Glottometrics 43
2018

In Remembrance of Fengxiang Fan, 1950–2018

A Pioneer of Quantitative Linguistics in China

RAM-Verlag

ISSN 1617-8351
e-ISSN 2625-8226

Glottometrics

Indexed in ESCI by Thomson Reuters and SCOPUS by Elsevier

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

External Academic Peers for Glottometrics

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. 43 (2018), Lüdenscheid: RAM-Verlag, 2018. 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 43 (2018)
online/ e-version ISSN 2625-8226 (print version ISSN 1617-8351)

Contents

History

Glottometrics 43, 2018, 1-10

Text Length and Vocabulary Size: Case of the Ukrainian Writer Ivan Franko

Alexei Vasilev[1] *and Ilona Vasileva*

Abstract. In the paper we study how the vocabulary size depends on the text length for the literary works of the famous Ukrainian writer Ivan Franko. We propose two models that explain growth of the vocabulary size due to increasing of the text length. In the core of the models are differential equations which allow us to obtain approximation functions. These functions contain phenomenological parameters whose values we estimate with the measured data.

Keywords: *text length, vocabulary size, function, approximation, Ukrainian*

1. Introduction

The problem of how vocabulary size (here we mean the number of different words in a text) depends on the text length has a long story and actually it is well studied. The examples of essential researches on this subject (including the synergetic approach by Köhler (2005), Altmann and Köhler (1996)) are the works by Herdan (1960), Tuldava (1987, 1993, 1995), Panas (2001), Panas and Yannacopoulos (2004), Wimmer (2005), Fan (2013), Kubát and Milička (2013), Panas et al. (2016), and many others (see, for example, references in the book by Tuldava (1987), in the paper by Panas (2001) or in the paper by Mitchell (2015)). Nevertheless, this problem is still of great theoretical and practical interest. Here we focus our attention on the particular case: we consider a set of works of the famous Ukrainian writer Ivan Franko (1856-1919) and study how the vocabulary size in his novels and short stories depends on the text length (to find more about Ivan Franko and quantitative research related to him, see the paper of Rovenchak and Buk (2017)). So strictly speaking, further theoretical analysis that we use are applicable to literary texts only. Nevertheless, the problem itself may be used also for characterization of text types.

Solving the problem, we can make some predictions even without any quantitative calculations. Indeed, it is obvious that the vocabulary size increases with increasing the text length. And it is also clear that the ratio of the vocabulary size to the text length decreases with increasing the text length. The reason for that is as follows. In general, there exists a limited number of words which potentially can be used in the text. On the other hand, the text length can be increased theoretically to infinity. If we deal with a literary text, then it is reasonable to assume that there exists some "core" of words which are related to the text subject and thus potentially can be used in the text. In other words, we can state that the text growth is provided by the restricted number of lexemes. And that could be the cause for decreasing the ratio of the vocabulary size to the text length. But there is still an unanswered

[1] Taras Shevchenko National University of Kyiv, Department of Theoretical Physics, 64/13, Volodymyrska Street, Kyiv 01601, Ukraine. e-mail: vasilev@univ.kiev.ua

question about how the number of different words in the text is related to the number of all words in the text.

If we denote the vocabulary size (types) by L and the text length (tokens) by V then the problem could be formulated as finding the dependence $L(V)$. The simplest dependence is of the form (Tuldava, 1987)

$$L = aV, \tag{1}$$

where a is a constant (that can be estimated from the measured data). Equation (1) means that the vocabulary size is proportional to the text length. But such dependence can be valid only for the texts of small lengths (Tuldava, 1987). So there exist more sophisticated models (one can find a great set of different dependences $L(V)$ in the paper of Mitchell (2015)). Say, very often we can use the more general dependence (Tuldava, 1987; Panas, 2001):

$$L = \frac{V_1 V}{V+V_0}, \tag{2}$$

which contains two parameters V_0 and V_1. This dependence gives the finite value V_1 for the vocabulary size L when the text length V growth to infinity. For small values of V it gives the approximate linear dependence $L \approx \frac{V_1}{V_0} V$.

Another dependence of the vocabulary size on the text length is given by the formula (Tuldava, 1987)

$$L = bV^n \tag{3}$$

with parameters b and n. This dependence can be obtained from the solution of the differential equation

$$\frac{dL}{L} = n \cdot \frac{dV}{V}. \tag{4}$$

This is the law of the constant relative growth (Tuldava, 1987): the relative change of the vocabulary size $\frac{dL}{L}$ is proportional to the relative change of the text length $\frac{dV}{V}$.

Two principal positions should be stressed here. The first one is that we can usually use several approximation functions to satisfy the same statistical data (Tuldava, 1987). In other words, there is more than one way to select a function which will be used to express the dependence of the vocabulary size on the text length. It is clear that some approximation functions fit the statistical data better than other ones. But we cannot state in principle that this function is correct, and that function is not. All we can do is to find the most general expression for the approximation function and this expression should be good enough to calculate acceptable approximation dependences for the wide range of measured data. If we have found (in some way) such a function then the problem is reduced to calculating (estimating) the parameters of this function for every particular set of data and test the function. What we get in this case, besides the calculated approximation dependence, is the possibility to classify the systems basing on the values of the distribution parameters (here we mean parameters of the approximation function) and join these values with other text properties.

The second conceptual position (Panas, 2001; Tuldava 1987; Vasilev et al., 2013) is that we do not just find some approximation function, but derive it from a background theory. We formulate a hypothesis about the creation of such a function. This hypothesis, after screening for a number of suitable measured data sets, can lead to the theory which characterizes the type-token relation. And such a theory is the way to get an approximation function. Besides, having a theory in the background allows us to predict characteristics of the system in the case, when the system satisfies the conditions of the theory. This is a qualitatively different situation as compared with the previous case (when we use only a general approximation function).

2. Mathematical Model

So our main goal in the paper is to set up an approximation dependence of the vocabulary size on the text length. We will construct two models and derive from them appropriately approximation functions. The question which we are faced with is how the size of the vocabulary L depends on the length of the text V. Using continuous models, we assume that the derivative $\frac{dL}{dV}$, which characterizes the change of the vocabulary size L due to the change of the text length V, is a function that depends on the current size of the vocabulary L. In other words, we consider the differential equation of the following form:

$$\frac{dL}{dV} = p(L). \tag{5}$$

Here $p(L)$ is a function that, as it was mentioned above, depends only on the parameter L. This is the first assumption about the function $p(L)$ (since in general case it also should depend on the parameter V). It is natural to ask a question about the reasons for making this assumption and its validity. Or, in other words, what is standing behind this assumption?

The assumption actually means that when the text length is increased then the increasing of the vocabulary size is determined by the number of different words in the text. When this could be true? We can assume that this could be true if the text is related to a specific theme or subject, and that is what occurs for literary texts. Actually, assuming that $p = p(L)$, we take into account that literary texts are considered.

If equation (5) is correct then we have the following equation

$$\frac{dL}{p(L)} = dV, \tag{6}$$

and next we can find

$$V = \int_0^L \frac{dx}{p(x)}, \tag{7}$$

where we have used the initial condition $L(0) = 0$ which obviously means that if there is no text so there is also no vocabulary. To find the explicit expression for the $L(V)$ dependence we have to know the function $p(L)$. The problem is that we know nothing about the function $p(L)$ except that it should be $\frac{dp}{dL} < 0$. Nevertheless, we propose two strategies for solving this problem.

The first strategy is based on using the Taylor expansion for the function $p(L)$. For example, if we take into account only the first term in the Taylor series then we get $p(L) \approx a$ and this yields formula (1). Taking the first and the second terms in the Taylor series for the function $p(L)$ means that $p(L) = k(L_0 - L)$, and we thus get the following equation:

$$\frac{dL}{dV} = k(L_0 - L). \tag{8}$$

In this equation we can identify the parameter L_0 as the total number of lexemes that are suitable for the text. The other parameter k is a phenomenological one and it determines the ratio of the unused words that is used while the text length is growing. Actually, in this case we assume that $p(L)$ is proportional to the number of words that potentially could be used in the text but are not used as yet. If so, then from equation (8) we get the expression for the function which determines the dependence of the vocabulary size L on the text length V:

$$L = L_0(1 - \exp(-kV)). \tag{9}$$

This formula was received firstly in the work of Thomson and Thompson (1915).

If we take three terms in the Taylor series for the function $p(L)$ then we can present this dependence as follows:

$$p(L) = k(L_0 - L - \gamma L^2), \tag{10}$$

where k, L_0 and γ are phenomenological parameters which can be calculated basing on the statistical data. Substituting formula (10) in equation (5) allows us to find this dependence:

$$L = \frac{2L_0 \tanh(\mu V)}{\rho + \tanh(\mu V)}, \tag{11}$$

where $\tanh(x)$ is the hyperbolic tangent of x, $\rho = \sqrt{1 + 4\gamma L_0}$ and $\mu = \frac{k\sqrt{1+4\gamma L_0}}{2}$. As we will show next, formula (11) is applicable for handling the measured data and it gives good results. Nevertheless, we can propose another idea about how vocabulary changes with changing the text length. And this is the second strategy.

We can make some propositions about $p(L)$. In particular, we assume that when the size of the vocabulary increases then the text subject broadens. So this, in turn, broadens the total number of words that potentially can be used in the text. Namely, we assume that the function $p(L)$ changes with changing L in the way that the change of $p(L)$ is proportional to the current value of $p(L)$. Mathematically this statement can be presented with the equation

$$\frac{dp}{dL} = -\alpha p(L), \tag{12}$$

where α is a phenomenological parameter of the model. If so, then considering equations (5) and (12) together we can easily get

$$L = \frac{1}{\alpha}\ln(1 + \beta V). \tag{13}$$

Here parameters α and β are to be found from the measured data.

So in fact, we have two approximation functions (11) and (13), which are supposed to give the dependence of the vocabulary size on the text length. Obviously, expressions (11) and (13) are different. Nevertheless, this is not too crucial as it might seem at first sight. In particular, for small values of V we can use the Taylor expansion for the functions. From equation (11) we have

$$L \approx kL_0 V. \tag{14}$$

Equation (13) gives the following approximation:

$$L \approx \frac{\beta}{\alpha} V. \tag{15}$$

We see that both expressions give the approximately linear dependence of the vocabulary size on the text length if the last one is small. But for big values of V expressions (11) and (13) differ qualitatively. According to expression (11) the value of L should be close to the value of $\frac{2L_0}{1+\rho}$ (the vocabulary size is restricted) meanwhile according to expression (13) the size of the vocabulary is not restricted (from the mathematical point of view): it grows (theoretically to infinity) when the text length increases. To examine the situation we use the measured data.

3. Numerical Results

To build the approximation dependences we use a set of data that contains the text lengths and vocabulary sizes for 57 novels and short stories of Ivan Franko. The list of titles, their lengths and the sizes of the vocabularies are presented in Table 1 (the data were obtained basing on the public resource www.mova.info).

Table 1.
Novels and short stories of Ivan Franko

N	Title	Text length	Vocabulary size	Formula (16)	Formula (17)
1	Bloščytsja	240	158	88	97
2	Vovk-staršyna	328	179	120	131
3	Iz Galytskoji Knygy Bytija	1022	549	366	397
4	Iz zapysok mučenyka	1135	576	405	439
5	Zvirjačyj budžet	1345	580	477	516
6	Istorija kožuha	1367	631	485	524
7	Budjaky	1388	677	492	532
8	Velyki rokovyny	1424	773	504	545
9	Grytseva škilna nauka	1491	883	527	569
10	Malyj Myron	1513	870	534	577
11	Doktor Besservisser	1568	801	553	596
12	Mavka	1588	665	560	603
13	Mij zločyn	1591	969	561	604

14	Muljar	1646	756	579	624
15	Dobryj zarobok	1682	649	591	637
16	Vivčar	1789	843	627	675
17	Z burlyvyh lit	1901	1461	665	714
18	Vilgelm Tell	2162	1015	751	804
19	Istorija mojeji sičkarni	2274	816	788	842
20	Dovbanjuk	2332	982	807	862
21	Domašnij promysel	2631	929	903	961
22	V tjuremnim špytali	2646	1051	908	966
23	Vugljar	2652	1121	910	968
24	Naša publika	2782	1126	952	1011
25	Dva pryjateli	2846	1059	972	1032
26	Zadlja praznyka	2905	1276	991	1051
27	Lesyšyna čeljad	3119	1383	1058	1120
28	Na loni pryrody	3120	1201	1058	1121
29	Lisy i pasovyska	3518	1113	1182	1246
30	Gutsulskyi korol	3585	1257	1202	1267
31	Mykytychiv dub	4129	1427	1367	1433
32	Gava i Vovkun	4160	1411	1376	1442
33	Borys Grab	4163	1464	1377	1443
34	Iz zapysok nedužogo	4259	1576	1406	1472
35	Driada	4677	1772	1528	1594
36	Girčyčne zerno	4991	1876	1619	1684
37	Moja strіča z Oleksoju	5107	1719	1652	1717
38	Do svitla!	6002	1835	1902	1963
39	Miž dobrymy ljudmy	6376	1740	2004	2062
40	Olivets	6491	1470	2034	2092
41	Gava	7841	2041	2385	2432

42	Irygatsija	8359	2762	2514	2556
43	Batkivščyna	10311	2752	2976	2997
44	Na veršku	10395	2900	2995	3015
45	Misija	10551	3259	3030	3049
46	Gutak	10797	3594	3085	3101
47	Geroj ponevoli	12752	3600	3503	3496
48	Na dni	17099	4135	4327	4273
49	Manipuljantka	17348	4186	4370	4314
50	Bez pratsi	20340	4298	4861	4780
51	Gryts i panyč	21541	4317	5044	4955
52	Boryslav	29902	6133	6140	6021
53	Velykyj šum	35647	7015	6748	6636
54	Dlja domašnjogo ognyšča	42940	7131	7397	7317
55	Ne spytavšy brodu	46846	7803	7698	7645
56	Zahar Berkut	47776	7213	7766	7720
57	Boryslav smijetsja	71839	9868	9119	9336

To build the approximation dependence based on formula (11) we, for the sake of simplicity, make the scaling transformation and present formula (11) in the following form:

$$\frac{L}{V_{max}} = \frac{\xi \tanh\left(\lambda \frac{V}{V_{max}}\right)}{\rho + \tanh\left(\lambda \frac{V}{V_{max}}\right)}, \tag{16}$$

where $\xi = \frac{2L_0}{V_{max}}$ and $\lambda = \mu V_{max} = \frac{k\sqrt{1+4\gamma L_0} V_{max}}{2}$, and here we have used the parameter $V_{max} = 71839$, which is the maximum text length among all the texts in Table 1. This presentation is convenient since we have $0 \le \frac{L}{V_{max}} \le 1$ and $0 \le \frac{V}{V_{max}} \le 1$. And the same scaling transformation we use when we build the approximation dependence basing on function (13). We can present that equation like this:

$$\frac{L}{V_{max}} = \eta \ln\left(1 + \kappa \frac{V}{V_{max}}\right), \tag{17}$$

where the parameters $\eta = \frac{1}{\alpha V_{max}}$ and $\kappa = \beta V_{max}$.

Thus, we use the measured data from Table 1 and formulae (16) and (17), and calculate the parameters for the corresponding approximation dependences. In particular, we find that $\xi \approx 0.1941$, $\lambda \approx 0.0068$, $\rho \approx 0.0036$ (for formula (16)) and $\eta \approx 0.0663$, $\kappa \approx$

6.1083 (for formula (17)). Table 1 also contains values for the vocabulary size (types) which were calculated according to formulae (16) and (17). The last two columns of the table present the approximation values of L (for corresponding values of V). Figure 1 illustrates the results of the approximation.

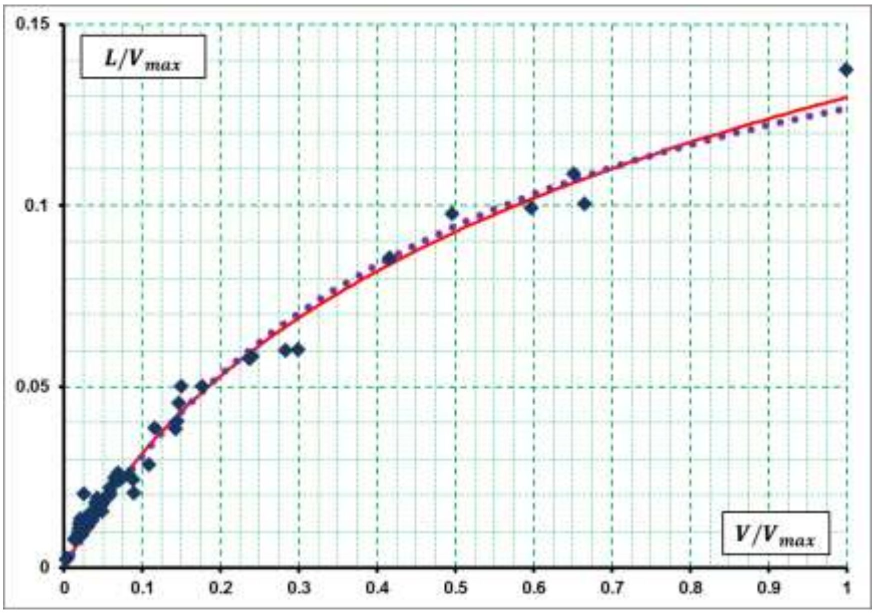


Fig. 1. The results of the approximation of the measured data (the squares). The dotted line is for approximation function (16). The solid line is for approximation function (17). The text length V and the vocabulary size L are scaled by the parameter $V_{max} = 71839$

As we can see the results of the both approximations are good enough and give almost identical dependences (the dotted and the solid lines almost coincide in Figure 1, and the determination coefficient $R^2 \approx 0.9812$ for the first dependence and $R^2 \approx 0.9848$ for the second one). This must not be surprising. The reason is that despite the functions

$$f_1(x) = \frac{\xi \tanh(\lambda x)}{\rho + \tanh(\lambda x)} \tag{18}$$

and

$$f_2(x) = \eta \ln(1 + \kappa x) \tag{19}$$

are formally different, but for the above mentioned values of the parameters ξ, λ, ρ, η and κ, and for the $0 \leq x \leq 1$ they present almost the same dependences. Numerical calculations give that

$$\int_0^1 |f_1(x) - f_2(x)| dx \approx 1.1 \cdot 10^{-3}. \tag{20}$$

This value is the area between the curves that are determined by the functions $f_1(x)$ and $f_2(x)$. And it is only the 0.1% of the total area (which is 1) in Figure 1 where the data are presented.

4. Conclusion

First of all, we have found the dependence of the vocabulary size on the text length for the texts of the famous Ukrainian writer Ivan Franko. We believe that this may have an impact on studies in area of text classification and authorship identification. Also we have some modest expectations about ability to use the methodology, which was proposed above, for developing similar models for other authors and text types.

We have also shown that the both proposed approximation functions give almost the same dependences. This could be explained in the following way. The formulated model is based on the fact that the derivative of the approximation function is the decreasing function of the vocabulary size. If the model is topologically stable (and we expect that it is so) then the result should not qualitatively depend on the particular form of the derivative function. That is what we have actually got.

Acknowledgements

The authors are grateful to *Prof. Natalia Darchuk* (Taras Shevchenko National University of Kyiv, Ukraine) and to her colleagues for developing and supporting the public resource *www.mova.info*, which was used to obtain the measured data for this paper.

References

Altmann, G., & Köhler, R. (1996). "Language Forces" and synergetic modelling of language phenomena. *Glottometrika 15, 62-76.*

Fan, F. (2013). Text Length, Vocabulary Size and Text Coverage Constancy. *Journal of Quantitative Linguistics 20, 288-300.*

Fan, F., Yang, Y., & Yaqin W. (2016). The Probability Distribution of Textual Vocabulary in the English Language. *Journal of Quantitative Linguistics 23, 49-70.*

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. Mouton: Gravenhage.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook*: 760-775. Berlin, New York: Walter de Gruyter.

Kubát, M. & Milička, J. (2013). Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics 20, 339–349.*

Mitchell, D. (2015). Type-token models: a comparative study. *Journal of Quantitative Linguistics 22, 1–21.*

Panas, E. (2001). The Generalized Torquist: Specification and Estimation of a New Vocabulary-Text Size Function. *Journal of Quantitative Linguistics 8, 233-252.*

Panas, E., & Yannacopoulos, A.N. (2004). Stochastic Models for the Lexical Richness of a Text: Qualitative Results. *Journal of Quantitative Linguistics 11, 251-273.*

Rovenchak A., & Buk, S. (2017): Part-of-Speech Sequences in Literary Text: Evidence From Ukrainian. *Journal of Quantitative Linguistics*.
DOI: 10.1080/09296174.2017.1324601.

Thomson, G., & Thompson, J.R. (1915). Outline of a measure for the quantitative analysis of writing vocabularies. *British Journal of Psychology 8, 52-69.*

Tuldava, J. (1987). *Problemy i metody kvantitativno-sistemnogo issledovanija leksiki* [Problems and methods of quantitative and systematic investigation of lexica]. Tallin: Valgus.

Tuldava, J. (1993). The statistical structure of a text and its readability. In: L. Hřebíček & G. Altmann (Eds.), *Quantitative text analysis: 215-227.* Trier: Wissenschaftlicher Verlag Trier.

Tuldava, J. (1995). *On the relation between text length and vocabulary size*. In: J. Tuldava. (Ed.), *Methods in Quantitative Linguistics*: 131–149. (Quantitative Linguistics, vol. 54.) Trier: Wissenschaftlicher Verlag Trier.

Vasilev, A.N., Chalyi A.V., & Vasileva I.V. (2013). About "Exotic" Problems of Physics, Winnie the Pooh and Zipf's Law. *Journal of Physical Studies, 17, (1001)1-8.*

Wimmer, G. (2005). The type-token relation. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook*: 361-368. Berlin, New York: Walter de Gruyter.

Glottometrics 43, 2018, 11- 30

Quantifying the Importance of Stylometric Indicators: A Principal Component Approach to Czech Sonnets

Peter Zörnig[1], Brasilia
Michal Místecký[2], Ostrava

Abstract. We apply multivariate statistical analysis to stylometric indicators, calculated for sonnets of a prominent Czech poet, J.S. Machar. The study makes it possible to quantify the importance of linguistic indicators and to partition the sonnets into stylistic groups. Moreover, it introduces the index of weighted occurrence, which is a way for a quantitative assessment of the importance of the studied properties. The procedure applied in the paper may be of universal use.

Keywords: *sonnets, poetry, stylometry, principal component analysis, eigenvectors, Czech*

1. Introduction

Over the past two decades, applications of multivariate statistical analysis in quantitative and computational linguistics have become increasingly popular (see, e.g., Baayen (2013)). In the present paper, we propose a principal component approach to study sonnets written by Josef Svatopluk Machar (1864–1942), a prominent Czech poet of the first modern, 1890s generation. In general, sonnets have recently turned to be the focus of many studies, and more papers are being prepared to provide a deeper insight into this pan-European lyric form. The articles published so far (see, e.g., Místecký (2018a) and Místecký (2018b)) concentrated on characterizing the overall stylometric features of the poems, with more profound interpretations left for the publications to come. These include a paper on the manifestation of the Piotrowski Law in the development of activity in the studied poems, and a book covering the subject thoroughly and presenting some general trends in sonnets, too.

After the description of the data (Section 2), we present a principal component approach to weight the importance of stylometric indicators. The principal components are essentially given by the eigenvectors of the correlation matrix corresponding to 18 characteristics of 47 sonnets (Section 3). The magnitudes of the elements of the eigenvectors represent the importance of the respective stylometric indicators, and are studied in detail. The scores (values) of the principal components are calculated in Section 4. They serve mainly to classify the studied sonnets.

[1] Peter Zönig, University of Brasilia, e-mail: peter@unb.br
[2] Michal Místecký, University of Ostrava; e-mail: MMistecky@seznam.cz.

2. Description of Data

In order to provide the data for the analysis, 47 poems of Machar's collection of *Letní sonety* ("Summer Sonnets", written in 1890) have been used; their stylometric features have been measured via 18 indexes, the characteristics of which are given in Table 2.1. The detailed descriptions of the calculations are presented, for instance, in Kubát, Matlach, and Čech (2014).

Table 2.1
The stylometric indicators and their characteristics

No.	Index	Features
1	TTR	Type-token ratio: the proportion of word-types to the text length
2	H-point	The place where the rank of an expression equals its frequency; a potential border between synsemantics and autosemantics
3	ATL	Average tokens length: the average length of a word
4	R1	H-point-based indicator of vocabulary richness
5	RR	Repeat rate of a word in a text
6	RRmc	Repeat rate normalized by McIntosh
7	TC	H-point-based thematic concentration
8	STC	Secondary thematic concentration based on the 2h-point
9	PTC	Proportional thematic concentration
10	Lambda	A frequency-interval-based indicator of vocabulary richness
11	R4	The additive inverse of the Gini Coefficient, measuring dispersion of lexis
12	Hapax Percentage (HP)	The proportion of the expressions that appear only once in a text to its length
13	Writer's View (WV)	H-point-based indicator of the vocabulary span
14	CurveLength R Index (R)	H-point-based indicator of vocabulary richness
15	Belza-chain Associations (BCA)	A measure of text thematic intercomnectedness
16	Activity (Q)	The proportion of verbs to the sum of adjectives and verbs
17	Nominality (N)	The proportion of nouns to the text length
18	Adjectivity (ADJ)	The proportion of adjectives to the text length

Table 2.2 gives the data for the respective indexes in the studied poems, including the means and variances for the individual counts at the bottom.

