characters. This perspective can be useful since it allows us to relate the number of distinct words to the structure of the language as defined in information theory.

Take an imaginary language with L different letters. It is well known that the number of possible combinations (W) in a string of length N is given by $W=L^N$. Granted, these are called strings instead of words because this basic equation gives no rules or bounds on which letters appear together or how many times. Real languages are much more constrained. A more accurate estimate was introduced by (Weaver & Shannon 1963) where assuming an entropy of $H(l)$ in bits, the number of possible N length strings of L letters is now

(7) $$W = 2^{NH(l)}$$

It may seem puzzling that L appears nowhere in this equation but the fact that the base is 2 and the entropy is in bits obviates its necessity. If one wants to keep L, they can find the same result from the below equation calculating $H(l)$ with a logarithm of base L:

(8) $$W = L^{N\,H_L(l)}$$

Again, this estimate is much reduced given that this first order entropy is usually much less than the zero order entropy $H_0=log\ L$ that returns the base equation $W=L^N$. However, can we do even better? In particular, can we fine tune our approach so that we have a relatively accurate estimate for the number of possible strings, or words, for a given word length?

The relationships between entropy, mutual information, and conditional entropy can be clearly elucidated by Venn Diagrams (Reza 1961) where

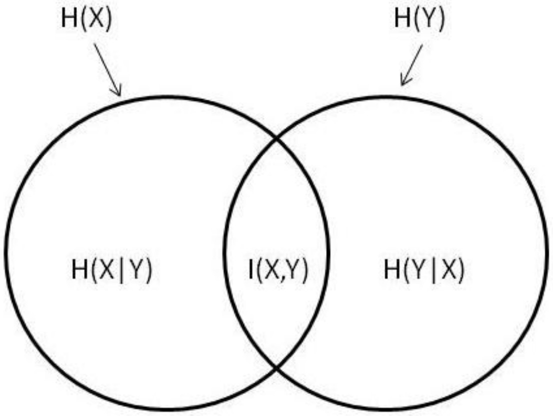


Figure 3: Venn Diagrams of entropies and mutual information

Given two symbols, the number of possible sequences of length N is the total possible number given the entropy divided by those by mutual information

(9) $$W = \frac{2^{NH(I)}}{2^{NI}} = 2^{NH(X|Y)}$$

This was an analysis first stated by Kolmogorov (Kolmogorov 1965). So the number of possible digrams given the second order conditional entropy is given by

(10) $$W_2 = 2^{2H(X|Y)}$$

This can be increased for any sequence of any length, however, it becomes correspondingly inaccurate as you use equation 9 for more than two symbols. For longer sequences, the total number of possible words can be more closely estimated by the higher order conditional entropies

(11) $$W_N = 2^{NH_N}$$

Of course, the “words“ will depend on how the entropies are defined. If you include spaces or punctuation as symbols, this can skew the entropy and the value of $W_{N.}$ However, for consistency with past studies, all entropies in this paper, including those in Table 5, are based only on the entropies of letters used in word formation plus the space character. Another issue, to be discussed later, is the accuracy of higher order entropies for a given text sample and how this can effect vocabulary size estimates.

Table 5
Conditional entropies up to the third order for a selection of languages.

			Entropy Order (Conditional Entropy)			
Language	**Source**	**Characters**	**0**	**1**	**2**	**3**
English	Shannon (1951)	27	4.75	4.14	3.56	3.30
English	Schürmann and Grassberger (1996)	27	4.75	4.08	3.32	2.73
Russian	Lebedev & Garmaš (1959)	32	5.00	4.35	3.52	3.01
Spanish	Guerrero & Perez (2008); Guerrero (2009)	33	5.04	4.15	3.56	3.09
German	Söder (1999)	31	4.95	4.06	3.62	3.25
French	Petrova (1964)	27	4.75	3.95	3.17	2.83
Portuguese	Gomes (2007)	27	4.75	3.94	3.56	3.27
Portuguese	Manfrino (1970)	23	4.52	3.91	3.35	3.20
Italian	Capocelli and Ricciardi (1980)	22	4.46	3.96	3.53	3.22
Italian	Manfrino (1960)	21	4.39	3.90	3.32	2.76

Swahili	Smith - this paper (2012)	25	4.64	3.95	3.33	2.82
Afrikaans	Smith - this paper (2012)	31	4.95	4.02	3.44	2.77
Classical Latin	Smith - this paper (2012)	24	4.58	3.90	3.24	2.79
Meroitic	Smith - this paper (2012)	24	4.58	4.24	3.10	

5. Results of the Entropy Method

In Table 6, a comparison of the predicted vocabulary given conditional entropy and actual numbers of distinct words of lengths two and three are given. The same corpus of Meroitic that provided the distinct words provides both the first and second order entropies. The corpus is too small to sample for third order entropy since the number of tokens is only slightly higher than 1,000 and much less than 24^3 = 13,824.

Table 6
The predicted number of distinct tokens of length two and three determined from the conditional entropies in Table 5 and the actual number of distinct words of length two and three from WinEdt iSpell spell check dictionaries.

			Calculated n-grams		**Dictionary n-grams**	
Language	**Source**	**Characters**	**2**	**3**	**2**	**3**
English	Shannon (1951)	27	139	955	93	754
English	Schürmann and Grassberger (1996)	27	100	292	93	754
Russian	Lebedev & Garmaš (1959)	32	132	523	87	995
Spanish	Guerrero & Perez (2008); Guerrero (2009)	33	139	617	64	300
German	Söder (1999)	31	151	861	164	546
French	Petrova (1964)	27	81	360	165	497
Portuguese	Gomes (2007)	27	139	898	76	338
Portuguese	Manfrino (1970)	23	104	776	76	338
Italian	Capocelli and Ricciardi (1980)	22	133	809	164	642
Italian	Manfrino (1960)	21	100	311	164	642
Swahili	Smith - this paper (2012)	25	101	352	95	557
Afrikaans	Smith - this paper (2012)	31	118	317	93	598
Classical Latin	Smith - this paper (2012)	24	89	331	73	465
Meroitic	Smith - this paper (2012)	24	74		45	121

By comparison of Table 6, it is clear that most estimates for the number of distinct tokens in the language are if not close, of the relatively same magnitude. One of the largest issues is that the predicted value is very sensitive to the value of NH_N so that variances in calculating the entropy, for example from the two Italian and Brazilian Portuguese examples, can lead to relatively large differences in the estimated numbers of distinct words.

Another possible revelation is that the higher order entropies can be equated with the expression from Section 2 as

$$(12) \quad W_N \approx 2^{NH_N} \approx L^{Np^N}$$

and by extension

$$(13) \quad H_N \approx p^N \frac{\ln L}{\ln 2}$$

Equation 13 is an estimate derived by reducing the structure of higher order entropies from *N* classes into 2, one with probability *p* and the other with probability *1-p*. It is obviously only an approximation and having a single value of *p* will neglect some of the conditional probability structure of letters and spaces. However, it is an interesting connection for future work.

6. Reverse Estimate of Text Entropies

While the size of texts may preclude calculating higher order entropies, the techniques outlined in the previous section allow us to back out estimates of these higher order entropies based on the vocabulary size by word length. In particular, applying the reverse transformation to equation 11 we have

$$(14) \quad H_N = \frac{log_2 W_N}{N}$$

Applying this to the distinct words by length of English and German, we can estimate the higher order entropies as shown in Table 7.

Table 7
Higher order entropies estimated from the distinct words by word length.

Order	English Words	Implied Entropy	German Words	Implied Entropy
2	93	3.27	164	3.68
3	754	3.19	546	3.03
4	3027	2.89	4323	3.02
5	6110	2.52	10486	2.67
6	10083	2.22	19092	2.37
7	14424	1.97	27574	2.11
8	16624	1.75	38933	1.91
9	16551	1.56	52212	1.74
10	14888	1.39	60596	1.59
11	12008	1.23	63115	1.45
12	8873	1.09	59232	1.32

13	6113	0.97	49708	1.20
14	3820	0.85	39908	1.09
15	2323	0.75	30678	0.99
16	1235	0.64	22897	0.91
17	707	0.56	16978	0.83
18	413	0.48	11883	0.75
19	245	0.42	8158	0.68
20	135	0.35	5584	0.62
21	84	0.30	3684	0.56
22	50	0.26	2398	0.51
23	23	0.20	1557	0.46
24	16	0.17	974	0.41
25	9	0.13	633	0.37
26	4	0.08	386	0.33
27	2	0.04	237	0.29
28	1	0.00	128	0.25
29	0	0.00	95	0.23
30	0	0.00	44	0.18

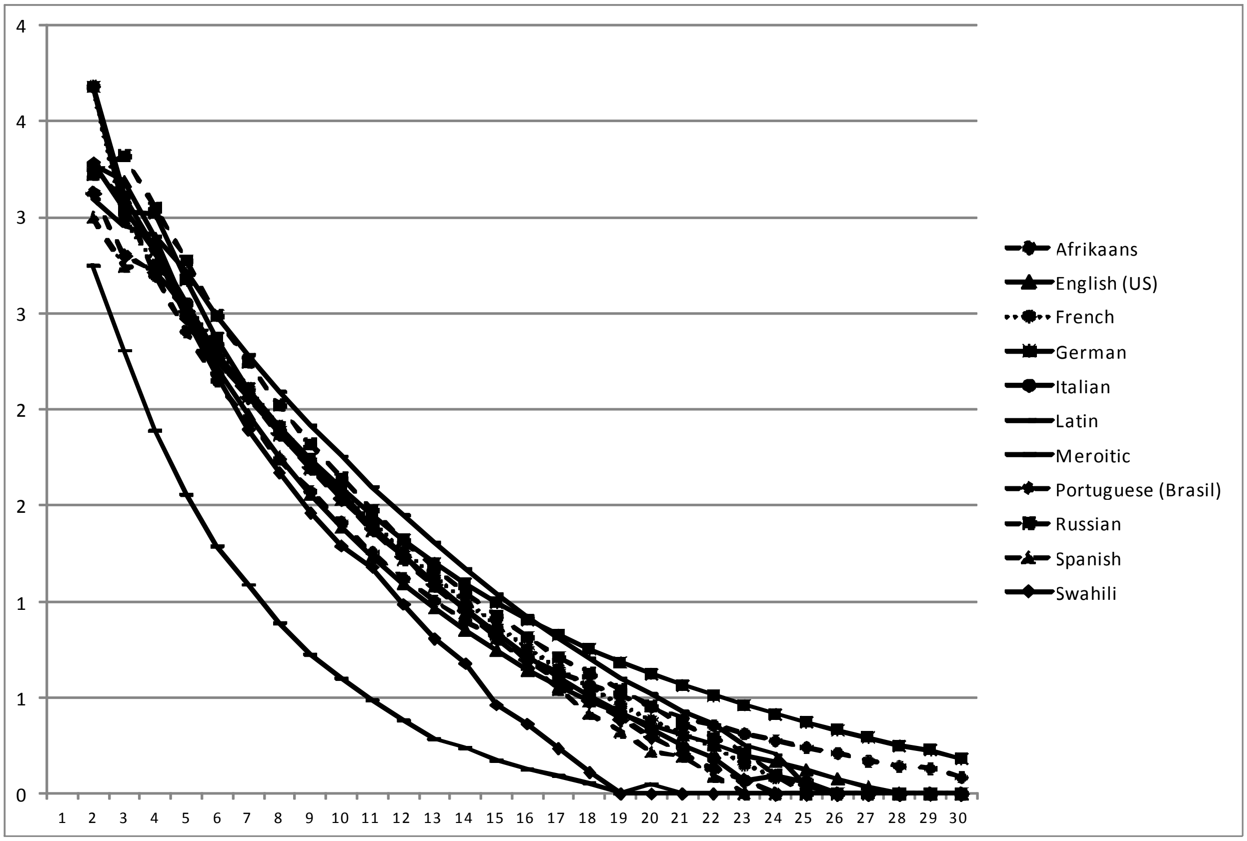


Figure 4: Graph of estimated higher order entropies by language. Meroitic is noticeably lower, likely due to the small sample size of the actual vocabulary.

Discussion

The results previously discussed have presented a two new lines of inquiry in the relationship between the number of distinct words of a given length and the underlying languages. First, we show that a simple distribution similar based on the number of distinct characters and the probability of a given character being a letter or a space can closely approximate the empirical distinct word length frequency distribution. Second, a relationshp between n-gram entropies and the number of distinct words of written languages was explained. Though it cannot be exact, the conditional entropy can provide a useful tool to estimate the scope of a given language by word length without a full sample of the vocabulary. Conversely, a knowledge of distinct words by length can possibly be used to estimate higher order entropies whose calculations are rendered difficult due to the massive size of the required corpus. Granted, the results can depend on which characters are used to calculate the n-gram entropy, whether spaces or punctuation are included, and other factors. One problem in the paper that is difficult to reconcile is error estimates since letter or word frequencies do not converge with a larger sample size but fluctuate as is typical in systems with large numbers of rare events (LNRE) (Baayen 2001).

References

Altmann, G. (1988). *Wiederholungen in Texten*. Bochum: Brockmeyer.

Baayen, R.H. (2001). *Word Frequency Distributions*. Dordrecht: Kluwer.

Best, K-H. (1998). Results and perspectives of the Göttingen project on quantitative linguistics. *Journal of Quantitative Linguistics* 5, 155 – 162.

Best, K.-H. (ed.) (2001). *Häufigkeitsverteilungen in Texten.* Göttingen: Peust & Gutschmidt Verlag.

Capocelli, R.M., Ricciardi, L.M. (1980). The entropy of written Italian. In: *A Selection of Papers from INFO II, the Second International Conference on Information Sciences and Systems, University of Patras, Greece, July 9-14, 1979, 96-100*. Dordrecht: D. Reidel.

Darrudi, E., Hejazi, M.R., Oroumchian, F. (2004). Assessment of a Modern Farsi Corpus. In: *Proceedings of the 2nd Workshop on Information Technology & its Disciplines (WITID'04), ITRC, Kish Island, Iran*.

Doyle, L.R. et. al. (2008). Applicability of Information Theory to the Quantification of Responses to Anthropogenic Noise by Southeast Alaskan Humpback Whales. *Entropy 10,* 33-46.

Ferrer-i-Cancho, R., McCowan, B. (2009). A Law of Word Meaning in Dolphin Whistle Types. *Entropy 11(4),* 688-701.

Gomes, L. (2007). Entropia e Automatização de Séries de Aproximação da Língua Portuguesa. Thesis *Departamento de Sistemas Computacionais - Escola Politécnica de Pernambuco*.

Grzybek, P. (ed.) (2006). *Word Length Studies and Related Issues*. Dordrecht: Springer.

Guerrero, F.G., Perez, L.A. (2008). A software for learning Information Theory basics with emphasis on entropy of Spanish. *Revista Energia y Computacion 16(1),* 58-64.

Guerrero, F.G. (2009). On the entropy of written Spanish. *preprint arXiv:0901.4784*.

Hailman, J.P., Ficken, M.S., Ficken, R.W. (1985). The “chick-a-dee” calls of parus atricapillus: A recombinant system of animal communication compared with written English. *Semiotica 56,* 191-224.

Kolmogorov, A.N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission 1,* 4-7.

Küpfmüller, K. (1954). Die Entropie der deutschen Sprache. *FTZ-Fermeldetechnische Zeitschrift 6* , 265-272.

Leclant, J. (2000). *Répertoire d'épigraphie méroïtique: corpus des inscriptions publiées, Volumes 1-3*. Paris: Diffusion de Boccard.

McCowan, B., Hanser, S., Doyle L.R. (1999). Quantitative tools for comparing animal communication systems: information theory applied to bottlenose dolphin whistle repertoires. *Animal Behaviour* 57, 409-419.

McCowan, B., Doyle, L.R., Hanser, S.F. (2002). Using information theory to assess the diversity, complexity and development of communicative repertoires. *Journal of Comparative Psychology 116,* 166-172.

Lebedev, D., Garmaš, V.A. (1959). Statističeskij analiz trechbukvennych sočetanij russkogo teksta. In: *Problemy peredači informacii 2, 78-80*. Moscow: Nauka.

Manfrino, R.L. (1960). L'entropia della lingua italiana ed il suo calcolo(The entropy of the Italian language and its computation). *Altafrequenza 29(1),* 4-29.

Manfrino, R.L. (1970). Printed Portugese (Brazilian) entropy statistical calculation. *IEEE Transactions in Information Theory IT-16,* 172.

Petrova, N.V., Piotrovski, R.G., Giraud, R. (1964). The entropy of written French. *Bulletin of Society of Linguistics of Paris 58,* 130-152.

Reza, F.M. (1961). *An Introduction to Information Theory*. New York: McGraw-Hill.

Roos, D. (2009). Translation features in a comparable corpus of Afrikaans newspaper articles. *Stellenbosch Papers in Linguistics PLUS 39,* 73-83.

Rothschild, L. (1986). The distribution of English dictionary wordlengths. *Journal of Statistical Planning and Inference 14*, 311–322.

Scannell, K. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In: Fairon, C., Naets, H., Kilgarriff, A., de Schryver G-M. (eds.) *"Building and Exploring Web Corpora", Proceedings of the 3rd Web as Corpus Workshop in Louvain-la-Neuve, Belgium, September 2007, 5-15.*

Schmidt, P. (ed.) (1996). *Glottometrika 15*: *Issues in General Linguistic Theory and the Theory of Word Length.* Trier: WVT.

Schürmann, T., Grassberger, P. (1996). Entropy estimation of symbol sequences. *CHAOS 6(3),* 414-427.

Shannon, C.E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal 30,* 50–64.

Smith, R. (2007). Investigation of the Zipf-plot of the extinct Meroitic language. *Glottometrics15,* 53-61.

Söder, G. (1999). *Wertdiskrete Informationstheorie.* Munich: Munich Technical University, pp. 23-29.

Wanas, M., Zayed, A., Shaker, M., Taha, E. (1976). First second- and third-order entropies of Arabic text. *IEEE Transactions on Information Theory 22 (1),* 123.

Weaver, W., Shannon, C.E. (1963). *The Mathematical Theory of Communication.* Urbana-Champaign: Univ. of Illinois Press.

Williams, C.B. (1970). *Style and Vocabulary: Numerical Studies.* London: Griffin.

Wimmer, G., Altmann, G. (1996). The theory of word length: some results and generalizations. *Glottometrika 15*, 112 – 133.

Yaglom, A.M., Yaglom, I.M. (1983). *Probability and Information*. Dordrecht: D. Reidel Pub. (Translated by V.K. Jain).

Glottometrics 23, 2012, 23-55

Aspects of nominal style

Sven Naumann, Trier
Ioan-Iovitz Popescu, Bucharest
Gabriel Altmann, Lüdenscheid

Abstract. Using the sequence of nouns and verbs different methods are applied to show some aspects of analysis of text nominality. Though nouns can be opposed also to adjectives in order to show the ornamentality of texts, we adhered to the classical duality. The article uses 85 texts in 9 languages, shows the commonalities of text-sorts, differences in style and authors, and presents regression, runs and their empirical distributions as well as the positioning of texts in the Ord-scheme. The study gives merely some stimuli for further research.

Keywords: sequences, style, text sort, ord-scheme

Style may be studied from as many points of view as we are able to devise. However, text-books on the art of composition usually contain only aspects that are conspicuous to the literary scientist avoiding any kind of testing the relevance of the given conspicuousness. A text may be ornamental, sentimental, active, lyric, formal, official, juridical, nominal, etc. because it contains a certain percentage of expressions that may be classified as such. One shows that a lyrical text “strongly differs” from a juridical text because it contains some lyric expressions not present in a juridical text. This way of study is, however, mostly a search for idiosyncrasies or mutual differences. But idiosyncrasies should be mirrored as a deviation from a common/neutral background; unfortunately, there is no common background of texts. All texts have something that is unique for them and distinguishes them from other texts.

In general, there are two possibilities of studying style:

(a) The *static approach* shows that the proportion of some entities significantly differs from that in other texts or from a neutral background. The artificial neutral background can be taken in some cases from a very mixed corpus and a test must show that the difference of the given proportion in the given text significantly differs from that in the corpus. One can set up intervals around the neutral background and classify texts according to their place in this ordering; or one can order the texts according to the degree of the given property and set up intervals *a posteriori*. The latter way does not yield a final result and the creation of intervals is a matter of decision.

(b) The *dynamic approach* considers the occurrence of the pertinent entities along the text. For example, an entity may occur more often in the beginning part of the text but its frequency may decrease towards the end. We may ask what kinds of sequences there are in texts, are there tendencies other than linear, is there a correlation with the contents and aim of the text, is the given trend linked with some other trends, etc.

The problem of nominal style has been studied from different points of view especially in German (for references cf. Ziegler et al. 2002). Different languages have different techniques to change a verbal expression into a nominal one or to add a noun in order to give the sentence a special form. Enumeration of objects increases nominality.

In the present article we shall scrutinize the nominal style defined in its static form as the ratio of nouns and verbs and in its dynamic form both as the number of nouns preceding the *i*-th verb and the development of the Busemann ratio. At last, we add an elementary analysis of runs of nouns and verbs.

1. The dynamic view

With the *dynamic view* we proceed as follows: using a text, the sequence of nouns (N) and verbs (V) will be extracted (all other parts-of speech are omitted). This can be done also using a lemmatizing program. We obtain for example a sequence NNVNVNNV where there are two *N* up to the first *V*, three *N* up to the second *V*, and five *N* up to the third *V*. The sequence can be written as

V	1	2	3
N	2	3	5

If the last symbol is *N*, we set behind it a virtual *V* in order to take the last nouns into account. The sequence NNVNVNNVN would then have the form

V	1	2	3	4
N	2	3	5	6

The sequence displays a non-decreasing trend. The variable *V* begins with 1, the variable *N* may begin with 0.

Let us consider first a prosaic text by Eva Bachletová "Moja Dolná zem" written in Slovak, in which the author describes the impressions of her visit to her native village in Hungary. We obtain the symbolic sequence

N,N,V,N,N,V,N,N,N,N,N,N,N,N,N,N,V,N,N,N,N,N,V,N,V,V,N,N,N,V,N,N,V,N,V,N,
N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,N,V,
N,N,N,V,N,N,N,N,N,N,N,N,N,N,N,V,N,N,V,N,N,V,N,N,V,V,V,V,V,N,N,V,V,N,V,V,
N,N,N,V,N,V,N,N,V,N,N,N,V,N,N,V,N,N,N,N,V,N,V,N,V,N,V,N,N,N,N,N,V,N,N,N,
N,N,V,N,N,N,V,N,V,N,N,V,N,N,V,N,N,N,V,N,N,N,V,N,V,N,V,N,N,N,N,V,N,N,N,N,
N,V,N,V,N,N,N,N,N,N,N,V,V,N,N,V,N,V,V,N,N,N,N,N,V,N,N,N,V,N,V,V,N,N,N,N,
N,N,V,V,N,V,N,V,N,V,V,V,N,N,V,N,V,N,N,N,V,V,N,N,V,N,V,N,N,N,N,V,N,V,N,N,
N,N,N,V,N,N,V,N,V,N,V,N,V,N,N,N,V,N,N,V,N,V,N,N,N,V,N,V,N,N,N,V,N,V,N,V,
N,N,N,V,N,N,N,N,V,N,N,V,N,V,N,V,N,N,N,N,V,N,V,N,N,V,N,N,N,V,N,N,V,N,V,V,
N,N,V,N,V,N,N,N,V,N,N,V,V,N,N,V,V,N,V,N,N,V,V,N,N,N,V,N,N,V,V,N,N,V,N,N,
N,N,N,N,V,N,V,N,N,V,V,N,N,V,N,V,N,N,N,N,N,N,N,N,V,V,N,N,N,V,N,V,N,V,N,N,
N,V,V,N,V,N,N,N,V,N,V,N,N,V,N,

from which a numerical sequence presented in Table 1 can be constructed. It is shown in Figure 1 in graphical form. As can be seen, the sequence is in its full length nominal: all points are situated over the bisector $x = y$ shown as a full line. In order to capture the given course one can use different formulas. Here, the simplest way is to fit the power function $y = ax^b$ yielding in our case

$$N = 9{,}8349V^{0{,}6941}$$

with the determination coefficient R^2 = 0,9934. Needless to say, there are slightly "better" functions but all contain more than two parameters. The fact that the empirical data steadily move away from the bisector shows the strengthening of the nominalization from beginning to the end of the text. The great jump after *V* = 9 is a characteristic feature of this text. At this place there is a long enumerative passage consisting of nouns.