Table 2.2
The data for the analysis based on the counts of the selected stylometric indexes

No.	**Sonnet**	**x_1 TTR**	**x_2 h-Point**	**x_3 ATL**	**x_4 R1**	**x_5 RR**	**x_6 RRmc**	**x_7 TC**	**x_8 STC**	**x_9 PTC**
1	E. Zolovi	0.846	3.000	4.795	0.929	0.019	0.984	0.000	0.000	0.000
2	Matce	0.828	3.000	4.517	0.937	0.017	0.987	0.000	0.000	0.000
3	Sonet cynický	0.711	4.000	4.052	0.825	0.028	0.948	0.000	0.000	0.000
4	Sonet de vanitate	0.765	4.000	4.435	0.906	0.021	0.978	0.250	0.196	0.250
5	Sonet elegický	0.767	3.000	4.044	0.894	0.021	0.971	0.000	0.000	0.000
6	Sonet ironický	0.821	3.000	4.345	0.935	0.018	0.985	0.000	0.150	0.000
7	Sonet k sociální otázce	0.819	3.500	3.830	0.938	0.017	0.983	0.000	0.000	0.000
8	Sonet k teorii; Boj o život	0.732	3.000	4.474	0.902	0.022	0.968	0.000	0.000	0.000
9	Sonet materialistický	0.793	3.333	4.043	0.908	0.019	0.977	0.000	0.000	0.000
10	Sonet mystický	0.882	2.000	4.447	0.961	0.017	0.992	0.000	0.000	0.000
11	Sonet na Chopinovu melodii	0.620	4.000	3.843	0.870	0.023	0.965	0.000	0.013	0.000
12	Sonet na sentenci z Goetha	0.814	3.000	4.453	0.948	0.018	0.985	0.000	0.000	N/A
13	Sonet na sklonku století	0.798	3.000	4.079	0.904	0.019	0.977	0.000	0.000	0.000
14	Sonet nad verši z mládí	0.766	3.000	4.138	0.910	0.020	0.975	0.000	0.000	0.000
15	Sonet noční	0.833	2.500	4.014	0.960	0.019	0.989	0.000	0.000	0.000
16	Sonet o antice a vlasech	0.736	3.000	4.055	0.929	0.019	0.982	0.000	0.000	0.000
17	Sonet o bídě	0.814	3.500	4.031	0.939	0.016	0.982	0.000	0.000	0.000
18	Sonet o hodinách	0.773	3.000	4.364	0.915	0.020	0.976	0.000	0.083	0.000
19	Sonet o lásce	0.756	3.000	4.100	0.872	0.025	0.959	0.667	0.333	0.692
20	Sonet o minulosti	0.676	3.000	4.095	0.885	0.028	0.968	0.000	0.000	0.000
21	Sonet o Panně Marii	0.790	3.000	4.049	0.944	0.018	0.988	0.000	0.000	0.000
22	Sonet o rokoku	0.802	4.000	4.011	0.890	0.020	0.973	0.000	0.021	0.000
23	Sonet o staré metafoře	0.679	3.500	4.222	0.903	0.026	0.969	0.000	0.129	0.000
24	Sonet o starém	0.716	4.000	3.990	0.833	0.025	0.952	0.000	0.000	0.000

	líci a rubu									
25	Sonet o třech metaforách	0.818	3.000	4.481	0.916	0.020	0.981	0.000	0.000	0.000
26	Sonet o třetí hodině v červenci	0.758	3.000	4.033	0.885	0.023	0.965	0.000	0.000	0.000
27	Sonet o vídeňských kosech	0.796	3.000	4.140	0.898	0.020	0.973	0.000	0.000	0.000
28	Sonet o západu slunce	0.835	4.000	4.010	0.893	0.017	0.974	0.000	0.030	0.000
29	Sonet o zlatém věku naší poezie	0.843	2.500	4.429	0.945	0.020	0.985	0.000	0.000	0.000
30	Sonet o životě	0.753	3.000	4.281	0.916	0.020	0.977	0.000	0.000	0.000
31	Sonet patologický	0.753	3.000	4.079	0.893	0.021	0.974	0.000	0.000	0.000
32	Sonet polední	0.802	4.000	4.180	0.901	0.017	0.974	0.000	0.000	0.000
33	Sonet sarkastický	0.753	3.500	4.034	0.923	0.020	0.977	0.000	0.000	0.000
34	Sonet svatební	0.815	3.000	4.304	0.929	0.017	0.984	0.000	0.000	0.000
35	Sonet úvodní	0.821	3.000	4.393	0.923	0.019	0.981	0.000	0.000	0.000
36	Sonet večerní	0.910	2.000	4.564	0.962	0.015	0.994	0.000	0.000	0.000
37	Sonet z dvacátého září	0.857	3.000	4.471	0.921	0.021	0.983	0.000	0.033	0.000
38	Sonet; apostrofa	0.756	2.500	3.859	0.925	0.023	0.977	0.000	0.000	0.000
39	Sonet; epilog čtenáři	0.798	3.500	4.096	0.895	0.021	0.966	0.000	0.000	0.000
40	Sonet; intermezzo(2)	0.744	3.500	4.067	0.890	0.023	0.967	0.000	0.000	0.000
41	Sonet; intermezzo	0.761	3.500	3.915	0.903	0.026	0.970	0.000	0.000	0.000
42	Sonety; Causerie I.	0.735	4.000	3.837	0.908	0.019	0.976	0.000	0.043	0.000
43	Sonety; Causerie II.	0.734	3.500	4.106	0.884	0.023	0.964	0.000	0.000	0.000
44	Sonety; Causerie III.	0.814	3.000	4.233	0.890	0.021	0.973	0.000	0.000	0.000
45	Sonety; Causerie IV.	0.689	4.333	3.822	0.849	0.028	0.954	0.000	0.000	0.000
46	Sonety; Causerie V.	0.747	3.000	4.022	0.863	0.025	0.960	0.000	0.000	0.000
47	Své ženě s předešlým sonetem	0.787	3.000	4.247	0.927	0.018	0.982	0.000	0.120	0.000
	MEAN	0.779	3.216	4.171	0.908	0.021	0.975	0.020	0.025	0.020
	VARIANCE	0.003	0.265	0.050	0.001	0.000	0.000	0.010	0.004	0.011

No.	**Sonnet**	x_{10} **Λ**	x_{11} **R4**	x_{12} **HP**	x_{13} **WV**	x_{14} **R**	x_{15} **BCA**	x_{16} **Q**	x_{17} **N**	x_{18} **ADJ**
1	E. Zolovi	1.607	0.859	0.756	2.710	0.948	2.330	0.470	0.269	0.103
2	Matce	1.611	0.848	0.701	2.707	0.947	2.670	0.600	0.230	0.276
3	Sonet cynický	1.477	0.735	0.608	2.157	0.899	1.400	0.600	0.165	0.227
4	Sonet de vanitate	1.490	0.797	0.624	3.092	0.920	1.800	0.380	0.282	0.153
5	Sonet elegický	1.530	0.790	0.656	2.189	0.934	1.670	0.500	0.222	0.111
6	Sonet ironický	1.586	0.842	0.702	2.708	0.951	1.780	0.420	0.214	0.131
7	Sonet k sociální otázce	1.621	0.836	0.713	2.978	0.937	1.330	0.670	0.277	0.064
8	Sonet k teorii; Boj o život	1.535	0.768	0.588	1.981	0.905	2.830	0.620	0.278	0.082
9	Sonet materialistický	1.581	0.815	0.674	2.555	0.937	2.330	0.400	0.293	0.130
10	Sonet mystický	1.654	0.894	0.776	2.372	0.964	1.750	0.520	0.276	0.132
11	Sonet na Chopinovu melodii	1.315	0.702	0.417	2.262	0.900	1.820	0.650	0.231	0.056
12	Sonet na sentenci z Goetha	1.571	0.837	0.698	3.112	0.957	1.780	0.600	0.279	0.070
13	Sonet na sklonku století	1.588	0.819	0.674	2.188	0.930	1.670	0.470	0.326	0.090
14	Sonet nad verši z mládí	1.573	0.798	0.617	2.063	0.918	2.220	0.530	0.277	0.074
15	Sonet noční	1.543	0.857	0.694	2.846	0.960	1.400	0.610	0.319	0.097
16	Sonet o antice a vlasech	1.448	0.788	0.549	2.709	0.949	1.800	0.300	0.231	0.176
17	Sonet o bídě	1.623	0.832	0.711	2.977	0.944	2.130	0.690	0.289	0.052
18	Sonet o hodinách	1.547	0.801	0.636	2.190	0.926	2.250	0.760	0.239	0.045
19	Sonet o lásce	1.571	0.779	0.633	1.923	0.896	1.170	0.600	0.322	0.089
20	Sonet o minulosti	1.300	0.740	0.486	2.201	0.910	1.670	0.830	0.108	0.027
21	Sonet o Panně Marii	1.504	0.830	0.617	3.109	0.947	2.000	0.350	0.333	0.160
22	Sonet o rokoku	1.586	0.816	0.725	2.863	0.934	2.000	0.500	0.297	0.110
23	Sonet o staré	1.311	0.737	0.506	2.650	0.91	2.00	0.53	0.22	0.11

	metafoře					3	0	0	2	1
24	Sonet o starém líci a rubu	1.507	0.742	0.598	2.155	0.894	3.000	0.630	0.255	0.069
25	Sonet o třech metaforách	1.549	0.835	0.714	2.711	0.946	1.270	0.320	0.338	0.169
26	Sonet o třetí hodině v červenci	1.558	0.783	0.637	1.982	0.910	1.780	0.330	0.308	0.110
27	Sonet o vídeňských kosech	1.598	0.812	0.699	2.187	0.938	2.570	0.700	0.237	0.065
28	Sonet o západu slunce	1.710	0.843	0.777	2.590	0.931	1.750	0.680	0.359	0.068
29	Sonet o zlatém věku naší poezie	1.562	0.860	0.729	2.383	0.952	1.400	0.670	0.300	0.057
30	Sonet o životě	1.511	0.790	0.596	2.190	0.925	3.200	0.730	0.270	0.079
31	Sonet patologický	1.500	0.789	0.596	2.190	0.926	1.330	0.750	0.236	0.056
32	Sonet polední	1.684	0.816	0.712	2.391	0.928	1.270	0.480	0.351	0.099
33	Sonet sarkastický	1.482	0.787	0.607	2.641	0.929	2.130	0.560	0.191	0.079
34	Sonet svatební	1.624	0.838	0.685	2.384	0.939	5.670	0.760	0.217	0.054
35	Sonet úvodní	1.605	0.839	0.714	2.386	0.940	1.640	0.310	0.298	0.214
36	Sonet večerní	1.718	0.917	0.833	2.371	0.966	1.080	0.380	0.385	0.128
37	Sonet z dvacátého září	1.588	0.867	0.786	2.713	0.943	1.600	0.500	0.243	0.143
38	Sonet; apostrofa	1.479	0.801	0.577	2.002	0.925	1.670	0.670	0.244	0.077
39	Sonet; epilog čtenáři	1.645	0.811	0.713	2.113	0.899	1.400	0.570	0.213	0.096
40	Sonet; intermezzo(2)	1.487	0.774	0.611	2.396	0.911	2.800	0.470	0.211	0.100
41	Sonet; intermezzo	1.425	0.786	0.648	2.651	0.912	3.750	0.440	0.169	0.141
42	Sonety; Causerie I.	1.476	0.772	0.592	2.864	0.934	1.880	0.740	0.235	0.051
43	Sonety; Causerie II.	1.488	0.763	0.606	2.229	0.914	2.630	0.740	0.277	0.053
44	Sonety; Causerie III.	1.608	0.828	0.721	2.189	0.929	1.500	0.360	0.314	0.105
45	Sonety; Causerie IV.	1.416	0.725	0.556	2.366	0.887	1.800	0.640	0.211	0.056
46	Sonety;	1.533	0.772	0.626	1.982	0.91	1.40	0.58	0.29	0.08

	Causerie V.					0	0	0	7	8
47	Své ženě s předešlým sonetem	1.557	0.817	0.640	2.386	0.940	3.000	0.550	0.315	0.101
	MEAN	1.542	0.806	0.654	2.447	0.929	2.028	0.557	0.265	0.103
	VARIANCE	0.008	0.002	0.006	0.109	0.000	0.628	0.019	0.003	0.003

3. Eigenvectors: Computation and Interpretation

Principal component analysis is a very useful statistical tool with countless applications in diverse scientific areas. It is a data-reduction technique that aims to express the observed variables (whose number is usually large in real-life applications) via a small number of unobservable variables that represent important characteristics of the objects under study. These “new” variables can be used, e.g., for statistical testing and data visualization. In our linguistic study of sonnets, the p = 18 variables – x_1,…, x_{18}, representing individual indexes (see Tables 2.1 and 2.2) – will be converted into m more general linguistic quantities. We shall mention later how the number m, which must be considerably smaller than p, is adequately determined. The new variables, called principal components, are usually considered linear functions of the original variables x_1, …, x_p. Formally, we get

$$\begin{aligned} PC_1 &= a_{1,1}\, x_1 + a_{1,2}\, x_2 + \ldots\ a_{1,p}\, x_p, \\ PC_2 &= a_{2,1}\, x_1 + a_{2,2}\, x_2 + \ldots\ a_{2,p}\, x_p, \\ &\qquad\qquad\qquad\ldots\ldots\ldots\ldots\ldots \\ &\qquad\qquad\qquad\ldots\ldots\ldots\ldots\ldots \qquad\qquad (3.1) \\ PC_m &= a_{m,1}\, x_1 + a_{m,2}\, x_2 + \ldots\ a_{m,p}\, x_p. \end{aligned}$$

To any values of the x_i, these equations assign corresponding values of the principal components PC_i, called *principal component scores*. It is intuitively clear that a small number m of components cannot contain the same amount of “information” than the original set of p variables. However, the principal components are chosen in a manner that the “information loss” – statistically measured by means of the “proportion of variance explained” by the principal components – is minimized. The higher this proportion of variance, the smaller the loss of information is.

As the first step in principal component analysis, the mean values, variances, and covariances of the x_1, …, x_p are determined.

Let us present the data matrix by

$$X = \begin{pmatrix} x_{1,1} & \Lambda & x_{1,p} \\ M & & M \\ x_{n,1} & \Lambda & x_{n,p} \end{pmatrix} , \qquad (3.2)$$

whose columns correspond to the columns in Table 2.2. The number of objects (sonnets) is n = 47, and the number of variables is p = 18 (see above). We define the *k-th mean value* by

$$\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{i,k} \qquad \text{for k = 1, …, } p, \qquad (3.3)$$

and the *k-th variance* by

$$s_k^2 = \frac{1}{n}\sum_{i=1}^{n} (x_{i,k} - \bar{x}_k)^2 \qquad \text{for } k = 1, …, p. \qquad (3.4)$$

The *covariance* between x_k and x_l is given by

$$s_{k,l} = \frac{1}{n}\sum_{i=1}^{n} (x_{i,k} - \bar{x}_k)(x_{i,l} - \bar{x}_l) \qquad \text{for } k, l = 1, …, p. \qquad (3.5)$$

In the specific case with $k = l$, the covariance becomes the variance s_k^2 – i.e., we get $s_{kk} = s_k^2$, and therefore the symbol s_{kk} is frequently used to denote the variance of x_k.

In order to apply the convenient matrix notation, we define the *covariance matrix* corresponding to (3.2) as

$$S = \begin{pmatrix} s_{1,1} & \Lambda & s_{1,p} \\ \mathrm{M} & & \mathrm{M} \\ s_{n,1} & \Lambda & s_{n,p} \end{pmatrix}. \qquad (3.6)$$

The (linear) *correlation* (also known as *Pearson's correlation coefficient*) between the variables x_k and x_l is defined by

$$r_{k,l} = \frac{s_{k,l}}{\sqrt{s_{kk}s_{ll}}} \quad , \qquad (3.7)$$

(see, e.g., Zörnig (2016, p. 112–116)), and all possible correlations are put together in the *correlation matrix*

$$R = \begin{pmatrix} r_{1,1} & \Lambda & r_{1,p} \\ \mathrm{M} & & \mathrm{M} \\ r_{n,1} & \Lambda & r_{n,p} \end{pmatrix}. \qquad (3.8)$$

By definition, both matrixes S and R are symmetric, and the diagonal elements $r_{1,1}, ..., r_{p,p}$ of R are all equal to 1.

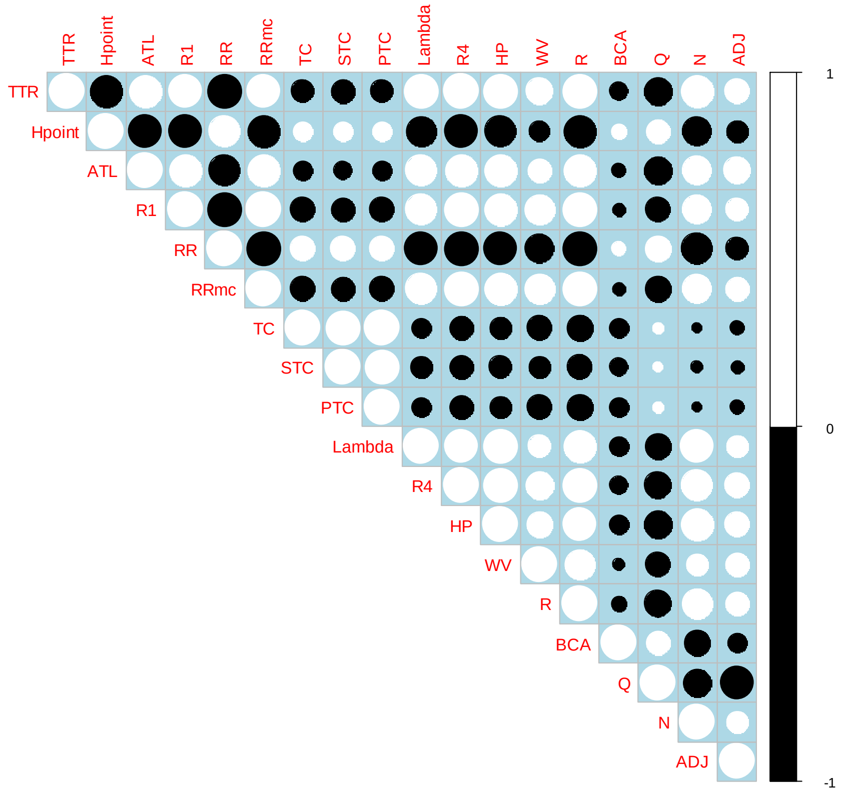
Example 3.1: Consider the variables x_4 and x_6 in Table 2.2, representing the vocabulary-richness indexes R1 and RRmc. The column 4 is the vector

$$\begin{pmatrix} x_{1,4} \\ \mathrm{M} \\ x_{47,4} \end{pmatrix} = \begin{pmatrix} 0.929487 \\ \mathrm{M} \\ 0.926966 \end{pmatrix}.$$

Thus $\bar{x}_4 = \frac{1}{47}\sum_{i=1}^{47} x_{i,4} = \frac{0.929487+\ldots+0.926966}{47} = 0.9080$, and $s_{4,4} = \frac{1}{47}\sum_{i=1}^{47}(x_{i,4}-\bar{x}_4)^2$ = $0.9181\cdot 10^{-3}$. In the same way, we determine mean and variance for the variable x_6 as $\bar{x}_6 =$ 0.9750 and $s_{6,6} =$ $0.1017\cdot 10^{-3}$, and the covariance between x_4 and x_6 is $s_{4,6} =$ $\frac{1}{47}\sum_{i=1}^{n}(x_{i,4}-\bar{x}_4)(x_{i,6}-\bar{x}_6) =$ $0.2951\cdot 10^{-3}$. Finally, we get the correlation coefficient $r_{4,6}$ = $\frac{s_{4,6}}{\sqrt{s_{4,4}\ s_{6,6}}} = \frac{0.2951}{\sqrt{0.9181\cdot 0.1017}}$ = 0.9656. Since both variables basically measure the same characteristics, this high correlation is expectable.

A complete presentation of all correlations is given in Fig. 3.1.

The visualized correlations among the indexes have given some results for interpretations. First, it is expectable that the vocabulary-richness indicators (such as TTR, R, R1, R4, RR, RRmc, h-point, etc.) are interrelated, as they are mostly based upon the h-point or other properties of the rank-frequency distribution; this is going to be paid more attention later. Second, thematic concentration counts (TC, STC, and PTC) seem to be disconnected from other indicators, forming thus a separate group of characteristics. The same holds for Belza-chain associations, which are, however, to a high extent linked to nomínality; this may account for the fact that the theme-carriers in the studied poems are mostly nouns. Third, the negative correlation between adjectivity and activity is easy to explain, as both counts are based upon the number of adjectives; the relation is inverted, since a high number of adjectives means a decrease in Q, but a rise in ADJ. Last but not least, it is to be noted that ATL, despite not being a vocabulary-richness index, behaves in accordance with many of these; for instance, it manifests a negative correlation with the h-point. This accounts for the fact that the h-point marks the border between very frequent words and those of lower



occurrence; and as the former are usually shorter than the latter, the smaller the h-point figure gets, the higher the proportion of long words in a text is, which raises the value of its ATL. However, a thorough commentary on the correlations would deserve a study of its own.

Fig. 3.1. Correlogram of the studied properties. The sizes and colours of the circles express the ranges of correlation values. Big, medium, and small white circles represent positive correlations over 0.5, near to 0.5 and below 0.5, respectively. Accordingly, the black circles indicate negative correlations.

Now, the count of principal components will be paid due heed. They are designed in such a way that the variance (contained "information") of PC_i is maximized and that each component is uncorrelated with the previously constructed ones. Principal components, which are optimal in this sense, are generated by means of the eigenvectors of the matrix *S* or *R* (see (3.6) and (3.8)). An *eigenvector v* of a square matrix *M* is characterized by the property

$$Mv = \lambda v\,, \qquad (3.9)$$

where λ is a real number, called the *eigenvalue* corresponding to *v*. Interpreting the matrix multiplication in (3.9) as a mapping v α Mv, the eigenvectors are the vectors that do not change their direction under this mapping. If the matrix *M* is symmetric, the eigenvalues are nonnegative, and eigenvectors of distinct eigenvalues are orthogonal. Since the variables x_1, …, x_{18} are specified on different scales, we extract the eigenvectors from the correlation matrix (3.8), as this is a usual statistical practice in such a case. This matrix can also be interpreted as the covariance matrix of the *standardized variables*

$$z_i = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}\,. \qquad (3.10)$$

There are several criteria for determining the number *m* of principal components (see (3.1)). A popular, but "fuzzy" criterion is to identify the "elbow" in the scree plot (see Fig. 3.2), according to which one could restrict the analysis to three components. However, in order to gain greater insights into the problem, we have decided to construct five principal components.

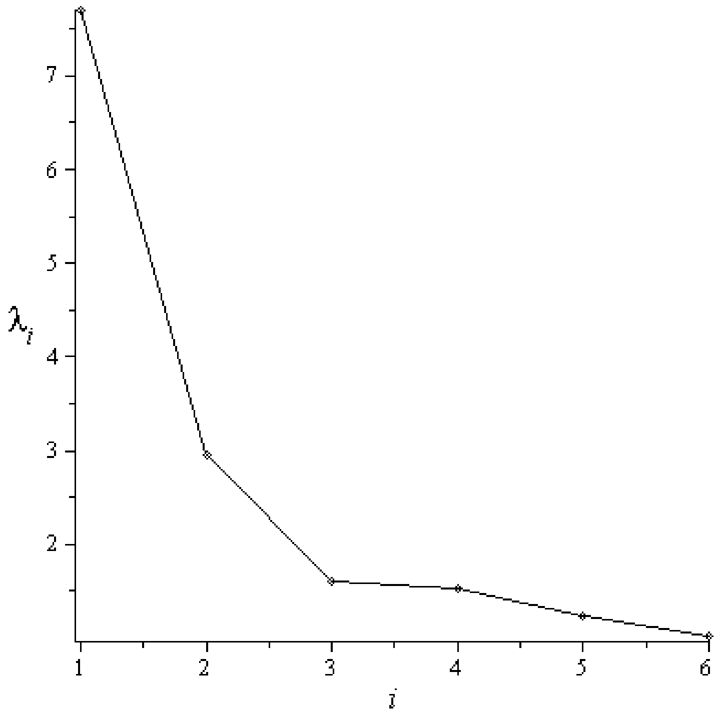


Fig. 3.2. Screen plot of the first six eigenvalues

The largest five eigenvalues of the correlation matrix R (see (3.8)) are collected in Table 3.1 in the decreasing order. The eigenvalue λ_i represents the variance of the i-th principal component PC_i. The last column gives the proportion of variance explained by the first i components – i.e., the ratio $\frac{\lambda_1 + ... + \lambda_i}{\lambda_1 + ... + \lambda_{18}} = \frac{\lambda_1 + ... + \lambda_i}{18}$ (i = 1, …, 5). In particular, the table shows that the first five principal components explain 83.5% of the variance.

Table 3.1
The eigenvalues and the respective proportions of variance explained

i	**Eigenvalue** λ_i	**Proportion of Variance Explained**
1	7.7008	0.428
2	2.9574	0.592
3	1.6027	0.681
4	1.5258	0.766
5	1.2419	0.835

Table 3.2
The first five eigenvectors of the studied indexes

Index	e_1	e_2	e_3	e_4	e_5
x_1 = TTR	-0.34	-0.0575	0.0694	-0.148	0.0109
x_2 = h-Point	0.213	0.0235	-0.199	-0.0835	0.583
x_3 = ATL	-0.217	-0.092	-0.0299	0.0822	-0.422

x_4 = R1	-0.303	0.0672	0.0723	0.354	-0.0265
x_5 = RR	0.312	-0.0339	-0.101	-0.0325	-0.241
x_6 = RRmc	-0.321	0.0731	0.0255	0.313	0.0025
x_7 = TC	0.0601	-0.55	0.0912	0.132	0.0153
x_8 = STC	0.0486	-0.493	0.0255	0.286	0.0507
x_9 = PTC	0.0603	-0.55	0.0941	0.13	0.0136
x_{10} = Lambda	-0.272	-0.110	0.176	-0.391	0.134
x_{11} = R4	-0.352	-0.0356	0.0677	-0.0464	-0.0197
x_{12} = HP	-0.301	-0.0769	0.051	-0.295	0.0812
x_{13} = WV	-0.155	0.0623	-0.313	0.409	0.508
x_{14} = R	-0.328	0.0871	-0.0194	0.203	0.0215
x_{15} = BCA	0.0323	0.182	0.209	0.234	-0.105
x_{16} = Q	0.128	0.125	0.569	0.148	0.114
x_{17} = N	-0.201	-0.207	0.0407	-0.308	0.227
x_{18} = ADJ	-0.107	-0.0851	-0.619	-0.0352	-0.249

In the principal component analysis, the eigenvectors are used to calculate the components as $PC_i = e_{i1} x_1 +...+ e_{ip} x_p$, where the e_{ij} denote the elements of the *i*-th eigenvector (see the next section). Thus, the larger is the *absolute value* of e_{ij}, the more strongly PC_i is influenced by this element. Generally, e_{ij} can be considered as an "indicator of importance" of the variable x_j, which is a stylometric index in our linguistic application.

Table 3.3 shows the five counts that score highest in the respective eigenvector (the columns of Table 3.2).

Table 3.3
Top five stylometric indexes in the first five eigenvectors

e_1	e_2	e_3	e_4	e_5
R4	TC	Adjectivity	Writer's View	H-point
TTR	PTC	Activity	Lambda	Writer's View
R	STC	Writer's View	R1	ATL
RRmc	Nominality	Belza-chain Associations	RRmc	Adjectivity
RR	Belza-chain Associations	H-point	Nominality	RR

Concerning the first eigenvector e_1, the listed indexes cover the sphere of vocabulary richness; RR and RRmc are mutually dependent, which explains the fact that they change together throughout the poems. As to TTR, it is noticeable that it is an essential factor in the composition of sonnets, though it considerably depends upon text length; as it still plays a role in the spatially-circumscribed poems, it thus means that lexical structure is what sonnets probably most differ in.

The second eigenvector e_2 yields a completely different shape. Now, all three types of thematic-concentration counts (TC, STC, and PTC) belong to the most important indexes; they tend to change together, which confirms their interdependence. As to N and BCA, it makes some sense, as there are spottable distinctions as to thematic organizations of the poems, and the proportion of nouns to the whole of the lexis speaks about vocabulary richness, reflecting, therefore, the results for the first eigenvector. Moreover, N and BCA are correlated, as it was indicated in the commentary to Figure 3.1; and since thematic words in the sonnets are mostly nouns, the common feature of the indexes in the present eigenvector may be the distribution of substantives in the researched sonnets.

The third column of Table 3.3 presents a miscellaneous mix of various features; there are h-point-based indicators (writer's view and the h-point itself), thematically-focused Belza-chain associations, and two POS indicators (adjectivity and activity – see, e.g., Zörnig, Altmann (2016)). The occurrence of the latter, which took place also in the preceding set, may underline the important position of POS distributions in the studied texts. The other outcomes confirm the tendencies mentioned above.

A coherent overview of the preferences is presented by the fourth eigenvector; here, the dominant role is played by vocabulary-richness indexes (R1. RRmc, and lambda), and those that are loosely linked to them (writer's view and nominality). The same trend is confirmed by the fifth set of measures, where the lexis-counting formula (RR) is accompanied by the h-point-based calculations (writer's view and h-point), style-marking average tokens length, and POS-governed adjectivity. It seems that this column provides a display of what matters in the style of the poems in question.

To put the previous ideas on a mathematical ground, we have devised a simple tool to assess the prominence of the indexes throughout the rankings. The ranks were given weights according to their decreasing importance – the first place is endowed with the weight of 5, the second one by 4, etc. For each index, the sum of its weighted rankings was divided by the total number 5 of characteristics. The numbers, which are to be found in the interval [0.2, 5], are useful for comparison when one needs to evaluate the importance of the individual counts. If an index scores 5, it came first in all the studied sets; if it yields the value of 0.2, it scores fifth in one of them only.

The calculation of our index, which we will call *weighted occurrence index* (WO), is exemplified upon the writer's view. Within the given scores, writer's view occurs three times – once in the third position (weight 3), once in the first position (weight 5), and once in the second position (weight 4). This gives the WO of the writer's view as –

$$WO = \frac{3+5+4}{5} = 2.4\,.$$

The results of other indexes are listed in Table 3.4.

Table 3.4
Indexes of weighted occurrence of the studied stylometric indicators

Index	**WO**
Writer's View	2.4
Adjectivity	1.4
H-point	1.2
TC	1
R4	1
TTR	0.8

PTC	0.8
Lambda	0.8
Activity	0.8
RRmc	0.8
ATL	0.6
R1	0.6
STC	0.6
Belza-chain Associations	0.6
Nominality	0.6
R	0.6
RR	0.4

The chart demonstrates as the most “weighted” indexes h-point and writer’s view, which are mutually dependent; moreover, the h-point is the basis for vocabulary-richness counts, so one might expect that it scores very high. As to the bottom part of the table, it seems noteworthy that in the researched poems, adjectivity is much more prominent than nominality, which appears in the lower ranks. This may mean that Machar’s sonnets differ much as to their usage of adjectives, whereas the number of nouns do not change so often; this can be employed as both a style indicator, and a check of the usefulness of the proposed counts.

Last but not least, a formula has been sketched for measuring the WO of groups, too. To this purpose, the indexes were divided into the classes according to the following key, the previous ideas on the correlations (see Fig. 3.1) having been considered:

1) Vocabulary-richness indexes (TTR, h-point, R, R1, R4, RR, writer’s view, lambda, and RRmc);

2) Thematic-concentration indexes (TC, STC, and PTC);

3) Lexical-style indexes (ATL);

4) Thematic-organization indexes (BCA);

5) Indexes of POS distributions (N, Q, and ADJ).

After the weighted values of the individual groups had been summed up, they were divided by the number of indexes in the group multiplied by five. For instance, in case of POS indexes, the count yields –

$$WO_{POS} = \frac{2+1+4+5+2}{3*5} = 0.93\,.$$

The remaining values are listed in Table 3.5.

Table 3.5
The indexes of weighted occurrence in the groups of the studied stylometric indicators

Index Group	**WO**
Vocabulary Richness	0.96
POS Features	0.93
Thematic Concentration	0.8
Lexical Style	0.6
Thematic Organization	0.6

The chart corroborates the vital character of vocabulary-richness counts for the studied poems; these are accompanied by POS-based indexes, which account for the distribution of word-classes in the sonnets. On the other hand, neither thematic concentration, nor thematic organization seem to say much about the samples; this may be indicative of the general nature of the poems, which do not revolve around one particular topic, and of the similar structural treatments of the themes. The sonnets thus tend to be the same in cohesion structure, to have no determined topics, and to vary mostly according to the range of the words used.

4. Principal Components: Computation and Interpretation

By means of Table 3.2, we can calculate the *factor scores* – i.e., the values of the principal components for the sonnets in Table 2.2. Let $z = (z_1, ..., z_p)^{\mathrm{T}}$ be the vector of standardized variables (see (3.10)), and $e_i = (e_{i1}, ..., e_{ip})^{\mathrm{T}}$ the i-th eigenvector – where $p = 18$, and "T" denotes the transpose of a vector –; then the principal components are computed as

$$PC_i = e_i^T z = (e_{i1},...,e_{ip}) \begin{pmatrix} z_1 \\ \mathrm{M} \\ z_p \end{pmatrix} = e_{i1}\, z_1 + ... + e_{ip}\, z_p \qquad (4.1)$$

(see, e.g., Johnson, Wichern (2007, p. 451)). For example, taking the values of e_i from Table 3.2, equation (4.1) yields the formulas

$$\begin{aligned} PC_1 &= -0.34\, z_1 \quad + 0.213 \quad z_2 - ... -0.107 \;\; z_{18} \\ PC_2 &= -0.0575\, z_1 + 0.0253\, z_2 - ... -0.0851\, z_{18} \\ &\quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \qquad\qquad\qquad\qquad (4.2) \\ &\quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ PC_5 &= 0.0109\, z_1 \quad + 0.583\, z_2 - ... \;\; -0.249\, z_{18} \end{aligned}$$

for the scores of a random vector Z. The factor scores of the sonnets are now obtained, substituting the values z_i in (4.2) by the numerical values obtained for the sonnets (see Table 2.2). In this way we obtain, for example, the following scores for the **first sonnet** (*E. Zolovi* – "To É. Zola"):

$$PC_1 = -0.34\ \frac{x_1 - \bar{x}_1}{\sqrt{s_{1,1}}} + 0.213\ \frac{x_2 - \bar{x}_2}{\sqrt{s_{2,2}}} -\ldots -0.107\ \frac{x_{18} - \bar{x}_{18}}{\sqrt{s_{18,18}}}$$

$$= -0.34\ \frac{0.8462 - 0.7791}{\sqrt{0.003}} + 0.213\ \frac{3 - 3.2163}{\sqrt{0.2647}} -\ldots -0.107\ \frac{0.1026 - 0.1026}{\sqrt{0.0025}}$$

$$= -0.34 * 1.225 \quad + 0.213 * (-0.420)\ -\ldots\ -0.107*0\ = -3.395 \qquad (4.3)$$

In the same way, we get:

PC_2 = -0.0575 *1.225 + 0.0253* (-0.420) -… -0.0851*0 = 0.091,
PC_3 = -0.112, PC_4 = 0.274, PC_5 = -0.805.

By analogy to (4.3), the scores of all other 47 sonnets are calculated in Table 4.1. They have to be interpreted as "new" variable values "summarizing the information" given by the original 18 variables.

Table 4.1
Component scores for 47 sonnets in Table 2.2

Sonnet	PC_1	PC_2	PC_3	PC_4	PC_5
1	-3.395	0.091	-0.112	0.274	-0.805
2	-3.136	0.381	-1.574	0.924	-1.157
3	4.864	0.165	-2.451	-1.829	-0.877
4	**0.329**	**-4.01**	-2.043	2.098	1.179
5	0.644	0.409	-0.304	-0.553	-0.683
6	-2.404	-0.755	-0.836	1.348	-0.191
7	-1.739	0.601	0.674	0.214	2.636
8	1.511	0.377	0.972	-0.446	-1.62
9	-0.943	0.232	-0.924	-0.477	0.545
10	**-5.129**	**0.037**	0.701	0.096	-1.563
11	5.825	1.123	-0.341	1.003	0.65
12	-2.878	0.567	0.274	1.413	0.816
13	-0.829	-0.069	0.212	-1.302	-0.006
14	0.354	0.338	0.75	-0.68	-0.586
15	-3.057	0.442	0.396	0.974	0.583
16	-0.285	0.691	-2.524	1.58	-0.447
17	-1.984	0.737	1.101	0.583	2.304
18	0.289	0.086	1.935	0.65	-0.631
19	**3.204**	**-10.19**	1.818	0.666	-0.305
20	5.092	1.618	1.078	1.921	-1.616
21	-2.183	0.339	-2.012	1.261	0.865
22	-0.467	0.046	-0.888	-0.726	2.09
23	3.586	-0.053	-1.517	2.272	-0.522
24	4.556	0.521	0.306	-1.668	0.392
25	-2.519	-0.46	-2.07	-0.562	-0.47
26	1.238	-0.271	-0.777	-2.043	-0.885
27	-0.244	0.671	1.666	-0.471	-0.262
28	-1.601	-0.355	1.113	-1.992	2.768

29	-2.796	0.211	1.518	0.153	-0.765
30	0.619	0.898	1.508	0.666	-0.834
31	1.358	0.677	1.27	-0.037	-0.269
32	-1.266	-0.386	-0.256	-2.116	1.751
33	0.889	1.067	-0.196	1.206	0.532
34	-1.505	1.617	2.941	1.524	-0.464
35	-2.583	-0.397	-2.093	-0.902	-1.058
36	**-6.759**	**-0.849**	0.299	-1.324	-1.104
37	-2.811	-0.242	-0.715	-0.038	-0.503
38	1.104	0.748	1.169	0.248	-1.09
39	0.811	0.019	0.332	-1.997	-0.022
40	2.065	0.735	-0.607	0.015	-0.247
41	2.175	1.146	-1.38	0.987	-0.44
42	1.578	0.909	0.449	1.419	2.421
43	2.524	0.717	1.226	-0.371	0.154
44	-1.064	-0.424	-0.502	-1.99	-0.494
45	5.788	0.751	-0.438	-0.829	1.077
46	2.369	-0.089	0.262	-2.059	-0.697
47	-1.195	-0.411	0.589	0.914	-0.151

Restricting our attention to the first two principal components, the scores of all 47 sonnets are graphically displayed in Fig. 4.1.

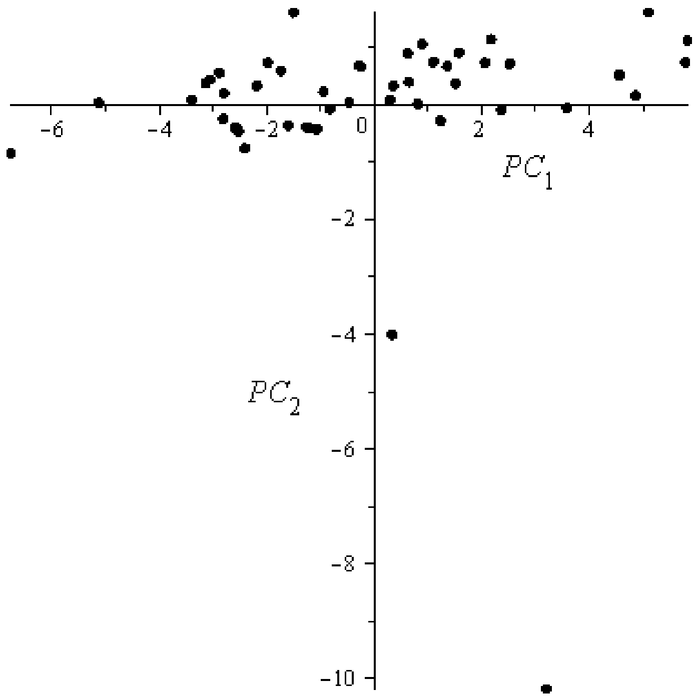


Fig. 4.1. Scatter plot of the first two principal components

Optically, it seems that there are four sonnets that escape the general trend of the principal components (see the bold values in Table 4.1); these are *Sonet de vanitate* (4 – "Sonnet on Vanity"), *Sonet mystický* (10 – "Mystical Sonnet"), *Sonet o lásce* (19 – "Sonnet on Love"), and *Sonet večerní* (36 – "Evening Sonnet"). The outliers, sonnets 4 and 19, are thematically concentrated (see Table 2.2), which is something unusual in the other samples; on the other hand, the Mystical and Evening Sonnets boast very high TTRs, which ranks them as lexically abnormally rich. This is attributable to the critical treatment of fashionable esoteric vocabulary in the former, and to the picture-like description of the eve in the latter.

In order to divide the sonnets into groups, we have used the k-means cluster algorithm supplied by the free software R. An illustration of the operation of this procedure can be found, e.g., in Zörnig et al. (2016). Given the shape of the elbow diagram, the optimal number of clusters appears to be three or four. As it is visible in Table 4.2, the three-cluster solution is primarily based on the values of PC_1, as these vary much more that the ones of PC_2. The sonnets belonging to cluster 3 usually deviate from the remaining ones by high TTRs (the average being 0.82), which is sometimes accompanied by other discrepancies, for instance in ATL (typically, in the first sonnet, which is a complicated dedication to Émile Zola, and contains a lot of long, abstract expressions). This causes that the vocabulary-richness indexes tend to be high as well; for instance, the Pearson Correlation Coefficient between TTR and R1 yields the value 0.53. Note that the correlations between variables of sonnets in the same cluster are calculated via the formulas (3.4), (3.5), and (3.7); however, the summation is only over the indexes i within the respective cluster. On the other hand, the scores of the poems in cluster 2 are lower, and the indexes are thus not prone to interconnected changes (the correlation of the above-mentioned equals as little as 0.25). As to cluster 1, the indexes of the poems manifest moderate values (the average of TTRs being 0.77), and the correlations are uneven, showing a middle figure in case of TTR and R1 (0.43) and a high one in comparing TTR and RR (-0.62). The differences among the clusters thus lie both in the figures of the indicators, and in their variegated interconnectedness.

We finally studied alternative clustering procedures and found out that a subdivision of the component values in clusters is essentially the same when four clusters are considered instead of three. The only difference compared with the previous result is that "Sonnet on Love" (no. 19) becomes a cluster of its own. This is due to its high score in thematic concentration, which is not paralleled in any other of the poems.

In the aforementioned algorithm, we have presented the sonnets as points in the plane, given by the scores of PC_1 and PC_2. For example, the sonnet 1 has been presented by the point $(-3.395, 0.091)^T$ (see Tab. 4.1). In order to make use of the information contained in the other columns of that table, we could present the sonnets by means of three, four, or five components. For example, if three columns were used, the sonnet 1 would be represented by the point $(-3.395, 0.091, -0.112)^T$ in the three-dimensional space. We have applied the clustering algorithm also for these higher-dimensional variants. However, the only consequence of presenting the sonnets via more than two components is that the sonnets 16 and 22 move from cluster 1 to cluster 3. This means that the additional information contained in the components PC_3, …, PC_5 does not change the results considerably.

The overall results are presented in Table 4.2, and visualized in Figure 4.2. Note that the positions of the points in Fig. 4.2 differ from that of Fig. 4.1 in so far that for a convenient graphical representation, the points have been reflected on both axes and scaled differently.

Table 4.2
Grouping the sonnets into three clusters

Cluster	PC$_1$	Sonnets	Properties
1	values < -2.5	4, 5, 8, 14, 16, 18, 22, 26, 27, 30, 31, 33, 38, 39, 40, 41, 42, 43, 46	lexical-richness values of medium size; varying correlations
2	-2.5 < values < 0	3, 11, 19, 20, 23, 24, 45	lexical-richness lower values; lower correlations
3	values > 0	1, 2, 6, 7, 9, 10, 12, 13, 15, 17, 21, 25, 28, 29, 32, 34, 35, 36, 37, 44, 47	lexical-richness higher values; higher correlations

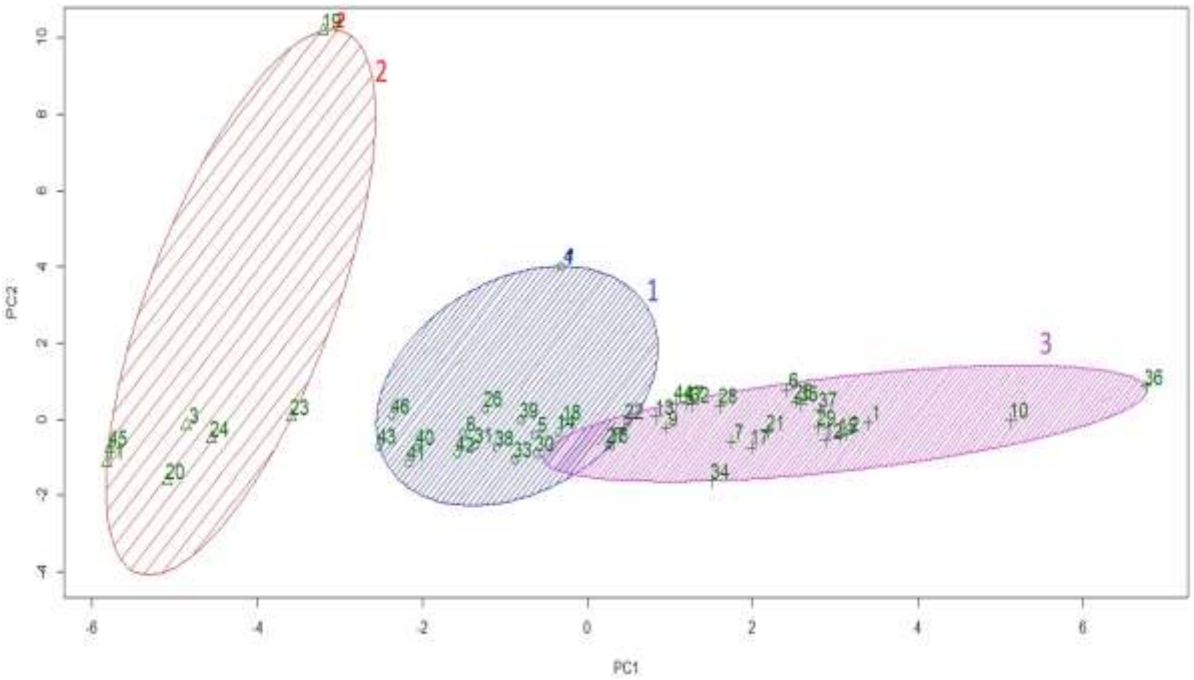


Fig. 4.2. The cluster plot of the studied sonnets on the basis of PC_1 and PC_2

5. Conclusions

The goal of the article was to apply the principal component analysis to the corpus of Czech sonnets; this having been done, it was possible to partition the studied indexes into groups according to their importance, and to sketch a couple of interpretations based on that approach. In order to check whether the results are generally valid, more texts from various genres must be studied, as the aforementioned outcomes are based solely on a specific collection of sonnets. So far, it seems that vocabulary-richness indexes are of the highest interpretational value, as they appear in many top positions in the first five eigenvectors. As to the ranking in the eigenvectors, the groups were presented in Table 3.3; the first set covers vocabulary-richness indexes, the second focuses on thematic concentration and connectedness, and the third introduces POS-based indicators. The other two combine the lexis-measuring counts with the POS-based ones.

As to the cluster analysis, which was elaborated on the basis of the PC_1 and PC_2 values, the poems can be divided into three groups, the main divisive factor being the scores of the lexical-richness indicators. There is a specific set – cluster 2 – which manifests higher repetitions of words and, in some cases, thematic concentration; it usually covers the topic of love (19 –"Sonnet on Love"), past (20 – "Sonnet on Past", 23 – "Sonnet on an Old Metaphor", 24 – "Sonnet on an Old Front and Reverse"), and capricious motifs (11 – "Sonnet on Chopin's Tune", 45 – "Sonnets – Causeries: IV"). On the other hand, there appear to be no linking topics in the other two groups of the sonnets; only a general statement can be made that the sonnets with high lexical-richness values are much more frequent than those with low ones (cluster 3 containing 22 poems out of the total of 47). This cluster also incorporates the poems on complex socio-political and that-day issues (1 – "To É. Zola", 7 – "Sonnet on a Social Question", 13 – "Sonnet at the Close of the Century", 29 – "Sonnet on the Golden Age of Our Poetry"); however, it does encompass the general-subject poems, too (35 – "Introductory Sonnet", 36 – "Evening Sonnet"). It is thus quite problematic to draw any conclusions as to its topical specificity.

To sum it up, the findings presented in the paper should thus be subjected to more analyses, and some topics – such as the index correlations – could form the basis for many more quantitative linguistic studies. More work also awaits the researcher concerning interpretations of the outcomes.

References

Baayen, R. H. (2013). Multivariate statistics. In: Podesva, R., Sharma, D. (eds.), *Research Methods in Linguistics*, 337–373. Cambridge: CUP.

Johnson, R. A., Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Upper Saddle River: Pearson Prentice Hall.

Kubát, M., Matlach, V., Čech, R. (2014). *QUITA – Quantitative Index Text Analyser.* Lüdenscheid: RAM-Verlag.

Machar, J. S. *Letní sonety.* Available at:
<http://www.rodon.cz/admin/files/ModuleKniha/1100-Ctyri-knihy-sonetu.pdf>.

Místecký, M. (2018a). Counting Stylometric Properties of Sonnets: A Case Study of Machar's *Letní sonety*. *Glottometrics* 41, 1–12.

Místecký, M. (2018b). Belza Chains in Machar's *Letní sonety*. *Glottometrics* 41, 46–56.

Zörnig, P. (2016). *Probability Theory and Statistical Applications: A Profound Treatise for Self-Study.* Berlin/Boston: De Gruyter.

Zörnig, P., Altmann, G. (2016). A sequential activity measure for texts and speeches. *Glottotheory* 7(2), 195–212.

Zörnig, P., Kelih, E., Fuks, L. (2016). Classification of Serbian Texts based on lexical characteristics and multivariate statistical analysis. *Glottotheory* 7(1), 41–66.

Glottometriocs 43, 2018, 31-57

Stylistic study of *Omnilingual* by H. Beam Piper

Tomas Melka[1]

Abstract: In order to capture stylistic features of H. Beam Piper's classical story "*Omnilingual*" (1957) a number of quantitative measures are drawn in. As the aid of such measures is enlisted, various possibilities present themselves. We shall restrict ourselves to a small choice of possible descriptions, focused on the vocabulary richness and the time-structured writing sequence.

This Piper-esque writing has entered the records of the sci-fi prose for the "Martian" periodic table of elements, being synonymous with a scientific "Rosetta-like stone" in the decipherment area. The work, while having a search potential in text analysis and stylistics, may incidentally add some luster to the validity of the science as a communicative channel in non-conventional circumstances.

Key words: archaeological decipherment, data interpretation, Gini's coefficient, rank-frequency distribution, "*Rosetta stone*" artifact, style, vocabulary richness

1. Introduction

The study of style is probably the most complex problem in text analysis.[2] There are numerous definitions and accounts but hardly any of them is thoroughly operational (cf. in chronological order, Coleridge, 1914 [1907, 1818]; Bain, 1877; Lanson, 1916 [1903]; Cooper, 1930 [1907]; Smith, 1916; Murry, 1922; Fulcher, 1927; Sebeok, 1960; Frye, 1963; Baker, 1966; Gray, 1969; Babb, 1972; Morton, 1978; C. Freeman, 1979; Guth, 1980; D. Freeman, 1981; Kane, Peters, Jackel, & Legris, 1981, and other past and recent scholars, too many to list here). In point of fact, one has to counter after all the elementary question "*what is it*?" As seen in the next formulations, the answers vary from *pithy* and *elegant* to *elaborate* and *sinuous*,

«Proper words in proper places make the true definition of a style» (Jonathan Swift, 1721, in Cooper, 1930 [1907], p. 161).

«Style is, of course, nothing else but the art of conveying the meaning appropriately and with perspicuity, whatever that meaning may be, and one criterion of style is that it shall not be translateable without injury to the meaning» (Samuel Coleridge, 1914 [1907, 1818], p. 325).

«Style is the mode of representation in language, conditioned partly by the psychological peculiarities of the one who represents, partly by the matter and purpose of what is represented» (Wilhelm Wackernagel, 1888 [1873, 1836-1837], in Cooper, 1930 [1907], p. 12).

«Style is a quality of language which communicates precisely emotions or thoughts or a system of emotions or thoughts, peculiar to the author. Where thought predominates, there the expression will be in prose; where emotion predominates, the expression will be indifferently

[1] Send correspondence to: tmelka@gmail.com

[2] See e.g. B. Gray (1969: 7), "*Few problems in literary scholarship continue to generate so much endeavor and so much conflict as the problem of style*."

in prose or poetry, except that in the case of overwhelming immediate personal emotion the tendency is to find expression in poetry. Style is perfect when the communication of the thought or emotion is exactly accomplished; its position in the scale of absolute greatness, however, will depend upon the comprehensiveness of the system of emotions and thoughts to which the reference is perceptible» (Murry, 1922, p. 71, Chapter IV).

«*Style is not a garment to be slipped on that the bare and shivering idea may be warmed, or decorated, or padded out. It is an integral part of the pure, warm, living, naked body of the thought, bone of its bone, muscle of its muscle, heart of its heart. Style is organic. Whosoever touches it, touches a man*» (Fulcher, 1927, p. vi).

Northrop Frye (1963, p. 60) is inclined to assess style in terms of *the individual conventions of the author*.

«*Whatever* "*styles*" *are, in language or elsewhere, they are part of a system of distinction, in which a style contrasts with other possible styles, and the social meaning signified by the style contrasts with other social meanings*» (Judith T. Irvine, 2001, p. 22).

«*Style is the intangible essence of what makes a person's writing unique. It is writer's voice that makes texts on the same theme written by different authors structured, verbalized and perceived in a different way*» (Nadia Yesypenko, 2008, p. 18).

«*Style is a latent property of language meaning it cannot be measured directly. The rarity and regularity of linguistic constructs can be quantified, and they provide an indirect indication of the underlying stylistic profile*» (Bell, Berridge, & Rayson, 2009, p. 3).

Despite the professed insights or the sharp / colorful devising, the above answers have nothing to do with an established theory. They may be viewed as the first steps on the ladder leading from intuitive, inductive definitions, to more theoretical answers concerning the questions: "*How does it behave*?" and "*Why does it behave* (*in*) *that way*?" The hypotheses set up in this domain will never be ideally answered / tested because style is a *property* not only of written texts but also of discourse situations, not to mention the fashion domain (cf. Sebeok, 1960; Eckert & Rickford, 2001). Nevertheless, one can try to approximate the theoretical level by setting up simple, quantified indicators or functions, test them and seek the answer to the "*why*?"-question by finding some other properties with which the given one is correlated. Style – just as any other structural phenomenon – is result of self-regulation, based on "…*its own internal and self-sufficient rules*" (Hawkes, 2004 [1977], p. 6). If an observed property *A* behaves in a certain manner, it may be thereby queried: *what is the cause of this behavior*? The other property associated with *A* must be quantified too, and their relation must be expressed in form of a testable hypothesis. Procedures of this kind are well known from the synergetic linguistics (cf. Köhler, 2005), already being applied in text analysis, too.

Methodologically, if only one text property is evaluated – *vocabulary richness*, in H. Beam Piper's case –, there must be a possibility of comparing at least two disparate texts, in order to show that there is a stylistic difference. Such a difference will relate to the personality of author, or one of her/his characters if the author is the same. But although the differences may be interpreted by and large intuitively, they must be supported statistically.

The number of statistical methods used for this purpose is sufficient and may be (should be) used for any examination. The question and the commitment to unlock the answers of "*what are the properties of a text*?" – on a microcosmic scale – are perhaps no less similar to those of "*what are the properties of the world*" – on a macrocosmic scale –? The identified text properties are in large numbers, but, as science develops, none of them should be considered definitive" (cf. Popescu et al., 2009, p. 250). They are merely more or less fundamental properties. The more properties are directly related to the examined one, the more fundamental it is – for us or for a theory. Further research can completely change its direction and study quite different properties. One should not be oblivious to the fact that every Inspection, however useful, is only a reflection of her/his restricted view of the text, not the capturing of "truth."