Table 1
The *V*-*N* sequence in E. Bachletová's "Moja Dolná zem"

V	N	V	N	V	N	V	N	V	N
1	2	26	91	51	154	76	196	101	240
2	4	27	94	52	157	77	197	102	241
3	14	28	96	53	158	78	200	103	243
4	19	29	100	54	158	79	201	104	243
5	20	30	101	55	164	80	204	105	246
6	20	31	102	56	164	81	205	106	248
7	23	32	103	57	165	82	206	107	248
8	25	33	108	58	166	83	209	108	250
9	26	34	113	59	167	84	213	109	256
10	62	35	116	60	167	85	215	110	257
11	65	36	117	61	167	86	216	111	259
12	76	37	119	62	169	87	217	112	259
13	78	38	121	63	170	88	221	113	261
14	80	39	124	64	173	89	222	114	262
15	82	40	127	65	173	90	224	115	270
16	82	41	128	66	175	91	227	116	270
17	82	42	129	67	176	92	229	117	273
18	82	43	133	68	180	93	230	118	274
19	82	44	138	69	181	94	230	119	275
20	84	45	139	70	186	95	232	120	278
21	84	46	146	71	188	96	233	121	278
22	85	47	146	72	189	97	236	122	279
23	85	48	148	73	190	98	238	123	282
24	88	49	149	74	191	99	238	124	282
25	89	50	149	75	194	100	240	125	285
								126	286

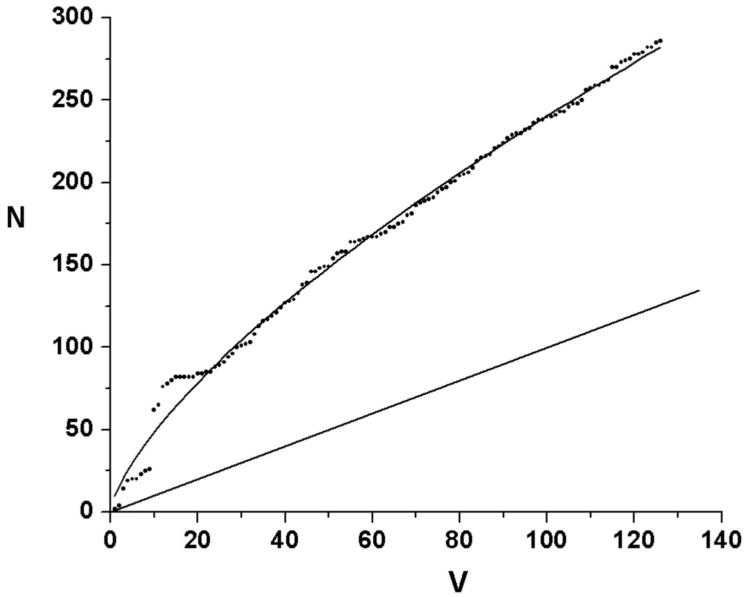


Figure 1. *N-V* sequence in *Moja Dolná zem* by E. Bachletová

In order to see whether the author prefers nominal style, we compare some of her other prosaic and poetic texts as shown in Table 2. The increase of nouns can in all cases be captured by the power function $y = ax^b$.

Table 2
Sequences in some prosaic and poetic text by E. Bachletová

Text	**a**	**b**	**R^2**
prose			
Moja Dolná zem	9,8349	0,6941	0,99
Jednoduché bytie	2,6218	0,9158	0,98
Nový obraz	0,9401	1,2152	0,95
Čas pre nádych	2,1040	0,8670	0,99
Jazvy prítomnosti	2.1130	1,0399	0,99
poetry			
Aby spriesvitnela	2,6552	0,9477	0,93
Čakáme šťastie	3,2962	0,8276	0,93
Dielo Stvoriteľa	2,4579	0,9851	0,98
Podobnosť bytia	1,5389	1,0920	0,98
Ťažko pokoriteľní	3,0726	0,3561	0,79
Z neba do neba	2,1021	0,9156	0,87

Parameter *a* shows that texts begin rather with nouns - a quite usual phenomenon in Slavic languages. The parameter *b* signals convexity if it is > 1. That means, nominality increases in the course of text. For *b* < 1 the curve is concave, but there is only one case in which it crosses the bisector (*b* = 0,3561).

The fact that there is a simple regularity - power function controlling the relation of *N* and *V* - automatically calls forth the question, whether this is a general phenomenon or merely a speciality of the given writer. In order to get a better image, we first consider poems of another Slovak writer and some press texts as well as some texts in German, Hungarian and Romanian. Thereby we obtain a wider view.

Table 3
Poetic and press texts

Text	a	b	R^2
Slovak poetry			
R. Dilong, Nostalgia	2,6042	0,8816	0,9981
R. Dilong, Jesenná	3,8357	0,7503	0,9682
R. Dilong, Žobrák	1,4749	1,0639	0,9770
R. Dilong, Minieme sa	3,1081	0,7459	0,9677
R. Dilong. Vo vetre myšlienky	0,9912	1,0559	0,9885
Slovak press texts			
Nenásytný (Plus 7 dní 22.12.2011)	3,7128	0,7642	0,9618
Národná vdova (Plus 7 dní 29.12.2011)	1,7025	1,1434	0,9726
German press texts			
Trittbrettfahrer der Arisierung (299 TAZ Nr. 5732 11.01.1999)	3.9824	0,8759	0,9930
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999)	1,5221	1,0556	0,9973
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999)	0,8633	1,1414	0,9960
Fortschritt findet anderswo statt TAZ Nr. 5732 11.01.1999	2,7993	0,9605	0,9877
Kontrolle der Kapitalströme emerging markets - Ein Gebot der Demokratie (Le Monde diplomatique Nr. 5736 15.01.1999)	2,3973	0,9734	0,9946
Ethnische Definitionen als Machtpolitik (Le Monde diplomatique Nr. 5736 15.01.1999)	2,6437	0,9492	0,9970
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999)	5,3854	0,8843	0,9944
IRAK : DIE UNO IM ABSEITS (Le Monde diplomatique Nr. 5736 15.01.1999)	1,5204	1,0115	0,9959
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999)	4.7248	0,8063	0,9959
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999)	4,1650	0,8054	0,9925
Die Bremer Kinotaz ... alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999)	5,1797	0,9193	0,9974
Strategische Spielhallen im Keller	1,2557	1,0927	0,9963

(TAZ Nr. 5735 14.01.1999)			
Klick, klick, klick (TAZ Nr. 5735 14.01.1999)	5,1081	0,8126	0,9957
German fable			
Pestalozzi, Stoffels Brunnen	1,8708	0,7376	0,9695
Hungarian poems			
Petöfi, S., Szeptember végén	3,3323	0,7418	0,9711
Ady E., Góg és Magóg fia vagyok én	1,5836	0,9203	0,9783
Romanian poetry			
M. Eminescu, De ce nu-mi vii?	0,4943	1,3092	0,9799
M. Eminescu, Mai am un singur dor	1,3523	1,0872	0,9881
M. Eminescu, Atât de fragedă...	3,9550	0,7893	0,9736
M. Eminescu, Pe lângă plopii fără soț...	0,4004	1,3868	0,9937
M. Eminescu, Floare-albastră	3,0769	0,8633	0,9926

This view is relatively uniform, there are no surprising exceptions, the texts lie almost on a straight line. The differences arise because the texts are short. Evidently, nouns and verbs are necessary parts of the sentence and the frequency of nouns is greater because they serve as subjects, objects and attributes of other nouns. Nevertheless, there are texts in which the empirical points may cross the bisector or be mostly under it. In J.H. Pestalozzi's short tale "Stoffels Brunnen" we can find the sequence $y = 1,8708x^{0,7376}$ with $R^2 = 0,9695$ which is concave and the majority of the sequence is placed below the bisector (cf. Popescu, Mačutek, Altmann 2009:176 f.). This construction arises evidently if *a* is small and *b* < 1. Hence the text is characterized by the relation of *a* to *b*. The result can be seen in Figure 2.

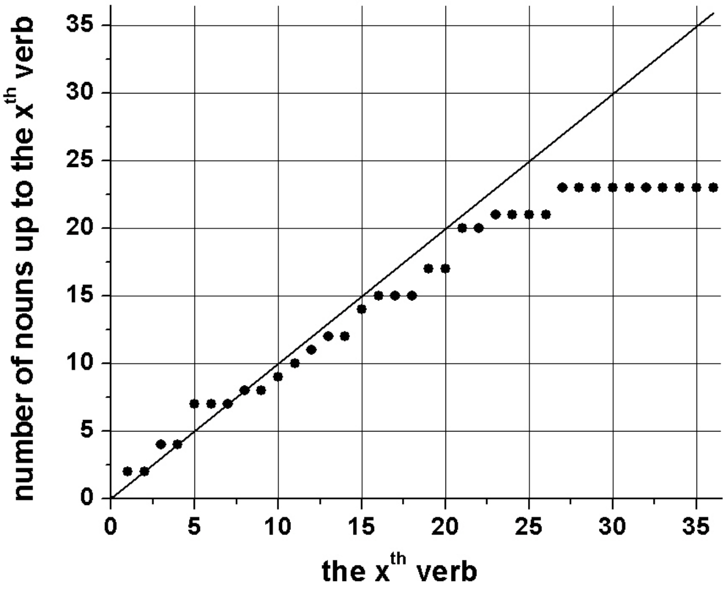


Figure 2. The N-V sequence in J.H.Pestalozzi's fable *Stoffels Brunnen* (from Popescu, Mačutek, Altmann 2009)

2. Static approach

However, this all has been done without a previous hypothesis concerning the occurrence of nouns and verbs in texts without their placing. Hence we scrutinize the situation statically. Looking at an Indo-European language we can expect that “every” sentence contains a noun and a verb, at least in written language. However, nouns *can* have attributes represented by other nouns joined with the main noun by a genitive or as a possessive or by means of a preposition, etc. And also verbs *may* have objects represented by nouns, or necessitate adverbial complements containing nouns, too. Since the existence of a noun after noun or after verb is not categorical, a simple conjecture leads to the conclusion that in a “normal” text the proportion of verbs in a NV-sequence can be about 0,3333. If it is greater, the text has the tendency to be more verbal; if it is smaller, the text has the tendency to be nominal. Considering 0,3333 as the expected value of p with variance $p(1-p)/K$ where K is the number of verbs and nouns ($K = N + V$) in the text, we can test the deviation of observed $p(V)$ from its expectation using the normal criterion

$$(1) \qquad u = \frac{p(V) - E(p)}{\sqrt{p(1-p)/(V+N)}} = \frac{p(V) - 0{,}3333}{\sqrt{0{,}2222/K}}$$

where p is the general proportion of verbs, $E(p)$ is its expectation and $p(V)$ is the observed proportion of V in text. The alternative hypothesis is $H_1 \neq 0{,}33$. Now, if $u < -1{,}64$, the text tends significantly to nominality; if $u > 1{,}64$ the text tends significantly to verbality. Of course, in different languages, the basic H_0-hypothesis might be quite different. Even texts of a unique author may differ in dependence on aim and theme. Here the hypotheses are: H_0: $p = 0{,}3333$; H_1: $p \neq 0{,}3333$.

For the above Slovak text we obtain $V = 125$ (with the dynamic approach, the last V has been added), $N = 286$, hence $p(V) = 125/411 = 0{,}3041$. Using the u-criterion we obtain

$$u = \frac{0{,}3041 - 0{,}3333}{\sqrt{0{,}2222/411}} = -1{,}2558$$

which shows that the text tends to heightened nominality (= too small proportion of verbs) but does not significantly deviate from the expectation.

Table 4
The verbal/nominal tendency of texts
(V - Number of verbs, N - number of nouns, Po - poetry, Pr - prose, J - press text)

Text	V	N	p(V)	u
Slovak texts				
E. Bachletová, Moja Dolná zem (Pr)	125	286	0,3041	-1,2558
E. Bachletová, Aby spriesvitnela (Po)	10	22	0,3125	-0,2496
E. Bachletová, Dielo Stvoriteľa (Po)	17	44	0,2786	-0,9049
E. Bachletová, Čakáme na šťastie (Po)	8	19	0,2963	-0,4079

E. Bachletová, Čas na nádych (Pr)	53	65	0,4492	2,6698*
E. Bachletová, Jazvy prítomnosti (Pr)	74	188	0,2824	-1.7463*
R. Dilong, Nostalgia (Po)	142	209	0,4029	2,7606*
R. Dilong, Jesenná (Po)	35	62	0,3608	0,5751
R. Dilong, Žobrák (Po)	34	60	0,3617	0,5841
R. Dilong, Minieme sa (Po)	25	33	0,4310	1,5790
R. Dilong, Ranený (Po)	14	20	0,4118	0,9706
R. Dilong, Vo vetre myšlienky (Po)	27	34	0,4426	1,8114*
Nenásytný. Plus 7 dni 22.12.2011: (J)	49	68	0,4188	1,9620*
Hungarian texts				
S. Petöfi, Szeptember végén (Po)	26	38	0,4063	1,2381
J. Arany, A lepke (Po)	35	44	0,4430	2,0692*
E. Ady, Góg és Magóg fia vagyok én…(Po)	17	23	0,4250	1,2303
E. Ady, A mi gyermekünk (Po)	21	22	0,4884	2,1572*
E. Ady, A könnyek asszonya (Po)	20	27	0,4255	1,3414
E. Ady, Ima Baál istenhez (Po)	50	90	0,3571	0,5985
German texts				
Ethnische Definitionen als Machtpolitik. Le Monde Diplomatique 5736, 15.11.1999 (J)	293	605	0,3263	-0,4462
Fortschritt findet woanders statt (TAZ 5732, 11.1.1999): (J)	193	417	0,3164	-0,8858
Trittbrettfahrer der Arisierung (TAZ Nr. 5732 11.01.1999) (J)	165	374	0,3061	-1,3385
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999) (J)	185	376	0,3298	-2,0528*
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999) (J)	234	436	0,3493	0,8760
Kontrolle der Kapitalströme emerging markets - Ein Gebot der Demokratie (Le Monde diplomatique diplomatique Nr. 5736 15.01. 1999) (J)	283	655	0,3017	-2,0528*
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	255	715	0,2629	-4,6523*
IRAK: Die UNO im abseits (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	268	435	0,3812	2,6956*
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999) (J)	215	364	0,3713	1,9413*
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999) (J)	264	366	0,4190	4,5658*
Die Bremer Kinotaz ... ... alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999) (J)	671	2083	0,2436	-9,9812*
Strategische Spielhallen im Keller (TAZ Nr. 5735 14.01.1999) (J)	172	339	0,3366	0,1580
Klick, klick, klick (TAZ Nr. 5735 14.01.1999) (J)	261	463	0,3605	1,5525
Geliebte Irrtümer (TAZ Nr. 5734 13.01.1999) (J)	146	404	0,2655	-3,3754*
Für uns Serben wird hier kein Platz sein (TAZ Nr. 5734 13.01.1999) (J)	283	373	0,4314	5,3304*

Es wird gesündigt auf Teufel komm raus (TAZ Nr. 5734 11 13.01.1999) (J)	231	329	0,4125	3,9760*
Der Auftrag (TAZ Nr. 5733 12.01.1999) (J)	244	406	0,3754	2,2762*
Ein Schiff wird kommen (TAZ Nr. 5733 12.01. 1999) (J)	219	322	0,4048	3,5283*
Von Frust und Lust im samtenen Sweat-shop (TAZ Nr. 5732 11.01.1999) (J)	270	444	0,3782	2,5424*
August von Platen**,** Sonett (Po)	21	19	0,5250	2,5720*
H.J. Pestalozzi, Stoffels Brunnen (Pr)	36	23	0,6102	4,6121*
Indonesian texts				
Sekolah Ditutup, Akibat Cuaca Dingin (J) (http://www.lihatberita.com/2012/02/sekolah-ditutup-akibat-cuaca-dingin.html**)**	47	96	0,3287	-0,1174
Lepaskan Depresi dengan Tertawa Lepas (J) http://www.lihatberita.com/2012/02/ lepaskan-depresi-dengan-tertawa-lepas.html	34	46	0,4250	1,7400 *
Merokok Sebabkan Otak Lemot (J) http://www.lihatberita.com/2012/02/merokok-sebabkan-otak-lemot.html	70	125	0,3590	0,7606
Romanian poetry				
M. Eminescu, De ce nu-mi vii? (Po)	23	29	0,4423	1,6676*
M. Eminescu, Mai am un singur dor (Po)	21	38	0,3559	0,3688
M. Eminescu, Atât de fragedă... (Po)	27	51	0,3462	0,2408
M. Eminescu, Pe lângă plopii fără soț... (Po)	31	49	0,3875	1,0284
M. Eminescu, Floare-albastră (Po)	40	73	0,3540	0,4664
Romanian prose E.Simion: Excerpts from “Imi place, nu-mi place”, Moartea lui Mercutio, Nemira, Bucharest, 1993				
Imi place inceputul toamnei…(Pr)	36	52	0,4091	1,5083
Nu-mi plac oamenii care vorbesc mult…(Pr)	25	59	0,2976	-0,6938
Imi place aceasta dimineata gri…(Pr)	30	45	0,4000	1,2254
Imi place sa asist la facerea bulionului…(Pr)	23	47	0,3286	-0,0839
Imi place sa privesc gutuile…(Pr)	24	51	0,3200	-0,2443
Romanian press texts				
C.T. Popescu, Roşia Montană (J) “Gândul”, February 2012	66	120	0,3548	0,6232
N. Djuvara, interview (J) "Adevărul", February 2012	252	359	0,4124	4,1499*
A. Pleşu, Despre Mihai-Răzvan Ungureanu (J) Excerpt from the book “Față către față”, Editura Humanitas, 2011	83	220	0,2739	-2,1925*
M. Dinescu, Opriți Istoria! (J) “Cațavencii”, November 2011	35	76	0,3153	-0,4020
M. Eminescu, În libera Engliteră… (J) “Timpul”, March 5, 1880	97	222	0,3041	-1,1073

As can be seen, several texts significantly deviate from the expectation. This deviation can be ascribed to the style, to the theme or even to the age of the author. However, every interpretation is a hypothesis of its own and must be tested separately. In epic texts, the proportion of verbs is greater than in lyric texts; the older the author, the smaller number of “active” verbs he uses. The field of investigations is not restricted.

In order to show another view, one can consider directly some sequences, for example three subsequent elements (trigrams) in the NV-sequence. We obtain the following possibilities: NNN, NNV, NVN, VNN, NVV, VNV, VVN and VVV. If there is no trend, the probability of each pattern can easily be computed. Using $p(N)$ = 0,6667 and $p(V)$ = 0,3333 as above, we obtain for the individual sequences

NNN	0,6667(0,6667)0,6667 = 0,2962
NNV	0,6667(0,6667)0,3333 = 0,1481
NVN	0,1481
VNN	0,1481
NVV	0,0740
VNV	0,0740
VVN	0,0740
VVV	0,0370

The sum of these (unconditional) probabilities yields 1. Now, if we count all trigrams from a text and state their frequency, we can obtain the expected values by multiplying the above probabilities by the number of trigrams in text (n-2, because the last trigram is at the position n-2). For the text by Eva Bachletová “Moja Dolná zem” shown above we obtain n-2 = 409 (trigrams); the observed trigrams are presented in Table 5 together with expected numbers computed by multiplying the above probabilities by n-2.

Table 5
N-V-trigrams in Bachletová’s text and their observed and expected values

Trigrams	Observed O_i	Expected E_i	$\frac{(O_i - E_i)^2}{E_i}$
NNN	117	121,15	0,14
NNV	65	60,57	0,32
NVN	85	60,57	9,85
VNN	64	60,57	0,19
NVV	18	30,27	4,97
VNV	38	30,27	1,97
VVN	18	30,27	4,97
VVV	4	15,13	8,19

One can directly see that trigrams containing two subsequent V are avoided, the other ones are preferred. We can perform a chi-square text

(2) $$X^2 = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i}$$

and obtain

$$X^2 = (117 - 121{,}15)^2/121{,}15 + \ldots + (4 - 15{,}13)^2/15{,}13 = 30{,}62$$

which is, with 7 degrees of freedom highly significant and tells us that there is some tendency, the sequence is not random. The most (significantly) preferred trigram is NVN, the avoided trigrams are NVV, VVN and VVV.

In Table 6 we present results obtained for 85 texts in 9 languages. Since in each case we have 7 degrees of freedom, the critical value at the 0,05 level is 14,1. Significant deviations from our theoretical expectation are marked with ***.