When modeling a property or a relation, some principles also used in other scientific disciplines may be followed. A number of them are mentioned ahead: (1) every modeling is a simplified version of the reality; hence we shall try to streamline the model[3] as much as possible. This can be partially achieved by allowing only a small number of parameters. In particular, the *richness of vocabulary variety*, as an aspect of style of H. Beam Piper (1957), becomes the main focus in the current study. Soon after, new parameters can be added due to different boundary conditions. However, the framework that includes additional properties / parameters needs to be carefully treated, as indicated via the so-called "*feature proliferation*" in Argamon et al. (2007). (2) The parameters should be defined in such a way that they are interpretable in textual, linguistic, typological, etc., terms. (3) One should see to the derivation of the resulting model by differential or difference equations, stochastic processes, etc. and in any case, modeling by polynomials should be avoided. Polynomials can grasp any sequence, but their parameters are not easily interpretable (cf. Zieffler, Harring, & Long, 2011, pp. 252-253). One should avoid also cyclic dependencies and rather redefine / re-quantify the properties in order to obtain some "*smooth*" relations. (4) Certain principal forces playing a role in the Zipfian/Köhlerian linguistics (Zipf, 1935; Köhler, 2005, p. 766, shows 22 requirements) can be implemented, but in text analysis, some of them need to be differently interpreted, or new ones will be added, with this course of action apparently never ending. Quite often, the ranking of phenomena according to their frequency is used. The ranks at hand mean an artificial, secondary quantification, but ranked phenomena mostly show a clear, smooth sequence which has some properties. As a matter of fact, *ranking* is a simple monitoring and capturing of our intuitive knowledge shaping the text under discussion, of our intuitive description of the authorial style, etc.

2. H. Beam Piper and his story "*Omnilingual*"

Technically, it would be not only appropriate but also prudent to enlighten some aspects about the author and "*Omnilingual*" (1957). While delving into the personal and historical background, the performed analyses/observations could be better understood and subject to further expansions/improvements in the future. Likewise, in terms of style, "*Omnilingual*" is perhaps not as (a) socially interesting and complex, (b) "mysterious" and suspenseful, or (c) witty, experimental and surreal, as several pieces written by (a) Charles Dickens, (b) Arthur Conan Doyle, or (c) Kurt Vonnegut Jr., but the author does his bit to bring a fresh look on the interpretation of unknown data (*deciphering*, in this case) using a cache of words that excelthe daily language output of many humans. In this sense, useful assumptions can be made and clues can be possibly found about the author and the deployed vocabulary richness.

H. Beam Piper (1904-1964) is a US fantasy and science-fiction writer. The first name assumed to be Henry or Horace is just another indication about the puzzle surrounding many facets of his life, and the untimely death by suicide. H. Beam Piper (HBP) did not have much literary success during lifetime: his first story *Time and Time Again* is published in 1947

[3] For additional statistical models, reviews, and problems in stylometry, authorship, and/or forensic linguistics, see Argamon et al. (2007); Bell, Berridge, & Rayson (2009); Holmes (1998); Juola (2008); Oakes (2009); udman (1998); Thisted & Efron (1987); Tuldava (2005); Tweedie (2005), and referenced literature thereof.

when he was forty-two years old. Having a full-time position at Pennsylvania Railways he was able to support his self-education and passion for writing during a non-negligible period of time. The mystery novel *Murder in the Gunroom* (1953); the sci-fi novels *Little Fuzzy* (1962), *Space Viking* (1963) and *Lord Kalvan of Otherwhen* (1965) are considered literary highlights of his career. Biography writers tend to think that a marriage gone awry, an unfriendly divorce, the death of his literary agent Kenneth White, together with dire economic woes, caused HBP to self-destruct (Hines, 2002; Carr, 2008). Scepticism about the value of his work has been supplanted in latter times by a cult-following movement, by several reprints of the major works, and by a growing acclaim of his originality and insightfulness.

*** *** ***

Omnilingual published as a novelette in *Astounding Science Fiction* magazine (February1957; Figure 1), subsequently known as *Analog*, was later collected in *Federation* (1981), a compilation of short stories by HBP. *Omnilingual* deals with a human survey party– archaeologists included –, looking for clues and/or indigenous relics among the ruins of a very ancient Martian city. Consider that we are in the realm of fiction and such things do occur accordingly. On the other hand, science-fact suggests that the red planet's surface-water likely disappeared before rocks formed about 3.9 billion to 4.6 billion years ago (cf. Carr & Head, 2010; Kramer, 2014). A local culture given to building Martian universities and complex research facilities could hardly thrive in the aftermath amidst the cold and barren environment and the absence of a protective atmosphere against UV showering and other radiations. It should also be stated at this point that Piper's is old-fashioned science-fiction, with the Mars theme and its nhabitants ("natives," and/or Earth immigrants), regularly used in several storylines by redecessors[4] or contemporaries. Unusual methods of traveling and mis/adventures, in concert with philosophical, technological, and socio-ethical quagmires of the various kinds are shown, e.g. in Percy Greg (1880); Gustavus William Pope (1894); Kurd Lasswitz (1971 [1897]); Herbert G. Wells (1898); Edwin Lester Arnold (1905); Alexander Bogdanov (1984 [1908]); Edgar Rice Burroughs (1917 [1912]); Aleksey Nikolayevich Tolstoy (1950 [1922]); Stanley G. Weinbaum (1934); John Beynon [a pen-name of John Wyndham] (1972 [1936]); for more, see Bleiler (1990).

The tradition of the "*Mars fever*" (Fergus, 2013) continues after WWII with Ray Bradbury (1950); Arthur C. Clarke (1951); Robert Heinlein (1961); Philip K. Dick (1964); Frederick Pohl (1976), who recount invented, spectacular, and polyvalent tales linked – one way or another – with our planetary neighbor. Ted E. Dikty (1966) also has an anthology of short stories dedicated to Mars, where HBP's *Omnilingual* is reissued (cf. Algis Budrys (1967). Later, the Mars-related literary patterns turn to more hard-science and specific topics. In such works, the human endurance and resourcefulness are strongly tested in one or more quandaries, with the nearly far-fetched outcomes tickling one's imagination, or with the realization of extraordinary feats, e.g. Kim Stanley Robinson (1992, 1993, 1996); Greg Bear (1993); Paul J. McAuley (1994); Geoffrey A. Landis with *Mars Crossing* (2000); Andy Weir with *The Martian* (2014 [2011]), or Terry Pratchett & Stephen Baxter (2014). And, to reconnect now with the tangible reality, it is no longer a question whether human species will set foot on Mars and initiate settlement projects, rather than *when* (see e.g. Smith, 1989; Mars One [M1], 2018).

[4] The quoted works, integrated in their time-space frame (1880-1936), ask for much scientific indulgence from the 21st century discerning readers.

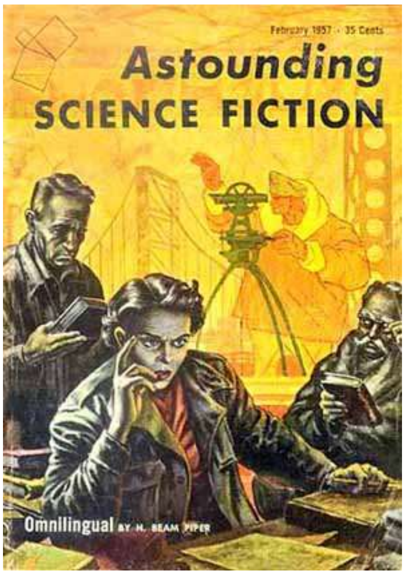


Figure 1: Cover of *Astounding Science Fiction* (February 1957; edited by John W. Campbell, Jr.) featuring the fictional characters M. Dane, H. Penrose, and S. von Ohlmhorst. In the far background, a mural found amidst the University ruins shows a "*heroic-sized Martian*" handling a "theodolite"-like apparatus. Illustration by Frank Kelly Freas; reprinted after Wikipedia's (2018) public domain image.

The human expedition, part of which is the protagonist of Piper's story, Martha Dane, uncovers some strange writings. In an act of scientific devotion Martha makes assumptions, as she confronts the assertive and ego-driven associate Anthony Lattimer. While the plot is situated in the '90-ies of the past century, "*Omnilingual*" itself was written in the '50-ies, with Anthony Latimer, PhD, not taking well to being upstaged by his young female colleague.With the "mystery" being pursued, the breakthrough occurs when an ancient Martian University is located. Within its confines, scores of books are discovered in what appears to be the *Mars University* library. At some moment, a diagram of a simple atom and a table of words and numbers are seen on one of the walls of the Department of Physics / Chemistry. Given the occurrence of ninety-two slots, Martha speculated that the structural layout corresponded to a *periodic table*. Hence, she builds up the chart in a piecemeal fashion, computing the best alignment between Earth and Martian tables: Hydrogen, No. 1:*Sarafaldsorn*; Helium No. 2: *Tirfaldsorn*; &c. The deconstruction of the Martian words in base roots and affixes helped her in grasping the meaning of extra words in a chain-like reaction. The fact is initially suggested in Section 4, where the archaeologist deduces some similarity between Martian and the German language, having the latter the penchant to generate new words by "pasting existing ones." Wikipedia (2018) keenly dwells on that notion, and suggests by analogy the string of chemical elements – hydrogen: *Wasserstoff*; carbon: *Kohlenstoff*; nitrogen: *Stickstoff*, and oxygen: *Sauerstoff*, each sharing the root word "*stoff*" [stuff, matter, substance], and a different prefix. The ensuing Greek-based patterns are comparable to some extent to the German-based ones: *Nitrogen*, *Hydrogen*, *Oxygen*, *Cyanogen*, etc., where the common root is "*gen*," i.e. *generating, producing, issuing*, with N:"*generating* nitrous gas, or nitrate substances;" H: "*generating* water;" O: "*generating* acid;" CN: "*generating* the gaseous compound of carbon and nitrogen," etc.

With the evidence getting stronger, the interpretation of a long lost language began to fall into place. As the known universe is mostly packed with the element *hydrogen*, it is – in retrospective – *hydrogen* on Earth, Jupiter, Pistol Star or elsewhere, so one tends to think that this kind of "*Rosetta stone*" is, if not fully warrantable, then conceivable. Without the assistance of an "*all-language*" text (the periodic table of chemical elements), it would have taken perhaps more than a lifetime to crack the Martian writings. Looking at *some* of the real historical decipherments, e.g. the names of Egypt-related rulers in the “*Rosetta stone*” (Pope, 1999 [1975]), or the names of Minoan towns for Linear B (Chadwick, 2000 [1958]), it must be concluded that – before penning “*Omnilingual*” –, H. Beam Piper was already familiar with the referents.

2.1 “*Rosetta stones*,” as “momentous” assistants in real-life settings

I should mention at the outset that the *discussions of decipherment* (subsections 2.1, 2.2) rather than a *distraction from understanding the story and the original idea of HBP* are intend- ed to come to their aid, especially for readers insufficiently familiar with the technicalities in the area. The related literature amounts to many articles and volumes of various degrees of depth and breadth, though here we concentrate on one basic criterion for a feasible decipherment of an unknown script/language: securing bilingual inscriptions that presumably encode speech in linearly arranged symbols (Gelb & Whiting 1975: 98-99).

The annals of decipherment have shown not infrequently records whose original content is duplicated, paralleled, or slightly paraphrased in other languages (see Daniels, 1996, 2013; Elliott, 2007; Friedrich, 1971 [1957], p. 153; Gelb & Whiting, 1975; Knight & Sproat, 2009). As language contact situations in pre-modern civilizations are considered matter-of-fact, the phenomenon of bi- and multilingualism should not come as surprise. For “dead” languages, however, the primary evidence about bilingualism is very different from that which modern linguists investigating bilingualism in spoken languages can call on (cf. Adams, 2003, p. 3; Campanile, Cardona, & Lazzeroni, 1988). For the decipherer of ancient languages, written data are the necessary medium. In order to tackle the problem, the precondition is that scholars should be acquainted *sine qua non* with one of the languages. Due to the nature of different and/or incomplete sound encodings in two or more scripts – on the iconic, morphemic, syllabic, segmental level –, consider that a bilingual text is not a type of external clue that instantly produces a unique solution to the retrieved portions of ancient writings (cf. Gelb & Whiting, 1975, p. 99). Overall, by virtue of a true (or quasi) bilingual – a duplicate, e.g. ENG. *Well done*! vs. SPA. ¡*Bien hecho*! –, or, of a virtual bilingual – place names, proper names, e.g. *a-mi-ni-so*, Amnisos; *ko-no-so*, Knossos; *tu-li-so*, Tu/ylissos, after the Linear B decipherment (Pope, 1999 [1975], p. 174; Robinson, 2002, p. 99) epigraphers/script analysts get a foothold allowing them to advance in the elucidation or decipherment of the available material.

Several instances corroborate the above: Abbé Jean-Jacques Barthélemy (1754) used a bilingual to crack the Palmyrene (a form of Aramaic) script (Parkinson, 1999, p.16); many deciphered words in the Luwian [Hittite] language were ascertained later through the Phoenician-Luwian bilingual records of Karatepe hill, aka *the Azatiwada inscription* (Gordon, 1987 [1968], pp. 100-101; Hawkins & Morpurgo Davies, 1978; Hawkins & Çambel, 1999); the Phoenician version of a Cypriot bilingual – provided the key to the Cypriot syllabary (Friedrich, 1971 [1957], pp. 136-139, Fig. 60; Gordon, 1987 [1968], pp. 114-118; Steele, 2013, p. 202); although no conclusive evidence has been proffered on the Etruscan language, the inscriptions on gold plaques found in 1964 in Phoenician and Etruscan at Pyrgi, on the Italian coast not far from Rome (Gordon, 1987 [1968], p. 102; Robinson, 2002, p. 170) hold promise for researchers; the Thugga (modern Dougga) bilingual text in Tunisia (Friedrich, 1971 [1957], pp. 118-121, Fig. 57; O’Connor, 1996, p. 113; R. A. Springer Bunk, 2010, pp. 150-151) also adds up to this list. The bitext is written in Punic and a variant of the “Numidian” (the so-called Lybico-Berber) alphabet, with the monument remembering king Massinissa ten years after his death (138 BCE). Still, the renowned multi-text inscription held responsible for starting the decipherment of a real-world script is the *Rosetta stone* – as is often known today –. The artefact of “*Rs*” shows three systems (Egyptian hieroglyphic, local

Egyptian demotic, Greek) coded in two languages (Egyptian and Greek), displaying a decree in honour of king Ptolemy V Epiphanes related to year 196 BCE. The literature on the artefact and its role in the understanding and interpretation of hieroglyphic writings is enormous; suffice to take note of some references at this juncture, Friedrich (1971 [1957], pp. 17-25); Pope (1999 [1975], p. 61); Quirke & Andrews (1988); Parkinson (1999); Solé & Valbelle (2002); Robinson (2002, pp. 56-60); Ray (2007). While the French scholar Jean-François Champollion is generally credited with using the "*Rs*" to decipher the ancient Egyptian script, the intricacies of the decipherment and the respective contributions are still open to debate (cf. Daniels, 1996, p. 145; Gordon, 1987 [1968]; Knight & Sproat, 2009; Pope, 1999 [1975]; Robinson, 2011; Rogers, 2005).

The NP-collocation *Rosseta stone* has come to indicate by antonomasia any artefact designed to convey parallel and repeat information about entities, events, or cultural phenomena, which assists in restoring their original structure and meaning.

2.2 Cosmic "*Rosetta stones*"

So far, unidentified scripts involving human-related settings have been mentioned. The cumulative knowledge brings on however challenges of a different nature: unknown signals, sign sequences, visual-like data streams derive similarly from other-than-human sources, whether Earth-bound, or not. With each passing year, establishing contact with non-human entities is viewed as more than plausible, raising scientific and philosophical concerns of the highest order.

The actuality of corresponding with other intelligences or sentient life-forms, plus its complex ramifications, are considered in different sources (cf. Callimahos, 1966; Cohen & Stewart, 2002; Dick, 1998; Drake & Sobel, 1992; Engdahl, 2006 [2001]; Golomb, 1968 [1961]; 1963: 17; Heidmann, 1995; Michaud, 2007). The co-ordinated efforts are mainly based hitherto on the current understanding of human needs, of the physics of space and of communication. Therefore, adjudging consciously or unconsciously human characteristics to the "contact language" or "channel" is regarded as a drawback (Baum, Haqq-Misra, & Domagal-Goldman, 2011; Michaud, 2007). The estimated languages and/or devices range from: naturally developed human languages, to mathematical ones, radio signals, visual-symbolic codes, or the dispatch of robotic space probes carrying messages in multiple ways. In this context, the most celebrated "*Rosetta stone*" – intended to be intercepted by any scientifically educated being in outer space – is the coded message of Pioneer 10 space probe of 1972 (see Chandler, 2007 [2002], p. 176; Davies, 1995, pp. 55-56; Gombrich, 1982, pp. 150-151). Despite the outcome of Pioneer 10's mission (e.g. falling short of achieving its goal for multiple reasons), the message stands for a deeply symbolic human effort in contacting intelligences beyond Earth. In a similar manner, it paved the way for additional communication experiments, each a reminder of the humankind's drive to expand the frontiers of knowledge in the physical and metaphysical sense.

3. Analysis of text

Omnilingual (1957) is an autonomous dataset of *c*.16.430 tokens (Figure 2) partitioned in twenty-two sections, with each of them far from being a "sufficiently large" text.[5] Certainly,the simple arithmetic mean would be *c*.747 words per section, but clearly Figure 2 highlights that the distribution is not uniform in nature. One realizes that comparing it in

[5] See e.g. Popescu et al. (2009, p. 3) on purported text lengths; for definitions of "text" in modern linguistics, refer to Yesypenko (2008, p. 18).

favorable terms with other popular works of H. Beam Piper (1962, 1963, 1965) adds up to a tenuous practice. Creative literary samples are (apparently) not written by smart automata with a quasi-perfect rhythm and disposition; they are not invariant and are characterized by an inherent lack of homogeneity (cf. Bell, Berridge, & Rayson, 2009; Strauß, Grzybek, & Altmann, 2007). The point does not suggest that intra-authorial analyses should not be made one day, rather than given the topic and the lexical specificity of *Omnilingual*, observations and results cannot be rid of (some) arbitrariness. A similar argument can be raised with regard to fiction works of other authors. If we hypothetically consider weighing the Mars-themed *Omnilingual* against another storybook, e.g. the *New Grub Street* of George Gissing (2016 [2008, 1891]), the discrepancy is pretty obvious: the reliance on sample size may cause bias-related conclusions (i.e. the sampling variation problem). Specifically, Mr. Gissing crafted a three-volume novel of *c*.220.000 tokens. Another concern is that the plot revolves around two contrasting characters in the late 19th century of London's literary world, and their hardships, attitudes and ethical (or not) choices regarding professional and social life. In this sense, attributions to these unlike spheres of action and location (Earth's London *vs*. Mars, plus the involved technicalities), seem to build more corpus-based gaps than bridges. In sheer size, paralleling or overlaying statistics between H. Beam Piper's and George Gissing's may lead to some implausible claims and encroachment of textual realities. Similar remarks are found in the thesis of Jack W. Grieve (2005, p. 21) when reviewing measures of vocabulary richness, «…*the vocabulary of a text depends far more on its subject than on its author*. *For while every word in a text must be drawn from its author*'*s vocabulary, different subjects will activate different sections of an author*'*s vocabulary, and different sections of any author*'*s vocabulary will not all be equally rich*…» The condition for a comparison can theoretically be "satisfied" if the range of lexis is near, or comparatively near Piper's, with the inter-authorial tests hammered out along some "vintage" sci-fi text. At first, the chosen language is English, though it would be both interesting and advantageous to explore texts in other languages too. The suggested experiment/s may be carried out in a future study. If relative and arguably (un)certain results emerge, they compel us to keep searching for ways and indicators that bear more validity.

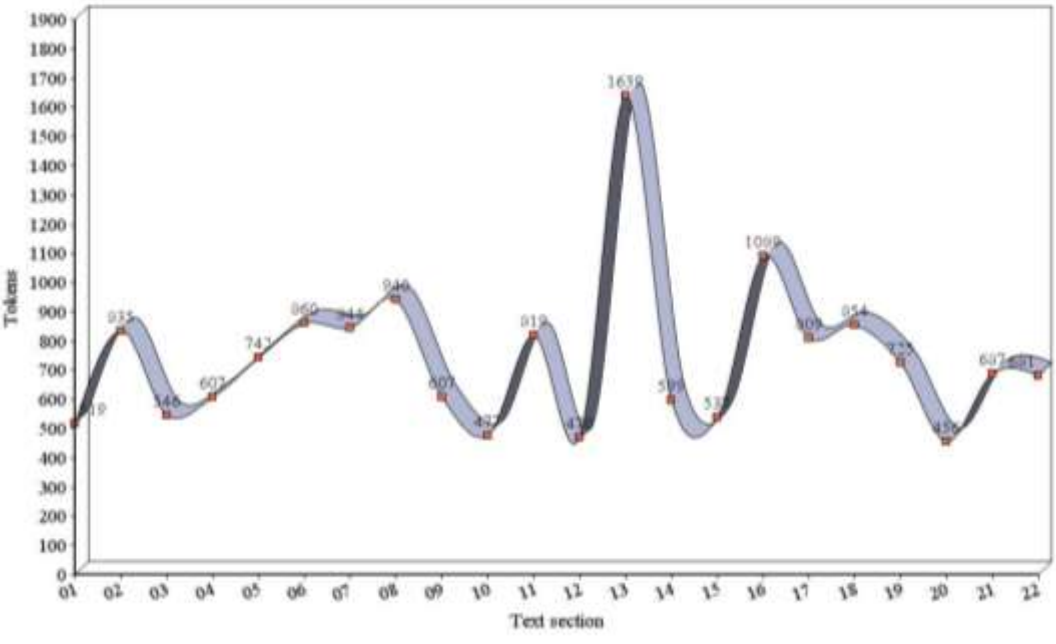


Figure 2: The strip-line conveys the number of tokens per individual section in HBP's *Omnilingual* (1957). The *N*-size shows various steep curves/slopes, especially through sections 10 to 20 [spiking at # 13], similar to a "rough ride" people experience in a roller coaster.

The raw data of *Omnilingual* (1957) are retrieved from the public domain webpage http://www.gutenberg.org/files/19445/19445-h/19445-h.htm. Once formatted in a Word.doc file and its 22 sections arranged in separate files, the language data are processed as per online statistical programs with know-how in such matters. The following programs appear to be resilient and suitable for the proposed tasks:

1- Frequency word-counter at http://www.writewords.org.uk/word_count.asp

2- Nonlinear Regression and Curve Fitting (Sherrod, 2018) at http://www.nlreg.com

3.1 Properties of the vocabulary/lexis

Vocabulary richness in an organized text can be measured: (a) directly, by naming the individual types – not tokens – because the number of tokens is automatically greater in synthetic languages. The types (i.e. distinct words) can be ordered according to some principle, e.g. ranked according to their frequency, their length, etc. (see e.g. Baayen, 2001; COCA, 2018; Herdan, 1964; Johnson, 2008; Köhler & Galle, 1993; Leech, Rayson, & Wilson, 2001; Malvern, Richards, Chipere, & and Durán, 2004; Sichel, 1986; Strauß, Grzybek, & Altmann, 2007). (b) Indirectly, by performing some classifications of the types and setting up new distributions: they can be classified according to the parts-of-speech (PoS) to which they belong, according to the role they play in the sentences. For instance, the case of *function words* vs. *content words* could be of some merit in more discussion of the complete set of *c*.16.430 tokens in terms of the varieties of lexical types. On the other hand, some special classes, e.g. adnominals, verb valencies… should be separated (cf. Fortis & Fagard, 2010; Herbst, Heath, Roe, & Götz, 2004). (c) By means of indicators which can be established either directly[6] or from the existing approaches (a) and (b). The situation in text analysis becomes continuously more complex. Some of the indicators are easy to interpret, other ones entail more omplexity. For example, the smaller is the *Repeat Rate* (*RR*) of words, the richer is the text; in contrast, the greater is the *entropy* of word types, the richer is the text (cf. Cover & Thomas, 2006 [1991]; Jurgens, 2016; Popescu, Čech, & Altmann, 2011). Many indicators are not that evident, for example, Gini's coefficient representing the plane between the Lorenz curve made up from cumulative relative frequencies and the linear function joining <0, 1> (Gini, 1921): the smaller is Gini's coefficient, the greater is the vocabulary richness, etc. (see e.g. Popescu & Altmann, 2006).

Evidently, the examples of indicators earlier mentioned can be used for other purposes too, but in that case a different interpretation takes precedence (cf. Gastwirth, 2017). It is re-emphasized that the aim in this paper is to study the vocabulary of H. Beam Piper (1957), and along this line we attempt to show some of its properties. As expected, the frequencies of distinct words in individual sections are counted, evaluated, and, in addition, the development of the text is studied. As stated earlier, H. Beam Piper's (1957) novelette has twenty-two sections/chapters, and for the sake of checking and further examination they are all brought in Table 1. Hereafter, only the frequency spectrum of tokens is examined because in English being an analytic language – the number of forms is not excessively large. The rows of the table, fashioned in a decreasing order, are to be read as follows:[7] in the first section, there are

[6] Examples include Brunet's W (1978); Orlov's Z (in Orlov & Chitashvili, 1983); Simpson's D index (1949) / Yule's characteristic K (2014 [1944]), or Entropy, as a measure of uniformity (Cover & Thomas, 2006 [1991]). Yet, the debate among experts over their real discriminative power is hardly complete.

[7] See e.g. Popescu et al. (2009, p. 10, Fig. 2.2).

206 tokens having the frequency <1>; 44 tokens having frequency <2>; 9 tokens with frequency <3>, etc.

Table 1
Word token frequencies in individual chapters of *Omnilingual* by H. Beam Piper

Section	**Frequencies of word tokens**
1	1 2 3 4 5 6 7 8 9 14 19 25 51 206 44 9 6 3 3 2 1 1 1 1 1 1
2	1 2 3 4 5 6 7 8 9 14 15 20 22 26 27 42 283 49 27 13 4 5 6 1 1 1 1 2 1 2 1 1
3	1 2 3 4 5 6 8 9 10 15 18 190 40 17 6 8 1 2 5 1 1 1
4	1 2 3 4 5 7 8 10 11 13 14 19 24 25 29 198 37 15 8 8 2 2 1 1 1 2 1 1 1 2
5	1 2 3 4 5 6 7 8 9 11 12 13 16 17 18 21 30 37 262 48 20 14 6 1 2 2 3 1 1 1 1 1 1 1 1 1
6	1 2 3 4 5 6 7 8 9 11 12 13 15 18 27 30 58 250 52 25 15 13 5 3 1 3 2 2 2 1 1 1 1 1
7	1 2 3 4 5 6 7 8 9 11 12 20 21 23 24 28 36 280 69 23 12 9 4 4 1 2 2 1 1 1 1 1 1 1
8	1 2 3 4 5 6 7 8 9 10 11 15 17 21 29 30 49 295 56 30 18 9 6 2 4 1 2 2 2 2 1 1 1 1
9	1 2 3 4 5 6 7 8 9 10 11 15 18 24 52 241 41 16 6 4 4 3 1 1 1 1 1 1 1 1
10	1 2 3 4 5 6 8 9 13 14 15 33 196 41 10 9 3 3 2 1 1 1 1 1
11	1 2 3 4 5 6 7 8 10 11 13 21 23 37 53 317 53 15 7 8 6 5 1 2 1 3 1 1 1 1
12	1 2 3 4 5 6 7 8 9 15 16 34 193 26 20 5 4 1 3 1 1 1 2 1

13	1 2 3 4 5 6 7 8 9 10 12 13 16 17 18 20 23 25 26 28
	409 99 44 19 13 6 11 3 7 1 3 1 3 1 2 2 1 1 1 1
	35 40 50 116
	1 1 1 1

14	1 2 3 4 5 7 14 16 24 33 63 231 45 13 8 4 3 1 2 1 1 1
15	1 2 3 4 5 6 7 8 13 14 25 32 222 44 10 4 4 1 2 1 1 3 1 1
16	1 2 3 4 5 6 7 8 9 10 11 12 13 15 18 20 21 25 27 31 36 304 77 36 14 15 7 5 2 3 2 2 2 1 1 1 1 1 1 1 1 1
17	1 2 3 4 5 6 7 8 9 12 13 15 16 28 31 35 56 288 63 17 7 6 4 2 3 2 1 1 1 1 1 1 1 1
18	1 2 3 4 5 6 7 8 9 10 11 14 16 17 18 20 23 36 65 297 62 18 6 10 3 3 1 2 2 1 1 1 1 1 1 1 1 1
19	1 2 3 4 5 6 7 8 9 10 12 17 18 22 54 228 55 27 16 5 1 3 1 1 1 1 1 2 2 1
20	1 2 3 4 5 6 8 9 10 11 13 32 147 27 19 9 5 3 2 3 2 1 1 1
21	1 2 3 4 5 6 7 8 9 12 13 32 39

	183 53 26 10 16 2 3 2 4 2 2 1 1
22	1 2 3 4 5 6 7 8 9 10 11 13 15 16 20 24 28 198 56 23 7 8 3 4 2 1 1 2 1 2 1 1 1 1

The token frequencies (spectrum) can be fitted by very simple functions. Here, the exponenttial function with added 1 is chosen being the solution of the differential equation,

$$\frac{y'}{y-1} = b \qquad (1)$$

in which the parameter b designates a constant relative change of frequencies. The result is given as

$$\mathrm{y} = 1 + a * \exp{(b * x)} \qquad (2)$$

Specifically, y' = dy/dx, and the formula is understood as a constant change of frequencies depending on the previous frequency (y – 1). The parameter *a* can be considered as a language constant.

Other functions are certainly open to the analysis (e.g. the transformation of the Zipf ranking function), but we abide by simplicity for now. Table 2 shows the results of fitting. the goodness-of-fit is given by the determination coefficient R^2. Usually, one tries to analyze the sequence through a probability distribution; yet, mathematical models are attributable to one's concept formation while aimed at determining the vocabulary richness, not to the ultimate real ity.

Table 2
Fitting the exponential function to the number of tokens in individual sections

Sec.	**Frequencies**	**Results**
1	1 2 3 4 5 6 7 8 9 14 19 25 51 205.98 44.05 10.04 2.90 1.40 1.08 1.02 1.0 1.0 1.0 1.0 1.0 1.0	*a* = 975.9063 *b* = 1.5605 R^2 = 0.9995

2	1 2 3 4 5 6 7 8 9 14 15 20 22 26
	282.25 56.21 11.84 3.13 1.42 1.08 1.01 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	27 42
	1.0 1.0
		a = 1432.7287
		b = 1.6281
		R^2 = 0.9942

3	1 2 3 4 5 6 8 9 10 15 18 189.56 43.61 10.63 3.18 1.49 1.11 1.01 1.00 1.00 1.00 1.00	*a* = 834.3604 *b* = 1.4873 R^2 = 0.9961

4	1 2 3 4 5 6 7 8 10 11 13 14 15 24 25
	197.64 40.32 8.86 2.57 1.31 1.01 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	29
	1.0
		a = 983.3113
		b = 1.6096
		R^2 = 0.9964

5	1 2 3 4 5 6 7 8 9 11 12 13 16 17
	261.46 52.90 11.34 3.06 1.41 1.08 1.02 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	18 21 30 37
	1.0 1.0 1.0 1.0
		a =
		1307.11.26
		b = 1.6131
		R^2 = 0.9962

	1 2 3 4 5 6 7 8 9 11 12 13 15 18	*a* = 1056.6261

6	249.04 59,28 14.69 4.22 1.76 1.18 1.04 1.01 1.0 1.0 1.0 1.0 1.0 1.0 27 30 58 1.0 1.0 1.0	*b* = 1.4483 R^2 = 0.9926

7	1 2 3 4 5 6 7 8 9 11 12 20 21 23
	279.44 72.62 19.42 5.74 2.22 1.31 1.08 1.02 1.0 1.0 1.0 1.0 1.0 1.0
	24 28 36
	1.0 1.0 1.0
		a = 1082.4863
		b = 1.3578
		R^2 = 0.9983

8	1 2 3 4 5 6 7 8 9 10 11 15 17 21
	293.93 64.93 14.95 4.04 1.66 1.14 1.03 1.01 1.0 1.0 1.0 1.0 1.0 1.0
	29 30 49
	1.0 1.0 1.0
		a = 1342.2959
		b = 1.5222
		R^2 = 0.9926

9	1 2 3 4 5 6 7 8 9 10 11 15 18 24 52 240.71 44.04 8.73 2.39 1.25 1.04 1.01 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0	*a* = 1335.0256 *b* = 1.7173 R^2 = 0.9982
10	1 2 3 4 5 6 8 9 13 14 15 33 195.86 42.05 9.64 2.82 1.38 1.08 1.0 1.0 1.0 1.0 1.0 1.0	*a* = 924.9248 *b* = 1.5575 R^2 = 0.9986

11	1 2 3 4 5 6 7 8 10 11 13 21 23 37
	316.80 55.17 10.29 2.59 1.27 1.05 1.01 1.0 1.0 1.0 1.0 1.0 1.0 1.0
	53
	1.0
		a = 1841.1649
		b = 1.7631
		R^2 = 0.9985

12	1 2 3 4 5 6 7 8 9 15 16 34 192.62 30.75 5.62 1.72 1.11 1.02 1.0 1.0 1.0 1.0 1.0 1.0	*a* = 1234.4029 *b* = 1.8628 R^2 = 0.9923

13	1 2 3 4 5 6 7 8 9 10 12 13 16
	407.41 109.62 30.03 8.76 3.07 1.55 1.15 1.04 1.01 1.0 1.0 1.0 1.0
	17 18 20 23 25 26 28 35 40 50 53 116
	1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
		a = 1520.6077
		b = 1.3195
		R^2 = 0.9959

14	1 2 3 4 5 7 14 16 24 33 63 230.79 46.84 10.15 2.82 1.36 1.01 1.0 1.0 1.0 1.0 1.0	*a* = 1151.8132 *b* = 1.6119 R^2 = 0.9989
15	1 2 3 4 5 6 7 8 13 14 25 32 221.95 44.42 9.53 2.68 1.33 1.06 1.01 1.0 1.0 1.0 1.0 1.0	*a* = 1124.2280 *b* = 1.6269 R^2 = 0.9993

16	1 2 3 4 5 6 7 8 9 10 11 12 13
	302.56 86.01 24.97 7.76 2.90 1.54 1.15 1.04 1.01 1.0 1.0 1.0 1.0
	15 18 20 21 25 27 31 36
	1.0 1.0 1.0 1.0 1.0 1.0 1.0 1.0
		a = 1069.7236
		b = 1.2662
		R^2 = 0.9950

17	1 2 3 4 5 6 7 8 9 12 13 15 16 28
	287.80 64.51 15.06 4.11 1.69 1.15 1.03 1.01 1.0 1.0 1.0 1.0 1.0 1.0
	31 35 56
	1.0 1.0 1.0
		a = 1295.2036
		b = 1.5076
		R^2 = 0.9994

18	1 2 3 4 5 6 7 8 9 10 11 14 16 17
	296.74 64.14 14.48 3.88 1.61 1.13 1.03 1.01 1.0 1.0 1.0 1.0 1.0 1.0
	18 20 23 36 65
	1.0 1.0 1.0 1.0 1.0
		a = 1385.1425
		b = 1.5441
		R^2 = 0.9988

19	1 2 3 4 5 6 7 8 9 10 12 17
	226.85 62.58 17.79 5.58 2.25 1.34 1.09 1.03 1.01 1.0 1.0 1.0
	18 22 54
	1.0 1.0 1.0
		a = 828.3191
		b = 1.2995
		R^2 = 0.9945

	1 2 3 4 5 6 8 9 10 11 13 32	*a* = 660.9737

20	146.30 32.94 8.02 2.54 1.34 1.07 1.0 1.0 1.0 1.0 1.0 1.0	*b* = 1.5149 R^2 = 0.9882
21	1 2 3 4 5 6 7 8 9 12 13 32 39 181.73 59.62 20.02 7.17 3.00 1.64 1.21 1.01 1.0 1.0 1.0 1.0 1.0	*a* = 557.1287 *b* = 1.1258 R^2 = 0.9910

22	1 2 3 4 5 6 7 8 9 10 11 13 16 20
	197.40 59.59 18.48 6.21 2.56 1.46 1.38 1.04 1.01 1.0 1.0 1.0 1.0 1.0
	24 28
	1.0 1.0
		a = 658.3425
		b = 1.2096
		R^2 = 0.9979

The determination coefficients R^2 are in all cases greater than 0.9, i.e. the fitting is very satisfactory.

The comparison of individual sections in spectral form does not lead to safe results because many classes have frequency 1, while several others have frequency zero hence a voluminous pooling of classes would be necessary.

The *ranking* may be fitted by means of the usual power function (Zipf-function, zeta-function). 1 is added to it because the most values are ones and the outcome would yield many values smaller than 1. The results for the first section are collected in Table 3. The differential equation is interpreted as the relative change of the frequency, being in direct proportion to the previous class, i.e. $y'/(y – 1) = b/x$. 1 is added anew to make the fitting better adapted. In the end, rather than distributions, functions are chosen for frequency ranking.

Table 3
Ranking the frequencies in Section 1 and fitting by the zeta function ($y = 1 + ax^b$); cf. Popescu et al. (2009, p. 191)

Rank	**Fr_x**	**f_x**	**Rank**	**Fr_x**	**f_x**	**Rank**	**Fr_x**	**f_x**
1	51	51.41	94	1	1.44	187	1	1.21
2	25	25.42	95	1	1.43	188	1	1.21
3	19	16.99	96	1	1.43	189	1	1.21
4	14	12.83	97	1	1.42	190	1	1.21
5	9	10.37	98	1	1.42	191	1	1.21
6	8	8.74	99	1	1.41	192	1	1.21
7	7	7.59	100	1	1.41	193	1	1.21
8	7	6.73	101	1	1.40	194	1	1.20
9	6	6.07	102	1	1.40	195	1	1.20
10	6	5.54	103	1	1.40	196	1	1.20
11	6	5.11	104	1	1.39	197	1	1.20
12	5	4.75	105	1	1.39	198	1	1.20
13	5	4.45	106	1	1.38	199	1	1.20
14	5	4.19	107	1	1.38	200	1	1.20
15	4	3.97	108	1	1.38	201	1	1.20
16	4	3.78	109	1	1.37	202	1	1.20
17	4	3.61	110	1	1.37	203	1	1.20
18	4	3.46	111	1	1.37	204	1	1.19
19	4	3.32	112	1	1.36	205	1	1.19
20	4	3.20	113	1	1.36	206	1	1.19
21	3	3.09	114	1	1.36	207	1	1.19
22	3	2.99	115	1	1.35	208	1	1.19

23	3	2.90	116	1	1.35	209	1	1.19
24	3	2.82	117	1	1.35	210	1	1.19
25	3	2.74	118	1	1.34	211	1	1.19
26	3	2.67	119	1	1.34	212	1	1.19
27	3	2.61	120	1	1.34	213	1	1.19
28	3	2.55	121	1	1.34	214	1	1.18
29	3	2.49	122	1	1.33	215	1	1.18
30	2	2.44	123	1	1.33	216	1	1.18
31	2	2.39	124	1	1.33	217	1	1.18
32	2	2.35	125	1	1.32	218	1	1.18
33	2	2.30	126	1	1.32	219	1	1.18
34	2	2.26	127	1	1.32	220	1	1.18
35	2	2.23	128	1	1.32	221	1	1.18
36	2	2.19	129	1	1.31	222	1	1.18
37	2	2.16	130	1	1.31	223	1	1.18
38	2	2.12	131	1	1.31	224	1	1.18
39	2	2.09	132	1	1.31	225	1	1.18
40	2	2.07	133	1	1.30	226	1	1.17
41	2	2.04	134	1	1.30	227	1	1.17
42	2	2.01	135	1	1.30	228	1	1.17
43	2	1.99	136	1	1.30	229	1	1.17
44	2	1.96	137	1	1.29	230	1	1.17
45	2	1.94	138	1	1.29	231	1	1.17
46	2	1.92	139	1	1.29	232	1	1.17
47	2	1.90	140	1	1.29	233	1	1.17
48	2	1.88	141	1	1.29	234	1	1.17
49	2	1.86	142	1	1.28	235	1	1.17
50	2	1.84	143	1	1.28	236	1	1.17
51	2	1.83	144	1	1.28	237	1	1.17
52	2	1.81	145	1	1.28	238	1	1.17
53	2	1.79	146	1	1.28	239	1	1.16
54	2	1.78	147	1	1.27	240	1	1.16
55	2	1.76	148	1	1.27	241	1	1.16
56	2	1.75	149	1	1.27	242	1	1.16
57	2	1.74	150	1	1.27	243	1	1.16
58	2	1.72	151	1	1.27	244	1	1.16
59	2	1.71	152	1	1.26	245	1	1.16
60	2	1.70	153	1	1.26	246	1	1.16
61	2	1.69	154	1	1.26	247	1	1.16
62	2	1.67	155	1	1.26	248	1	1.16
63	2	1.66	156	1	1.26	249	1	1.16
64	2	1.65	157	1	1.26	250	1	1.16
65	2	1.64	158	1	1.25	251	1	1.16
66	2	1.63	159	1	1.25	252	1	1.16
67	2	1.62	160	1	1.25	253	1	1.15
68	2	1.61	161	1	1.25	254	1	1.15
69	2	1.60	162	1	1.25	255	1	1.15
70	2	1.59	163	1	1.25	256	1	1.15
71	2	1.58	164	1	1.24	257	1	1.15
72	2	1.58	165	1	1.24	258	1	1.15
73	2	1.57	166	1	1.24	259	1	1.15

74	1	1.56	167	1	1.24	260	1	1.15
75	1	1.55	168	1	1.24	261	1	1.15
76	1	1.54	169	1	1.24	262	1	1.15
77	1	1.54	170	1	1.23	263	1	1.15
78	1	1.53	171	1	1.23	264	1	1.15
79	1	1.52	172	1	1.23	265	1	1.15
80	1	1.52	173	1	1.23	266	1	1.15
81	1	1.51	174	1	1.23	267	1	1.15
82	1	1.50	175	1	1.23	268	1	1.15
83	1	1.50	176	1	1.23	269	1	1.15
84	1	1.49	177	1	1.23	270	1	1.14
85	1	1.48	178	1	1.22	271	1	1.14
86	1	1.48	179	1	1.22	272	1	1.14
87	1	1.47	180	1	1.22	273	1	1.14
88	1	1.47	181	1	1.22	274	1	1.14
89	1	1.46	182	1	1.22	275	1	1.14
90	1	1.46	183	1	1.22	276	1	1.14
91	1	1.45	184	1	1.22	277	1	1.14
92	1	1.45	185	1	1.21	278	1	1.14
93	1	1.44	186	1	1.21	279	1	1.14
$\boldsymbol{a} = 50.4140$, $\boldsymbol{b} = -1.0455$, $\mathbf{R}^2 = 0.9914$								

For the other sections only the parameters and the determination coefficients are presented, see Table 4.

Table 4
Fitting the zeta-function to tokens in individual sections in *Omnilingual* (1957)

Section	***a***	***b***	**R^2**	**Section**	***a***	***b***	**R^2**
1	50.4140	-1.0455	0.9914	12	33.6966	-0.9008	0.9615
2	48.2709	-0.7844	0.9030	13	115.4085	-0.8560	0.9785
3	32.8355	-0.8004	0.9385	14	63.6117	-1.0540	0.9878
4	38.7999	-0.7699	0.8542	15	35.1885	-0.8572	0.9397
5	42.7064	-0.7931	0.9334	16	48.9021	-0.7045	0.8990
6	57.6461	-0.8466	0.9753	17	60.2801	-0.8848	0.9607
7	43.0180	-0.7652	0.9097	18	65.4717	-0.9073	0.9819
8	52.6031	-0.7927	0.9569	19	52.2918	-0.8906	0.9655
9	50.4911	-0.9794	0.9868	20	29.8525	-0.8262	0.9336
10	32.1040	-0.9012	0.9681	21	40.9215	-0.7981	0.9429
11	56.2035	-0.8812	0.9712	22	34.2479	-0.7290	0.9015

In light of the data, it can be said that the parameter *a* depends on the frequency of the first value, whereas the parameter *b* shows the strength of the decrease; nevertheless, the parameters are not codependent. The second parameter expresses the impact of the law. The results of the determination coefficients are in all cases greater than 0.85, and in the majority of cases greater than 0.9.

One can study the properties of the individual sections using the rank-frequency distributions. A number of studies have been produced on this matter (*v. supra*). Here we shalldeal with the question of richness by the repetition of words. George U. Yule's (2014 [1944])
"*characteristic K*" basically indicates through inversion that the richer is the text, the smalleris the repetition of words. Hence, the *Repeat Rate* (RR) is at present a sufficient indicator of richness. The formula is

$$RR = \frac{1}{\mathrm{N}^2}\sum_{x=1}^{\mathrm{V}} \mathrm{f}_{\mathrm{x}}^2 \qquad (3)$$

where N is the sum of frequencies, V is the number of different words, and f^2 are the squares of individual frequencies (cf. Popescu et al., 2009, p. 166). The formula can be relativized, transformed in entropy, or chi-squared (cf. Altmann & Köhler, 2015, p. 38), etc. In the richest lexically text all frequencies are 1, obtaining RR_{min} = 1/N thereof. If the whole text was concentrated in one uniformly replicated word, we would acquire RR_{max} = 1. Hence, a relative *Repeat Rate* yielding

$$RR_{rel} = \frac{1 - \mathrm{RR}}{1 - 1/\mathrm{N}} \qquad (4)$$

is set up. With the *Repeat Rate* (RR) values computed for the individual sections, the results are listed in Table 5.

Table 5
Repeat Rates (RR) of individual sections (tokens)

Section	RR	Section	RR	Section	RR	Section	RR
1	0.0178	8	0.0091	15	0.0118	19	0.0123
2	0.0103	9	0.0142	16	0.0070	20	0.0129
3	0.0113	10	0.0117	17	0.0129	21	0.0105
4	0.0139	11	0.0111	18	0.0128	22	0.0096
5	0.0100	12	0.0130				
6	0.0112	13	0.0112				
7	0.0086	14	0.0202				

The sequence of RR is not monotonic, with the steep jump revealingly shown in section 14. Evidently, the boundary conditions – in this case, the description of some "new" view – lead HBP to write section 14 in a slightly different manner.

3.2 Gini's coefficient

One of the many possibilities to account for the richness of text is Gini's coefficient (see Popescu et al., 2009, pp. 54-63). It is the space between the Lorenz curve and the straight line joining <0, 1> in the two-dimensional coordinate system. The Lorenz curve is the stepwise adding of relative frequencies beginning from the lowest up to the highest (Popescu et al., 2009, p. 56, Fig. 3.11). Since this constitutes an area, one is bound to sort out all individual areas between the two lines. Regardless of the fact, there are easily computable approximations at our disposal. One of them is given as

$$G = \frac{1}{V}\left(V + 1 - \frac{2}{N}\sum_{x=1}^{V} xf(x)\right) \quad (5),$$

to be simply rendered as 1 +1/V- 2* μ/V, where μ is the *mean of the frequencies*. For comparative purposes, one can use the variance of *G* consistent with

$$Var(G) = \frac{4\sigma^2}{V^2 N} \quad (6)$$

where σ^2 is the variance of the rank frequencies.[8] The values for individual sections are listed in Table 6.

Table 6
Gini's coefficient of individual sections

Section	*Gini*	Section	*Gini*	Section	*Gini*	Section	*Gini*
1	0.4117	8	0.4612	15	0.3972	19	0.4438
2	0.4583	9	0.4171	16	0.4668	20	0.4279
3	0.4255	10	0.3790	17	0.4498	21	0.4485
4	0.4719	11	0.4343	18	0.4588	22	0.4488
5	0.4409	12	0.3961				
6	0.4746	13	0.5345				
7	0.4366	14	0.4361				

The sequence of Gini's coefficients could be captured by a straight line, but section 13 involves a climactic value, plus the variation among the other chapters is clearly perceptible. This can mean, for example, that the individual values depend on some preliminarily not known property. Here, the data analysis begins.

4. Discussion

Most assuredly, interpretation of data is as good as the performed statistical measures, plus the authenticity and size of sample. Whilst not offering pat solutions, statistics can be symptomatic of underlying patterns. I am prepared to concede that HBP did not plan in advance the length of the whole story or that of each chapter (cf. Popescu et al., 2009, p. 70).

[8] The applied formulas on Gini's coefficient are based on Popescu et al. (2009, p. 57).

He was an ingenious and spontaneous writer and not a post-WWII journalist forced to write articles according to fixed instructions and pay fees, or space-constrained norms.

In lumping together the quantitative evidence, Gini's coefficient appears equally supportive of a qualitative approach of the text, i.e. close reading. As already noted, the smaller is Gini's coefficient, the greater is the vocabulary richness, which is expressly revealed in the sections 10, 12, and 15, with values below the 0.4-threshold. The quoted sections show a number a features bolstering that property: carefully described scenarios strewn with techno-parlance, quite often falling next to an "enumerative" style; sparingly used dialogues (mostly bearing the mark of a silent monologue); and brevity in terms of tokens.The substantial use of technologically / scientifically-related words in line with the size of text-section offsets the lexical dearth. On the other side, section 13 (the longest in the novelette) shows the highest jump in Gini's coefficient: 0.5345. The relative diminishing of richness in vocabulary could hint at more embedded dialogues, where colloquial / "normal" speech may "taint" to a degree the pool of scientific hapax legomena or dislegomena. The other sections do not exhibit the change noticed in section 13: fluctuations are *strong*, between 0.41- and 0.47-, but not that *striking* (see Table 6). The data collectively suggest that H. Beam Piper (1957) might have taken a respite[9] before writing the section in question. The outcome is a more "relaxed" and "protracted" text, or words to that effect. Nonetheless, generalizing on the basis of a single piece should be cautiously avoided as it may stand only for a subset or a frame of writer's linguistic skills. Overall, the stated indicator does not suggest a poor acquisition or management of English vocabulary. This can mean that H. Beam Piper was an avid consumer of historical / scientific material about archaeological decipherment, biochemistry, interplanetary travel & exploration, and gadget engineering. The following standard and non-standard words are spread across the chapters (neologisms, rare portmanteaus or not), and they come in different flavors: <*spraygun*>, <*Airdyne*>, <*airsealing*>, <*viviparous*>, <*gamogenetic*>, <*Photostat*>, <*stenophone*>, <*loess*>, <*oxyacetylene*>, <*spectroscope*>, <*vibratool*>, <*tarpaulins*>, <*radiophone*>, <*jetticopters*>, <*transuranics*>, <*beryllium*>, <*boron*>. One positive implication is that this range of words shows the diverse intellectual concerns and the inventive strain of the author. The observation finds justification in J. F. Carr (2008), with Piper's up-to-date information accomplished by dint of relatable literature or participation in sci-fi conventions.

The *Repeat Rate* data in Table 5 shows that the flow of narrative – in terms of vocabulary wealth – is unmarked by dull or relatively dull uniformity. The fact fits well with the non-homogenous nature of text samples (cf. Bell, Berridge, & Rayson 2009, p. 3) and may have to do with the time axis through which *Omnilingual* (1957) was written. It may be theorized that fluctuating values do not only act in response to the required situations/subplots along the sections, but also to the prevailing emotional mood of the author himself. An interesting observations relates to section 14, where the *Repeat Rate* (RR) value is doubled or nearly doubled in comparison e.g. with sections 2, 5, 8, 21, 22, pointing at less vocabulary richness. In opposition, Gini's coefficient registers 0.4361 for # 14, whereas the values for # 2, 5, 8, 21, 22, swing within the range 0.4409-0.4612.

The doubling in *Repeat Rate* occurs in section 13 – the longest in the story and the one with several instances of up-close dialogues. Section 14, in turn, is a dry and technical description of part of the Martian University and the measures taken by the deployed international team for camping and a better examination of it. Here, the writing, besides being "underprivileged" in number of tokens, lacks dialogues and is loaded with past tenses and passive constructions.

[9] The assumed break could have responded to any physical or personal recreational activity: sleeping, sipping coffee / drinking rum, smoking a cigarette, hiking for a non-determined period of time, hunting, and so forth. Considering the 1950-ies in rural Pennsylvania (USA), each of them does not seem a mis-entry rather than a proper "manly activity" for H. Beam Piper.

Whether consulting Gini's coefficient for sections # 12: 0.3961, # 13: 0.5345, # 14: 0.4361, or the *Repeat Rate* (RR), # 12: 0.0130, # 13: 0.0112, # 14: 0.0202, the figures reveal certain conspicuous and "anomalous" behavior nearby Section 13. Although there is divergence in the way these indicators perform: Gini's shows lexical richness for # 14, while counter-posed by the *Repeat Rate* showing decrease in richness, it may be stated with some confidence that section 13 (or the circumstances that led to its conception) act/s as a breaking point in the lexical set up. Again, I would tend to reconcile the observations with pauses breaks that the author took in the interim. Such pauses might have conditioned a slightly different creative rhythm in HBP's mind, or affecting him psychologically, with the result of a discrepant use of vocabulary.

A pair of observations that go beyond the tabulated data follow. First, if present-day readers have one "quibble" with Piper's story, that may regard the words «*smoking*» on more than one occasion and having libations on planet Mars by way of «*cocktail pitchers*» and «*Martinis*». Admittedly, these conspicuous nouns are far from accidental: they served to "cheer up" the atmosphere and may be adduced to the private baggage of the author and the time in which he lived. Qualitatively speaking, such lexical choices function as "shibboleths" (a peculiarity of speech / writing; cf. Juola, 2008, pp. 237-238), by which specific stylistic or broad social inferences can be made. Second, in section 16, the author puts into the mouth of Anthony Lattimer who was arguing with Martha Dane and Selim von Ohlmhorst about the decipherment of the Hittite language, that it was done "*when they found Hittite-Assyrian* bilinguals." This moment rings false as the decipherment was confirmed through the *Hittite* [*Luwian*]-*Phoenician bilinguals* of Karatepe hill, dated from the late part of 8th century BCE (see Friedrich, 1971 [1957], pp. 98-101). It so happens that in the selfsame section the referenced German epigrapher is under the appellation "… *that distinguished Hittitologist, Johannes Friedrich*."

5. Conclusions

With *Omnilingual*'s (H. Beam Piper, 1957) authorship proven, this study, rather than seeking attribution, attempts to make a corpus-based statistical inspection in order to extract style-related features. The feature of relevance of the written text – *vocabulary richness* – may shed light on characteristic traits of Piper's style and/or socio-psychological background. While it is agreed that «*In practice no single indicator can measure style in its entirety*»[10] (or *two indicators,* for that matter), this is a "first step" in the stylistic analysis of H. Beam Piper's work. One has to eventually acknowledge that further efforts and more robust statistical methods may add more objectivity to such analyses.