Table 6
N-*V*-trigrams in 85 texts in 9 languages
(all chi-squares have 7 DF)

Text	**X^2**	**P**	**n**	**Preferred trigrams**
Slovak				
E. Bachletová, Moja Dolná zem (Pr)	30,62	***	411	NVN
E. Bachletová, Aby spriesvitnela (Po)	8,61		32	
E. Bachletova, Dielo Stvoriteľa (Po)	6,37		61	
E. Bachletova, Čakáme na šťastie (Po)	1,26		87	
E. Bachletova, Čas na nádych (Po)	20,66	***	118	VNV
E. Bachletova, Jazvy prítomnosti (Po)	8,47		262	
R. Dilong, Nostalgia (Po)	87,74	***	350	NVN, VNV
R. Dilong, Jesenná (Po)	5,83		97	
R. Dilong, Žobrák (Po)	12,29		94	NVN,VNV
R. Dilong, Minieme sa (Po)	20,02	***	58	NVN, VNV
R. Dilong, Ranený (Po)	3,86		34	
R. Dilong, Vo vetre myšlienky (Po)	31,66	***	61	
Nenásytný. Plus 7 dni 22.12.2011: (J)	9,38		117	
Národná vdova. Pus 7 dní 29.12.2011	24,48	***	222	NNN
French poetry				
Ch.Baudelaire, Harmonie du soir (Po)	35,39	***	53	NVN, VNV
Ch.Baudelaire, Hymne à la beauté (Po)	8,20		96	-
Ch.Baudelaire, Le balcon (Po)	9,26		95	-
Ch.Baudelaire, Chant d’ automne (Po)	8,12		83	-
Ch.Baudelaire, Tristesse de la lune (Po)	5,47		40	-
Ch.Baudelaire, Madrigal triste (Po)	33,34	***	88	NVN, VNV
P.Verlaine, Nevermore (Po)	44,16	***	82	NVN, VNV
P.Verlaine, Dans les bots (Po)	8,54		69	-
P.Verlaine, À la promenade (Po)	3,39		43	
P.Verlaine, L’amour par terre (Po)	19,12	***	49	NNV
P.Verlaine, La chanson des ingenues (Po)	3,55		57	-
French press texts				
Le Monde, 21 février 2012. Vers des	9,21	***	415	-

pourparlers entre l'Iran et les Occidentaux (J)				
Le Figaro, 19 février 2012.Le discours de Nicolas Sarkozy pour retrouver les Français (J)	20,91	***	262	VVV
Le Parisien, 21 février 2012. Nicolas Sarkozy : "Violent, moi ?" (J)	6,74		193	-
La Tribune, 21 février 2012. En imposant son plan aux Grecs, l'Europe commet une faute politique historique (J)	13,05		308	-
La Croix, 21 février 2012, Le sens du Carême (J)	33,56	***	143	NNN
Hungarian poetry				
S. Petőfi, Szeptember végén (Po)	13,04		64	VNV
J. Arany, A lepke (Po)	21,01	***	79	VNV
E.Ady, Góg és Magóg fia vagyok én (Po)	10,44		40	
E.Ady, A mi gyermekünk (Po)	31,20	***	43	VNV
E. Ady, A könnyek asszonya (Po)	7,42		47	
E. Ady, Ima Baál istenhez (Po)	20,03	***	70	NVN, VNV
Indonesian press texts				
Sekolah Ditutup, Akibat Cuaca Dingin**.** (http://www.lihatberita.com/ 2012/02/sekolah-ditutup-akibat-cuaca-dingin.html) (J)	13,55		143	NVN
Lepaskan Depresi dengan Tertawa Lepas.. (http://www.lihatberita.com/ 2012/02/ lepaskan-depresi-dengan-tertawa-lepas.html**)** (J)	8,13		80	
Merokok Sebabkan Otak Lemot. http://www.lihatberita.com /2012/02/ merokok-sebabkan-otak-lemot.html (J)	15,13	***	195	NNN, NVN
Romanian press texts				
C.T. Popescu, Roşia Montană (J) “Gândul”, February 2012	6,04		186	-
N. Djuvara, interview (J) "Adevărul", February 2012	61,83	***	611	NNN,VNV
A. Pleşu, Despre Mihai-Răzvan Ungureanu (J). Excerpt from the book “Faţă către faţă”, Editura Humanitas, 2011	13,18		303	-
M. Dinescu, Opriți Istoria! (J) “Cațavencii”, November 2011	3,40		111	-
M. Eminescu, În libera Engliteră (J) “Timpul”, March 5, 1880	11,46		319	-
Romanian prose Eugen Simion: Excerpts from “Imi place, nu-mi place”, Moartea lui Mercutio, Nemira, Bucharest, 1993				
Imi place inceputul toamnei…(Pr)	14,92	***	88	-
Nu-mi plac oamenii care vorbesc mult… (Pr)	28,73	***	84	NVN

Imi place aceasta dimineata gri…(Pr)	12,52		75	VNV
Imi place sa asist la facerea bulionului…(Pr)	1,31		70	-
Imi place sa privesc gutuile…(Pr)	19,13	***	75	NVN, VNV
Romanian poetry				
M. Eminescu, De ce nu-mi vii? (Po)	6,71		50	
M. Eminescu, Mai am un singur dor (Po)	16,30	***	55	NVN
M. Eminescu, Atât de fragedă... (Po)	2,65		76	
M. Eminescu, Pe lângă plopii fără soț...(Po)	16,20	***	80	NNN
M. Eminescu, Floare-albastră (Po)	24,05	***	111	NNN
German Texts				
Ethnische Definitionen als Machtpolitik. Le Monde Diplomatique 5736, 15.11. 1999 (J)	5,45		905	
Fortschritt findet woanders statt (TAZ 5732, 11.1.1999): (J)	12,30		610	
Trittbrettfahrer der Arisierung (TAZ Nr. 5732 11.01.1999) (J)	4,65		539	
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999) (J)	24,40	***	541	NVN
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999) (J)	7,02		670	
Kontrolle der Kapitalströme emerging markets… (Le Monde diplomatique Nr. 5736 15.01. 1999) (J)	14,69	***	940	
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	60,82	***	977	NNN
IRAK: Die UNO im abseits (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	26,74	***	705	
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999) (J)	12,74		579	VVV
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999) (J)	59,94	***	630	NVV,VNV,VVN,VVV
Die Bremer Kinotaz ...alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999) (J)	48939	***	2757	NNN
Strategische Spielhallen im Keller (TAZ Nr. 5735 14.01.1999) (J)	26,62	***	514	NVN
Klick, klick, klick (TAZ Nr. 5735 14.01.1999) (J)	26,07	***	727	NVN
Geliebte Irrtümer (TAZ Nr. 5734 13.01.1999) (J)	42,08	***	551	NNN
Für uns Serben wird hier kein Platz sein (TAZ Nr. 5734 13.01.1999) (J)	82,20	***	657	NVV,VNV,VVN,VVV
Es wird gesündigt auf Teufel komm raus (TAZ Nr. 5734 11 13.01.1999) (J)	45,56	***	560	NVV, VVN, VVV
Der Auftrag (TAZ Nr. 5733 12.01.1999) (J)	13,25		650	
Ein Schiff wird kommen (TAZ Nr. 5733 12.01.	41,81	***	541	NVV,VNV,V

1999) (J)				VN,VVV
Von Frust und Lust im samtenen Sweat-shop (TAZ Nr. 5732 11.01.1999) (J)	33,59	***	714	VNV
August von Platen**,** Sonett (Po)	19,13	***	75	NVN, VNV
Russian poetry http://www.stihi-rus.ru/1/Esenin/				
S. Jesenin, Alyj mrak v nebesnoj černi (Po)	8,96	-	33	-
S. Jesenin, Sestre Šure (Po)	14,66	***	33	NVN
S. Jesenin, Puškinu (Po)	32,32	***	35	NVN, VNV
S. Jesenin, Pismo k ženščine (Po)	50,34	***	134	VVV
S. Jesenin, Glupoe serdce, nebejsja (Po)	21,70	***	53	NVN, VNV
English poetry				
Byron, To Caroline, Think'st thou... (Po)	21,22	***	63	VVV
Byron, To Mary, On receiving her picture (Po)	12,38	-	75	-
Byron, Remind me not (Po)	27,36	***	76	VNV
Byron, To my son (Po)	12,30	-	99	NVN
Byron, There was a time (Po)	17,87	***	44	VVV
Italian texts				
A.Negri, Nevicata (Po)	1,56	-	20	-
La violenza sotto traccia. Corriere dela sera, February 24, 2012 (J)	2,55	-	208	-
Riflessi condizionati. Corriere dela sera, February 22, 2012 (J)	18,74	***	290	-
I sotterranei del Vaticano.Corriere dela sera, February 19, 2012 (J)	25,60	***	150	NVN
L'immagine del potere. Corriere dela sera, February 18, 2012 (J)	8,73	-	204	-
L'orgoglio delle nazioni. Corriere dela sera, February 17, 2012 (J)	23,58	***	201	NVN

From the result one can draw the following conclusions.

(1) Out of 85 texts in 9 languages, 45 significantly differ from the hypothesis of $E(p)$ = 0,3333. That means, if no special style is involved, texts behave neutrally as to the attribute/complement representation. Texts deviating from this regular scheme are idiosyncratic, i.e. they display some style. One of the many possible aspects can be studied formally, namely the representation of noun and verb sequences. Here we studied only trigrams but the same can be done also with higher-order-grams

(2) If testing individual trigrams (last column in Table 6), we find that not all are significantly represented to the same extent. Significant cases found are NVN (22), VNV (22), VVV (9), NNN (7), VVN (4), NVV (4), NNV (1) - the number in parentheses denotes the number of texts in which they occurred significantly frequently. There is a certain regularity: alternating trigrams > non alternating trigrams > asymmetric trigrams. This fact has surely some association with the theme, aim and style of the texts but in order to find it, one must analyze the same texts for other properties.

(3) N-V-trigrams display an aspect of syntax but do not display at once the complete sequence. The accumulation of nouns or verbs in an uninterrupted sequence may tell us more about the narrative or active character of the text. This can be studied either using the theory of runs or Markov chains. We shall use the first possibility.

3. Runs of N and V

Runs are sequences of identical entities in a sequence of different entities. If we have two kinds of entities, here N and V, we may study the placing of symbols and decide whether it is random or hides a tendency. In texts, the runs may be random but the grammar of the language which creates sequential dependencies usually does not allow it. In texts, non random placing of N and V may appear, depending on style, theme, text sort, etc. The research is not quite advanced, and N-V sequences are only one of the many possibilities to combine parts-of speech. In some languages the same form may belong to different word classes (e.g. E. *running*) thus even the study of conversion in text would be of interest.

Here, we may simply show that texts differ and randomness is only one of the possible states. Consider e.g. the sequence NNNNNVVVVV which nobody would consider random. But the same holds also for the quite regular sequence NVNVNVNVNV. There are tables for finding the boundaries of randomness but only for small number. Since our numbers are usually greater than 30, we test the randomness using the normal test

$$(3) \qquad u = \frac{r - \dfrac{2NV}{n}}{\dfrac{2NV}{n\sqrt{n}}}$$

where N = number of N, V = number of V, $n = N+V$, r = number of runs (cf. Gibbons 1971:57f.). If u > 1,96, the sequence tends to greater order in alternating N and V, most probably dictated by the grammar but it may be also the personal style. In this case, usually the number of V is greater than expected and verbs occur after short runs of N. The text tends to the verbal (active) style. If u < -1,96, the number of runs is too small, i.e. mostly longer runs of nouns are present and the text strives for a special nominal style. Though we analyzed 9 languages, the number of texts is too small to give a definite answer. The results of testing are presented in Table 7.

An expressed nominal style has been found only in one sole case, namely in the text *Le Parisien, 21 février 2012. Nicolas Sarkozy : "Violent, moi ?"*

In Italian, the first three texts display a random number of runs, the last three a rather regular alternation.

In Slovak 8 texts display randomness, 6 texts a tendency to regularity.

With Byron, there are 3 random cases and 2 tendencies to regularity

Table 7
Runs of N and V in 85 texts in 9 languages
(*N* = number of nouns, *V* = number of verbs, *r* = number of runs,
u = normal test; J - press text, Po - poetry, Pr - prose)

Text	**N**	**V**	**r**	**u**
Italian texts				
A.Negri, Nevicata (Po)	14	6	10	0,85
La violenza sotto traccia. Corriere dela sera, February 24, 2012 (J)	133	75	97	0,16
Riflessi condizionati. Corriere dela sera, February 22, 2012 (J)	181	110	147	1,27
I sotterranei del Vaticano. Corriere dela sera, February 19, 2012 (J)	122	68	109	3,42
L'immagine del potere. Corriere dela sera, February 18, 2012 (J)	139	47	98	5,59
L'orgoglio delle nazioni. Corriere dela sera, February 17, 2012 (J)	133	68	108	2,83
Slovak texts				
E. Bachletová, Moja Dolná zem (Pr)	286	125	207	3,85
E. Bachletová, Aby spriesvitnela (Po)	22	10	19	2,16
E. Bachletova, Dielo Stvoriteľa (Po)	44	17	30	1,74
E. Bachletova, Čakáme na šťastie (Po)	19	8	14	1,26
E. Bachletova, Čas na nádych (Po)	65	53	66	1,42
E. Bachletova, Jazvy prítomnosti (Po)	188	74	109	0,43
R. Dilong, Nostalgia (Po)	209	141	220	5,73
R. Dilong, Jesenná (Po)	62	35	45	0,06
R. Dilong, Žobrák (Po)	60	34	54	2,37
R. Dilong, Minieme sa (Po)	33	25	38	2,56
R. Dilong, Ranený (Po)	21	14	16	-0,28
R. Dilong, Vo vetre myšlienky (Po)	34	27	43	3,35
Nenásytný. Plus 7 dni 22. 12.2011 (J)	68	50	61	0,64
Národná vdova. Pus 7 dní 29.12.2011 (J)	168	54	89	1,32
English poetry				
Byron, To Caroline, Think'st thou... (Po)	32	31	35	0,88
Byron, To Mary, On receiving her picture (Po)	45	30	43	0,26
Byron, Remind me not (Po)	38	38	47	2,06
Byron, To my son (Po)	67	32	56	2,91
Byron, There was a time (Po)	21	23	25	0,92
Russian poetry http://www.stihi-rus.ru/1/Esenin/				
S. Jesenin, Alyj mrak v nebesnoj černi (Po)	25	8	17	2,31
S. Jesenin, Sestre Šure (Po)	20	13	23	2,64
S. Jesenin, Puškinu (Po)	20	15	27	3,40
S. Jesenin, Pismo k ženščine (Po)	70	64	62	-0,84
S. Jesenin, Glupoe serdce, nebejsja (Po)	31	23	38	3,23

German texts				
Ethnische Definitionen als Machtpolitik. Le Monde Diplomatique 5736, 15.11.1999 (J)	612	293	415	1,42
Fortschritt findet woanders statt (TAZ 5732, 11.1.1999): (J)	417	193	274	0,95
Trittbrettfahrer der Arisierung (TAZ Nr. 5732 11.01.1999) (J)	374	165	227	-0,20
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999) (J)	376	185	282	3,25
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999) (J)	436	234	285	-1,66
Kontrolle der Kapitalströme emerging markets - Ein Gebot der Demokratie (Le Monde diplomatique diplomatique Nr. 5736 15.01. 1999) (J)	657	283	407	0,88
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	722	255	395	1,50
IRAK: Die UNO im abseits (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	437	268	349	1,34
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999) (J)	364	215	260	-0,99
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999) (J)	366	264	316	0,76
Die Bremer Kinotaz ... ... alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999) (J)	2086	671	983	-1,67
Strategische Spielhallen im Keller (TAZ Nr. 5735 14.01.1999) (J)	342	172	263	3,38
Klick, klick, klick (TAZ Nr. 5735 14.01.1999) (J)	466	261	366	2,53
Geliebte Irrtümer (TAZ Nr. 5734 13.01.1999) (J)	405	146	199	-1,71
Für uns Serben wird hier kein Platz sein (TAZ Nr. 5734 13.01.1999) (J)	374	283	341	1,50
Es wird gesündigt auf Teufel komm raus (TAZ Nr. 5734 11 13.01.1999) (J)	329	231	255	-1,40
Der Auftrag (TAZ Nr. 5733 12.01.1999) (J)	406	244	308	0,27
Ein Schiff wird kommen (TAZ Nr. 5733 12.01. 1999) (J)	322	219	262	0,12
Von Frust und Lust im samtenen Sweat-shop (TAZ Nr. 5732 11.01.1999) (J)	444	270	368	2,56
August von Platen, Sonett (Po)	19	21	21	0,33
French poetry				
Ch.Baudelaire, Harmonie du soir (Po)	34	19	38	4,07
Ch.Baudelaire, Hymne à la beauté (Po)	65	31	50	1,87
Ch.Baudelaire, Le balcon (Po)	71	24	43	1,93
Ch.Baudelaire, Chant d’automne (Po)	57	26	44	2,11
Ch.Baudelaire, Tristesse de la lune (Po)	31	9	17	1,38
Ch.Baudelaire, Madrigal triste (Po)	56	32	56	3,51
P.Verlaine, Nevermore (Po)	53	29	56	4,47
P.Verlaine, Dans les bots (Po)	47	22	38	2,23

P.Verlaine, À la promenade (Po)	29	14	23	1,43
P.Verlaine, L'amour par terre (Po)	30	19	26	0,82
P.Verlaine, La chanson des ingenues (Po)	38	19	30	1,39
French press texts				
Le Monde, 21 février 2012. Vers des pourparlers entre l'Iran et les Occidentaux (J)	293	123	179	0,67
Le Figaro, 19 février 2012.Le discours de Nicolas Sarkozy pour retrouver les Français (J)	156	106	127	0,10
Le Parisien, 21 février 2012. Nicolas Sarkozy : "Violent, moi ?" (J)	129	64	59	-3,93
La Tribune, 21 février 2012. En imposant son plan aux Grecs, l'Europe commet une faute politique historique (J)	203	105	159	2,61
La Croix, 21 février 2012, Le sens du Carême (J)	115	28	51	1,58
Romanian press texts				
C.T. Popescu, Roşia Montană (J) "Gândul", February 2012	120	66	93	1,25
N. Djuvara, interview (J) "Adevărul", February 2012	359	252	332	1,51
A. Pleşu, Despre Mihai-Răzvan Ungureanu (J). Excerpt from the book "Față către față", Editura Humanitas, 2011	220	83	131	1,51
M. Dinescu, Opriți Istoria! (J) "Cațavencii", November 2011	76	35	53	1,11
M. Eminescu, În libera Engliteră (J) "Timpul", March 5, 1880	222	97	149	1,85
Romanian prose Eugen Simion: Excerpts from "Imi place, nu-mi place", Moartea lui Mercutio, Nemira, Bucharest, 1993				
Imi place inceputul toamnei…(Pr)	52	36	52	1,08
Nu-mi plac oamenii care vorbesc mult… (Pr)	59	25	50	3,88
Imi place aceasta dimineata gri…(Pr)	45	30	42	1,44
Imi place sa asist la facerea bulionului…(Pr)	47	23	30	-0,23
Imi place sa privesc gutuile…(Pr)	51	24	44	3,01
Romanian poetry				
Eminescu, De ce nu-mi vii?	29	23	27	0,38
Eminescu, Mai am un singur dor	38	21	36	2,54
Eminescu, Atât de fragedă...	51	27	37	0,42
Eminescu, Pe lângă plopii fără soț...	49	31	47	2,13
Eminescu, Floare-albastră	73	40	68	3,35
Hungarian texts (Po)				
S. Petöfi, Szeptember végén (Po)	38	26	39	2,11
J. Arany, A lepke (Po)	44	35	48	2,05
E.Ady, Góg és Magóg fia vagyok én (Po)	23	17	21	0,47
E.Ady, A mi gyermekünk (Po)	22	21	29	2,29
E. Ady, A könnyek asszonya (Po)	27	20	28	1,50
E. Ady, Ima Baál istenhez (Po)	45	25	44	3,09
Indonesian press texts				

Sekolah Ditutup, Akibat Cuaca Dingin**.** (http://www.lihatberita.com/ 2012/02/sekolah-ditutup-akibat-cuaca-dingin.html) (J)	96	47	78	2,82
Lepaskan Depresi dengan Tertawa Lepas. (http://www.lihatberita.com/ 2012/02/ lepaskan-depresi-dengan-tertawa-lepas.html**)** (J)	46	34	40	0,21
Merokok Sebabkan Otak Lemot. http://www.lihatberita.com /2012/02/ merokok-sebabkan-otak-lemot.html (J)	125	70	105	2,37

In German, 4 texts tend to regularity, 16 to randomness, there are only some cases in which the number of runs is smaller than the expectation.

In French poetry, the ratio is fifty-fifty. In press texts, one is significantly nominal, one is significantly regular and 3 are random.

All Romanian press texts display randomness, in the prose there are 4 significant regularities and 3 random run placings, in poetry 3 texts are “regular”, 2 are “random”.

In Hungarian poetry, we have 3 regularities and 2 random runs policy.

Out of 3 Indonesian texts, two are regular, one is random.

This is merely the first look showing that texts may differ in this respect. In any case, one can measure nominality/verbality also in other form, our aim was to attract attention at this phenomenon.

4. Distribution of runs

Though the distribution of runs is well known (cf. text-books on non-parametric statistics), we have to do with empirical cases which may display their own regime. Comparing the empirical cases with the theoretical ones we would mostly obtain a difference. Hence, avoiding this work, we ask at once what is the distribution of N-V runs in real texts.

We start from the theoretical framework established by Wimmer and Altmann (2005) from which a number of linguistic laws can be derived. We reduce and generalize it to obtain the basic difference equation

$$(4) \qquad P_x = z\left(1 - \frac{1}{b+x}\right)^a P_{x-1}$$

where parameter *z* can be considered a constant factor of language, *b* is the control factor of the hearer and *a* is a generalization of the original approach. The solution of (2) yields the Lerch distribution which is known as the generalization of the Zipf-Mandelbrot law (cf. Zörnig, Altmann 1995), i.e.

$$(5) \qquad P_x = \frac{z^x}{(b+x)^a \Phi_d(z,a,b)}, \; x = 1,2,3,...$$

where $\Phi_d(z,a,b)$ is the displaced Lerch function $\Phi_d(z,a,b) = \sum_{j=1}^{\infty} \frac{z^j}{(b+j)^a}$. Since our distributions are frequently very short and three parameters of the distribution would require at least five length classes, we shall use only some special classes of (3), viz. substituting *z* = *θ, a* = 1, *b* = 0 we obtain the difference equation

$$(4) \qquad P_x = \theta\left(1-\frac{1}{x}\right)P_{x-1}$$

leading to

$$(5) \qquad P_x = \frac{\theta^x}{-x\log_e(1-\theta)},\ x = 1,2,3,...$$

representing the *logarithmic distribution*. Substituting *z* = *q* and *a* = 0 we obtain

$$(6) \qquad P_x = qP_{x-1}$$

whose solution yields

$$(7) \qquad P_x = pq^{x-1},\ x = 1,2,3,...$$

where *p* = 1-*q*, representing the one 1-displaced *geometric distribution*, and, finally substituting *z* = 1 and *b* = 0 we obtain

$$(8) \qquad P_x = \left(1-\frac{1}{x}\right)^a P_{x-1} = \left(\frac{x-1}{x}\right)^a P_{x-1}$$

yielding the *zeta (Zipf) distribution*

$$(9) \qquad P_x = \frac{x^{-a}}{T},\ x = 1,2,3,...$$

where T = $\sum_{j=1}^{\infty} j^{-a}$, with *a* > 1.

Of course, there are also other special cases but for our purposes the above ones are sufficient. In some cases the right truncated version, i.e. an alternative with finite support would be more appropriate but we use the infinite support because the trunkation brings a further parameter, and often, the number of classes is not sufficient.

The results of fitting are presented in Table 8. Not all the three alternatives were appropriate in all cases but at least one of them did in almost all cases. There were only few exceptions: for the Russian text *Puškinu* by Jesenin the only adequate distribution was *Johnson-Kotz* (cf. Wimmer, Altmann 1999); for the German text *Die Bremer Kinotaz ... ... alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999)* a *mixture of two geometric distribution* had been adequate because the text is too long and possibly consisting of several parts; for the French text *L'amour par terre* by P. Verlaine and for the Hungarian text E.Ady, *Góg és Magóg fia vagyok én* the Poisson distribution had been the correct alternative; and for the Romanian text: A. Pleşu, *Despre Mihai-Răzvan Ungureanu* (excerpt from the book "Față către față", Editura Humanitas, 2011) the right truncated zeta distribution was the only one yielding acceptable results.

Table 8
Distributions of runs of N and V in 85 texts
(G = geometric, L = logarithmic, Z = zeta)

Text	**P_x**	**Parameters**	**X^2**	**DF**	**P**
Italian texts					
A.Negri, Nevicata (Po)	L G Z	0,7737 0,4919 1,8507	0,10 0,59 1,08	2 2 2	0,95 0,72 0,58
La violenza sotto traccia. Corriere dela sera, February 24, 2012 (J)	L G	0,7674 0,4355	10,93 6,17	7 6	0,14 0,40
Riflessi condizionati. Corriere dela sera, February 22, 2012 (J)	G	0,4937	7,35	5	0,20
I sotterranei del Vaticano. Corriere dela sera, February 19, 2012 (J)	L G	0,6603 0,5656	4,12 0,78	4 4	0,39 0,94
L'immagine del potere. Corriere dela sera, February 18, 2012 (J)	G	0,4716	1,94	5	0,86
L'orgoglio delle nazioni. Corriere dela sera, February 17, 2012 (J)	L G	0,7019 0,5157	3,78 4,89	5 5	0,58 0,43
Slovak texts					
E. Bachletová, Moja Dolná zem (Pr)	L Z	0,7073 2,1378	6,03 18,79	8 13	0,64 0,13
E. Bachletová, Aby spriesvitnela (Po)	L G Z	0,6576 0,6794 2,4856	1,07 0,01 0,39	2 1 1	0,59 0,94 0,53
E. Bachletova, Dielo Stvoriteľa (Po)	L	0,7255	7,15	3	0,07
E. Bachletova, Čakáme na šťastie (Po)	L G Z	0,7025 0,4830 2,0951	0,63 1,91 0,00	2 2 2	0,73 0,38 1,00
E. Bachletova, Čas pre nádych (Po)	L G	0,6877 0,5485	2,88 1,04	3 3	0,41 0,79
E. Bachletova, Jazvy prítomnosti (Po)	L Z	0,7988 1,8857	8,51 12,81	8 9	0,39 0,17
R. Dilong, Nostalgia (Po)	L G	0,5993 0,6225	7,38 2,29	4 4	0,12 0,68
R. Dilong, Jesenná (Po)	G	0,4336	6,45	4	0,17
R. Dilong, Žobrák (Po)	L G	0,6041 0,5760	1,56 1,83	3 3	0,67 0,61

	Z	2,2425	3,68	3	0,30
R. Dilong, Minieme sa (Po)	L	0,5784	0,79	2	0,67
	G	0,6391	1,24	2	0,54
	Z	2,2940	2,68	2	0,26
R. Dilong, Ranený (Po)	L	0,7512	0,46	3	0,93
	G	0,4256	0,76	3	0,86
	Z	1,8719	1,31	3	0,73
R. Dilong, Vo vetre myšlienky (Po)	L	0,5305	0,69	1	0,41
	G	0,6842	0,08	1	0,78
	Z	2,3425	3,65	1	0,06
Nenásytný. Plus 7 dni 22. 12.2011: (J)	L	0,7196	5,73	5	0,33
	G	0,5082	2,76	4	0,63
Národná vdova. Pus 7 dní 29.12.2011 (J)	L	0,8035	3,74	8	0,88
	G	0,4054	7,83	6	0,25
	Z	1,7840	13,16	9	0,16
English poetry					
Byron, To Caroline, Think'st thou... (Po)	L	0,7108	4,24	3	0,24
	G	0,5096	4,23	3	0,24
Byron, To Mary, On receiving her picture (Po)	L	0,6746	7,16	3	0,07
	G	0,5597	3,02	3	0,39
Byron, Remind me not (Po)	L	0,6072	4,70	3	0,20
	G	0,6190	2,03	2	0,36
Byron, To my son (Po)	L	0,6922	2,20	4	0,70
	G	0,5571	0,76	3	0,86
	Z	2,0292	7,90	4	0,10
Byron, There was a time (Po)	L	0,7012	0,24	2	0,89
	G	0,5339	0,41	2	0,82
	Z	2,0246	1,73	2	0,42
Russian poetry http://www.stihi-rus.ru/1/Esenin/					
S. Jesenin, Alyj mrak v nebesnoj černi (Po)	L	0,7285	1,53	2	0,46
	G	0,5133	1,80	2	0,41
	Z	2,4259	1,75	1	0,19
S. Jesenin, Sestre Šure (Po)	L	0,5429	0,89	1	0,35
	G	0,6752	0,27	1	0,60
	Z	2,3051	3,11	1	0,08
S. Jesenin, Pismo k ženščine (Po)	L	0,7753	4,85	6	0,56
	G	0,4671	2,25	5	0,81
S. Jesenin, Glupoe serdce, nebejsja (Po)	L	0,5079	0,92	2	0,63
	G	0,6979	0,21	2	0,90
	Z	2,4010	4,47	2	0,11
German texts					
Ethnische Definitionen als Machtpolitik. Le Monde Diplomatique 5736, 15.11.1999 (J)	L	0,7668	15,33	11	0,17
Fortschritt findet woanders statt (TAZ 5732, 11.1.1999) (J)	G	0,4419	9,78	7	0,20

Trittbrettfahrer der Arisierung (TAZ Nr. 5732 11.01.1999) (J)	L	0,7909	6,62	10	0,76
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999) (J)	G	0,4967	7,44	6	0,28
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999) (J)	L	0,7809	17,95	11	0,08
	G	0,4137	13,10	8	0,11
Kontrolle der Kapitalströme emerging markets - Ein Gebot der Demokratie (Le Monde diplomatique diplomatique Nr. 5736 15.01. 1999) (J)	L	0,7788	16,19	12	0,18
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	L	0,7972	6,86	12	0,87
IRAK: Die UNO im abseits (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	G	0,4798	9,28	6	0,16
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999) (J)	G	0,4632	3,28	3	0,35
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999) (J)	G	0,5059	4,52	3	0,21
Strategische Spielhallen im Keller (TAZ Nr. 5735 14.01.1999) (J)	L	0,7126	7,01	3	0,07
	G	0,5043	4,51	3	0,21
Klick, klick, klick (TAZ Nr. 5735 14.01.1999) (J)	L	0,7203	1,61	4	0,81
Geliebte Irrtümer (TAZ Nr. 5734 13.01.1999) (J)	L	0,8387	0,44	4	0,98
Für uns Serben wird hier kein Platz sein (TAZ Nr. 5734 13.01.1999) (J)	G	0,5233	4,11	3	0,25
Es wird gesündigt auf Teufel komm raus (TAZ Nr. 5734 11 13.01.1999) (J)	G	0,4595	0,58	3	0,90
Der Auftrag (TAZ Nr. 5733 12.01.1999) (J)	L	0,7480	5,91	4	0,21
	G	0,4813	1,03	4	0,90
Ein Schiff wird kommen (TAZ Nr. 5733 12.01. 1999) (J)	L	0,7398	1,10	3	0,78
	G	0,4875	4,68	3	0,20
Von Frust und Lust im samtenen Sweatshop (TAZ Nr. 5732 11.01. 1999) (J)	G	0,5110	3,67	4	0,45
August von Platen**,** Sonett (Po)	L	0,6846	4,14	2	0,13
	G	0,5350	2,46	2	0,29
French poetry					
Ch.Baudelaire, Harmonie du soir (Po)	L	0,4683	0,05	2	0,97
	G	0,7289	0,05	1	0,82
	Z	2,5875	1,37	2	0,50
Ch.Baudelaire, Hymne à la beauté (Po)	L	0,7098	1,80	1	0,18
	G	0,5405	0,27	1	0,60
Ch.Baudelaire, Le balcon (Po)	L	0,7321	2,59	4	0,63
	G	0,4657	5,61	4	0,23
	Z	2,0796	4,10	5	0,54

Ch.Baudelaire, Chant d’automne (Po)	L	0,6556	0,43	3	0,93
	G	0,5309	4,70	3	0,19
	Z	2,0686	5,04	4	0,28
Ch.Baudelaire, Tristesse de la lune (Po)	L	0,8325	3,25	3	0,35
	G	0,3930	2,58	3	0,46
	Z	1,6538	6,79	3	0,08
Ch.Baudelaire, Madrigal triste (Po)	Z	2,7345	4,85	2	0,09
P.Verlaine, Nevermore (Po)	L	0,5444	2,13	2	0,34
	G	0,6666	2,61	2	0,27
	Z	2,6355	5,04	2	0,08
P.Verlaine, Dans les bots (Po)	L	0,6850	0,23	3	0,97
	G	0,5434	0,40	3	0,94
	Z	2,0478	4,02	3	0,26
P.Verlaine, À la promenade (Po)	L	0,7160	1,37	3	0,71
	G	0,5434	0,48	2	0,79
	Z	1,9656	4,35	2	0,11
P.Verlaine, La chanson des ingenues (Po)	L	0,6453	1,53	3	0,68
	G	0,6024	1,95	2	0,38
	Z	2,1651	3,59	3	0,31
French press texts					
Le Monde, 21 février 2012. Vers des pourparlers entre l'Iran et les Occidentaux (J)	L	0,7922	0,98	4	0,91
Le Figaro, 19 février 2012. Le discours de Nicolas Sarkozy pour retrouver les Français (J)	G	0,4499	5,06	2	0,08
Le Parisien, 21 février 2012. Nicolas Sarkozy : "Violent, moi ?" (J)	L	0,7270	0,26	1	0,.61
La Tribune, 21 février 2012. En imposant son plan aux Grecs, l'Europe commet une faute politique historique (J)	L	0,7239	9,05	7	0,25
	G	0,5045	7,22	6	0,30
La Croix, 21 février 2012, Le sens du Carême (J)	Z	1,7716	0,33	1	0,56
Romanian press texts					
C.T. Popescu, Roşia Montană (J)	L	0,7134	6,45	5	0,26
“Gândul”, February 2012	G	0,4951	3,91	5	0,56
N. Djuvara, interview (J)	L	0,6888	12,09	8	0,15
"Adevărul", February 2012	G	0,5637	2,07	5	0,84
M. Dinescu, Opriți Istoria! (J)	L	0,7627	4,08	5	0,54
“Cațavencii”, November 2011	G	0,4756	2,55	4	0,64
M. Eminescu, În libera Englireră (J)	L	0,7534	8,49	8	0,39
“Timpul”, March 5, 1880	G	0,4650	5,24	6	0,51
Romanian prose Eugen Simion: Excerpts from “Imi place, nu-mi place”, Moartea lui Mercutio, Nemira, Bucharest, 1993					

Imi place inceputul toamnei…(Pr)	L G	0,6506 0,5728	5,49 2,72	3 3	0,14 0,44
Nu-mi plac oamenii care vorbesc mult…(Pr)	L G Z	0,6303 0,6239 2,2759	3,47 5,80 2,35	3 2 4	0,33 0,06 0,67
Imi place aceasta dimineata gri…(Pr)	L G	0,6744 0,5476	4,96 2,88	4 3	0,29 0,41
Imi place sa asist la facerea bulionului…(Pr)	L G Z	0,7655 0,4272 1,6878	2,92 0,35 9,74	4 4 5	0,57 0,99 0,08
Imi place sa privesc gutuile…(Pr)	L Z	0,6674 2,1972	4,26 3,09	3 3	0,23 0,38
Romanian poetry					
M. Eminescu, De ce nu-mi vii?	L G Z	0,7273 0,5137 1,9253	0,96 0,52 4,93	3 3 3	0,81 0,92 0,18
M. Eminescu, Mai am un singur dor	L G	0,6611 0,5794	4,37 2,57	2 2	0,11 0,28
M. Eminescu, Atât de fragedă...	L G	0,7460 0,4521	6,42 3,66	4 4	0,17 0,45
M. Eminescu, Pe lângă plopii fără soț...	L G	0,6504 0,5789	4,60 1,57	3 3	0,20 0,67
M. Eminescu, Floare-albastră	L G Z	0,6291 0,5941 2,1496	0,68 0,48 7,04	3 3 3	0,88 0,92 0,07
Hungarian poetry					
S. Petöfi, Szeptember végén (Po)	L G	0,6361 0,6071	3,49 1,35	3 2	0,32 0,51
J. Arany, A lepke (Po)	L G Z	0,6480 0,6201 2,2101	2,39 1,98 4,53	3 2 3	0,49 0,37 0,21
E.Ady, A mi gyermekünk (Po)	L G Z	0,5236 0,6833 2,4100	0,16 0,02 1,45	2 1 2	0,92 0,90 0,48
E. Ady, A könnyek asszonya (Po)	L G	0,6378 0,5894	1,30 0,46	3 2	0,73 0,80
E. Ady, Ima Baál istenhez (Po)	L G Z	0,6269 0,6117 0,3121	1,04 2,76 1,97	3 2 3	0,79 0,25 0,58
Indonesian press texts					
Sekolah Ditutup, Akibat Cuaca Dingin. (http://www.lihatberita.com/ 2012/02/ sekolah-ditutup-akibat-cuaca-dingin. html) (J)	L G Z	0,6788 0,5345 2,0833	0,64 2,90 8,39	5 4 6	0,99 0,58 0,21
Lepaskan Depresi dengan Tertawa Lepas. (http://www.lihatberita.com/ 2012/02/	L G	0,7175 0,4593	2,77 2,25	4 4	0,60 0,69

lepaskan-depresi-dengan-tertawa-lepas.html) (J)	Z	2,1209	9,16	4	0,06
Merokok Sebabkan Otak Lemot. (J) http://www.lihatberita.com /2012/02/ merokok-sebabkan-otak-lemot.html	L G	0,6930 0,5396	3,87 2,00	5 4	0,57 0,74

The study of further texts and languages will surely result in extending this family of distributions but a more extensive research requires consequent distinguishing of text sorts, time of creation, and paying regard to some background grammatical boundary conditions. In any case, the start has been made.

Now, knowing that the runs abide by some theoretical distributions belonging to the same family, we can represent the results graphically. We consider the first three moments of the empirical distributions and set up Ord's (1972) <I,S>-scheme used very frequently in linguistics (cf. Popescu et al. 2009). Here $I = m_2/m'_1$ and $S = m_3/m_2$ where m'_1 is the mean and m_2 (variance), m_3 are the central moments defined as $m_r = (1/\mathrm{N})\sum_{x=0}^{\infty}(x-\bar{x})^r f_x$. The complete computation is presented in Table 9.

Table 9
Ord's coordinates for 85 texts in 9 languages

Slovak	***N***	***M₁***	***M₂***	***M₃***	***I***	***S***
E. Bachletová, Moja Dolná zem (Pr)	210	1,97	7,73	196,55	3,92	25,43
E. Bachletová, Aby spriesvitnela (Po)	19	1,68	2,11	8,35	1,26	3,96
E. Bachletova, Dielo Stvoriteľa (Po)	30	2,03	2,9	7,34	1,43	2,53
E. Bachletova, Čakáme na šťastie (Po)	14	1,93	3,49	16,11	1,81	4,62
E. Bachletova, Čas na nádych (Pr)	66	1,79	1,17	1,63	0,65	1,39
E. Bachletova, Jazvy prítomnosti (Pr)	109	2,4	6,86	50,7	2,86	7,39
R. Dilong, Nostalgia (Po)	220	1,59	0,86	1,48	0,54	1,72
R. Dilong, Jesenná (Po)	45	2,16	1,95	2,97	0,90	1,52
R. Dilong, Žobrák (Po)	54	1,74	1,49	3,47	0,86	2,33
R. Dilong, Minieme sa (Po)	38	1,53	0,78	1,05	0,51	1,35
R. Dilong, Ranený (Po)	16	2,19	3,4	12,89	1,55	3,79
R. Dilong, Vo vetre myšlienky (Po)	43	1,42	0,43	0,36	0,30	0,84
Nenásytný. Plus 7 dni 22.12.2011 (J)	61	1,93	1,83	5,61	0,95	3,07
Národná vdova.Plus 7 dní 29.12.2011 (J)	89	2,49	5,67	32,42	2,28	5,72
French poetry						
Ch.Baudelaire, Harmonie du soir (Po)	38	1,39	0,66	1,45	0,47	2,20
Ch.Baudelaire, Hymne à la beauté (Po)	50	1,92	2,19	8,97	1,14	4,10
Ch.Baudelaire, Le balcon (Po)	43	2,21	7,05	77,26	3,19	10,96
Ch.Baudelaire, Chant d' automne (Po)	44	1,89	2,42	9,99	1,28	4,13
Ch.Baudelaire, Tristesse de la lune (Po)	17	2,35	2,58	3,58	1,10	1,39
Ch.Baudelaire, Madrigal triste (Po)	56	1,57	2,49	7,72	1,59	3,10
P.Verlaine, Nevermore (Po)	56	1,47	0,68	0,92	0,46	1,35

P.Verlaine, Dans les bots (Po)	38	1,82	1,52	3,19	0,84	2,10
P.Verlaine, À la promenade (Po)	23	1,87	1,33	1,96	0,71	1,47
P.Verlaine, L’amour par terre (Po)	26	1,88	0,49	0,05	0,26	0,10
P.Verlaine, La chanson des ingenues (Po)	40	1,68	1,22	2,45	0,73	2,01
French press texts						
Le Monde, 21 février 2012. Vers des pourparlers entre l'Iran et les Occidentaux (J)	179	2,32	3,82	14,14	1,65	3,70
Le Figaro, 19 février 2012.Le discours de Nicolas Sarkozy pour retrouver les Français (J)	127	2,06	3,29	18,05	1,60	5,49
Le Parisien, 21 février 2012. Nicolas Sarkozy : "Violent, moi ?" (J)	97	1,99	3,1	14,62	1,56	4,72
La Tribune, 21 février 2012. En imposant son plan aux Grecs, l'Europe commet une faute politique historique (J)	159	1,94	1,96	5,55	1,01	2,83
La Croix, 21 février 2012, Le sens du Carême (J)	51	2,8	8,82	48,77	3,15	5,53
Hungarian poetry						
S. Petöfi, Szeptember végén (Po)	39	1,64	1	2,3	0,61	2,30
J. Arany, A lepke (Po)	48	1,65	1,19	2,58	0,72	2,17
E.Ady, Góg és Magóg fia vagyok én (Po)	21	1,9	0,66	0,38	0,35	0,58
E.Ady, A mi gyermekünk (Po)	29	1,48	0,8	1,69	0,54	2,11
E. Ady, A könnyek asszonya (Po)	28	1,68	1	1,75	0,60	1,75
E. Ady, Ima Baál istenhez (Po)	44	1,59	1,2	2,9	0,75	2,42
Indonesian press texts						
Sekolah Ditutup, Akibat Cuaca Dingin. (http://www.lihatberita.com/ 2012/02/sekolah-ditutup-akibat-cuaca-dingin.html) (J)	78	1,83	2,06	7,56	1,13	3,67
Lepaskan Depresi dengan Tertawa Lepas.. (http://www.lihatberita.com/ 2012/02/ lepaskan-depresi-dengan-tertawa-lepas.html) (J)	40	2	2,4	8,4	1,20	3,50
Merokok Sebabkan Otak Lemot. http://www.lihatberita.com /2012/02/ merokok-sebabkan-otak-lemot.html (J)	105	1,86	1,65	4,39	0,89	2,66
Romanian press texts						
C.T. Popescu, Roşia Montană (J)	93	2	1,94	4,65	0,97	2,40
“Gândul”, February 2012						
N. Djuvara, interview (J)	332	1,84	1,66	5,66	0,90	3,41
"Adevărul", February 2012						
A. Pleşu, Despre Mihai-Răzvan Ungureanu (J). Excerpt from the book “Faţă către faţă”, Editura Humanitas, 2011	131	2,31	4,64	21,07	2,01	4,54

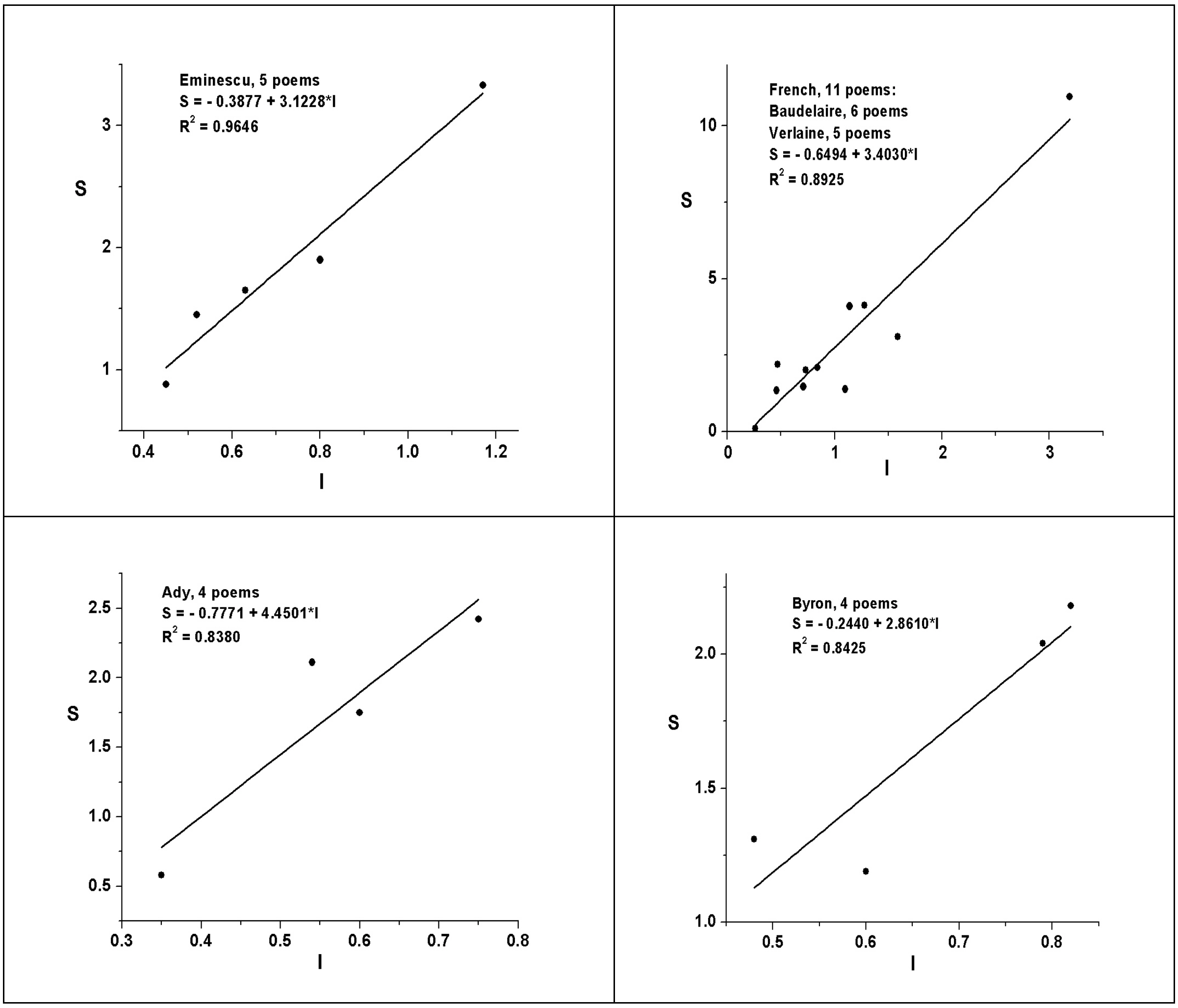
M. Dinescu, Opriți Istoria! (J)	53	2,09	2,58	7,63	1,23	2,96
“Cațavencii”, November 2011						
M. Eminescu, În libera Engliteră (J)	149	2,14	3,17	15,43	1,48	4,87
“Timpul”, March 5, 1880						
Romanian prose Eugen Simion: Excerpts from “Imi place, nu-mi place”, Moartea lui Mercutio, Nemira, Bucharest, 1993						
Imi place inceputul toamnei…(Pr)	52	1,69	0,94	1,63	0,56	1,73
Nu-mi plac oamenii care vorbesc mult…(Pr)	50	1,68	2,1	8,21	1,25	3,91
Imi place aceasta dimineata gri…(Pr)	42	1,79	1,31	2,92	0,73	2,23
Imi place sa asist la facerea bulionului…(Pr)	30	2,33	2,96	10,54	1,27	3,56
Imi place sa privesc gutuile…(Pr)	44	1,7	1,57	3,17	0,92	2,02
Romanian poetry						
M.Eminescu, De ce nu-mi vii?	27	1,93	1,55	2,94	0,80	1,90
M.Eminescu, Mai am un singur dor	36	1,64	0,73	0,64	0,45	0,88
M. Eminescu, Atât de fragedă...	37	2,11	2,47	8,22	1,17	3,33
M. Eminescu, Pe lângă plopii fără soț...	47	1,7	0,89	1,29	0,52	1,45
M. Eminescu, Floare-albastră	68	1,66	1,05	1,73	0,63	1,65
German Texts						
Ethnische Definitionen als Machtpolitik. Le Monde Diplomatique 5736, 15.11.1999 (J)	415	2,18	3,64	18,24	1,67	5,01
Fortschritt findet woanders statt (TAZ 5732, 11.1.1999): (J)	274	2,23	2,6	5,98	1,17	2,30
Trittbrettfahrer der Arisierung (TAZ Nr. 5732 11.01.1999) (J)	227	2,37	5,19	36,72	2,19	7,08
Tendenz zur Lästigkeit (TAZ Nr. 5732 11.01.1999) (J)	282	1,99	1,85	3,94	0,93	2,13
Wo Es war, soll Wir werden (TAZ Nr. 5732 11.01.1999) (J)	285	2,35	4,61	35,54	1,96	7,71
Kontrolle der Kapitalströme emerging markets - Ein Gebot der Demokratie (Le Monde diplomatique diplomatique Nr. 5736 15.01. 1999) (J)	407	2,31	6,7	116,68	2,90	17,41
Bündnisse und Rivalitäten im Mittleren Afrika (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	395	2,47	6,56	64,22	2,66	9,79
IRAK: Die UNO im abseits (Le Monde diplomatique Nr. 5736 15.01.1999) (J)	351	2,03	2,95	5,36	1,45	1,82
"Liebling Kreuzberg" oder "Hallo, Herr Kaiser" (TAZ Nr. 5736 15.01.1999) (J)	260	2,23	3,8	25,62	1,70	6,74
In den Niederungen der Statistik (TAZ Nr. 5736 15.01.1999) (J)	318	2	2,53	18,15	1,27	7,17