• Textual linguistic patterns directing us toward extraction of statistical knowledge about style, mean also *vulnerabilities* which are exploited and assist in extracting meta-knowledge, i.e. about the author and his psychological and sociological inclinations (cf. Daelemans, 2013). The degrees of meta-knowledge are still variable and undecided, especially when the subconscious features that shape and drive literary creativity are deliberated. It is explicitly admitted that current univariate statistics that probe vocabulary size or richness are far from capturing the entire complexities of the human brain. Further developments in quantitative methods that accurately correlate with some intuitions, together with cutting-edge break-throughs in neurosciences and AI, will help in gaining a significant advantage on meta-knowledge.

• Both statistical measures, the *Repeat Rate* and *Gini*'*s coefficient*, show that H. Beam Piper (1957) has on the whole an estimable level of vocabulary richness. The observation suggests that notwithstanding his wanting academic training, Piper was an assiduous reader of

[10] See Bell, Berridge, & Rayson (2009, p. 3).

fiction and non-fiction literature. When recombined with his unique talent to develop ideas about off-world events and adventures, the concerned readership is fortunate to have and enjoy his creations.

• The vocabulary richness in various sections of *Omnilingual* appears sensitive to a time axis. The novelette – very likely – was not written at one sitting, rather than over different sessions, some more non-linear than the others. However, fixing the temporal gap (hours, days, weeks …) among sessions, i.e. building a time-structured succession, is far beyond the capability of the applied quantitative measures. All that said, we only can speculate about such perceived distances and the reasons behind them.

• While the quantitative approach is regarded as a potential discriminant for meta-knowledge, I am unwilling to dismiss salient qualitative aspects. Overplaying or downplaying the importance of each approach is not advisable for my part; for all practical purposes, a complementary methodology may be more useful.

• Comparative studies involving *Omnilingual* (1957) and other stories of the author may be scheduled in the future. Their target could be on the level of lexical richness, and further expanded into the syntactic class distributions and patterns. Another possible approach is considering the word class, or the content / function word contrast in search of stylistic and thematic propensities.

Acknowledgements

The multi-author book, Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., et al. (2009). *Word Frequency Studies*. Berlin/New York: de Gruyter, offered valuable information on several quantitative measures.

The online repository sites *Project Gutenberg*, *Archive.org*, and *The Library Service* of Parkland College, Champaign, IL. (USA) have been of assistance with several reference sources.

References

Adams, J. N. (2003). *Bilingualism and the Latin Language*. Cambridge, UK: Cambridge, University Press.

Altmann, G. & Köhler, R. (2015). *Forms and Degrees of Repetitions in Texts*. Berlin: de Gruyter.

Argamon, S., Whitelaw, C., Chase, P., Dhawle, S., Hota, S., Garg, N., and Levitan, S. (2007).Stylistic Text Classification Using Functional Lexical Features. *Journal of the American Society of Information Science, 7,* 91-109.

Arnold, E. L. (1905). *Lieut. Gullivar Jones*: *His Vacation* [later misnamed *Gulliver of Mars*, 1964]. London: S. C. Brown, Langham & Co. Retrieved from https://archive.org/details/ gulliver_of_mars_0905_librivox

Baayen, R. H. (2001). *Word frequency distributions*. Text, Speech and Language Technology, Vol. 18. Series editors, Nancy Ide and Jean Véronis. Dordrecht, Netherlands: Kluwer Academic Publishers.

Babb, H. S. (Ed.). (1972). *Essays in Stylistic Analysis*. New York: Harcourt Brace Jovanovich, Inc.

Bain, A. (1877). *English Composition and Rhetoric*. London: Longmans, Green and Co. Retrieved from https://archive.org/details/englishcompositi01bain

Baker, S. (1966). *The Complete Stylist*. New York: Thomas Y. Crowell Company.

Baum, S. D., Haqq-Misra, J. D., & Domagal-Goldman, S. D. (2011). Would Contact with Extraterrestrials Benefit or Harm Humanity? A Scenario Analysis. *Acta Astronautica*, *68*(11-12): 2114-2129.

Beam Piper, H. (1947). Time and Time Again. *Astounding Science Fiction*, *39*(2), April 1947. Retrieved from http://www.gutenberg.org/ebooks/18831

Beam Piper, H. (1953). *Murder in the Gunroom*. New York: Alfred A. Knopf. Retrieved from https://www.gutenberg.org/files/17866/17866-h/17866-h.htm

Beam Piper, H. (1957). *Omnilingual*. Originally published in *Astounding Science Fiction*, *58* (6), February 1957, pp. 8-46, with cover and interior illustration by Frank Kelly Freas. Retrieved from http://www.gutenberg.org/files/19445/19445-h/19445-h.htm

Beam Piper, H. (1962). *Little Fuzzy*. New York: Avon. Retrieved from http://www.teleread.org/blog/

Beam Piper, H. (1963). *Space Viking*. New York: Ace Books. Retrieved from https://www.gutenberg.org/files/20728/20728-h/20728-h.htm

Beam Piper, H. (1965). *Lord Kalvan of Otherwhen*. New York: Ace Books. Retrieved from https://gutenberg.ca/ebooks/piperhb-lordkalvanofotherwhen/piperhb-lordkalvanofotherwhen-00-h.html

Beam Piper, H. (1981). *Federation*. Preface by Jerry Pournelle. New York: Ace Books.

Bear, G. (1993). *Moving Mars*. New York: Tor Books.

Bell, E. J. L., Berridge, D., & Rayson, P. (2009). *Measuring style with the authorship ratio*: *An invariant metric of lexical similarity*. Retrieved from http://ucrel.lancs.ac.uk/publications/ cl2009/280_FullPaper.pdf

Bleiler, E. F. (1990). *Science-Fiction*: *The Early Years*. Kent, OH: Kent State University Press.

Bogdanov, A. (1984 [1908]). *Red Star*. *Engineer Menni*. *A Martian Stranded on Mars*. (Ch. Rougle, Trans.). Bloomington and Indianapolis: Indiana University Press. Retrieved from https://archive.org/details/BogdanovRedStar

Bradbury, R. (1950). *The Martian Chronicles*. New York: Doubleday.

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux*: *Structure et évolution*; *Statistique et informatique appliquées à l'étude des textes, à partir du Trésor de la langue française*. Paris: Slatkine.

Budrys, A. (1967). Review of Great Science Fiction Stories about Mars. In T. E. Dikty (Ed.), *Galaxy Bookshelf*. *Galaxy Magazine*, *April 1967* (pp. 166-169).

Burroughs, E. R. (1917 [1912]). *A Princess of Mars* [original title, *Under the Moons of Mars*]. Chicago, IL: A. C. McClurg & Co. Retrieved from https://www.gutenberg.org/files/62/62-h/ 62-h.htm

Callimahos, L. D. (1966). Communication with Extraterrestrial Intelligence. Lecture at the "Cosmos Club" in Washington D.C. (1965). *NSA Technical Journal*, *11*(1): 79-86. Retrieved from http://www.nsa.gov/public_info/_files/tech_journals/communications_extraterrestrial_ intelligence.pdf

Campanile, E., Cardona, G. R., & Lazzeroni, R. (Eds.). (1988). *Bilinguismo e Biculturalismo nel Mondo Antico. Atti del Colloquio interdisciplinare tenuto a Pisa il 28 e 29 settembre 1987*. Testi Linguistici 13. Pisa: Giardini Editori e Stampatori.

Carr, J. F. (2008). *H. Beam Piper*: *A Biography*. Series Editors, Palumbo, D. E. & Sullivan III, C. W. Critical Explorations in Science Fiction and Fantasy, 8. Jefferson, NC: McFarland & Company, Inc.

Carr, M. H., & Head, J. W. (2010). Acquisition and History of Water on Mars. In Cabrol, N. A., & Grin, E. A. (Eds.), *Lakes on Mars* (pp. 31-67). Amsterdam: Elsevier Science. Retrieved from http://www.planetary.brown.edu/pdfs/3757.pdf

Chadwick, J. (2000 [1958]). *The Decipherment of Linear B*. Cambridge: The Press Syndicate of the Cambridge University.

Chandler, D. (2007 [2002]). *Semiotics*: *The Basics*. London and New York: Routledge, Taylor & Francis Group.

Clarke, A. C. (1951). *The Sands of Mars*. London: Sidgwick & Johnson.

COCA (2018). *Word frequency data – Corpus of Contemporary American English* (*COCA*). Retrieved from http://www.wordfrequency.info/free.asp?s=y

Cohen, J., & Stewart, I. (2002). *Evolving the Alien*. London: Ebury Press

Coleridge, S. (1914 [1907, 1818]). *Coleridge*'*s Essays* & *Lectures on Shakspeare* & *Some Other Old Poets & Dramatists*. London: J. M. Dent & Sons / New York: E. P. Dutton & Co. Retrieved from

https://ia902303.us.archive.org/0/items/coleridgesessays00cole/coleridges essays00cole.pdf

Cooper, L. (1930 [1907]). *Theories of Style*, *with Especial Reference to Prose Composition*. New York – London: The Macmillan Company.
Retrieved from https://archive.org/details/ theoriesofstylew00coop

Cover, T. M. & Thomas, J. A. (2006 [1991]). *Elements of Information Theory*. New York: John Wiley, & Sons, Inc. Retrieved from http://www.cs-114.org/wp-content/uploads/2015/ 01/Elements_of_Information_Theory_Elements.pdf

Daelemans W. (2013). Explanation in Computational Stylometry. In Gelbukh, A. (Ed.), *Computational Linguistics and Intelligent Text Processing*. *CICLing 2013*. *Lecture Notes in Computer Science*, *7817* (pp. 451-464). Berlin-Heidelberg: Springer. Retrieved from https://www.clips.uantwerpen.be/~walter/papers/2013/d13.pdf

Daniels, P. T., & Bright, W. (Eds.). (1996). *The World*'*s Writing System*. Oxford, NY: Oxford University Press.

Daniels, P. T. (1996). Methods of Decipherment. In Daniels, P. T., & Bright, W. (Eds.), *The World*'*s Writing Systems* (pp. 141-159). New York-Oxford: Oxford University Press.

Daniels, P. T. (2013). Decipherment. In Bagnall, R. S., Erskine, A., Brodersen, K., Champion, C. B., & Huebner, S. R. (Eds.), *The Encyclopedia of Ancient History*. First edition (pp. 1949-1950). Oxford: Blackwell Publishing Ltd.

Davies, P. (1995). *Are We Alone*?: *Philosophical Implications of the Discovery of Extraterrestrial Life*. New York: Basic Books / Harper Collins Publishers.

Dick, P. K. (1964). *Martian Time-Slip*. New York: Ballantine Books / Random House.

Dick, S. J. (1998). *Life On Other Worlds*: *The 20th century Extraterrestrial Debate*. Cambridge, UK: Cambridge University Press.

Dikty, T. E. (Ed.). (1966). *Great Science Fiction Stories about Mars*. Chicago, IL: Fredrick Fell.

Drake, F. & Sobel, D. (1992). *Is Anyone Out There*? *The Scientific Search for Extraterrestrial Intelligence*. New York: Delacorte Press.

Eckert, P., & Rickford J. R. (Eds.). (2001). *Style and Sociolinguistic Variation*. Cambridge, UK: Cambridge University Press.

Elliott, J. (2007). A post-detection decipherment matrix. *Acta Astronautica*, *61*(7-8): 712-715.

Engdahl, S. (Ed.). (2006 [2001]). *Extraterrestrial Life*. Contemporary Issues • Companion. Detroit: Greenhaven Press / An imprint of Thomson Gale.

Fergus, C. (2013). *Beyond Earth*: *Mars Fever*. Penn State News. Pennsylvania State University. May 1, 2013. Retrieved from http://news.psu.edu/story/140745/2003/05/01/ research/beyond-earth

Fortis, J-M., & Fagard, B. (2010). *Space in Language*. *Part IV*: *Adnominals*. *Adnominals: Topological-Functional Adpositions, Spatial Phrases and Spatial Cases*. DGfS-CNRS Summer School on Linguistic Typology. Leipzig, August 15-September 3, 2010. Retrieved from https://www.eva.mpg.de/lingua/conference/2010_summerschool/pdf/course_materials/ Fortis_4.ADNOMINALS.pdf

Freeman, D. C. (1970). *Linguistics and Literary Style*. New York: Holt, Rinehart and Winston, Inc.

Friedrich, J. (1971 [1957]). *Extinct Languages*. (F. Gaynor, Trans.). Westport, Connecticut: Greenwood Press, Publishers.

Frye, N. (1963). *The Well-Tempered Critic*. Bloomington, IN: Indiana University Press.

Fulcher, P. M. (1927). *Foundations of English Style*. New York: F. S. Crofts & Co.

Gastwirth, J. L. (2017). Is the Gini Index of Inequality Overly Sensitive to Changes in the Middle of the Income Distribution? *Statistics and Public Policy*, *4*(1): 1-11.

Gelb, I. J., & Whiting, R. M. (1975). Methods of Decipherment. *Journal of the Royal Asiatic Society of Great Britain and Ireland*, 95-104.

Gini, C (1921). Measurement of Inequality of Incomes. *The Economic Journal*, *31* (121): 124-126. https://www.jstor.org/stable/pdf/2223319.pdf?seq=1#page_scan_tab_contents

Gissing, G. (2016 [2008, 1891]). *New Grub Street*. In Three Volumes. Second Edition. London: Smith, Elder, & Co. Retrieved from https://www.gutenberg.org/files/1709/1709-h/1709-h.htm

Golomb, S. W. (1968 [1961]). Extraterrestrial Linguistics. *Word Ways*, *1*(4/5): 202-205. Retrieved from http://digitalcommons.butler.edu/wordways/vol1/iss4/5

Golomb, S. W. (1963). When is Extra-Terrestrial Life Interesting? *Engineering and Science*, *26*(5): 15-17. Retrieved from http://calteches.library.caltech.edu/2209/1/golomb.pdf

Gombrich, E. H. (1982). *The Image and the Eye*: *Further Studies in the Psychology of Pictorial Representation*. Ithaca, New York. Cornell University Press / Phaidon Books.

Gordon, C. H. (1987 [1968]). *Forgotten Scripts*. New York: Dorset Press / Marboro Books.

Greg, P. (1880). *Across the Zodiac*: *The Story of a Wrecked Record*. London: Trübner & Co. / Ludgate Hill. Retrieved from https://archive.org/details/acrosszodiacstor01greg

Grieve, J. W. (2005). *Quantitative Authorship Attribution*: *A History and an Evaluation of Techniques*. Master’s thesis, Simon Fraser University (Burnaby, BC, Canada), 2005. Retrieved from http://www.summit.sfu.ca/system/files/iritems1/8840/etd1721.pdf

Grzybek, P. (Ed.). (2007). *Contributions to the Science of Text and Language*. Dordrecht, Netherlands: Springer.

Guth, H. P. (1980). *Words and Ideas*: *Handbook for College Writing*. 5th Revised edition. Belmont, CA: Wadsworth Publishing Co.

Irvine, J. T. (2001). “Style” as Distinctiveness: The Culture and Ideology of Linguistic Differentiation. In Eckert, P., & Rickford, J. R. (Eds.), *Style and Sociolinguistic Variation* (pp. 21-44). Cambridge, UK: Cambridge University Press.

Hawkes, T. (2004 [1977]). *Structuralism and Semiotics*. London and New York: Routledge / Taylor & Francis Group.

Hawkins, J. D., & Morpurgo Davies, A. (1978). On the Problems of Karatepe: The Hieroglyphic Text. *Anatolian Studies* (British Institute at Ankara), *28*: 103-119. Retrieved from http://www.ling-phil.ox.ac.uk/files/hawkins-amd_karatepe._ the_ hieroglyphic _text_ anatolian_studies_28_1978.pdf

Hawkins, J. D., & Çambel, H. (1999). *Corpus of hieroglyphic Luwian inscriptions*: *Karatepe-Aslantaş*: *The inscriptions*. Berlin: de Gruyter.

Heidmann, J. (1995). *Extraterrestrial intelligence*. 2nd Ed. Cambridge: Cambridge University Press.

Herbst, Th., Heath, D., Roe, I. F., & Götz, D. (2004). *A Valency Dictionary of English*: *A Corpus Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Berlin: Mouton de Gruyter.

Herdan, G. (1964). *Quantitative Linguistics.* London: Butterworths.

Heinlein, R. (1961). *Stranger in a Strange Land*. New York: G. P. Putnam's Sons

Hines, D. (2002). *H. Beam Piper*. Retrieved from http://www.mib.org/~hradzka/piper/

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing*, *13*: 111-117.

Hřebíček, L., & Altmann, G. (Eds.). (1993). *Quantitative Text Analysis*. Trier: Wissenschaftlicher Verlag Trier.

Johnson, K. (2008). *Quantitative Methods in Linguistics*. Malden, MA - Oxford: Blackwell,

Juola, P. (2008). Author attribution. *Foundations and Trends in Information Retrieval*, *1*(3): 233-334. Retrieved from http://www.mathcs.duq.edu/~juola/papers.d/fnt-aa.pdf

Jurgens, A. (2016). *Entropy in Written English*. Retrieved from http://csc.ucdavis.edu/~chaos/courses/ncaso/Projects2016/Jurgens/paper.pdf

Kane, Th. S., Peters, L. J., Jackel, D., & Legris, M. R. (1981). *Writing Prose*: *Techniques and Purposes*. Oxford: Oxford University Press.

Kerr, R. M. (2010). *Latino-Punische Epigraphik*: *Eine Beschreibung der Inschriften* [Latino-Punic epigraphy: A Descriptive Study of the Inscriptions]. *Forschungen zum Alten Testament 2, Reihe 42*. Tübingen: Mohr Siebeck.

Knight, K., & Sproat, R. (2009). *Writing Systems*, *Transliteration and Decipherment*. Retrieved from http://www.isi.edu/natural-language/people/naac109-print-1x2.pdf

Köhler, R., & Galle, M. (1993). Dynamic aspects of text characteristics. In Hřebíček, L., & Altmann, G. (Eds.), *Quantitative Text Analysis* (pp. 46-53). Trier: Wissenschaftlicher Verlag Trier.

Köhler, R., Altmann, G., & Piotrowski, R. G. (Eds.). (2005). *Quantitative Linguistik* / *Quantitative Linguistics*: *Ein Internationales Handbuch* / *An International Handbook*. *Handbücher zur Sprach- und Kommunikations-wissenschaft*, *Band 27.* Berlin – New York: Walter de Gruyter.

Köhler, R. (2005). Synergetic linguistics. In Köhler, R., Altmann, G. & Piotrowski, R. G. (Eds.), *Quantitative Linguistik* / *Quantitative Linguistics*: *Ein Internationales Handbuch* /*An International Handbook*. *Handbücher zur Sprach- und Kommunikations-wissenschaft*, *Band 27* (pp. 760-774). Berlin - New York: Walter de Gruyter.

Kramer, M. (2014). *Curiosity Rover Drills Into Mars Rock*, *Finds Water*. In space.com, December 16, 2014. Retrieved from https://www.space.com/28030-mars-water-curiosity-rover.html

Landis, G. A. (2000). *Mars Crossing*. New York: Tor Books.

Lanson, G. (1916 [1903]). *Conseils sur l'Art d'écrire*: *Principes de Composition et de Style*. [Advices on the Art of Writing: Principles of Composition and Style]. Neuvième Edition. Paris: Librairie Hachette et Cie.
Retrieved from http://gallica.bnf.fr/ark:/12148/bpt6k5452634t

Lasswitz, K. (1971 [1897]). *Auf Zwei Planeten* [Two Planets]. Weimar: Emil Felber. (H. H. Rudnick, Trans.). Carbondale: Southern Illinois University Press. Retrieved from http://www.gasl.org/refbib/Lasswitz_Auf_2_Planeten.pdf

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*: *Based on the British National Corpus*. London: Longman.

Lüdeling, A., & Kytö, M. (Eds.). (2009). *Corpus Linguistics*: *An International Handbook*. *Handbooks of Linguistics and Communication Science*, *29/2*. Berlin: Mouton de Gruyter.

M1. (2018). *Human Settlement on Mars*. Retrieved from https://www.mars-one.com/

Malvern, D., Richards, B.J., Chipere, N. and Durán, P. (2004). *Lexical diversity and language development*. Basingstoke, UK: Palgrave Macmillan.

McAuley, P. J. (1994). *Red Dust*. New York: Avon / HarperCollins.

Michaud, M. A. G. (2007). *Contact with Alien Civilizations*: *Our Hopes and Fears about Encountering Extraterrestrials*. New York: Copernicus Books / Springer Science + Business Media LLC.

Middleton Murry, J. (1922). *The Problem of Style*. London: Humphrey Milford / Oxford University Press. Retrieved from https://archive.org/details/problemofstyle00murruoft

Sherrod, Philip H. (2018). NLREG -- Nonlinear Regression and Curve Fitting. Retrieved from http://www.nlreg.com/

Oakes, M. P. (2009) Corpus Linguistics and Stylometry. In Lüdeling, A., & Kytö, M. (Eds.), *Corpus Linguistics*: *An International Handbook*. *Handbooks of Linguistics and Communication Science*, *29/2* (pp. 1070-1090). Berlin: Mouton de Gruyter.

O'Connor, M. (1996). The Berber Scripts. In Daniels, P. T., & Bright, W. (Eds.), *The World's Writing Systems* (pp. 112-116). New York-Oxford: Oxford University Press.

Orlov, J. K. & Chitashvili, R. Y. (1983). Generalized z-distribution generating the well-known 'rank-distributions.' *Bulletin of the Academy of Sciences of Georgia*, *110*: 269-272.

Parkinson, R. B. (1999). *Cracking Codes*: *The Rosetta Stone and Decipherment*. Berkeley - Los Angeles: The University of California Press.

Pohl, F. (1976). *Man Plus*. New York: Random House.

Pope, G. W. (1894). *Journey to Mars the Wonderful World*: *Its Beauty and Splendor*; *Its Mighty Races and Kingdoms*; *Its Final Doom*. New York: G. W. Dillingham.

Pope, M. (1999 [1975]). *The Story of Decipherment*: *From Egyptian Hieroglyphs to Maya script*. Rev. ed. London: Thames & Hudson.

Popescu, I.-I., & Altmann, G. (2006). Some aspects of word frequencies. *Glottometrics*, *13*: 23-46.

Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., (2009). *Word Frequency Studies*. Berlin – New York: de Gruyter.

Popescu, I-I., Čech, R., & Altmann, G. (2011). *The Lambda-structure of Texts*. Lüdenscheid: RAM-Verlag.

Pratchett, T. & Baxter, S. (2014). *The Long Mars*. Series *The Long Earth*. New York: Doubleday.

Ray, J. (2007). *The Rosetta stone and the Rebirth of Ancient Egypt*. London: Profile.

Robinson, A. (2002). *Lost Languages*: *The Enigma of the World's undeciphered Scripts*. New York: McGraw-Hill.

Robinson, A. (2011). Styles of Decipherment: Thomas Young, Jean-François Champollion and the Decipherment of the Egyptian Hieroglyphs. *SCRIPTA*: *International Journal of Writing Systems* (The Hunmin jeongeum Society), *3*: 123-132. Retrieved from http://scripta.kr/scripta2010/kr/scripta_archives/scripta_v03_a006.pdf

Robinson, K. S. (1992). *Red Mars*. New York: Spectra/Bantam, Dell/Random House.

Robinson, K. S. (1992). *Green Mars*. New York: Spectra/Bantam, Dell/Random House.

Robinson, K. S. (1992). *Blue Mars*. New York: Spectra/Bantam, Dell/Random House.

Rogers, H. (2005). *Writing Systems*: *A Linguistic Approach*. Malden, MA: Blackwell Publishing, Ltd.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities*, *31*: 351-365.

Sebeok, T. A. (Ed.). (1960). *Style in Language*. Cambridge, MA: The Technology Press of MIT / New York - London: John Wiley & Sons, Inc.

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

Simpson, E. H. (1949). Measurement of Diversity. *Nature*, *163*: 688. Retrieved from https://www.nature.com/articles/163688a0

Smith, L. W. (1916). *The Mechanism of English Style*. New York – London: Oxford University Press. Retrieved from https://archive.org/details/mechanismofengli00smitrich

Smith, A. E. (1989). *Mars*: *The Next Step*. Boca Ratón, FL: Chapman & Hall/CRC Press.

Solé, R., & Valbelle, D. (2002). *The Rosetta Stone*: *The Story of the Decoding of Hieroglyphics*. (S. Rendall, Trans.). London: Profile.

Springer Bunk, R. A. (2010). Los orígenes de la escritura líbico-bereber. Estudios Canarios: *Anuario del Instituto de Estudios Canarios* (La Laguna, Tenerife, Islas Canarias) *56*: 141-165.

Steele, P. M. (2013). *A linguistic history of Ancient Cyprus*: *The Non-Greek Languages and their Relations with Greek*, c.*1600-300 BC*. Cambridge: Cambridge University Press.

Strauß, U., Grzybek, P., & Altmann, G. (2007). Word Length and Word Frequency. In Grzybek, P. (Ed.), *Contributions to the Science of Text and Language* (pp. 277-294). Dordrecht, Netherlands: Springer.

Swift, J. (1721). A Letter to a Young Clergyman, Lately entered into Holy Orders. By a Person of Quality. London, 1721. The "Letter" is dated "Dublin, January the 9th, 1719-1720." In Cooper, L. 1930 [1907] (Ed.), *Theories of Style, with Especial Reference to Prose Composition* (pp. 161-169). New York – London: The Macmillan Company.

Thisted, R., & Efron, B. (1987). "Did Shakespeare write a newly-discovered poem?", *Biometrika*, *74*: 445-455. Retrieved from https://www.researchgate.net/publication/30962620_Did_Shakespeare_Write_a_Newly-Discovered_Poem

Tolstoy, A. N. (1950 [1922]). *Aelita*. (L. Flaxman, Trans.). Moscow: Foreign Languages Publishing House.

Tuldava, J. (2005). Stylistics, author identification. In Köhler, R., Altmann, G. & Piotrowski, R. G. (Eds.), *Quantitative Linguistik* / *Quantitative Linguistics*: *Ein Internationales Handbuch*/ *An International Handbook. Handbücher zur Sprach- und Kommunikations-wissenschaft, Band 27* (pp. 368-387). Berlin - New York: Walter de Gruyter.

Tweedie, F. J. (2005). Statistical models in stylistics and forensic linguistics. In Köhler, R., Altmann, G. & Piotrowski, R. G. (Eds.), *Quantitative Linguistik* / *Quantitative Linguistics*: *Ein Internationales Handbuch* / *An International Handbook*. *Handbücher zur Sprach- und Kommunikations-wissenschaft, Band 27* (pp. 387-395). Berlin - New York: Walter de Gruyter.

Wackernagel, W. (1888 [1873]). *Poetik*, *Rhetorik und Stilistik* [Poetics, Rhetoric, and Stylistics]. Academische Vorlesungen, 2. Sieber, L. (Ed.). Halle: Verlag der Buchhandlung des Weisenhauses. Retrieved from http://reader.digitale-sammlungen.de/de/fs1/object/display/ bsb11159934_00005.html

Weinbaum, S. G. (1934). *A Martian Odyssey*. *Wonder Stories*, *July 1934*. Gernsback Publications. Retrieved from http://gutenberg.net.au/ebooks06/0601191h.html

Weir, A. (2014 [2011]). *The Martian*. New York: Crown Publishing / Penguin Random House.

Wells, H. G. (1898). *The War of the Worlds*. London: William Heinemann. Retrieved from https://archive.org/details/warofworlds00welluoft

Wikipedia. (2018). *Entry page*: *Omnilingual*. Retrieved from https://en.wikipedia.org/wiki/Omnilingual

Write Words. (2002-2018). *Frequency word counter*. Retrieved from http://www.writewords. org.uk/word_count.asp

Wyndham, J. [aka John Beynon]. (1972 [1936]). *Planet Plane*. Later republished as *Stowaway to Mars* (1972). London: Newnes Limited.

Yesypenko, N. (2008). Writer's voice in the texts of "Peace and War" themes. *Glottometrics,* (RAM-Verlag) *16*: 17-26.

Yule, G. U. (2014 [1944]). *The Statistical Study of Literary Vocabulary*. Cambridge, UK: Cambridge University Press.

Zieffler, A. S., Harring, J. R., & Long, J. D. (2011). *Comparing Groups*: *Randomization and Bootstrap Methods Using R*. Hoboken, NJ: John Wiley & Sons, Inc. Publication.

Zipf, G. K. (1935). *The Psycho-Biology of Language*: *An introduction to dynamic philology*. Boston: Houghton Mifflin.

Glottometrics 43, 2018, 58-67

The Lexicon and the Noisy Channel: Words are shaped to avoid confusion

Adam King[1]
University of Arizona

Abstract. Language exists in a noisy channel and so an optimized lexicon should be structured to avoid possible confusion among words. For a code to be optimal in a noisy channel, it should maximize the mutual information between what is sent and what is received through noise. Using confusion matrices for English phonemes, this article investigates the extent to which the English lexicon is structured under pressures of a noisy channel. I find that the relative frequency values and phonological make up of English words cause the lexicon to be less likely to suffer confusion than a comparable baseline. I discuss the results with respect to the growing body of literature on the lexicon as an optimal code.

Keywords:
Information theory, lexicon, optimization, word length, noisy channel, synergetic linguistics, phonetics

There has been much recent (and not so recent) discussion of the lexicon as being optimized for efficient communication (Zipf 1935, Piantadosi et al. 2009; 2011; 2012, Mahowald et al. 2013, Dautriche et al. 2017). The primary direction of this work involves the investigation into the relationship between the probability of a word and its length. Coined by Zipf (1935) as *Zipf's law of abbreviation*, the length of a word is inversely proportional to its frequency and this pattern is apparent in a diverse array of languages (Bentz and Ferrer-i-cancho 2013). More recent work has suggested that word length is more closely correlated with a word's average contextual probability than its frequency and frequency functions as heuristic for measuring this (Piantadosi et al. 2011). All in all, the relationship between word probability and length causes the lexicons of natural languages to be similar to an optimal non-singular code (Cover and Thomas 2012) and this has been used to drive the argument that the lexicon is optimized for efficient communication. If there were no pressure for language to be efficient then the lexicon would not show this effect. Yet, minimizing length is only one part of creating a communicatively efficient lexicon.

Language is a system of communication and as such involves a sender and a receiver, a speaker and a listener. Because of this, a truly efficient lexicon should benefit both speaker and listener. That is, the lexicon should be organized to minimize length while ensuring that words maintain large perceptual distance from each other, thus minimizing the chance that words are confused (Köhler 1987; 1993). At this point, many of the arguments for an optimized lexicon are focused on word length and thereby focused on optimization to benefit the speaker. All things being equal, longer words involve more effort to produce and so minimizing length minimizes effort for the speaker. Because of this focus on length, much of the existing work on lexical optimization is also compatible with an argument for a solely

[1] Adam King, University of Arizona (USA), e-mail: adamking@email.arizona.edu

speaker-optimized lexicon, i.e. compression (see Ferrer-i-cancho 2016; 2017, section 3 for discussion). To this end, a convincing argument must also show benefits to the listener. If the lexicon is indeed structured for communication, words should be shaped in a way that minimizes the chance of misperception. In other words, the ideal lexicon should also be robust to noise.

Language exists in a noisy channel. Whatever form this noise takes (e.g. physical noise, differences in dialects, idiosyncrasies in an individual's pronunciation, etc.), it will probabilistically cause the sequence of phonemes said by the speaker to be misperceived as a different sequence by the listener. This can cause ambiguity and be a potential problem for communication. For example, noise may cause the [h] in the English word *hat* to be perceived as a [v], causing a listener to mistakenly identify the word as *vat*. The converse can be said of the [v] in *vat* which can cause the word to be misperceived as *hat*. Considering this, an optimized lexicon should assign phonological forms to word meanings such that the likelihood of any misperception due to noise is minimal. As defined in Information Theory (Shannon 1948; 1949, MacKay 2005, ch. 8), the optimal code in a noisy channel should aim to maximize the mutual information between the input (what a speaker says) and output (what a listener perceives). Thus, an optimized lexicon should yield a greater value for the following equation compared to a lexicon that is not:

(1)

$$I(X;Y) = H(Y) - H(Y|X)$$

$$= \sum_{y \in Y} p(y) \log \frac{1}{p(y)} - \sum_{x \in X,\ y \in Y} p(x)p(y|x) \log \frac{1}{p(y|x)}$$

Here, p(x) represents the probability that a word will be produced by a speaker, p(y) represents the probability a word will be perceived by a listener and p(y|x) represents the conditional probability a form will be perceived given the form that was produced. To achieve a maximally informative channel, the most frequently produced word forms should be the least confusable with others. Continuing the earlier example, [h] is less likely to be confused with [v] than vice versa for English speakers (see Methods for more detail). Because [h] is more robust to noise than [v], the form [hæt] should be assigned to a more frequent meaning than the form [væt]. Doing this minimizes the probability that miscommunication will occur at all. This is, in fact, the case in English and *hat* is a more frequent word than *vat*. The important question is whether this pattern carries beyond an anecdotal example. If it is the case that the lexicon does assign the most robust forms to the most frequent meaning, it will be evidence that the lexicon is indeed closer to the optimum for communicative efficiency beyond what would be expected in compression.

In this work, I will provide evidence that this is the case, at least for English. I will show that the English lexicon assigns word forms to meanings in a way that causes English to be more robust to noise than a comparable baseline. Doing so, I argue that the lexicon is, in fact, closer to the optimum for communication than might be expected from random chance.

Methods and Results[2]

To test this, I used phonetic confusion matrices for American English from Cutler et al. (2004). A confusion matrix represents the probability that a particular phoneme would be confused with another phoneme by a speaker of a given language. For example, the probability an English-speaker would mistake [h] for [v]. These matrices were constructed from data from English-speakers listening to tokens of English words in artificial noise. The participants were then asked to report what they heard. From these data, Cutler et al. (2004) were able to compile the likelihood a phoneme in English would be perceived as another. Doing so, these matrices represent an estimation of the likelihood any phoneme would be subject to noise in speech. Though these confusion matrices may not be perfect representations of the noise that affects everyday speech, they are a good representation of the relative odds that one phoneme will be confused for another (for more, see Pisoni 1996).

Individual words

With these matrices, I computed the probability that words from English would avoid confusion with other words. To do so, I first collected the phonetic transcriptions for English word lemmas via the Carnegie Mellon Pronouncing Dictionary (Weide 2005), ignoring any word with a consonant cluster or adjacent vowels in the same syllable. I ignored consonant and vowel clusters as the confusion matrices do not cover clusters and it would not be possible to estimate confusion for these words. Given the phonetic transcriptions, I grouped the words by their CV skeleton, the pattern of consonants and vowels in the word. As was done in the confusion matrices, I distinguish between syllable-initial and syllable-final consonants as well as vowels in open and closed syllables in the construction of the CV skeletons. In total, this resulted in 140 total CV skeletons and 9210 lemmas overall.

To calculate the probability that a word would avoid noise, I use the product of the probabilities that each phoneme from the input maps to the same phoneme in the output. For example, for the word *hat*, p(y = *hat*|x = *hat*) = p(y = [h]|x = [h]) * p(y = [æ]|x = [æ]) * p(y = [t]|x = [t]). To avoid the effects of word length, I only compared words within the same CV skeleton for the bulk of the analysis. While this makes the assumption that the noise that affects each phoneme in the word is independent, this method's intention is to provide a means to compare the odds a word is understood correctly given the other words in the lexicon. By doing this, I was able to estimate the probability each word would be perceived correctly or the *probability of noiselessness* for that word. Crucially, the probability of noiselessness does not include any measure of the word's frequency in its calculation. If the lexicon is structured to be robust to noise, there should be a positive correlation between word frequency and the probability that the word avoids noise. That is, more frequent words should be more likely to avoid misperception in order to minimize the likelihood of misperception as a whole.

After calculating the probability of noiselessness for all relevant words in English, I fit a linear model in R (R team 2013) to test the effect of word frequency on the robustness to noise for that word. The dependent variable of the model was the probability of noiselessness and the independent variables were the frequency of the word and the length of the word in phonemes. I used word frequency values from a 40 million word corpus of spoken American English (COCA, Davies 2011) and natural log transformed them to mirror the logarithmic effect of word probability on word recognition times (see Levy 2008, Smith and Levy 2013). I included word length as a factor to offset the fact that longer words would naturally have a

[2] All data and code are available at: https://github.com/AdamKing11/Noisy_Lexicon

lower probability of noiselessness because there were more phonemes that must avoid noise. Because the dependent variable was the result of the product of multiple probabilities, I log transformed the word length factor to account for the negative sublinear effect that word length was likely to have.

The results of the model are in Table 1. There was a significant effect of word frequency on the probability of noiselessness (t = 4.63, p < .01), independent of the length factor. This indicates that frequent words are more likely to avoid the effects of the noise. Because the model showed an effect of log word frequency independent of the length factor, this suggests that this is not solely due to the fact that more frequent words tend to be shorter.

Table 1
Results of linear model, testing the effect of word frequency on the probability a word will avoid noise. The positive effect of word frequency indicates that more frequent words are more likely to avoid noise.

Fixed Effects	*β*	**Std. Error**	**t-value**
Intercept	0.20925	0.00257	81.29
Log Word Freq.	0.00088	0.00019	4.63
Log Word Length	-0.10154	.001257	-80.76
AIC	-31244.36	R^2	0.4473

Testing the lexicon as a whole

Following this, I tested the lexicon as a whole for its robustness to noise. Recall, the optimal communication system in a noisy channel is one that maximizes the mutual information between what is sent and what is received. If the English lexicon is optimized for communication, then it should yield a significantly higher amount of mutual information than similar randomized alternatives. To create these randomized alternatives, I randomly shuffled the frequency values for words of the same CV skeleton. This created alternate lexicons that differed only in the assignment of word form to frequency. Because of the similarity between the original English lexicon and these shuffled alternatives, any difference between them must derive from the assignment of word form to that word's frequency. A crucial thing to notice here is that any of these alternatives is equally optimized as the original English lexicon with respect to word length. If the words in the lexicon are optimized for compression and not by pressures for efficient communication, the original English lexicon should not contain greater mutual information than these shuffled alternatives. However, if the original lexicon has a significantly higher amount of mutual information, then this is evidence that the lexicon is closer to the optimum to avoid noise than would be expected from chance.

Using Eq. 1, I calculated the mutual information for the original English lexicon and 1000 randomly shuffled alternatives. Because the confusion matrices only covered consonant-consonant and vowel-vowel confusion and frequency values were shuffled within skeletons, I restricted the analysis to only consider words being misperceived as other words from the same CV skeleton. This also ensured that comparisons were between words of the same length. To avoid over-representing certain word lengths or particular CV skeletons in the analysis, I only included the 30 most frequent words from the English lexicon for both the original and shuffled alternatives, leaving 74 CV skeletons and 1600 word lemmas.

For the unmodified English lexicon, the mutual information was equal to 4.899667. Numerically, the calculated value for the English lexicon is unusually small and as such it is an unlikely estimate for English in use. However, the goal here is to compare the lexicon against possible baselines and not to accurately estimate the efficiency of English. Using the same method, I calculated the mutual information for all alternative lexicons. The original English lexicon had a greater value than 95% of the shuffled alternatives (see Fig. 1). This suggests that the lexicon as a whole is structured to be robust to noise. Combined with the results in Table 1, this is strong evidence that the distribution of word frequencies in English is closer to optimal and this is unlikely due to chance. Considering that all shuffled alternatives are equally optimized for length, this effect goes beyond what might be expected from compression. This further supports the theory that the lexicon is optimized for both the speaker and listener and thus for communication.

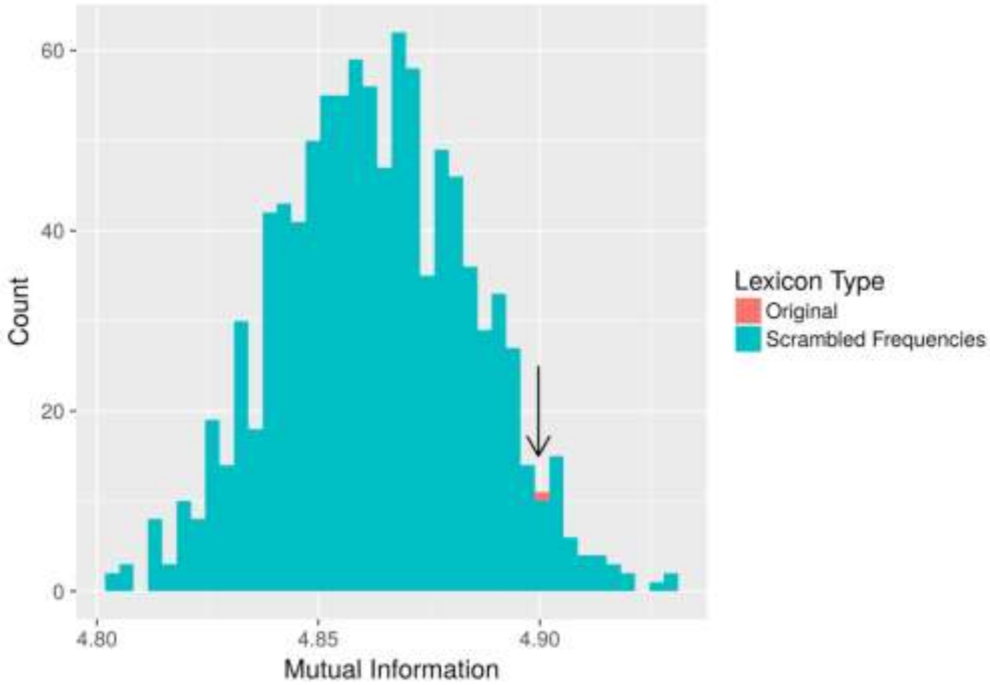


Figure 1: Histogram of mutual information of original English lexicon compared to 1000 randomly shuffled variants. The English lexicon contains a greater amount of mutual information than 95% (956 of 1000) of the variants.

Subsequent Tests

However, a possible confound lies in the confusion matrices themselves. When given a token that is ambiguous between multiple words, listeners rely on their knowledge of lexical statistics and more often perceive the ambiguous form as the most frequent alternative (Ganong 1980, Hilpert 2008). For example, when listeners are given a token that is ambiguous between a high and low frequency word, they are more likely to report hearing the high frequency word. Because of this, it is possible that the results thus far are a result of biases of the English-speaking participants which were used to create the confusion matrices.

That is, a possible reason that [v] is more likely to be misperceived as [h] and not vice versa is due to the fact that more English words contain [h] than [v] and the words themselves are more frequent.

To circumvent this confound, I used another set of confusion matrices from Cutler et al. (2004). These were constructed from the data of Dutch-speakers given the same input as the English-speaking participants. Because the Dutch participants had less knowledge of the English lexicon than their English-speaking counterparts, these matrices avoid the earlier confound. Constructing a similar model as in Table 1, I tested the effect of word frequency on the probability of noiselessness for English words using the confusion matrices of the Dutch participants (see Table 2). As with the previous model, there is a significant effect of word frequency ($t = 3.49$, $p < .01$). However, the magnitude of the effect and the overall R^2 for the model are not as great. This indicates that the effect is affected by speakers' knowledge of lexical frequencies but not totally so.

Table 2
Results of linear model using Dutch-speaker confusion matrices. Word frequency shows a smaller though still significant effect compared to the English-speaker confusion matrices.

Fixed Effects	*β*	**Std. Error**	**t-value**
Intercept	0.09591	0.00139	68.87
Log Word Frequency	0.00036	.00010	3.49
Log Word Length	-0.04732	.00068	-69.56
AIC	-42557.98	R^2	0.3740

As a final test, I constructed 1000 shuffled lexicons as I did previously with the English-speaker confusion matrices. I then used the Dutch confusion matrices to calculate the mutual information for the original frequency values of the English lexicon and for the shuffled variants. Using the Dutch-speaker confusion matrices, the original lexicon had greater a mutual information than 84% of the randomized lexicons (see Fig. 2). Though this is a drop from the previous comparison, the fact that word frequency showed a positive effect on the probability of noiselessness in a linear model suggests that the effect is not solely a result of using the English-speaker confusion matrices. Combined with the earlier tests, this suggests that the English lexicon is organized to be robust to noise and these effects are not likely due to confounds in the creation of the confusion matrices.

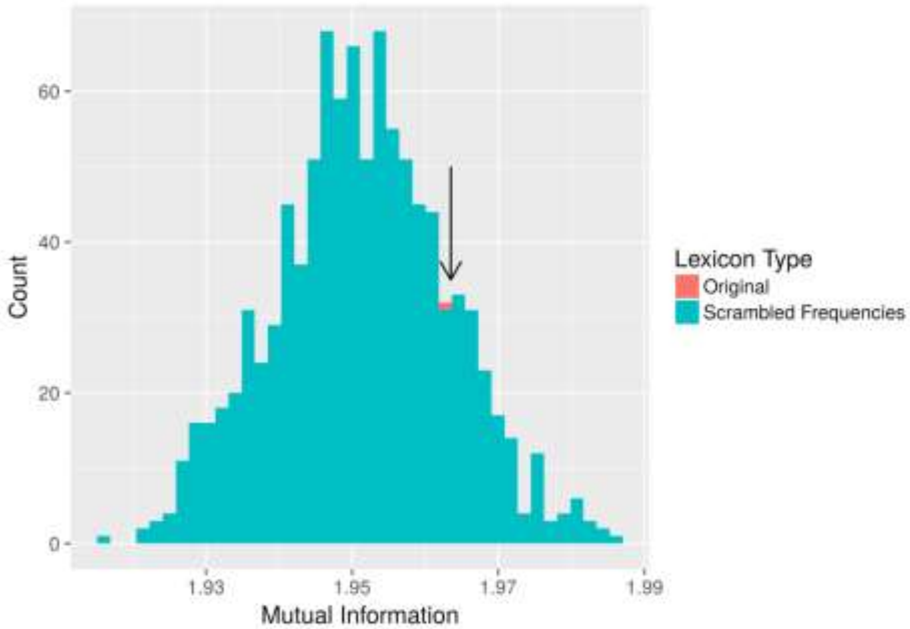


Figure 2: Histogram of mutual information of original English lexicon compared to 1000 randomly shuffled variants. Here, the mutual information has been calculated using confusion matrices constructed from Dutch speakers. The English lexicon contains a greater amount of mutual information than 84% (843 of 1000) of the variants.

Discussion

This work seeks to address whether the lexicon is optimized for efficient communication or whether the relationship between word frequency and length can be attributed to solely speaker-oriented optimization, e.g. compression. I find that the relationship between word forms and frequency in the English lexicon makes the language significantly more robust to noise than random. This effect goes beyond optimization of length and so is less likely to be compatible with a solely speaker-based story. If lexical optimization was only focused on reducing word length, then the lexicon should show no effect of avoiding ambiguity among words of the same length. However, the fact that words of the same length are organized such that the most robust word forms are assigned to the most probable meanings indicates that the lexicon is also structured to minimize potential ambiguity. When considered with work that shows that the shortest word forms are assigned to most probable meanings (e.g. Piantadosi et al. 2011), this makes a strong argument that the lexicon is optimized for efficient communication.

Primarily, this supports the predictions of synergetic linguistics (Köhler 1987; 1993). Namely, multiple factors affect the shape of words in the lexicon and the ultimate structure of the lexicon is a balance between these factors. Speakers desire word forms that are easy to produce, i.e. short lengths. Listeners prefer word forms that are easy to disambiguate from other words, i.e. forms which have little or no potential ambiguity. In the terminology of synergetic linguistics, this is a balancing of minP (word length) and minD (phonetic distinctiveness). To reach an optimal state, the lexicon should balance these two pressures and create a system that contains short word forms while still being robust to noise. A possible strategy for this is to structure the lexicon such that the most frequent words are also the least likely to be confused with other words. This can be done while simultaneously assigning short forms to probable words. In this way, the lexicon is able to balance the desires of both the speaker and listener. Because the English lexicon shows the effects of both pressures, this suggests that the lexicon is structured for efficient communication, beyond what would be expected of solely-speaker oriented optimization.

This balancing between speaker and listener aligns with the predictions of Ferrer-i-cancho and Solé (2003) and Ferrer-i-cancho (2018). There, they find that the distributions of word frequencies in natural languages benefit both speaker and listener. However, in their experiments use a more simplified means to estimate the likelihood a word will be confused. Here, I used words from the lexicon of a natural language and an empirically motivated estimation of potential ambiguity between words. Doing so, I find similar results that the lexicon is structured to benefit both speaker and listener. This further suggests that the lexicon is indeed closer to the optimum code than random for efficient communication.

The results presented here are also similar to those of Piantadosi et al. (2009). Piantadosi et al. (2009) showed that the parts of words which are pronounced with the greatest effort, i.e. the stressed syllables, are the most informative for identifying the containing word. Because the most informative syllables are pronounced with more effort, these syllables are less likely to be misperceived due to noise. In this way, individual words are structured to be robust to noise. Here, I show that the assignment of word form to meaning is such that the expected effect of noise is minimized across the entire lexicon. Taken together, these show that words are shaped individually for efficient communication and this pattern holds across the entirety of the lexicon.

A noteworthy result of this work is the difference between the English-speaker and Dutch-speaker confusion matrices and their predictions for how well English words avoid confusion. As stated previously, this could be the result of English speakers' implicit knowledge of the lexical statistics of English. On the other hand, this may be evidence that the perceptual system of English-speakers is also optimized to minimize confusion between words of their language. Lab studies have shown that listeners are more aware of the phonetic cues that are particularly useful given the phonemic contrasts in their language (Boomershine et al. 2008, Johnson and Babel 2010). That is, listeners are aware of the specific phonetics of the phonemes of their language and use this knowledge to aid them in distinguishing contrastive sounds. Thus, the difference between the experiments based off the two sets of confusion matrices may in fact be evidence that the phonetic perceptual system for the speakers of a language is structured such that listeners pay strongest attention to the relevant cues for their language. This, in turn, causes the words to be robust to noise as listeners are cued to the most relevant phonetic properties. This offers another avenue for a noise-resistant lexicon. In addition to the lexicon being structured to take advantage of noise-resistant phonemic contrasts, the phonetic system of the language – how those contrasts are realized – may be structured to take advantage of the contrasts in the lexicon. That is, just as phonetics affect the lexicon, the lexicon may affect phonetics. This predicts that the phonemic system of a language should also be structured to maximize the perceptual difference between its phonemes.

Interestingly, this may coincide with the results found in Everett (2013) and Everett et al. (2015; 2016). Together, these show that languages are more likely to lose phonemic contrasts in climates where those contrasts are more difficult to produce or perceive. This is further evidence that lexicons aim to minimize misperception. By avoiding constructing words with ambiguous phonemes and then eventually losing these phonemes altogether, the lexicon is less likely to yield ambiguous word forms. However, for these works, this occurs at a structural level rather than at a lexical level. This suggests that both phonemic inventories and lexicons evolve in tandem in order to cause language to be an efficient code in the presence of noise. In other words, maximizing communicative efficiency is a driving force behind many aspects of natural languages and optimizing the form of words is a single part of this. However, future work is merited to unravel the effects of phonemic inventory on the lexicon and how it is robustness to noise. As well, future work is merited to confirm the results found here for other languages. That being said, the results presented here still offer strong support for a lexicon that is efficient for communication, beyond the effects that would be expected for solely speaker-oriented optimization.

Acknowledgements

I would like to thank Andy Wedel for his feedback and mentorship for this project. This work was supported by a National Science Foundation Graduate Research Fellowship grant to the author.

Works Cited

1. Bentz, C., Ferrer-i-Cancho, R. (2016). Zipf's law of abbreviation as a language universal. In: Bentz C, Jäger G, Yanovich I. (eds.). *Proceedings of the Leiden Workshop on Capturing Phylogenetic Algorithms for Linguistics*. Tübingen: University of Tübingen. https://publikationen.uni-tuebingen.de/xmlui/handle/10900/68558.
2. Boomershine, A., Hall, K. C., Hume, E., & Johnson, K. (2008). The impact of allophony versus contrast on speech perception. In: *Contrasts in phonology: Theory, perception, acquisition*, *145-171.*
3. Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.
4. Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6)*, 3668-3678.*
5. Davies, M. (2011). Word frequency data from the Corpus of Contemporary American English (COCA).
6. Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive science, 41(8), 2149-2169.*
7. Everett, C. (2013). Evidence for direct geographic influences on linguistic sounds: The case of ejectives. *PloS one, 8(6), e65275.*
8. Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences, 112(5), 1322-1327.*
9. Everett, C., Blasí, D. E., & Roberts, S. G. (2016). Language evolution and climate: the case of desiccation and tone. Journal of Language *Evolution, 1(1), 33-46.*
10. Ferrer-i-Cancho, R., & Solé, R. V. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, *100*(3), 788-791.
11. Ferrer-i-Cancho, R. (2016). Compression and the origins of Zipf's law for word frequencies. arxiv.org/abs/1605.01326
12. Ferrer-i-Cancho, R. (2017). The Placement of the Head that Maximizes Predictability. An Information Theoretic Approach. *Glottometrics 39*, 38 – 71.
13. Ferrer-i-Cancho, R. (2018). Optimization models of natural communication. *Journal of Quantitative Linguistics*, *25*(3), 207-237.
14. Ganong, W.F. (1980) Phonetic Categorization in Auditory Word Recognition. *J. of Exp. Psychol. Hum. Percept.Perform., 6, 110-125.*

15. Hilpert, M. (2008). New evidence against the modularity of grammar: Constructions, collocations, and speech perception. *Cognitive linguistics*, *19*(3), 491-511.
16. Johnson, K., & Babel, M. (2010). On the perceptual basis of distinctive features: Evidence from the perception of fricatives by Dutch and English speakers. *Journal of phonetics*, *38*(1), *127-136.*
17. Köhler, R. (1987). System theoretical linguistics. *Theoretical linguistics*, *14*(2-3), *241-258.*
18. Köhler, R. (1993). Synergetic linguistics. In: *Contributions to quantitative linguistics* (pp. 41-51). Springer, Dordrecht.
19. Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126-1177.
20. MacKay, D. (2005). *Information theory, inference, and learning algorithms*. Cambridge University Press.
21. Mahowald, K., Fedorenko, E., Piantadosi, S. and Gibson, E. (2013). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition 126: 313–318. doi:10.1016/j.cognition.2012.09.010.*
22. Piantadosi, S., Tily, H. and Gibson, E. (2009). The Communicative Lexicon Hypothesis. *The 31st Annual Meeting of the Cognitive Science Society (CogSci09). Austin, TX. Cognitive Science Society. pp. 2582–2587.*
23. Piantadosi, S., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. Proceedings of the National Academy of Sciences 108: 3526–3529. doi:10.1073/pnas.1012551108.
24. Piantadosi, S., Tily, H. and Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition 122: 280–291*. doi:10.1016/j.cognition.2011.10.004
25. Pisoni, D. B. (1996). Word identification in noise. *Language and cognitive processes, 11*(6*), 681-688.*
26. R Team. (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.
27. Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal 27: 623–656. doi:10.1002/j.1538-7305.1948.tb00917.x*
28. Shannon, C. E. (1949). Communication in the Presence of Noise. *Proceedings of the IRE 37, 10–21*. doi:10.1109/jrproc.1949.232969.
29. Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), *302-319.*
30. Weide, R. (2005). *The Carnegie mellon pronouncing dictionary* [cmudict. 0.6].
31. Zipf, G. K. (1935). *The psycho-biology of language: an introduction to dynamic philology*. M.I.T. Press. Cambridge, MA.