Die Bremer Kinotaz ... alle Filme, Termine (TAZ-BREMEN Nr. 5735 14.01.1999) (J)	982	2,77	14,49	211,52	5,23	14,60
Strategische Spielhallen im Keller						
(TAZ Nr. 5735 14.01.1999) (J)	263	1,95	2,05	6,19	1,05	3,02
Klick, klick, klick (TAZ Nr. 5735 14.01.1999) (J)	366	1,99	2,55	9,48	1,28	3,72
Geliebte Irrtümer (TAZ Nr. 5734 13.01.1999) (J)	199	2,77	7,24	43,94	2,61	6,07
Für uns Serben wird hier kein Platz sein (TAZ Nr. 5734 13.01.1999) (J)	341	1,93	1,95	6,41	1,01	3,29
Es wird gesündigt auf Teufel komm raus (TAZ Nr. 5734 11 13.01.1999) (J)	255	2,2	2,39	6,38	1,09	2,67
Der Auftrag (TAZ Nr. 5733 12.01.1999) (J)	308	2,11	2,77	11,64	1,31	4,20
Ein Schiff wird kommen (TAZ Nr. 5733 12.01. 1999) (J)	292	2,06	2,55	7,79	1,24	3,05
Von Frust und Lust im samtenen Sweat-shop (TAZ Nr. 5732 11.01.1999) (J)	368	1,94	1,79	4,79	0,92	2,68
August von Platen**,** Sonett (Po)	21	1,9	1,9	6,45	1,00	3,39
Russian poetry						
http://www.stihi-rus.ru/1/Esenin/						
S.Jesenin, Alyj mrak v nebesnoj černi	17	1,94	1,94	4,17	1,00	2,15
S. Jesenin, Sestre Šure	27	1,44	0,42	0,33	0,29	0,79
S. Jesenin, Puškinu	27	1,3	0,43	0,55	0,33	1,28
S. Jesenin, Pismo k ženščine	62	2,16	2,68	9,6	1,24	3,58
S. Jesenin, Glupoe serdce, nebejsja	38	1,42	0,51	0,65	0,36	1,27
English poetry						
Byron, To Caroline, Think'st thou...	35	1,8	1,42	2,89	0,79	2,04
Byron, To Mary, On receiving her picture	43	1,74	0,84	1,1	0,48	1,31
Byron, Remind me not	47	1,62	0,92	2,13	0,57	2,32
Byron, To my son	55	1,8	1,47	3,2	0,82	2,18
Byron, There was a time	25	1,76	1,06	1,26	0,60	1,19
Italian texts						
A.Negri, Nevicata (Po)	10	2	1,6	2,4	0,80	1,50
La violenza sotto traccia. Corriere dela sera, February 24, 2012 (J)	97	2,14	2,7	11,03	1,26	4,09
Riflessi condizionati. Corriere dela sera, February 22, 2012 (J)	147	1,98	1,58	3,42	0,80	2,16
I sotterranei del Vaticano.Corriere dela sera, February 19, 2012 (J)	109	1,74	1,16	2,25	0,67	1,94
L'immagine del potere. Corriere dela sera, February 18, 2012 (J)	98	2,08	1,85	3,67	0,89	1,98
L'orgoglio delle nazioni. Corriere dela sera, February 17, 2012 (J)	108	1,86	1,8	4,95	0,97	2,75

It can easily be shown that the <I,S> points for individual authors and text sorts are placed on straight lines in Ord’s scheme. The deviations of individual points are very small and we may say that the texts have a well equilibrated sentence structure. Some of the results are presented in Figure 3. Though the German press texts were written by several authors, they display the same tendency (with one outlier)

Since the schemes have different scaling of both axes, we show the result in a common presentation, without the measured points displaying merely the linear trend.

The resulting hypothesis is somewhat premature but in virtue of the present results we can conjecture that every future study must take into account at least the triad: author/text-sort/language.



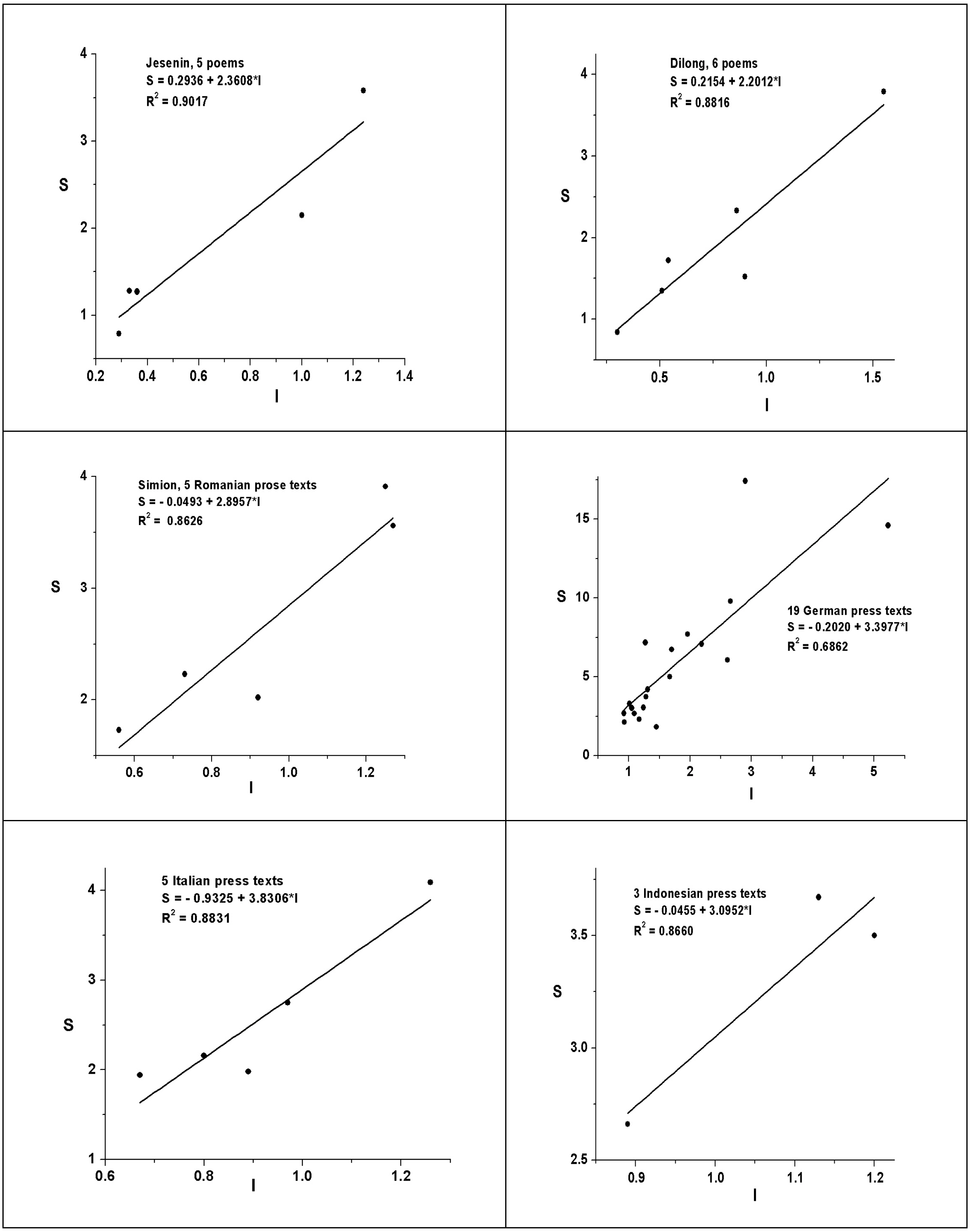


Figure 3. Some text-sort graphs in Ord's <I,S> scheme

The language-factor controls the grammar, the text sort-factor (aim) controls the style and the author-factor controls the repetition of individual patterns. Separating the factors one secures at least a part of the *ceteris paribus* condition which is itself the factor warranting data homogeneity. Peculiar enough, only the study of individual texts can help us to reveal the general mechanisms that were active during their creation. The mixing of texts, as it is often done in corpus linguistics, cannot lead to the discovery of relevant phenomena.

Ord’s linear regression *S* = *a* + *b*I* for 35 poems in 6 languages considered above is illustrated in Table 10 and plotted in Figure 4, similarly to Figures 9.4 and 9.6 from Popescu et al.(2009: 160, 165).

Table 10
linear I,S fitting of 35 poems in 6 languages

Language	**Poet**	**# poems**	**slope b**	**R^2**
Hungarian	Ady	4	4.45	0.84
French	Baudelaire	6	3.43	0.85
French	Verlaine	5	3.17	0.87
Romanian	Eminescu	5	3.12	0.96
English	Byron	4	2.86	0.84
Russian	Jesenin	5	2.36	0.90
Slovak	Dilong	6	2.20	0.88

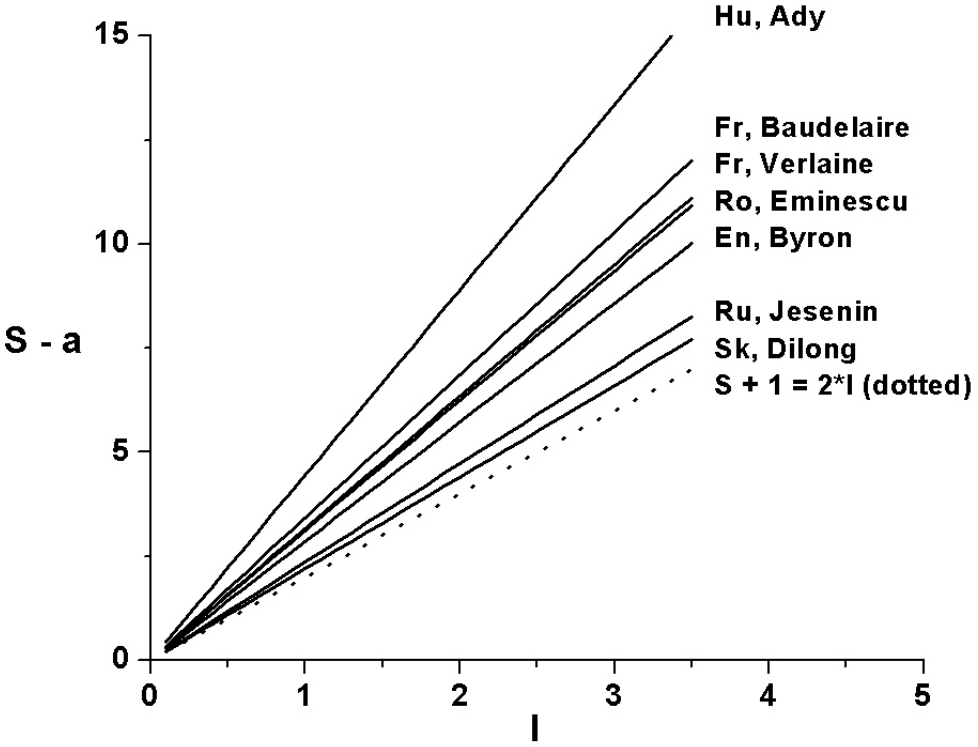


Figure 4. I,S lines of poetry in 6 languages

Though this is only poetry, i.e. bound language, the result is surprising: at the one end we have a strongly agglutinating language (Hungarian), at the other end strongly inflectional languages (Russian and Slovak). The two Roman languages (French and Romanian) and English, all tending towards analytism, are posited between these extremes. All works are placed in the domain S > 2I - 1, i.e. above the negative hypergeometric domain.

Further study in different languages and text sorts would surely shed more light on this phenomenon based only on nouns and verbs. Neverthless, Figure 4 yields a starting point.

References

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

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

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-Verlag

Wimmer, G., **Altmann, G.** (2005). Towards a unified derivation of some linguistic laws. In: Grzybek, P. (ed.), *Contributions to the Science of Language. Word Length Studies and Related Issues: 207-316*. Boston: Kluwer.

Ziegler, A., Best, K.-H., Altmann, G. (2002). Nominalstil. *ETC – Empirical Text and Culture Research 2, 72-85.*

Zörnig, P., Altmann, G. (1995). Unified representation of Zipf distributions**.** *Computational Statistics & Data Analysis 19, 1995, 461-473.*

Glottometrics 23, 2012, 56-61

Zur Verslänge bei G. A. Bürger

Karl-Heinz Best

Abstract. In this contribution the distribution of word numbers in poetic texts by G.A. Bürger is tested. The displaced binomial distribution seems to be the best model. There was only one case in which the empirical findings deviate from this model. Evidently some boundary conditions valid in this case must be found.

Keywords: verse length, poetry, German

1. Zum Thema

Angeregt durch den Hinweis: „Man kann fragen: Wieviel Silben oder auch wieviel Wörter hat jeder einzelne Vers?“ (Fucks 1968: 78) wurde in Best (2012) am Beispiel von Heinrich Heines *Atta Troll* untersucht, welche Gesetzmäßigkeit die Verteilung der Wörter auf Verszeilen in diesem Werk steuern könnte. Der Beitrag stand unter der Hypothese, dass die Häufigkeit, mit der unterschiedlich viele Wörter in den Versen eines Textes vorkommen, einem Sprachgesetz unterliegt. Die Theorie dazu wurde den Arbeiten von Altmann (1988a, b), Wimmer u.a. (1994) sowie Wimmer, Altmann (1996) entnommen, in der Annahme, dass die Häufigkeit von unterschiedlichen Verslängen in Texten der gleichen Gesetzmäßigkeit folgt wie die von Satz- oder Wortlängen. Es hat sich herausgestellt, dass von den Verteilungen, die hierfür in Frage kommen, die verschobene Binomialverteilung am ehesten geeignet erscheint. Dies gilt auch für die wenigen Erhebungen, die von Muller (1972) und Grotjahn (1979) genannt werden (Tests dazu in Best 2012).

Im vorliegenden Beitrag geht es nun darum, die bisher noch geringe Datenbasis zu erweitern. Als Beispiel wurden 20 Gedichte von G. A. Bürger bearbeitet. Die Gedichte wurden vor allem danach ausgewählt, dass sie eine gewisse Mindestlänge aufweisen sollten; ansonsten war eher der Zufall das zugrunde gelegte Auswahlprinzip.

2. Bearbeitung der Texte

Für die Bearbeitung der Texte galten folgende Prinzipien: Das „Wort“ wird als ununterbrochene Graphemkette definiert; Bindestriche und Apostrophe werden als Schriftzeichen gewertet, die Graphemketten zu einem Wort vereinen. Enklitika werden also nicht als eigenständige Wörter aufgefasst. Die Verszeile ergibt sich aus dem Druckbild der Gedichte. Es wurde immer das ganze Gedicht ausgewertet, aber ohne die Überschrift (also nur der laufende Text).

3. Zur Frage nach einem Modell für die Verslängenverteilung

Anknüpfend an Best (2012) wurde die Hypothese geprüft, dass auch im Fall der Gedichte von G. A. Bürger die Binomialverteilung

$$P_x = \binom{n}{x} p^x q^{n-x}, \qquad x = 0, 1, \ldots n$$

sich als geeignetes Modell für die Verteilung der Wörter auf Verszeilen erweisen wird. Die Binomialverteilung ist hier in der unverschobenen Form angegeben, muss aber immer in verschobener Form angewendet werden, da es keine Verse mit null Wörtern gibt. Allerdings kann auch keine einheitliche Form der verschobenen Binomialverteilung angegeben werden, da die kürzesten Verslängen in den Gedichten unterschiedlich ausfallen. Die Ergebnisse finden sich im folgenden Abschnitt 4.

4. Anpassung der verschobenen Binomialverteilung an die Gedichtdateien

Die Ergebnisse der Anpassung der Binomialverteilung an die Gedichte G. A. Bürgers finden sich in den folgenden Tabellen. Die Anpassungen wurden mit einer geeigneten Software, dem Altmann-Fitter (1997), durchgeführt.

In den Tabellen sind folgende Angaben enthalten: *x* ist die Zahl der Wörter pro Verszeile, n_x die Zahl der Verszeilen mit *x* Wörtern, NP_x die aufgrund der Binomialverteilung zu erwartende Anzahl der Verszeilen mit *x* Wörtern; *n* und *p* sind die Parameter der Binomialverteilung; X^2 ist das Chiquadrat, *P* die Überschreitungswahrscheinlichkeit für das berechnete Chiquadrat; *FG* gibt die Zahl der Freiheitsgrade an. Eine Anpassung mit $P \geq 0.05$ gilt als zufriedenstellend; Ergebnisse mit $0.05 \geq P > 0.01$ gelten nicht als zufriedenstellend, aber auch nicht als völlig misslungen. Diese Bedingungen sind mit einer Ausnahme in allen Fällen erfüllt.

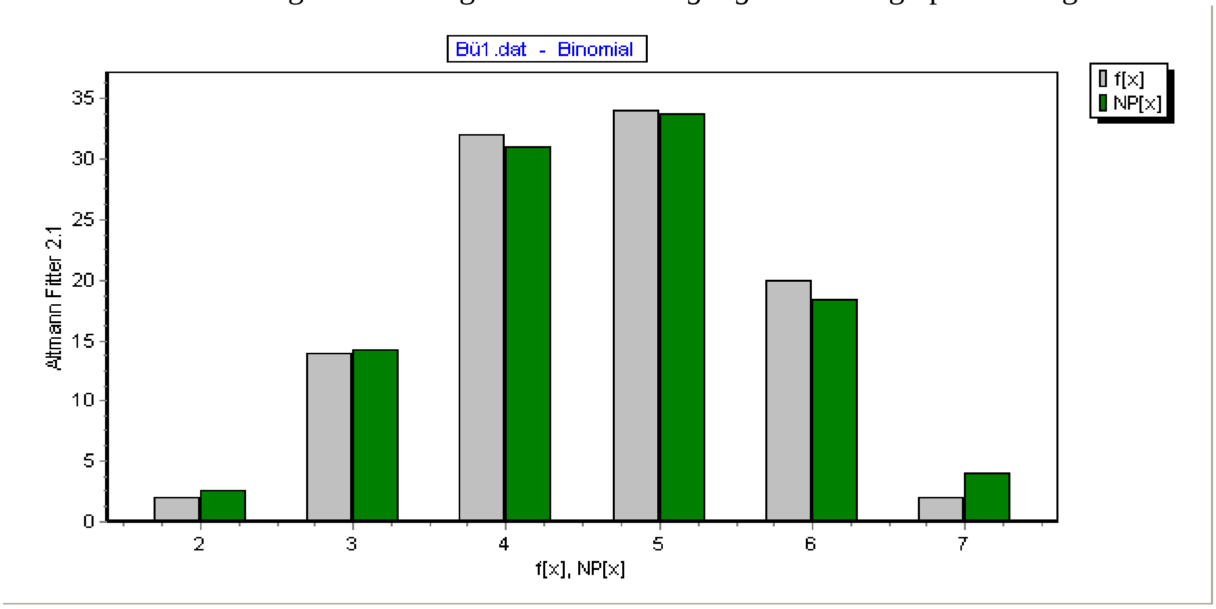
Die Binomialverteilung wird, wie bereits erwähnt, in verschobener Form angepasst, da kein Vers mit x = 0 Wörtern existiert. In der angegebenen Formel muss dazu lediglich statt *x* nun bei 1-verschobener Form, nämlich dann, wenn ein Vers nur ein Wort enthält, $x - 1$ gesetzt werden, bei 2-verschobener Form $x - 2$, wenn die Datei mit $x = 2$ beginnt, etc. Entsprechend verschiebt sich auch die obere Grenze des Definitionsbereiches auf $n + 1$, $n + 2$ usw.

Zu den Tabellen: Die Angaben im Kopf der Tabellen (Titel der Gedichte mit Seitenangabe) beziehen sich auf die Ausgabe: Bürger, *Sämtliche Werke.* 1987. Die Gedichte stammen alle aus der Sammlung *Gedichte 1789. Erster Teil* und *Gedichte 1789. Zweiter Teil.*

Anmerkung zu *Der Liebesdichter*: die senkrechten Striche in der Datei zeigen eine Zusammenfassung der betreffenden Klassen an; dies gilt auch für alle folgenden Dateien.

	1. *Huldigungslied*, 38-42		2. *An die Hoffnung*, 45-49		3. *Der Liebesdichter*, 52-54		4. *An Agathe*, 54-56	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2	2	2.61	3	9.58	1	1.00|	1	1.13
3	14	14.22	39	33.78	9	6.49|	3	6.67
4	32	30.99	48	44.64	11	16.83	17	15.80
5	34	33.77	27	26.22	25	21.84	27	18.71
6	20	18.39	3	5.78	14	14.17	7	11.08
7	2	4.01			4	3.68	1	2.62
n =	5		4		5		5	
p =	0.5214		0.4684		0.5647		0.5422	
*X*² =	1.327		6.941		3.353		8.306	
FG =	3		2		2		3	
P =	0.72		0.03		0.19		0.04	

Zur Veranschaulichung wird das Ergebnis zum *Huldigungslied* auch graphisch dargestellt:



Graphik zum *Huldigungslied* (Der jeweils linke Balken steht für die beobachteten Werte, der rechte für die Werte, die man erhält, wenn man an die Datei des Gedichts die verschobene Binomialverteilung anpasst.)

	5. *Danklied*, 57-59		6. *Das Lob Helenens*, 61-63		7. *Die beiden Liebenden*, 67-71		8. *Die Menagerie der Götter*, 77-80	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2							2	1.26
3	2	2.68	9	7.97	6	4.40	7	7.14
4	12	12.98	17	18.64	18	22.68	12	16.85
5	26	23.57	18	18.70	45	46.76	25	21.20
6	22	19.02	13	10.42	58	48.20	14	15.00
7	2	5.76	2	3.49|	21	24.84	8	6.55
8			1	0.78|	4	5.12		
n =	4		7		5		6	
p =	0.5476		0.2506		0.5076		0.4854	
X^2 =	3.416		1.319		4.446		2.901	
FG =	2		2		3		3	
P =	0.18		0.52		0.22		0.41	

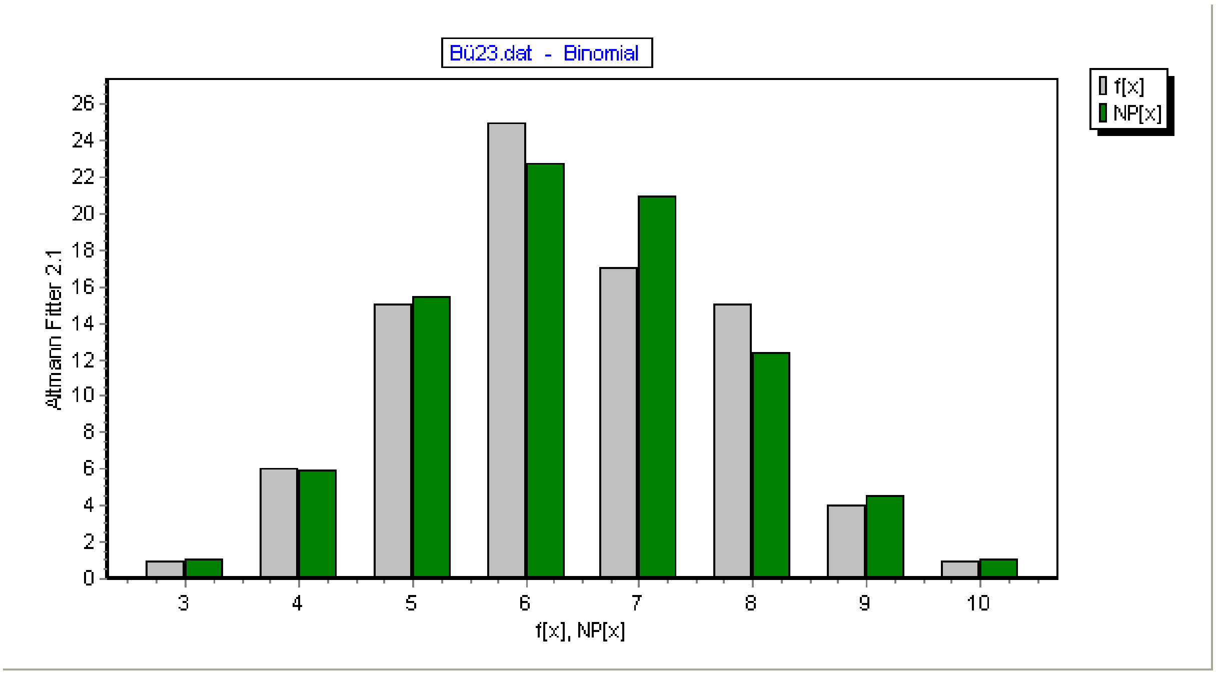
	9. *Das Mädel, das ich meine*, 84-86		10. *Die Elemente*, 89-93		11. *Auch ein Lied an den lieben Mond*, 101-103		12. *Elegie*, 105-114	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2							1	2.02
3			4	2.62			17	16.97
4	6	7.57	12	12.73	1	0.68|	57	57.08
5	23	21.76	22	24.72	7	4.54|	96	96.02
6	25	23.44	25	24.01	8	12.58	86	80.75

7	11	11.22	13	11.66	16	18.60	23	27.16
8	1	2.01	2	2.26	21	15.47		
9					6	6.86		
10					1	1.27		
n =	4		5		6		5	
p =	0.4180		0.4927		0.5259		0.6271	
*X*² =	1.018		1.293		5.654		1.493	
FG =	2		3		3		3	
P =	0.60		0.73		0.13		0.68	

	13. *Fortunens Pranger*, 116-120		14. *Das hohe Lied von der Einzigen...*, 131-143		15. *Gesang am heiligen Vorabend...*, 146-151		16. *Das Blümchen Wunderhold*, 158-160	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
1			1	1.09				
2			10	11.07	8	5.07	3	1.00|
3	1	1.91	32	47.04	19	20.44	7	6.84|
4	6	11.51	112	106.57	33	37.42	19	19.49
5	34	28.91	170	135.80	38	41.10	30	29.62
6	43	38.75	91	92.30	37	30.10	22	25.32
7	29	29.21	4	26.14	17	15.43	14	11.54
8	9	11.74			4	5.65	1	2.19
9	2	1.97			1	1.78		
n =	6		6		11		6	
p =	0.5013		0.6295		0.2680		0.5327	
*X*² =	5.072		32.576		5.111		2.220	
FG =	4		4		5		3	
P =	0.28		0.00		0.40		0.53	

	17. *Lenore*, 178-188		18. *Der Raubgraf*, 188-192		19. *Die Weiber von Weinsberg*, 193-97		20. *Die Kuh*, 271-274	
x	n_x	NP_x	n_x	NP_x	n_x	NP_x	n_x	NP_x
2					1	0.90|		
3	18	17.41	11	8.92	6	5.94|	1	1.01
4	55	59.04	29	29.67	18	16.41	6	5.97
5	89	83.43	36	43.18	21	24.19	15	15.40
6	62	62.88	42	35.91	21	20.06	25	22.72
7	25	26.65	20	18.67	11	10.50	17	20.94
8	7	6.59	6	7.66			15	12.35
9							4	4.56
10							1	1.05
n =	6		8		6		8	
p =	0.3611		0.2937		0.5250		0.4244	
*X*² =	0.809		3.183		0.646		1.618	
FG =	3		3		2		5	
P =	0.85		0.36		0.72		0.90	

Die Graphik zu *Die Kuh* sieht wie folgt aus:



5. Ergebnis und Perspektive

Die folgende Tabelle gibt eine Übersicht über die Ergebnisse der Anpassung der verschobenen Binomialverteilung an die Gedichte:

Gedicht	*P*	Gedicht	*P*	Gedicht	*P*	Gedicht	*P*
1	0.72	6	0.52	11	0.13	16	053
2	0.03	7	0.22	12	0.68	17	0.85
3	0.19	8	0.41	13	0.28	18	0.36
4	0.04	9	0.60	14	0.00	19	0.72
5	0.18	10	0.73	15	0.40	20	0.90

Es ist zu konstatieren, dass von den 20 Gedichten G.A. Bürgers 19 der verschobenen Binomialverteilung unterliegen; in zwei dieser Fälle ist das Ergebnis nicht wirklich befriedigend, muss aber auch nicht ganz verworfen werden. An Gedicht 2, *An die Hoffnung*, kann man statt der Binomialverteilung die verschobene Hyperbinomialverteilung mit *P* = 0.15 erfolgreich anpassen. Bei Gedicht 4, *An Agathe*, bewährt sich die verschobene Morse-Verteilung mit *P* = 0.33, die in der Linguistik bisher aber keine theoretische Begründung hat. (Zu beiden Verteilungen siehe die entsprechenden Kapitel in Wimmer & Altmann 1999.). Nur Gedicht 14 folgt der Binomialverteilung nicht, aber anscheinend auch keiner anderen.

Die Befunde bestätigen die Ergebnisse der Untersuchung von Best (2012). Es bleibt jedoch die Frage, ob die Binomialverteilung eine Form des Verteilungsgesetzes ist, die sich bei beliebigen Verstexten bewährt, oder ob je nach Sprache, Autor, Zeit, Textsorte mit anderen Verteilungen gerechnet werden muss? Denkbar wäre auch, dass sich weitere Texte finden lassen, bei denen keine der bisher vorgeschlagenen Verteilungen als Modell dienen kann. Bisher wurden in beiden Untersuchungen zusammen zwei unter den insgesamt 43

Texten/Textdateien gefunden, bei denen die verschobene Binomialverteilung versagt. Bisher sieht es so aus, als ob keine andere Verteilung besser geeignet sei, als Verteilung für die Zahl der Wörter je Vers zu dienen.

Bestimmt man die Verslänge anders, als es hier erfolgt ist, also durch die Zahl der Buchstaben, Laute, Moren, Morphe, Phoneme, Silben oder Versfüße, muss in jedem dieser Fälle damit gerechnet werden, dass das Sprachgesetz, das die Verteilung sprachlicher Einheiten regelt, andere Formen annimmt.

Als Ergebnis kann konstatiert werden, dass die Häufigkeit der Verslängen nicht chaotisch ist, sondern von einem Sprachgesetz gesteuert wird, das mit der Theorie der Verteilung von Satz- oder Wortlängen (und anderer sprachlicher Einheiten) übereinstimmt.

Literatur

Altmann, **Gabriel** (1988a). Verteilungen der Satzlängen. In: Schulz, Klaus-Peter (ed.), *Glottometrika 9* (S. 147-169). Bochum: Brockmeyer.

Altmann, **Gabriel** (1988b). *Wiederholungen in Texten.* Bochum: Brockmeyer.

Best, Karl-Heinz (2012). How many words are in a verse? An exploration. In: Naumann, S., Grzybek, P., Vulanović, R., Altmann, G. (eds.), *Synergetic linguistics. Text and language as dynamic systems: 13-22*. Wien: Praesens.

Bürger, Gottfried August (1987). *Sämtliche Werke.* Hrsg. von Günter und Hiltrud Häntzschel. München/Wien: Carl Hanser Verlag.

Fucks, Wilhelm (1968). *Nach allen Regeln der Kunst. Diagnosen über Literatur, Musik, bildende Kunst – die Werke, ihre Autoren und Schöpfer.* Stuttgart: Deutsche Verlagsanstalt.

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

Muller, Charles (1972). *Einführung in die Sprachstatistik*. München: Hueber.

Wimmer, Gejza, & Altmann, Gabriel (1996). The Theory of Word Length Distribution: Some Results and Generalizations. In: Schmidt, Peter (Hrsg.), *Glottometrika 15* (S. 112-133). Trier: Wissenschaftlicher Verlag Trier.

Wimmer, Gejza, & Altmann, Gabriel (1999). *Thesaurus of univariate discrete probability distributions*. Essen: Stamm.

Wimmer, Gejza, Köhler, Reinhard, Grotjahn, Rüdiger, & Altmann, Gabriel (1994). Towards a Theory of Word Length Distribution. *Journal of Quantitative Linguistics 1, 98-106.*

Software

Altmann-Fitter (1997). *Iterative Fitting of Probability Distributions.* Lüdenscheid: RAM-Verlag.

Glottometrics 23, 2012, 62-69

Literal vs. Liberal Translation – Formal Estimation

Sergey Andreev[1]

Abstract. The article tackles the problem of measuring the degree of correspondence of verse translation from one language into another. To this end a number of formal features are taken into consideration and different tests are performed using Coleridge's texts and their translations into Russian.

Keywords: translation, formal feature, discriminant analysis

Introduction

One major problem confronting any translator (of poetry, in particular) is the problem of balancing two approaches one of which is directed at recreating in translation the exact formal feature system of the original (or, at least, following it as closely as possible) whereas the other consists in sacrificing accuracy of translation in order to fit the original feeling of the piece of work. The manner in which a translator balances the accuracy of translation and the rules of the language into which the text is translated ("resistance of material"), these Scylla and Charybdis on the way to the reader, may serve as a marker of the creative manner of translator.

This problem is further aggravated by the fact that strictly speaking there is no instrument which could show the degree of alienation of the translation, so one and the same translation can be considered as "too literal" or "too free, too liberal." Assessment is usually very subjective. Corinne McKay in her paper "We just translate? or do we?" gives examples when one and the same her translation was rejected in different offices on quite contradictory grounds – as being "too faithful to the original" in one case, and because it "sounded great in English, but it strayed too far from the original" – in the other (McKay 2008).

M. Gasparov proposed to measure the degree of liberty of the translation by counting nouns, adjectives and verbs and adverbs in translation which are absent in the original. And vice versa, the number of notional words which are present in both the original and translation will show the degree of exactness of translation (Gasparov 2001). This approach seems to be very fruitful, but by no means easy to use because of the problems of equivalence, synonymy and polysemy.

Hypothesis

The degree of correspondence of translations of verse to the original can be established on the basis of their formal features.

[1] Address correspondence to: Sergej Andreev, Bakunin str. 13, apt. 3, 214000 Smolensk, Russia.
E-mail: smol.an@mail.ru

Data and properties

Verse is a much more organized text than prose which is why translation of verse presents additional difficulties and requires both the talent of a poet and the technique of a translator.

The data source for the study is the poem *The Rime of the Ancient Mariner* by Samuel Taylor Coleridge (1772 – 1834) – an English poet, Romantic, literary critic and philosopher who was a founder of the Romantic movement in England and a member of the Lake Poets. *The Rime of the Ancient Mariner* is one of his best poems (written in 1797-1798). It has many semantic layers and very many interpretations such as allusion to the coming of Christ and the public reversion from adoration to hatred, crime and its atonement, co-existence of two worlds – real and spiritual, and many others. Due to the complexity of senses its formal structure becomes of special importance. It should ne noted that, of course, in verse formal features have much semantic relevance.

To determine the degree of similarity of the original and translations on a formal level the following features were examined.

Syntax:

1. the number of complex sentences (COMPL);
2. the number of simple sentences (SIMPLE);
3. the number of clauses (CLS);
4. the number of sentences with complete inversion (INV-F);
5. the number of sentences with partial inversion (INV-P);

Poetical syntax:

6. the number of enjambements (ENJ);
7. the number of lines broken with syntactic pauses (PAUSE);
8. the number of emphatic lines (EMPH-LINE);
9. the number of emphatic sentences (EMPH-SENT).

Stanza types (according to the type of rhyming):

10. the number of stanzas with the rhyme type ABCB (ABCB);
11. the number of stanzas with the rhyme type ABAB (ABAB);
12. the number of stanza with the rhyme type ABCBDB (ABCBDB).

Parts of speech in rhyme:

13. the number of nouns (N-R);
14. the number of verbs (V-R);
15. the number of adjectives (ADJ-R);
16. the number of pronouns (PRN-R);
17. the number of adverbs (ADV-R).

Parts of speech in the final strong position of lines which are not rhymed:

18. the number of nouns (N-UNR);
19. the number of verbs (V-UNR);
20. the number of adjectives (ADJ-UNR);
21. the number of pronouns (PRN-UNR);
22. the number of adverbs (ADV-UNR).

The features which we use in the study were chosen from those that are considered as very relevant for the verse text, are often used in many studies of verse. They reflect vertical and horizontal integrity of the text. These are formal characteristics, but it should be mentioned that all the so-called formal features of verse (rhythm, rhyme, stanza organization, syntax) have semantic relevance. There is nothing formal in poetry which has no meaning.

It is generally accepted that semantic flavour fills all structures and all deviations from the standard.

The calculations were performed separately for each chapter of the three texts – the original and two translations (see Supplement).

Analysis and results

The analysis consisted of several stages. At the first stage characteristics which discriminated the three texts were revealed, that is the characteristics, which have a relatively higher differential capacity for the analyzed classes. Then the degree of similarity of the texts was defined. In both cases discriminant analysis was used which was proved effective in a number of linguistic studies (Andreev 2010; Micros, Carayannis 2000). At the third stage the degree of similarity of the texts according to their structural organization was studied. In this case cluster analysis was applied.

Discriminant analysis showed that the main features which introduce difference between the texts are ENJ, INV-F, INV-P, N-R, V-UNR, ADJ-R, PRN-UNR, ABAB, COMPL, CLS, V-R, SIMPLE.

Post-hoc test gave a very high degree of correctness – 100% (Table 1). In Table 1 lines contain expected groups and columns – automatically received.

Table 1
Post hoc test

	Correct	Coleridge	Gumilev	Levik
Coleridge	100%	7	0	0
Gumilev	100%	0	7	0
Levik	100%	0	0	7
Total	100%	7	7	7

Fig. 1 demonstrates the position of the objects (parts of three texts) in the space of two discriminant functions (roots). It should be underlined that both functions are relevant and statistically significant (see Supplement).

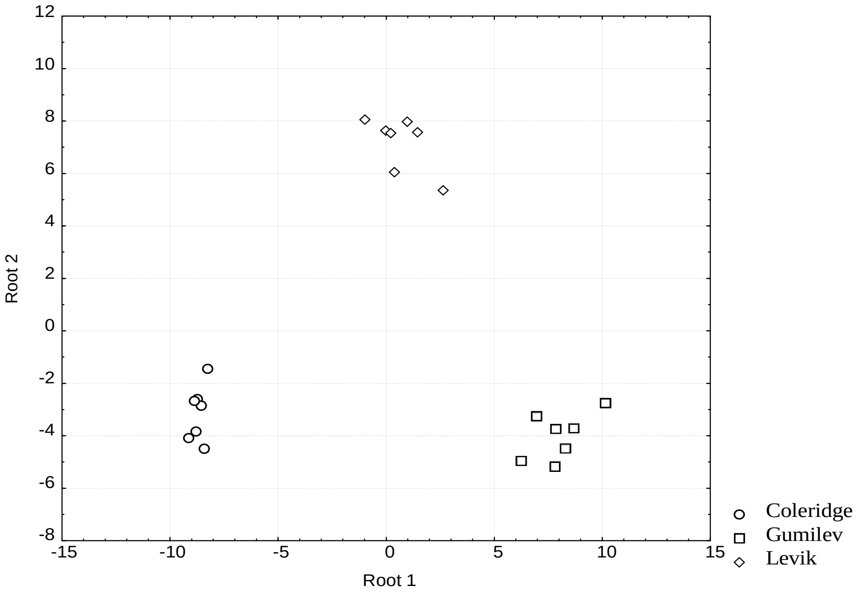


Fig. 1. Objects in the space of two discriminant functions

The first function differentiates the original from two translations, especially from the translation of Gumilev. Judging by the standardized coefficients (Table 2) the greater contribution to the discrimination between the groups are observed for the characteristics ENJ, INV-F, INV-P and V-UNR, ADJ-R, ABAB, SIMPLE.

The second function discriminates between the original text and Gumilev's translation, on the one hand, and Levik's translation, on the other hand. The biggest contributions are made by N-R, CLS and COMPL, ABAB, PRN-UNR, INV-P, V-R, SIMPLE.

Both translations seem to be equally distant from the original, but there is some difference between them too. It seems that "the resistance of the material" (i.e. the influence of the system of the Russian language on the texts of the translations) is not strong enough to subdue the differences of the manner of translators when compared to the original.

The results show that all three characteristics, unique in English for their changing the standard word-order in sentences (inversions) and the structure of a verse text (enjambement in the verse) are used differently in the original and translations. If such difference in the use of inversion can to some extent be explained by less constrains on the word-order in Russian, difference in the number of enjambments can not be ascribed so readily to language factors.

Levik differs in one important feature – rhymed nouns and verbs. Nouns are the main source of forming images, verbs – of motifs. In the final strong position in rhyme their meanings are intensified.

Table 2
Standardized coefficients for the characteristics in discriminant functions

	Function 1	Function 2
ENJ	2,04	0,01
INV-F	2,77	-0,29
INV-P	1,11	-1,13
N-R	-0,66	2,09
V-UNR	-1,03	0,93
ADJ-R	-1,39	-0,62
PRN-UNR	0,17	-1,74
ABAB	-1,25	-2,36
COMPL	-0,80	-7,01
CLS	0,56	6,15
V-R	0,23	-1,20
SIMPLE	-1,01	-1,00
Eigenvalue	54,39	30,09
Cumulative Prop.	0,64	1,00

The results obtained give a general estimation of the degree of correspondence of the translations to the original. Now we shall try and see the correspondence of their structures. The structure of a text is represented by the type of relationship of its parts which in its turn is characterized by the representation of syntactic features in different parts. In order to compare the structure of these three texts cluster analysis (hierarchical, complete linkage) was used. All the parts of the original and translations were grouped in the space of their syntactic characteristics: COMPL, SIMPLE, CLS, INV-F, INV-P, ENJ, PAUSE, EMPH-LINE.

The results of this analysis are shown in Fig. 2–4.

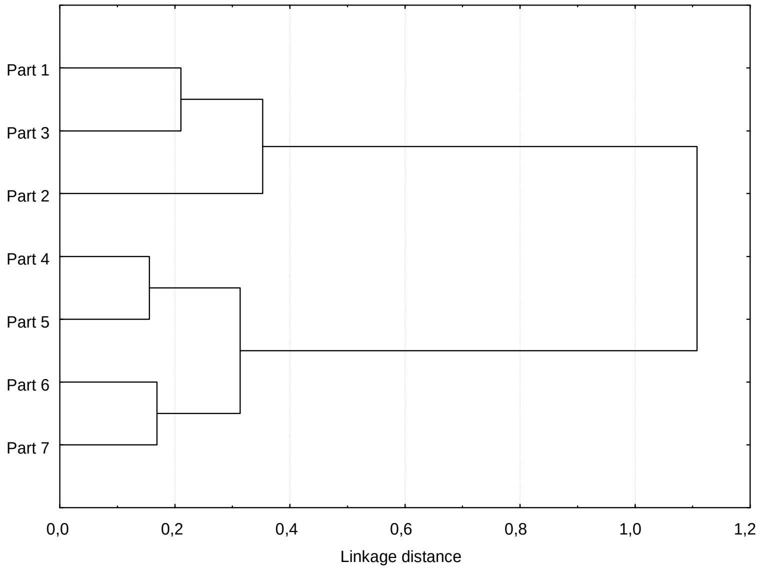


Fig. 2. Grouping of the parts of the text by Coleridge.

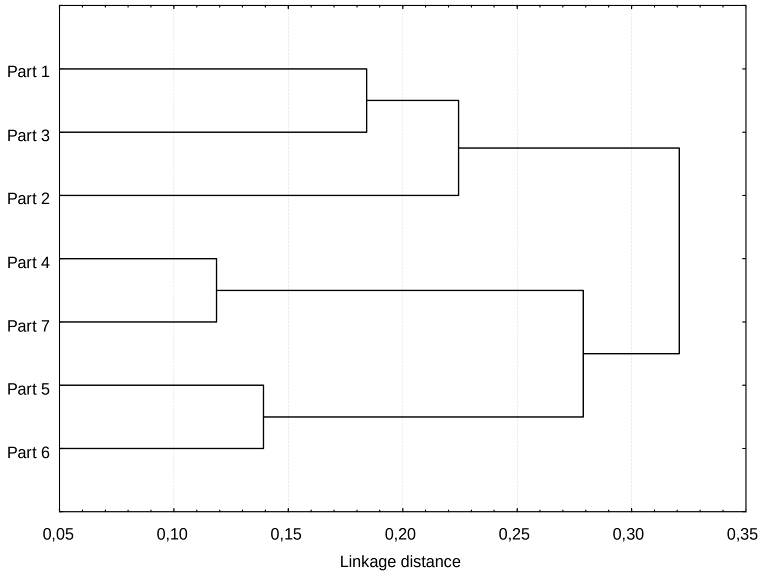


Fig. 3. Grouping of the parts of the text by Gumilev.

Both translations are rather homogeneous. Judging by the linkage distances all their parts are relatively close to one another whereas the original is more clearly separated into clusters.

The structural organization of the original and translations is rather similar. Fig. 2 shows that the text of the original is subdivided into 2 main clusters: Parts 1–3 and 4–7. The closest to each other are Parts 1 and 3, 4 and 5, 6 and 7. It is interesting to note that syntactically the closest in most cases are the neighboring parts. Parts 1–3 deal with the transition of the ship from the real into spiritual world and with crime and punishment, 4 and 5 – speak about the way to spiritual redemption and Parts 6 and 7 – of partial redemption and at

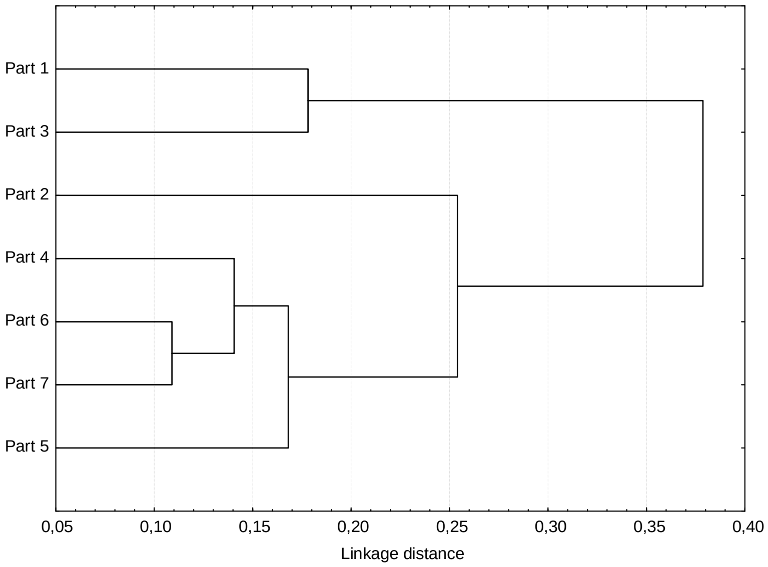


Fig. 4. Grouping of the parts of the text by Levik.

the same time of further torment of the Mariner in the real world by the necessity (which is, in fact, a curse) of telling his story to selected people.

Gumilev's text (Fig. 3) follows the main division of the original, falling into two main clusters which coincide with the original. The first cluster also consists of Parts 1 – 3, and the order of their joining is the same. The inner structure of the second cluster is to some extent different from the original. Here Parts 4 and 7 (partial redemption in the spiritual and real world) are closest.

Levik's translation structure (Fig 4) deviates a little from the original. His text also shows the opposition of Parts 1 and 3 to Parts 4–7. But the difference is in the separation of Part 2 (the idea of responsibility) from the rest, forming thus an opposition of severe punishment to other events.

On the whole it is possible to state that there is a high degree of similarity between the original and both translations in their syntactic organization. Both translators preserved the main opposition between the beginning and the end of the story (themes of the beginning and the end of the journey, of crime and punishment and of physical and moral sufferings, followed by partial redemption). In both translations there is certain separation of Part 2 (in Levik's translation – more, in Gumilev's – less).

Summing up the results it is possible to state that the multivariate analysis made it possible to carry out the measurement of the similarity of the original poem and verse translations displaying the structural peculiarities of the translations.

It should be stressed that we do not intend to give any kind of esthetic or literary estimation of translations.

References

Andreev S. (2010). Quantitative analysis of Keats' style: genre differences, In: Grzybek, P., Kelih, E., Mačutek, J. (eds.), *Text and Language: Structures. Functions. Interrelations. Quantitative perspectives: 1-11.* Wien: Praesens Verlag.

Gasparov M.L. (2001). Podstrochnik i mera tochnosti [Word for word translation and the degree of laterality] In: *O russkoj poezii. Analizi. Interpretatsii. Kharakeristiki: 361-372*. Moscow.

McKay, C. (2008), “We just translate”…or do we? In: *Thoughts on Translation.* http://thoughtsontranslation.com/2008/12/11/we-just-translateor-do-we/ (accessed 21.10.2011).

Mikros G., Carayannis G. (2000). Modern Greek corpus taxonomy. In: *Proceedings of the 2nd International Conference on Language Resources and Evaluations. Athens, Greece (31 May – 2 June). 2000. Vol. 3. P. 129–34.*

SUPPLEMENT

Data-base

	COMPL	SIMPLE	CLS	INV-F	INV-P	ENJ	PAUSE	EMPH-LINE	EMPH-SENT	ABCB	ABAB	ABCBDB
CO-1	1,33	0,27	0,35	0,17	0,13	0,07	0,29	0,13	0,17	0,22	0	0
CO-2	1,55	0,15	0,23	0,12	0,22	0,15	0,47	0,10	0,18	0,18	0,02	0,03
CO-3	1,31	0,23	0,27	0,14	0,17	0,11	0,42	0,26	0,36	0,11	0,01	0,02
CO-4	0,76	0,07	0,09	0,09	0,15	0,15	0,34	0,10	0,15	0,12	0,02	0,03
CO-5	0,69	0,07	0,07	0,03	0,20	0,12	0,23	0,09	0,11	0,13	0,01	0,03
CO-6	0,47	0,07	0,10	0,04	0,12	0,13	0,36	0,19	0,27	0,18	0,04	0,02
CO-7	0,62	0,04	0,04	0,03	0,13	0,19	0,37	0,16	0,20	0,13	0,02	0,03
GU-1	0,71	0,01	0,01	0,32	0,22	0,24	0,30	0,06	0,06	0,18	0,05	0
GU-2	0,67	0,05	0,10	0,20	0,32	0,23	0,30	0,03	0,03	0,15	0,03	0,02
GU-3	0,73	0,04	0,04	0,22	0,21	0,23	0,42	0,16	0,20	0,12	0	0,02
GU-4	0,79	0,10	0,10	0,25	0,22	0,25	0,29	0,10	0,10	0,12	0	0,01
GU-5	0,68	0,08	0,11	0,29	0,17	0,27	0,14	0,06	0,05	0,11	0,02	0,03
GU-6	0,75	0,09	0,13	0,24	0,16	0,22	0,15	0,15	0,14	0,17	0	0,02
GU-7	0,84	0,12	0,15	0,26	0,16	0,21	0,33	0,16	0,17	0,14	0	0,02
LE-1	0,98	0,05	0,10	0,37	0,08	0,10	0,36	0,12	0,16	0,16	0,01	0
LE-2	0,77	0,11	0,12	0,32	0,20	0,17	0,34	0,09	0,08	0,11	0	0,02
LE-3	0,94	0,03	0,03	0,29	0,05	0,17	0,35	0,23	0,32	0,02	0	0,01
LE-4	0,79	0,04	0,04	0,22	0,10	0,16	0,25	0,18	0,19	0,09	0	0,04
LE-5	0,76	0,03	0,06	0,21	0,11	0,21	0,36	0,07	0,08	0,13	0	0,03
LE-6	0,70	0,04	0,06	0,16	0,08	0,20	0,29	0,16	0,17	0,19	0	0,02
LE-7	0,77	0,04	0,09	0,13	0,11	0,21	0,31	0,12	0,14	0,08	0,01	0,03

	N-R	ADJ-R	V-R	ADV-R	PRN-R	N-UNR	ADJ-UNR	V-UNR	ADV-UNR	PRN-UNR
CO-1	0,27	0,04	0,09	0,05	0,06	0,34	0,00	0,07	0,02	0,02
CO-2	0,27	0,05	0,13	0,02	0,02	0,25	0,00	0,15	0,03	0,02
CO-3	0,36	0,02	0,16	0,04	0,02	0,17	0,02	0,07	0,06	0,00
CO-4	0,26	0,10	0,12	0,07	0,03	0,22	0,06	0,06	0,01	0,03
CO-5	0,33	0,05	0,14	0,05	0,01	0,25	0,03	0,03	0,03	0,02
CO-6	0,27	0,03	0,18	0,05	0,01	0,19	0,03	0,04	0,12	0,02
CO-7	0,27	0,08	0,15	0,06	0,03	0,20	0,03	0,10	0,02	0,03
GU-1	0,39	0,04	0,12	0,01	0,05	0,29	0,00	0,07	0,01	0,01
GU-2	0,38	0,02	0,20	0,00	0,02	0,22	0,00	0,08	0,00	0,05
GU-3	0,26	0,05	0,15	0,05	0,10	0,22	0,01	0,05	0,04	0,04
GU-4	0,28	0,10	0,09	0,01	0,09	0,19	0,01	0,07	0,01	0,03
GU-5	0,31	0,06	0,15	0,03	0,06	0,22	0,00	0,06	0,04	0,03
GU-6	0,20	0,04	0,03	0,01	0,16	0,28	0,00	0,06	0,07	0,09
GU-7	0,33	0,04	0,13	0,03	0,04	0,20	0,03	0,06	0,01	0,07
LE-1	0,45	0,04	0,12	0,02	0,01	0,27	0,01	0,07	0,00	0,02
LE-2	0,52	0,05	0,02	0,02	0,02	0,08	0,02	0,22	0,00	0,06
LE-3	0,47	0,06	0,16	0,01	0,10	0,06	0,00	0,01	0,02	0,00
LE-4	0,43	0,03	0,16	0,00	0,01	0,18	0,01	0,06	0,00	0,01
LE-5	0,32	0,06	0,10	0,03	0,06	0,27	0,00	0,07	0,03	0,03
LE-6	0,30	0,02	0,09	0,02	0,08	0,23	0,02	0,12	0,05	0,03
LE-7	0,36	0,03	0,13	0,02	0,09	0,23	0,02	0,09	0,00	0,02

Means of Canonical Variables

	Root 1	Root 2
Coleridge	-8,68	-3,158
Gumilev	8,01	-4,02
Levik	0,67	7,16

Chi-Square Tests with Successive Roots Removed

	Eigenvalue	Canonical R	Wilks' Lambda	Chi-Square.	df	p-level
0	54,39	0,99	0,00	93,14	24,00	0,00
1	30,09	0,98	0,03	42,96	11,00	0,00

Glottometrics 23, 2012, 70-80

History of Quantitative Linguistics

Since a historiography of quantitative linguistics does not exist as yet, we shall present in this column short statements on researchers, ideas and findings of the past – usually forgotten – in order to establish a tradition and to complete our knowledge of history. Contributions are welcome and should be sent to Peter Grzybek, peter.grzybek@uni-graz.at.

Michail Lopatto: *Attempt at an Introduction into the Theory of Prose* (1918)

Peter Grzybek

ПУШКИНИСТЪ
ИСТОРИКО-
ЛИТЕРАТУРНЫЙ
СБОРНИКЪ

подъ редакціей
проф. С.А. Венгерова

III

ПЕТРОГРАДЪ
1918

ПУШКИНИСТЪ.

ИСТОРИКО-ЛИТЕРАТУРНЫЙ
СБОРНИКЪ

подъ редакціей профессора С. А. Венгерова.

III.

This article presents an early scientific paper, published by Michail Osipovič Lopatto (1892-1982) in 1918. It goes back to a presentation he gave in a 1915 seminar on Puškin, held at Sankt Petersburg University. This seminar, led by the renowned Puškin expert S.A. Vengerov, has been referred to as one of the off-shots of the Russian Formalist School. Lopatto's article has been largely neglected in works on the history of Russian Formalism as well as on the history of Quantitative Linguistics. In it, the author attempts to lay the foundations for a general theory of prose, distinguishing between material, form and content, on the one hand, and postulating quantitative methods, on the other. The analyses

offered, and the ideas presented as to the need of quantification, are far from being what might be called 'ripe', from a modern point of view. Yet, they are not only far ahead of their time, but largely wait for an answer still today.

Introduction: Historical Context

Lopatto's paper emerged in the context of Russian Formalism or, to be more correct, on the late pre-formalist threshold to a radically new understanding of literature and its analysis. Generally speaking, Russian Formalism is an umbrella term for a group of scholars who, in the second and third decades of the twentieth century, set out to lay the foundations for modern structuralism and semiotics. Specifically they attempted to develop radically new concepts for a theory of literature, art and, in fact, culture. Literature, specifically realistic prose, had lost its dominant ideological function at the turn of the century; the first Russian revolution in 1905, the beginning of World War II in 1914, and the October Revolution in 1917 meant significant changes not for the Russian society as a whole, but for academia, where younger scholars and students increasingly started to question traditional concepts. Two groups of young scholars are known to have formed the theoretical breeding ground and organizational centers of this movement: In 1915, the Moscow Linguistic Circle (MLK[1]) was founded, followed by the Society for the Study of Poetic Language (OPOJAZ[2]) in 1916, in Sankt Peterburg. The MLK, whose founders were Pëtr G. Bogatyrëv, Roman O. Jakobson, and Grigorij O. Vinokur, existed from 1915-1924; its main interest were language and linguistic approaches to literature and folklore, particularly analytical and methodological approaches to the distinction between what they termed practical vs. poetic language. OPOJAZ, too, was founded by linguists, such as Lev P. Jakubinskij and Evgenij D. Polivanov (both disciples of Baudouin de Courtenay), but from the very beginning, they joined by trained literary scholars such as Viktor B. Šklovskij, Sergej I. Bernštejn, Boris M. Ėjchenbaum, Jurij N. Tynjanov, and Boris V. Tomaševskij. As Victor Erlich in his detailed survey on *Russian Formalism* (1955) reconstructed, many of them had been students of Semën A. Vengerov (1855-1920), an outstanding expert on Puškin, who had gathered young students around him an involved them in theoretical and methodological discussions; among these students was Michail Lopatto, whose name is not mentioned by Erlich, however.

The Russian Formalists shaped the academic atmosphere of the late 1910s and the 1920s in the field of the humanities although. And despite all differences in their personal interests and profiles, all these scholars may be said to have concentrated on literature as verbal art, trying to work out the devices [приемы] (i.e., processes, or structures) by which a verbal message is rendered into a form which is likely to be received as a work of art. In their concentration on form, they did not understand the latter to be in opposition, or juxtaposition, to content. In fact, it was some kind of a priori assumption from the very beginning that form and content can only be conceived of in a dialectical interrelation: any change of form would result in a change of content. Therefore, it was rather the 'formation' of some material (e.g., language, or linguistic material) by way of specific techniques, which was in the center of their interest, and their whole orientation was to determine these specifics and regularities – an approach, which has correctly been related to nomothetic concepts in their orientation (cf. Striedter 1969: XIII). But 'formalist' soon became some kind of derogatory term, since the dominant (political and ideological) tendencies were rather interested in "theories" focusing aspects of content. The formalists would counter these tendencies with an elaboration of their concepts, which was fruitful, in the beginning: after concentrating on the determination of

[1] Московский лингвистический кружок

[2] Общество по изучению поэтического языка

specific devices, seeing an individual work of art first seeing as the sum, then as a complex system of such devices, they extended their view on the whole ‘row’ of literature (as well as other rows, such as music, religion, painting, etc.), finally integrating them into a whole system and thus anticipating what systems-oriented (semiotic) theory of culture, half a century later. At that time, however, starting with the mid-1920s, it became increasingly clear that there could be no peaceful and harmonious co-existence with Soviet ideology and politics.

Nevertheless, despite all differences in the individual scholars’ personal fates (ranging from being exiled as political victims to Roman Jakobson’s international success story), Russian Formalism as a whole became one of the most important and influential theoretical movements of the 20th century, at least in the long run. Starting with Viktor Erlichs above-mentioned book on *Russian Formalism* (1955), and Jurij M. Lotman’s *Lectures on Structural Poetics* (1962), accounts and (re-)interpretations of the history of Russian formalism started, which will hardly ever be finished. Yet, over the last half century, an almost complete picture has been reconstructed of the discussions which took place almost 100 years ago. Only accidentally, individual stones can be added to the overall mosaique patchwork arising, and one such stone seems to be an almost forgotten early work by Michail Lopatto.

◊ ◊ ◊ ◊ ◊

Michail Osipovič Lopatto was born in Vilnjus, in 1892, as the son of a Livonian squire. One of Lopatto’s student mates and life-long friends was Nikolaj M. Bachtin (1894-1950; cf. Christian 1992), a brother of Michail M. Bachtin (1895-1975), who was later to become an outstanding philosopher, whose works are very much appreciated still today. Lopatto was able to finish the Historical-Philological Faculty of Petersburg University in 1917.

During his time in Sankt Petersburg and Odessa, Lopatto was not only interested in literary theory, rather he turned out to be an active (and obviously talented) poet himself. In fact, before he left Russia in 1920, to become a successful merchant, he edited two volumes of his poetry: *Избыток* [The Loss] in Petrograd (1916), and *Круглый стол* [Round Table] in Petrograd/Odessa (1919); the second volume was published in his own publishing house, Omphalos (1916-19). After his emigration, Lopatto first moved to Berlin, later settled in Italy (and received Italian citizenship after Word War II), where he had no (or only very reduced) contact with the Russian literature scene any more. Lopatto mainly worked as a fur trader of. Only at a later period of his life, he would return to poetry: in 1959, he edited a volume of *Стихи* [Verses] in Paris, which included both earlier and later works of his; from his memorial novel, written later, only the first part remained in Italian translation. In Italy, Lopatto’s archive is kept by slavicist Stefano Garzonio.

◊ ◊ ◊ ◊ ◊

Lopatto’s article *Повести Пушкина* [The Tales of Puškin] was published in 1918, in the third volume of the historical-literary series *Пушкинист* [Puškinist]. This series was edited by renowned professor S.A. Vengerov, a literary scholar and a famous expert in Puškin. The title of Lopatto’s article is an allusion, of course, to Puškin‘s *Повести Белкина* [The Tales of Belkin], which, in 1834, represented a milestone in the development of Russian prose literature. But detailed analyses of T*he Tales of Belkin* represent only part of Lopatto’s interest: more important is the sub-title of Lopatto’s text, *Attempt at an Introduction into the Theory of Prose* [Опыт введения в теорию прозы]. In fact, Lopatto repeatedly points out that it is just a *theory* of prose, he wants to lay the foundations for, and that for this purpose, the establishment of *laws* is indispensable – a point of view which is very familiar with everyone who has ever read Mario Bunge’s (1967) *Scientific Research*[3], and which clearly contradicts Viktor Šklovskij’s usage of this term in his important and highly influential article “Art as Device“ [Искусство как прием] from 1917, which was later repeatedly reprented[4], and in which the author, too, attempts to present objective criteria for the description of prose texts, part of these criteria even calling ‘laws’, but using the term ‘law’ in a rather everyday understanding of this word. For Lopatto, as opposed to this, a law cannot be established without quantificational methods, and although his own methods in achieving this goal are rather poor, it is just the claim and postulation, brought forth by a 22-year old student in 1915, which makes this text quite attractive for the history of quantitative linguistics and text analysis.

Although the text was published only in 1918, it goes back to a presentation Lopatto made in Vengerov’s famous Puškin seminar, at the meeting held on February 12, 1915. We know this date from the protocols of the meeting. Half a year later, in fall 1915, after Vengerov’s seminar had ended, a number of students approached Vengerov and suggested to him to continue their work, devoted to the study of 19th century in generally, and to Puškin specifically, by founding a historical-literary circle – one of these students was Michail Lopatto. In accordance with the university rules about student organizations, this society indeed started its work in on December 5, 1915. In the first one and a half year of its existence, that is until summer 1917, this Puškin Circle organized 22 meetings, which took place on Saturday evenings in the Museum of Ancient Times [Музей древности / Muzej drevnosti], until in fall 1917, due to restrictions of tramway traffic in the evenings, the activities came to an ending. Among the participants and speakers of this circle were some of the most outstanding literary scholars of the time, people like poet and literary theorist Andrej Belyj, or literary scholars like Viktor M. Žirmunskij (later an important person in literature and folklore research) and Boris M. Ėjchenbaum, and others. In 1918, the work was continued, but now not any longer in form of a specific student organization; instead, the organizational form had turned into an academic society by the name of Historical-literary Puškin Society at Petrograd University [Историко-литературное общество имени Пушкина при Петроградском университете / Istoriko-literaturnoe obščestvo imeni Puškina pri Petrogradskom Universitete]. The difference was that now, not only students, but any researcher in the field of history and theory of literature, could become a member of this society.

Lopatto had been one of the members of Vengerov’s famous 1915 seminar on Puškin. This seminar was one of the decisive off-shots of the Russian Formalist School, as has been shown by Erlich (1955) in his survey on *Russian Formalism*. On February 12, 1915, Lopatto, delivered a presentation on Puškin’s prose. Although this presentation and its publication three years later became almost forgotten in later times, Lopatto’s contribution to Vengerov’s seminar must be qualified as one of the earliest formalist presentations at large. In fact, Lopatto’s work was well known and highly appreciated (though not uncontradicted) at that time.

As we know from the meeting’s protocol, Vengerov himself is reported to have reacted to Lopatto’s presentation by objecting that one should not talk about sugar and forget to say that it is sweet, thus obviously criticizing Lopatto’s rigorous removal of literary theory from the field of esthetics. A second spontaneous reaction goes back to Jurij N. Tynjanov, himself still a student at that time, who was a participant of the seminar, too, and who was present at Lopatto’s presentation. in his contribution, Tynjanov shifted the attention of what Lopatto had observed and presented into a slightly different direction: whereas Lopatto had mainly focused on issues of frequency, Tynjanov suggested to interpret and elaborate his

[3] Cf. vol. I, p. 318: “No laws, no science.”

[4] This article was later repeatedly reprented, among others, in his book *О теории прозы* [Theory of Prose], where further relevant contributions can be found.

results in terms of sequences, thus raising the question of specific rhythms of prose – a topic which he would later pick up in the early 1920s, elaborating it in context of the then dominating discussion on verse and prose, and on possibilities to theoretically distinguish between these two species of literature. According to Tynjanov's later opinion, the qualification of a text as being written in verse or prose cannot be made on the basis of inherent features; rather, reference must be made to a given literary system regulating the relationship between these two types of artistic speech (as well as various violations of the borderlines between them). At the end of the 1920s, this view would result in the principal distinction of form vs. function.

Lopatto's interest was not, however, to find a basis for a theoretical distinction between verse and prose; rather, his focus was on the description of specific traits of Puškin's prose, by way of this trying to provide a basis from which one would then be able to elaborate and extend. Lopatto sets out with a general critique of the current state the theory of literature.[5]

> Our classical theory of verbal art is in the poorest state, not even containing a remark on scholarly method. Some shreds by ancient rhetorics, historical and philological particulars about the terminology of forms, somehow put together, are declared to be a 'theory'. It goes without saying that such a theory does not explain anything and cannot detect any laws. Its terms are indefinite and random, and the form of works is explained by their content (e.g., ballad, poem). (4f.)

Thus, right from the beginning, Lopatto sees it as a necessary condition that there can be no theory without laws. In this context, he excludes scholars like Potebnja and Ovsvjaniko-Kulikovskij from the general reproach of literary 'theory' of his time being avoid of scientific character; considering the works of these two scholars to be of high scientific value, he classifies their interest, however, primarily in the fields of what we would call today cognitive psychology ("psychology of thinking"), and reception of art, as a conseqeuence attributing them to psychology of language (or psycholinguistics). According to Lopatto, these authors do not concentrate on the specifics of art itself. This, in fact, is Lopatto's major concern: a theory of literature which, according to him (p. 5), "must be autonomous" ["Теория же литературы должна быть автономна"], i.e. independent from other branches of science. With this claim, Lopatto fully sides with his contemporaries, as can be seen, for example, from the famous passage of Roman Jakobson's[6] 1919/21 comparison between literary scholars and police who, interested in a particular person, would arrest everybody around.

Generally speaking, the Russian formalists, who understood themselves to be 'specificators' [спецификаторы], detecting and describing the specifics of art, would strive for an autonomy of a theory of art. But they would dismiss linguistics as a relevant discipline for literary theory to a lesser degree as Lopatto does, and not in such a categorical way – rather, they were literary scholars who would understand (and study) literature as a branch of verbal art und therefore need linguistics as a helpful discipline, or who would conceive of themselves as being linguists understanding the field of linguistics in a broad sense, covering all aspect of verbal products, among them literary works.

In this sense, Lopatto is even more radical, clearly setting apart his interest in founding a theory of literature from the field of linguistics and grammar; for him, the sphere of

[5] In detail, Lopatto terminologically distinguishes between the traditional 'Theory of Verbal Art' [Теория словесности] and what he terms a 'Theory of Literature' [Теория литературы] which, according to him, still remains to be established.

[6] Jakobson's contribution, published in 1921, goes back to a presentation in 1919: „Между тем, до сих пор историки литературы преимущественно уподоблялись полиции, которая, имея целью арестовать определенное лицо, захватила бы на всякий случай всех и всё, что находилось в квартире, а также случайно проходивших по улице мимо."

grammar "includes language and the laws of language", and is not in charge of a theory of prose: "The laws of language are not important for us […]. We need art only" (p. 39). In fact, he restricts the responsibility of linguistics to the sphere of stylistics, only:

> "The gravitation center of my work is in its methods. The conclusions of a theory of prose belong to future, when the methods will be sufficiently elaborated, and when special studies will appear on separate fields of artistic prose." (p. 50)

As to the method of analysis, Lopatto discusses and distinguishes the concepts of material, content, and form, as well as the relevance of these categories for art and a theory thereof. In this, he is in full accordance with the discussions of his time, albeit he is, in fact, slightly ahead of his time. As to literature, its material is language, of course. However, just like no sound in music has any artistic value in isolation, but appears to harmonious or disharmonious only in combination with others, no word has artistic value in itself: again, only in combination with other words, a given word adopts take some value. This usage of the term 'value', however, should not be confounded with the discipline of esthetics, which, according to Lopatto, is concerned with the objectives of art and the laws of its emergence (p. 6): a form, according to Lopatto, cannot be nice in itself, although the analysis of art products may eventually establish specific laws of taste (p. 7): "A theoretician's task is not the evaluation of a work […], but the study of its form" (p. 27).

So, what exactly is such a form, what does it consist of? Given that it is ultimately the combination, i.e. the internal interrelations and successive order of elements, which represent what may be called the *form* of a work, the crucial tasks of analysis can be considered to be (p. 8):

1. to find the formal elements of artistic prose, and
2. to study their interrelation and succession.

Whereas, at first sight, the definition of what a 'form' is seems to be clear, in some preliminary or elementary sense, the situation becomes more complex and difficult with regard to the analysis of a piece of art as a whole: first, it must be decided which elements are to be analyzed and how they can usefully be defined; and second, due to the fact that the relation between these elements is not only qualitative, but also quantitative, a quantitative analysis is necessary along with qualitative approaches, in order to specify the relation between form and content. Taken together, both aspects represent an extraordinary complexity, and both exact and systematic approaches have to be pursued. Again, Lopatto's view turns out to be very modern: as we assume to know since Saussurian times, linguistic elements and entities are given as positive in some a priori sense, but a matter of definition, and as we also know, the study of linguistics structures without paying due attention to the frequency of their occurrence is but the description of one side of the complex linguistic Rubik dice.

With regard to prose texts, which are in the center of Lopatto's approach, there are three basic categories to be studied: these are:

1. the chapter,
2. the paragraph,
3. the phrase.

With regard to all of them, Lopatto offers a number of calculations which he presents in various tables; unfortunately, not only are most data in these tables incomplete, but also are the descriptions of these tables (i.e., their captions or their description in the running text) partly incomprehensible and cannot, as a consequence, be reproduced. Therefore, there is no sense in presenting them here, the more since no reliable re-analyses are possible. Yet, Lopatto's way of stating the problem seems to be clear, in each single case, and it seems reasonable to focus here on his basic ideas.

With regard to the chapter, Lopatto concentrates on chapter length as one basic characteristic. According to him, a chapter represents something like a 'main caesura' in a writer's structuring of a text: by distinguishing chapters, pauses are introduced (not so much for the production, than for the reception process), which result in a specific distribution of information and thus alleviate the reader's reception. Analyzing Puškin's *Повести Белкина*, *Дубровский*, *Капитанская дочка*, and *Пиковая дама*, Lopatto calculates the length of each chapter for each of the stories. From a modern point of view, the results obtained and reported) may be considered to be disappointing – there are no further calculations as, e.g. averages, no questions as to some theoretical distribution model, no question as to specifics of the sequence of longer or shorter chapters (what might be called 'rhythm', in a broad understanding of this term). But one should not forget that these are the very first ideas developed in the direction outlined, some kind of a sketch for a whole research paradigm, completely new at the time of its publication.

For Lopatto, the unit of the paragraph is even more important than the chapter, at least with regard to Puškin's prose. According to Lopatto, the separation of paragraphs is an attempt to convey some kind of discontinuity to continuous thinking. For him, a paragraph in a prose text corresponds to what a stanza [строфа] is in poetry (p. 18). Again concentrating on length, Lopatto raises the question how to measure the length of a paragraph, i.e. in which measure units: whereas it is common to calculate the length of a stanza in the number of verse lines – what might an adequate entity for measuring paragraph length? To measure paragraph length in the quantity of pages, as does Lopatto in case of the chapter, does not seem reasonable to him, not only because most paragraphs are much shorter than one page, but also because the format of the edition and the font chosen will influence the results, particularly when comparing different authors. Likewise refusing to calculate paragraph length in the number of words, and looking for an adequate cognitive unit as the basic measuring unit, Lopatto suggests choosing the phrase as an adequate entity.

A phrase [фраза], according to Lopatto, can further be subcategorized into simple and complex, into direct and indirect ones, etc. In any case, a phrase must not be equated with sentence, and it should not be confused with it: whereas a sentence is a grammatical term, or concept, and as such determines laws of language, a phrase has semasiological meaning – it is a cognitive unit which cannot further be subdivided; a phrase can eventually be expressed by a sentence, but is not the same as a sentence (p. 21).

Thus, calculating the number of phrases per paragraph, Lopatto is well aware of the shortcomings of his study: "True, my figures do not cover the huge material of phrases, but yet, they are large, and the 'law of large numbers' may well be applied to them" (p. 23).

Given this restriction, it is not so important, in our context, that Lopatto must admit that he is not able to establish some law about the pauses between paragraphs – it is much more important that he raises this question. Analyzing the four above-mentioned texts written by Puškin, Lopatto finds the result to show that the average length of a paragraph is always one and the same across the four texts (p. 20). Table 1 reproduces the data as given by Lopatto; in addition to his original data, mean value (x) and standard deviation (s) are presented in the last two columns.

Text	**I**	**II**	**III**	**IV**	**V**	$\bar{x}$	*s*
Povesti Belkina	17	22	118	31	53	48,20	37,02
Dubrovskij	46	34	72	30	8	38,00	20,98
Kapitanskaja dočka	29	47	38	12	80	41,20	22,59
Pikovaja dama	47	4	20	16	80	33,40	27,21

Given the fact that no more data are offered by Lopatto, and re-analyzing these spare data, it can clearly be seen, that paragraph length does not seem to constant; rather, length ranges from 33.40 to 48.20, with an overall mean of $\bar{x} = 40.20$ (*s* = 28.18).[7]

Furthermore comparing these results to those from other authors (Gogol', Tolstoj, and Kuz'min), Lopatto (p. 20) arrives at the conclusion "that for each author there are specific constant extents of paragraph" [свои постоянные размеры абзаца], with a specific variation from minimum to maximum: whereas, according to Lopatto, Puškin's paragraphs are characterized by a mean length of ca. 30 phrases, for Tolstoj, paragraphs concentrate on a length of ca. 15-20 phrases, that is it is two times less as compared to Puškin.

With regard to phrases, Lopatto offers tables with frequencies of phrases, separately for simple phrases and phrases with different degrees of complexity, for direct and indirect ones, for Puškin's texts and those of the other authors mentioned. Again, Lopatto interprets his results in terms of author-specific characteristics.

We will not go into further details here, fading out questions as far as, for example, the possible generalization of Lopatto's observations is concerned, or the assumed author-specific characteristics. These differences might as well be genre-dependent, or vary according to the differing proportions of descriptive, narrative or dialogical passages in the individual texts. In fact, Lopatto himself discusses such factors in detail towards the end of Lopatto's treatise (p. 33ff.), as well as differences with regard to perspicuity vs. abstractness (p. 28ff.); in his ruminations on style, he additionally discusses frequencies of sensitive [чувственные] vs. cognitive [мыслимые] words (the first directly relating to sensations, the second to abstract ideas), and of different parts of speech, concentrating on nouns, verbs, and epithets.

What seems worthwhile mentioning, however, is a remark by Lopatto (p. 21), which he himself does not pursue at any length, but which anticipates ideas which we today use to treat in terms of Menzerathian concepts. Pointing out the fact that Tolstoj's paragraphs consist of significantly less phrases, Lopatto mentions the fact that these phrases in turn are more complex and consist of multi-word fusions – thus, in a way, compensating, or counter-balancing, the lesser extent of paragraphs (p. 21).

◊ ◊ ◊ ◊ ◊

Lopatto's work has been largely neglected in the history of science. But we know that the rare reactions to it have always been extremely controversial, as could already be seen from the two immediate contributions to the discussion by Vengerov and Tynjanov, which were mentioned above.

[7] As has been mentioned above, one should be very cautious with re-analyzing the data given by Lopatto, not only because they tend to be incomplete, but also because their overall amount is rather scarce, resulting in very few data points. Anyway, a Kruskall-Wallis test over these data shows the differences between the four texts not to be significant (X^2 = 0.66, *d.f.* = 3; *p* = 0.88), thus corroborating Lopatto's finding in this case.

In the following years, too, there were contradictory statements on Lopatto's work which, however, had not been forgotten at that time. In 1925, for example, Tomaševskij presented a synopsis of Puškin studies of his time; in it, he explicitly mentioned Lopatto's study as an exception to the overall beginning status [зачатное состояние[8]] (ibd., 90). Interestingly enough, Tomaševskij also made reference to Lopatto in his 1929 article[9] on the question of prose rhythm, illustrated on Puškin's *Пиковая дама [Pique Dame*]. In this article, basically concentrating on regularities of rhythmic structures in cola and sentences, Tomaševskij also mentioned A.M. Peškovskij's (1928) suggestion as to the extended structural analysis of paragraphs; in this context, he mentioned Lopatto's relevant study, and he regretted that Lopatto did not relate his findings to the question of rhythm (Tomaševskij 1929: 301).

As compared to these rather positive reactions, a clearly negative view of Lopatto's work was later taken by D.P. Jakubovič who, in his 1936 synopsis on the study of Puškin's prose, wrote:

> Far-reaching conclusions were claimed in M.O. Lopatto's work about Puškin's short stories. The prematurity and subjectivity of his generalizing constructions about Puškin's style can be seen from the fact alone that Lopatto arrives at his conclusions – which postulate to be exact, on the basis of counting pages, phrases, paragraphs – not according to Puškin's original writings, but on the basis of printed editions […]. Thus, in addition to works, which were published by Puškin, also unpublished drafts are among the works analyzed. In all of them, "the number of pages containing dialogues and descriptions" are equally calculated. According to the very same method, Puškin's prose is compared to the prose of other writers. Accept for some disseminated clever remarks […], nothing from that work entered the academic world and will not enter it.[10]

Subsequently, Lopatto's work became largely forgotten; only in 1975, in his retrospective view on the study of poetics in the 1920s, Vinogradov (1975) would mention Lopatto's contribution, qualifying them as a precursor of Tomaševskij's subsequent studies on prose rhythm.[11]

Notwithstanding the overall neglect of Lopatto's paper as such, it seems to be much more important that many, if not most, questions raised in his text, wait for an answer still

[8] Quoted from Garzonio (2006b: 138; fn. 11)

[9] This article goes back to a presentation which Tomaševskij gave in the Moscow Linguistic Circle in 1920 and, in some elaborated form, on June 26, 1921, in the Russian Institute for the History of Arts. (RIII). Parts of that enlarged version were published separately in 1920/21 and eliminated from the 1929 version, mainly critical remarks on theoretical positions held by Andrej Belyj and Valerij Brjusov.

[10] „На широкие выводы претендовала работа М.О. Лопатто о повестях Пушкина. Преждевременность и субъективность обобщающих построений о стиле Пушкина видны хотя бы из одного того, что Лопатто делает свои претендующие на точность выводы на основании подсчета страниц, фраз, абзацев не по рукописям Пушкина, а по изданиям (между прочим, неизвестно каким). В число анализируемых вещей, рядом с напечатанными Пушкиным, попадают и черновые, неоконченные. Во всех одинаково подсчитывается „число страниц диалога и описания". Таким же методом проза Пушкина сравнивается и с прозой других писателей. Кроме разве отдельно брошенных умных замечаний […] ничего из работы не вошло и не сможет войти в научный оборот."

[11] Most recently, Italian Slavicist Stefano Garzonio, the keeper of Lopatto's archive, made references to him in two publications (Garzonio 2006a,b).

today. In this respect, it seems worthwhile pointing out, that insights most recently obtained on both chapter length and paragraph length point into the direction outlined by Lopatto. Neumann (2009), in her extensive work on paragraph length, analyzing 57 German texts of different kinds, is able to show that obviously there are indeed regularities in the organization of paragraph length which, however, seem to be text-type specific; she also proves there to be regular relations between paragraph and sentence length, which can be interpreted in Menzerathian terms. Likewise, Grzybek (2012a,b), analyzing chapter length of Tolstoj's *Война и мир* [War and Peace] for all 336 chapters separately, shows that not only follows the frequency distribution of chapter length follows a well-known distribution model (in detail, the hyper-Pascal distribution), but also corresponds the dependence of sentence length on chapter length the well-known Menzerath-Altmann law. It seems that with quite some temporal distance, we slowly arrive at answers to questions, which were asked almost a century ago. One of those, who raised these questions, was Michail Lopatto.

References

Bunge, Mario (1967). *Scientific Research. Vol. I: The Search for System.* New York: Springer.

Christian, R. (1992). Nicolas Bachtin and Mikhail Lopatto. Friendship Renewed. In: Dunn, John A. (ed.), *The Wider Europe. Essays on Slavonic Languages and Cultures. In Honour of Professor Peter Henry on the Occasion of His Retirement.* Nottingham: Astra; 1-14.

Erlich, Victor (1955). *Russian Formalism. History – Doctrine.* S'Gravenhage: Mouton.

Gardzonio, Stefano (2006a). Novoe o Lopatto. In: Ibd. (ed.), *Stat'i po russkoj poėzii i kul'lture XX vek*a. Moskva: Vodolej Publishers; 122-135.

Gardzonio, Stefano (2006b). Michail Lopatto – prozaik i issledovatel' puškinskoj prozy. In: Ibd. (ed.), *Stat'i po russkoj poėzii I kul'lture XX vek*a. Moskva: Vodolej Publishers; 136-147.

Grzybek, Peter (2012a). Der Satz und seine Beziehungen. II: Kapitellänge und Satzlänge (Am Beispiel von L.N. Tolstojs «*Война и мир*»). *Anzeiger für slawische Philologie, XIX*; 39-74.

Grzybek, Peter (2012b). The sentence and its relatives: Chapter length ↔ sentence length ↔ word length. *Paper presented to the International Quantitative Linguistics Conference* (QUALICO 2012). Belgrade, April 26-29.

Grzybek, Peter; Kelih, Emmerich (2005). „Zur Vorgeschichte quantitativer Ansätze in der russischen Sprach- und Literaturwissenschaft." In: Köhler, Reinhard; Altmann, Gabriel; Piotrowski, Rajmund G. (eds.), *Quantitative Linguistics. An International Handbook. Quantitative Linguistik.·Ein internationales Handbuch.* Berlin: Walter de Gruyter; 23-64. [= Handbücher zur Sprach- und Kommunikationswissenschaft; 27]

Jakobson, Roman O. (1919/21). Novejšaja russkaja poėzija. In: *Jakobson, R.O, Selected Writings. Vol. V: On Verse, Its Masters and Explorers.* The Hague etc.: Mouton, 1979; 299-354.

Jakubovič, Dmitrij P. (1936). Obzor statej i issledovanij o proze Puškina s 1917 po 1935 goda. In: *Puškin. Vremennik Puškinskoj Komissii.* Moskva/Leningrad; 297-298. [Online: http://feb-web.ru/feb/pushkin/serial/vr1/vr12295-.htm]

Lotman, Jurij M. (1964). *Lekcii po struktural'noi poėtike. Vvedenie, teorija sticha.* Tartu.

Neumann, Susanne (2009). *Das Menzerath-Altmann-Gesetz als Textcharakteristikum.* M.A. thesis, Trier University.

Peškovskij, Aleksandr M. (1928). Ritmika «Stichotvorenij v proze» Turgeneva. In: Ščerba, Lev V. (ed.), *Russkaja reč'. Tom 2.* Leningrad: Academia; 69-83.

Striedter, Jurij (1969). Zur formalistischen Theorie der Prosa und der literarischen Evolution. In: Striedter, Jurij (ed.), *Texte der russischen Formalisten. Band I: Texte zur allgemeinen Literaturtheorie und zur Theorie der Prosa.* München: Fink; IX-LXXIII.

Šklovskij, Viktor B. (1917). Iskusstvo kak priem. In: *Sbornik po teorii poėtičeskogo jazyka. Vyp. II.* Petrograd: Acedemia; 3-14.

Tomaševskij, Boris V. (1925). *Puškin. Sovremennye problemy istoriko-literaturnogo izučenija.* Leningrad.

Vinogradov, V.V. (1975). Iz istorii izučenija poėtiki (20-e gody). In: *Izvestija AN SSSR, ser. literatury i jazyka, 3,* 259-272.

Reviews

Peter Grzybek, Emmerich Kelih, Ján Mačutek (Eds.),
Text and Language: Structures · Functions · Interrelations. Quantitative Perspectives. Wien: Praesens Verlag, 2010. ISBN 978-3-7069-0625-8, 251pp.
Reviewed by **Haitao Liu** ((Zhejiang University, lhtzju@gmail.com)

This volume contains 23 presentations of the Qualico-2009, organized by IQLA (International Quantitative Linguistics Association). As revealed by the complicated title, these contributions explore structures, functions and interrelations of text and language from quantitative perspectives.

Following the short preface of the three editors is *Sergej Andreev's* quantitative analysis of Keats' style, which is concerned with the factors influencing genre differences.

Solomija Buk and others have attempted to spot word-length-related parameters in the Ukrainian language for the sake of automatic differentiation of text genres, arriving at the discovery that phoneme distribution can be a useful supplement to word-length-related parameters in genre classification.

Radek Čech and Ján Mačutek have examined the distribution of valency frames in Czech and tested the hypothesis of a relationship between word length and the number of valency frames Their work shows that the distribution of valency frames well fits the Good distribution and "the shorter the verb, the more valency frames ".

Łukasz Debowski announces recent developments of a new probabilistic explanation of the Zipf law and the Herdan law, and reports, in particular, his work in correlating the number of set phrases in a text and the number of described facts.

On the basis of 120 Slovenian texts of four different genres, *Gordana Đuraš* and *Ernst Stadlober* have modelled word length frequencies according to the Singh-Poisson distribution.

With their experience in the application of Multidimensional Scaling (MDS) to geolinguistic data, *Sheila Embleton* and others investigate approaches to testing quantitative hypotheses.

Peter Grzybek's article *covers* the measurement of text difficulty. He has found that, at least for German, text difficulty can be measured without any language-specific adaptation; and probably no text typological specifics need to be considered when Tuldava's TD is employed to measure text difficulty.

Emmerich Kelih presents empirical Serbian evidences of the Menzerath law and advances linguistic interpretations of the usually iteratively determined parameters which can be replaced by the mean syllable length of one-syllable words.

Based on textual properties which are purely formal and automatically determinable, *Reinhard Köhler* and *Sven Naumann* propose a syntagmatic approach to automatic text classification, which facilitates not only automatic classification of text but also the linguistic understanding of text properties.

Jan Králík probabilistically explains Zipf's law. According to the author, the utility hierarchy, which is introduced in his paper, can accout for the working mechanism of the principle of least effort.

Jerónimo Leal and *Giulio Maspero* quantitatively probes into Tertullian's authorship of the *Passio Perpetuae,* showing the advantage of entropic distance over bigram distance in classifying the texts studied in their research.

Sylvain Loiseau explores the hypothesis that the plurality of axes of variation may be useful in textual typology. This paper introduces some variationist viewpoints into text typology and corpus-based analyses of variation.

Ján Mačutek illustrates, with an example whose data well fits the chi square goodness of fit test, the ways to avoid the pitfall that may occur in the investigation of rank-frequency distributions.

Gregory Martynenko argues that a Weibull-approximation of the empirical dependencies "vocabulary size - sample size" is useful in both establishing hypothetical borders of the lexical diversity and seeking a rule for the harmonious organization in vocabulary size.

Ivan Obradovic and his colleagues contribute two papers to this volume. The first paper, which focuses on wordnets as a means for refining queries in IR tasks, advances a set of simple and natural relevance indices for tuning the query formulation process. The second paper, which is concerned with the distribution of canonical syllable types in Serbian, points to, in spite of the negative results, some interesting directions worthy of further investigation.

Vasilij V. Poddubnyj and *Anastasija S. Kravcova,* with their investigation into statistical reduction of the feature space of text styles, argues that the transformation of the feature space is helpful to find a minimal set of statistically independent latent features.

Andrij Rovenchak and *Valentin Vydrin* quantitatively explores properties of Nko, an indigenous writing system of Manding languages of West Africa, including script's complexities, their interrelations with frequencies, the grapheme-phoneme correspondence and phoneme distribution.

Haruko Sanada's exploration into the distribution of motifs and the relationship between length and frequency of motifs in Japanese texts registers that motifs, which are defined as word-length sequences following time-series, are correct conceptual abstractions subject to the same laws as all other linguistic units.

Tatiana Sherstinova reports the quantitative data processing in the ORD speech corpus of Russian everyday communication. This kind of corpus benefits both the investigations into Russian communication strategies and the descriptions of the vocabulary and grammar of modern spoken Russian.

O.G. Shevelyov and *V.V. Poddubnyj* present how to conduct complex investingations of texts with "StyleAnalyzer", a software tool enabling researchers to carry out various investigations in the fields of quantitative and computational linguistics. With the objective of creating a dictionary of Japanese collocations,

Tadaharu Tanomura discusses several issues of the retrieval of collocational information from Japanese corpora.

Nicolas Turenne examines the influence of time on the distributions of both content-word occurrences (or named entities) in texts and clusters of content-words where these words are linked due to sharing contexts.

At the end of this volume are listed, in addition to subject and author indexes, addresses of all authors — a thoughtful arrangement facilitating not only retrieving needed information in this volume but contacting the authors concerned.

In summary, this volume includes both the topics of purely traditional quantitative linguistics and contents of computational linguistics, corpus linguistics, literary stylistics, etc. Even those papers concerning topics of purely quantitative linguistics are somewhat oriented toward practical applications such as text classification. Perhaps, this trend reflects that quantitative linguistics involves in fact interdisciplinary research which can assist significantly practical studies such as text processing. This volume is very useful for readers to understand the latest developments of quantitative linguistics.