Glottometrics 43, 2018, 68-76

Belza-chain Analysis: Weighting Elements

Michal Místecký[1], Jiang Yang[2], Gabriel Altmann

Abstract. In the article, it will be shown that weighting of elements of Belza-chains is possible using a classification of relations of elements and the main words. Uninterrupted sequences of sentences/lines form a chain, which has a main word; the other elements are some lexical or grammatical references. By counting the number of different classes, one obtains a possible frequency ranking which can be modelled. The main model is the Zipf-Alekseev function, but one can also employ the Lorentzian function. The analysis can be of use in multifarious investigations of styles.

Keywords: Belza-chain, weighting, ranking, Zipf-Alekseev function, Lorentzian function, English, Chinese, French, Czech, Slovak, Hungarian, Indonesian, German, Italian.

Belza-chains show how a text is linearly structured. As a matter of fact, it is a variant of the famous Skinner hypothesis (cf. Skinner 1957), conjecturing that entities lying nearer to one another are more similar than those in a greater distances, This fact can be measured in various ways and using various linguistic entities (cf. Altmann 2018). Here, we restrict ourselves to concepts and will measure the conceptual inertia of the texts, not the distances.

A Belza-chain is an uninterrupted sequence of sentences or poem lines containing the same concept. Here, too, there are several possibilities, both qualitative and quantitative, to obtain data and test them. The simplest way is the evaluating of the chain lengths in the whole text using an indicator (cf., e.g., Chen, Altmann 2015). The longer, say, the mean chain length in the text is, the stronger the conceptual inertia is. Another way is the weighting of the variants of the same concept. The concept itself need not appear directly, but in the form of words or morphemes which reflect it. There are great differences between languages; hence, a translation can serve as a means for determining types. For example, in Slavic languages, the first person is always expressed by a pronoun and/or a verbal ending; on the other hand, in English "I speak", the first person is expressed only with "I", while the third person "he speaks" has two morphemes related to it. In strongly agglutinating languages, for each person there is a special morpheme in the verb. In analytic languages, there is only the pronoun.

Now, here we want to weight the inertia and utilize the scaling proposed by Roelcke, Popescu and Altmann (2017) used for the "hreb-analysis" of texts. Hrebs – defined by Hřebíček (1997) and elaborated by Ziegler and Altmann (2002) – differ from chains because they consider all occurrences of the given concept in text, not only in subsequent sentences. The above authors propose the weighting as presented in Table 1, but for other types of languages, one could construct it differently. And, needless to say, other linguists would set up such a table differently. More thought should thus be paid to the classification before a stylometric analysis is started; the one presented here serves for modelling purposes only.

[1] Michal Místecký: mmistecky@seznam.cz.
[2] Jiang Yang: yangjiang@hnust.edu.cn.

Table 1
Weighting/classifying of hreb/chain elements

Weight/ Class	Hreb/chain element
1	Main word, head of the hreb/chain
2	Synonym, metaphor, variant
3	Hyponym (= specification)
4	Hypernym (= generalization, class)
5	Relative pronoun, relative phrase, rhetoric question, rhetoric answer, article, interrogative pronoun, demonstrative pronoun
6	Personal pronoun
7	Possessive pronoun
8	Grammatical affix or introflection referring to the head
9	Derivation or composition containing the head; conversion of head to other POS
10	Suppletion

Since in different languages the situation is different, one should consider the above table only as a kind of possible classification otherwise there will be many problems in individual languages. One can reduce the types or increase their number. What is important is the very distinguishing of classes. Sometimes, a chain begins with a personal pronoun, and all sentences in the given chain contain only the pronoun. In that case, we consider the pronoun as the main element of the hreb.

The main word of the concept need not be a noun, but a word which is repeated under various forms in the text. The scale shows the increasing weight with an increasing "grammatical" distance. This means that the higher the weight is, the more distant the elements of a chain are to be considered, the greatest being the suppletion – e.g., "good" and "best". It has been defined as the greatest weight because even the meaning is somewhat modified. In the present paper, we limit the notion to various forms of adjectives, and the forms of irregular verbs (e.g., "be" – "was" – "were" in English). If the same concept occurs twice in the same sentence/line, it is taken into account only once, and one considers the variant with the lowest weight.

When one analyzes a text, one can mark the members of the same chain by a colour or a typographic sign (cf. Chen, Altmann 2015), then exchange the words/morphemes by their weigh, and characterize the text by some indicator. First of all, one can set up the distribution of weights and capture it by means of a simple function. The properties of the function may be used for characterization.

Another way is the squaring of individual lengths and dividing the sum by the number of sentences/lines. Here, every sentence/line not belonging to any chain has the value 1. The respective indicator can be defined as

$$(1) \qquad CL = \frac{\sum f(W_i)^2}{N}$$

where $f(W_i)$ is the weight and N is the number of sentence/lines in the text. Since $f(W_i)$ is a quite usually measured value (it is not the usual second moment), we can consider it – for the sake of implicity – as a usual variable, hence, the definition of the variance and a possible normal test are quite simple. However, if we consider the squared values and divide *CL* by *N*, we obtain the usual repeat rate of the weights.

Let us consider a German example (*Der Erlkönig* by Goethe) presented by Chen and Altmann (2015), in which we weight the variants of the concepts. Here, e.g. the concept *Vater* is represented in the first six lines by *wer*, (reit)*et*, *ist*, *der*, *Vater*, *er*, *mein* *du*. As aforementioned, if two of these words occur in a line, one should consider only the one that has the smaller weight. This rule holds generally.

As a matter of fact, examining the chains, one weights the whole line, not the individual representatives of the given concept.

Table 1
Inertia in a German text (Goethe, Der Erlkönig)

Wer reitet so spät durch Nacht und Wind?	5,1,6,6,7,1 (Vater)
Es ist der *Vater* mit seinem *Kind*;	
Er hat den *Knaben* wohl in dem Arm.	1,3,6,3 (Kind)
Er fasst *ihn* sicher, er hält ihn warm.	
Mein *Sohn*, was birgst du so bang dein Gesicht?	1,1 (Erlkönig)
Siehst, *Vater*, du den *Erlkönig* nicht?	
Den *Erlkönig* mit Kron und Schweif?	
Mein *Sohn*, es ist ein Nebelstreif.	3,1,6 (Sohn,Kind, dir)
Du, liebes *Kind*, komm, geh mit *mir*!	6,6 (mir,ich)
Gar schöne Spiele spiel *ich* mit *dir*;	1
Manch bunte Blumen sind an dem Strand,	1
Meine Mutter hat manch gülden Gewand.	
	2,6,1 (mein, mir, Kind)
Mein Vater, mein Vater, und hörest du nicht,	
Was Erlenkönig *mir* leise verspricht?	1
Sei ruhig, bleibe ruhig, mein *Kind*	
In dürren Blättern säuselt der Wind	
	3,6 (Knabe,dich)
Willst, feiner *Knabe*, du mit *mir* gehn?	6,6,6 (mir, meine, meine)
Meine *Töchter* sollen *dich* warten schön;	1,1,8 (Töchter, Töchter, wiegen)
Meine *Töchter* führen den nächtlichen Reihn	
Und *wiegen* und tanzen und singen *dich* ein.	1
	1
Mein Vater, mein Vater, und siehst du nicht dort	
Erlkönigs Töchter am düstern Ort?	6,6 (es, es)
Mein Sohn, mein Sohn, ich sehe es genau:	
Es scheinen die alten Weiden so grau.	6,6,6,6,1 (ich, ich, mir, er, Erlkönig
	6,6 (dich, du, mich, mir)
Ich liebe *dich*, mich reizt deine schöne Gestalt;	
Und bist *du* nicht willig, so brauch *ich* Gewalt.	
Mein Vater, mein Vater, jetzt faßt *er* *mich* an!	1,6,8,7 (Vater, er, erreicht, seinen)
Erlkönig hat *mir* ein Leids getan!	
Dem *Vater* grausets, er reitet geschwind,	
Er hält in Armen das ächzende Kind,	

Erreicht den Hof mit Mühe und Not: In *seinen* Armen das Kind war tot.	

If we consider the frequency of individual values, we do not obtain any regular monotonic sequence, but rather an oscillating one. Hence, it is simpler to rank the weights in the usual way – i.e., consider them to be classes. In any case, one should work with longer texts. A special case that must be taken into consideration is the change of the perspective – in one sentence/line, one can speak about “I”, in the next one, the same person is “you”.

By ordering the frequencies of weights/classes in the above text, we obtain the results presented in Table 2. This simple sequence can be captured by several functions; here, we use here merely the Zipf-Alekseev and the Lorentzian ones. The former is of the shape

$$(2) \qquad y = 1 + c * x^{a+b*\ln(x)} \,,$$

whereas the latter is expressed by formula –

$$(3) \qquad y = \frac{a}{1+\left(\frac{x-b}{c}\right)^2} \,.$$

The choice of the function will be explained later; analysing a large amount of texts will show which is more stable. The ranks represent here individual weight classes of the above classification.

Table 2
Fitting two functions to the ranked weights of Belza elements
(German, Goethe, *Der Erlkönig*)

Rank	Frequency	Zipf-Alekseev f.+1	Lorentzian f.+1
1	20	20.01	20.00
2	16	15.96	16.00
3	4	4.26	3.95
4	2	1.60	2.15
5	2	1.11	1.60
6	1	1.03	1.37
7	1	1.00	1.05

		a = 1.8073
		b = -3.1055
		c = 19.0065
		R^2 = 0.9973
			a = 37.1230
			b = 1.4458
			c = 0.4564
			R^2 = 0.9990

The indicator in formula (1) would yield

$$CL = \frac{16^2+1^2+4^2+1^2+20^2+2^2+2^2}{46} = 14.98 \,.$$

The usual repeat rate would be

$$RR = \frac{14.98}{46} = 0.3257 \,;$$

the minimum of the repeat rate being

$$\frac{1}{N} = \frac{1}{46} = 0.0217\,,$$

and the maximum 1, we have a relatively high repeat rate (almost 15 times greater than the minimum). As can be seen, the parameter *c* of the Zipf-Alekssev formula yields an almost perfect f₁-value, while the parameter *a* of the Lorentzian function does not reflect it, attaining very high figures. As this becomes even more prominent in the texts to come, the Zipf-Alekseev function is considered better for this type of fitting.

Another possibility would be as follows: after identifying a chain, one should consider all concepts of the main concept. Thus, for example, in the first two chains of the above poem we have

Wer reitet so spät durch Nacht und Wind? Es ist der Vater mit seinem Kind; Er hat den Knaben wohl in dem Arm. Er fasst ihn sicher, er hält ihn warm. Mein Sohn, was birgst du so bang dein Gesicht? Siehst, Vater, du den Erlkönig nicht?	Wer, (reit)et, ist, der, Vater, sein(em), er, (ha)t, er, (fass)t, mein, (sieh(st, Vater, du Kind, den, Knaben, ihn, ihn, Sohn, (birg)st, du, dein

Weighting all words, one would obtain quite a different result, both in form of an indicator and in form of a function. However, in the present paper, we adhere to the first variant.

In Slovak, we will take a prosaic text, namely one by E. Bachletová (see Sources). The text comments on the need to create an artistically demanding audience. We weight the selected members of the chains and obtain the ranked results presented in Table 2.

Table 2
The ranking of weights of Belza-chain elements in a Slovak press text

Rank	Frequency	Zipf-Alekseev f.+1	Lorentzian f.
1	95	95.12	95.02
2	34	32.30	33.58
3	9	13.67	11.95
4	8	6.94	6.19
5	6	2.73	3.99
6	5	2.03	2.93
7	3	1.64	2.35
8	2	1.52	2.00
9	2	1.42	1.76

		a = 1.1836
		b = -0.5838
		c = 94.1226
		R^2 =0.9951
			a = 94.0255
			b = 1.0062
			c = 0.7237
			R^2 = 0.9972

Here, too, the first two ranks concern “weight”/classes 1 and 8. This means that in the chains, the core word manifests itself mostly in the form of inflection endings of the expressions grammatically linked to it.

For Indonesian, we took the pages 5 to 11 of the book *Burung api* by Pak Ojik (Djakarta: Pustaka Jaya 1971) and obtained the weights as presented in Table 3.

Table 3
The ranking of weights of Belza-chain elements in an Indonesian fairy-tale

Rank	Frequency	Zipf-Alekseev f.+1	Lorentzian f.
1	68	67.89	67.74
2	11	13.31	15.15
3	10	7.18	6.49
4	9	5.34	3-59
5	4	4.57	2.27
6	1	4.19	1-57

		a =-2.9122,
		b = 0.6776
		c = 66.8914
		R^2 = 0.9883
			a = 358374.904
			b = 0.1030
			c = 0.0123
			R^2 = 0.9804

As to English, we chose the text of the picture book *Miss Rumphius* written by Barbara Cooney, which narrates the life story of how Alice Rumphius has been planting flowers to make the world more beautiful. The results are shown in Table 4.

Table 4
The ranking of weights of Belza-chain elements in an English text

Rank	Frequency	Zipf-Alekseev f.+1
1	67	67.04
2	60	59.78
3	10	11.77
4	8	2.56
5	2	1.23
6	2	1.04

	a = 2.3676, b = -3.6578
	c = 66.0368, R^2 = 0.9875	

In order to investigate the situation in French, a press text on the use of air-conditioning devices in the scorching heat of summer 2018 has been utilized. The results, presented in Table 5, corroborate the good fit provided by both functions; moreover, the parameter *c* of the Zipf-Alekseev function expresses f_1 again. The most numerous weights in the text are 1 and 9, which points out the richness of derivational devices in French; this is also linked to the fact that in the article, the topic of electricity is treated in manifold ways.

Table 5
The ranking of weights of Belza-chain elements in a French press text

Rank	Frequency	Zipf-Alekseev f.+1
1	27	26.93
2	5	5.94
3	4	2.87
4	3	1.94
5	1	1.55

	a = 2.9587, b = 0.6051
	c = 25.9818, R^2 = 0.9944	

For Chinese, a widespread fairy tale titled *Grandmother Wolf* has been analyzed for the same purpose. The corresponding results are shown in Table 6.

Table 6
The ranking of weights of Belza-chain elements in a Chinese text

Rank	Frequency	Zipf-Alekseev f.+1
1	134	133.72
2	32	36.94
3	30	17.61
4	8	10.57
5	3	7.23
6	2	5.38

	a = -1.8728, b = -0.0172
	c = 132.7178, R^2 = 0.9831	

As to the Czech language, a press text concerning the tenth anniversary of a railway accident has been chosen; once again, the functions have shown satisfactory fits, with the objection to the Lorentzian one being the same as above. The top-scoring weights – 1 and 9 – confirm the synthetic tendencies present in Czech word-formation.

Table 7
The ranking of weights of Belza-chain elements in a Czech text

Rank	Frequency	Zipf-Alekseev f.+1
1	69	66.97
2	12	12.81
3	8	5.25
4	2	3.06
5	1	2.17
6	1	1.74

	a = -2.5274, b = 0.0033
	c = 67.9710., R^2 = 0.9968	

For Hungarian we analyzed the short story *Beszéd a kígyóròl meg más szörnyüségek* by G. Gárdonyi from his book *Az én falum.*

Table 8

The ranking of weights of Belza-chain elements in a Hungarian short story

Rank	Frequency	Zipf-Alekseev f.+1
1	86	85.99
2	79	79.05
3	15	14.46
4	1	2.80
5	1	1.24
6	1	1.04

	a = 2.5345, b = -3.8336
	c = 84.9896, R^2 = 0.9996	

For Italian, we analyzed the first chapter of the book *Intrigo* (*Un Incontro alle Grazie*) by Gianpaolo Pansa.

Table 9
The ranking of weights of Belza-chain elements in an Italian text

Rank	Frequency	Zipf-Alekseev f.+1
1	45	44.79
2	14	16.59
3	14	8.62
4	5	5.37
5	3	3.75
6	1	2.86
7	1	2.31
8	1	1.96

	a = -1.3168, b = -0.2497
	c = 43.7900, R^2 = 0.9730	

The results attained by this procedure indicate the following conjectures:

(1) The texts – in any language – exhibit a regularity which can be substantiated by the usual state of the language, the requirements of the author and the control by the reader, namely

(4) $$\frac{dy}{y} = \frac{A + B\ln x}{Dx}dx,$$

where the relative change of the frequency (*dy/y*) of weights behaves according to the state of the language (*A*), the requirements of the writer (*B ln x*) and is controlled by the requirements of the reader/hearer (*Dx*). Solving and reparametrizing the function, one obtains the Zipf-Alekseev function (see formula 2). All the above results confirm this relationship. Since (2) and (3) are parts of the usual theory (cf. Köhler 2005; Wimmer, Altmann 2005), it can be considered a preliminary law. Needless to say, it must be tested in many languages and many texts to obtain a better confirmation. One usually considers the change of *y* in relation to the *(y-1)* class. This is also done because the last classes – especially in short texts – contain, several times, the frequency 1, and the function would underestimate it.

(2) The elementary classification shown in Table 1 shows that texts are formed also according to unconscious lexico-grammatical relations. The text consists of chains, but the

chains have different length and are represented by relations of variegated weights – which are considered here merely as classes.

(3) Chains automatically express the Skinner hypothesis, but taking into account the distance between elements of chains conceals many complicated definitions and decisions. They are not solved as yet, but the above outcomes give the first hints. Evidently, it will be necessary to bring the whole into a lexico-grammatical harmony, to re-investigate the Latin based conceptions, and set up a quite general one containing more classes.

(4) One automatically asks: what are the other Zipf-Köhlerian properties with which the chains are associated? – Here, we are at the border of the Köhlerian system, and after some time, one will be able to insert the above relationship into it.

(5) One should not forget that mathematical modeling is a formal expression of our views, not the truth. Hence,exchanging a distribution for a function or for a sequence is no theoretical error, but a technical means. The simplest way of finding an appropriate function is the mechanical testing of some functions contained in the general theory (cf. Wimmer, Altmann 2005).

References

Altmann, G. (2018). The nature and hierarchy of Belza-chains. *Glottometrics 42, 75–85.*

Chen, R., Altmann, G. (2015). Conceptual inertia in texts. *Glottometrics 30, 73–88.*

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

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (eds.), *Quantitative Linguistics. An International Handbook: 760–774.* Berlin: de Gruyter.

Roelcke, Th., Popescu, I.-I., Altmann, G. (2017). Aspects of text concentration. *Glottometrics 36, 70–89.*

Skinner, B. F. (1957). *Verbal Behavior.* Acton: Copley Publishing Group.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In. Köhler, R., Altmann, G., Piotrowski, R,G (eds.), *Quantitative Linguistics. An Interntional Handbook: 791–807*. Berlin: de Gruyter.

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

Sources

Bachletová, E. Formovanie náročného publika – prežitok či nevyhnutnosť? *Otázky žurnalistiky 2006(3-4), 274-276).*

Wakim, N. Canicule : la climatisation accroît la consommation électrique.
Available at: < https://www.lemonde.fr/economie/article/2018/08/06/canicule-la-climatisation-accroit-la-consommation-electrique_5339763_3234.html>.

Honus, A. Deset let po pádu mostu ve Studénce: Viník je dosud bez trestu.
Available at: <https://www.novinky.cz/domaci/479954-deset-let-po-padu-mostu- ve-studence-vinik-je-dosud-bez-trestu.html>.

Pak Ojik, *Burung api.* Djakarta: Pustaka Jaya 1971.

Gárdonyi, G. *Beszéd a kígyóròl meg más szörnyüségek* from the book *Az én falum*. Available at: <http://mek.oszk.hu/00600/00657/html>.

Pansa, G. *Intrigo* (*Un Incontro alle Grazie*). Sperling & Kupfer 1993.

Cooney, B. *Miss Rumphius*. New York: Viking Juvenile 1982.

Wei, Y. *Lang Waipo*. Beijing: 21st Century Press 2013.

Glottometrics 43, 2018, 77-90

Some Properties of Polysemy

Sergey Andreev[1], Smolensk
Giuseppe G. A. Celano[2], Leipzig
Jiang Yang[3] , Xiangtan
Gabriel Altmann, Lüdenscheid

Abstract. In the article, the polysemy of individual words in some poems (English, Chinese, Russian, German, Italian and Slovak) is examined. Polysemy has its distribution, one can compute quantitative motifs omitting the line boundaries, positional representation, and a number of other problems mentioned below. The results can be modeled. We propose three models but the examination is not finished. A number of texts in every language are necessary, in order to show the differences between text types, authors, languages. It is rather a program for further research.

Keywords: Polysemy, motifs, modeling, positional trend, English, Chinese, Russian, German, Italian, Slovak

Introduction

Polysemy is usually restricted to words though it could be considered beginning with morphemes and ending somewhere in the domain of motifs or Belza-chains. It simply means the diversification of the meaning of words. It is well known that words increase their polysemy with age. This is caused by the fact that in the development of language, a word gets in contact with other words and the increasing polytextuality causes the increase of polysemy. In order to reduce the polysemy of a word the speaker specifies it by attaching to it some other specifying word elements, a process leading to increase of derivation and composition. Hence polysemy is one of the most central properties of the Köhlerian synergetic circuit (cf. Köhler 2005).

Monolingual dictionaries usually distinguish individual meanings of a word marking them by letters (a,b,c,…) or by numbers (1,2,3,… I,II,III,…) or both. Since in the text, words should be monosemic, with regard to the hearer, the study of polysemy in texts means rather the study of polysemy reduction.

There is a great number of studies concerning individual words and their polysemy, the polysemy of certain parts of speech, the distribution of polysemy in individual texts, constructing motifs of polysemy in texts, etc. Needless to say, the continuation of this study is infinite. Not only individual languages should be analyzed but also individual texts – a possibility given only for well documented languages. The number of vistas is infinite.

Here we shall restrict ourselves to some poetic texts in English, Chinese, Russian, Italian, and Slovak. We omit translations. The polysemy of individual words will be stated on the basis of the respective dictionaries.

The method used is as follows: For each word and line in the poem a numerical sequence of polysemies will be stated on the basis of the given dictionary. The sequence

[1] Sergey Andreev, Smolensk University_ smol.an@mail.ru
[2] Giuseppe Celano, Universität Leipzig, Institute for Computer Science: celano@informatik.uni-leipzig.de
[3] Jiang Yang, Hunan University of Science and Technology, Xiangtan, 411201, China, yangjiang@hnust.edu.cn

represents a vector and vectors can be evaluated in several ways. The individual polysemies will be counted and their frequencies yield a distribution which can be captured by a discrete or continuous function. Both motifs (in their numerical form) or their frequencies can be used for constructing polysemy motifs proposed by Köhler (2015). Further, it can be shown whether Skinner's hypothesis holds also in this domain: the lines (or strophes) placed nearer to one another exhibit greater polysemic similarity. That means, again, that the polysemic similarity of lines can be captured by a monotonically decreasing function. Further, the polysemies of a poem in the given order may be evaluated also vertically. Either ones takes the smallest/greatest polysemy of the column or the mean of the column. According to Zörnig (2016), there is a tendency in the resulting vector which can be modeled.

If a poem exhibits different behavior, then boundary conditions must be taken into account. The trends – if well corroborated - can be derived from a background theory, or a theory may be constructed basing on the resulting functions. Modeling the relations one should omit polynomials. If possible, one should interpret the parameters of the functions using synergetic arguments. One should not believe that one discovered the truth: the results are merely mathematical descriptions which can be mechanically used in further research.

Studying polysemy one must be aware of the fact that every dictionary marks them differently; that the results using two dictionaries may be different because of their extent; that every researcher has his own method of counting polysemy depending for example on the linguistic school, etc., hence to find a commonality is a problem which cannot be solved.

We considered the gollowing poems:
English; E.A.Poe, *A dream within a dream*
Chinese: *Zai bie kangqiao* (*Saying Good-bye to Cambridge Again*)
German: J.W.v. Goethe: *Der Erlkönig*
Russian: Nikolay Gumilev; *Devochka*
Slovak: A.B. Sládkovič_*Dcérka a mať*
Italian: A. Dante: *A ciascun'alma presa e gentil core*

1. The vectors

Taking a short poem, e.g. in English *A dream within a dream* by Edgar Allan Poe we obtain the following result. The polysemies stated from the dictionary are written under each word.

Take this kiss upon the brow!
19 7 9 1 8 3
And, in parting from you now,
6 33 3 5 3 12
Thus much let me avow—
4 6 9 5 2
You are not wrong, who deem
3 10 1 14 3 3
That my days have been a dream;
15 5 11 17 10 33 14
Yet if hope has flown away
7 5 7 17 22 14
In a night, or in a day,

33 33 12   3 33 33 11
In a vision, or in none,
33 33 8   3 33   5
Is it therefore the less gone?
10 8   1   8 9   31
All that we see or seem
13 15   6 14   3   3
Is but a dream within a dream.
10 13 33 14   4   33 14
I stand amid the roar
11 33   1   8   11
Of a surf-tormented shore,
21 33   1   6
And I hold within my hand
6   11 27   4   5   24
Grains of the golden sand—
19   21   8   7   9
How few! yet how they creep
13   5   7   13   3   12
Through my fingers to the deep,
21   5   12   21   8 18
While I weep--while I weep!
6   11 8   6   11 8
O God! can I not grasp
15 4   13 11 1   9
Them with a tighter clasp?
7   28 33   20   6
O God! can I not save
15 4   13 11 1   15
One from the pitiless wave?
16   5   8   1   14
Is all that we see or seem
10 13 15 6   14 3   3
But a dream within a dream?
13 33 14   4   33 14

Considering the line a unit, one can characterize the vector, e.g. by the mean polysemy of individual words, by other statistical properties (variance, Ord’s criterion, repeat rate, entropy, etc.). The indicators of lines may show certain tendencies and the tendency may be used as a characterization of the poem.

It can be conjectured that some words have a smaller polysemy than other ones. But that means that the vector of polysemies in a line automatically represents an oscillating movement. One may compute the differences between neighboring words, add them and divide by “number of words in the line minus 1” (= number of differences). In this way one obtains a characteristic serial polysemy indicator of the line. Considering the whole poem we may obtain a monotonic function of the mean line polysemy or an oscillating course. In any case we may construct an indicator of the poem polysemy. New questions arise that can be solved merely by a very intensive investigation: (a) Is the polysemy technique of an author constant? (b) Is the polysemy formation of a certain text type constant? (c) Is the polysemy formation in a language constant? (d) How is the evolution of polysemy distribution in the history of written texts? (e) Do special poem types, e.g. hexameter, have some "normed" polysemy distribution or do they differ from text to text, author to author, language to language? The number of questions will increase if one begins to perform research in this domain. Here we simply show some possibilities.

2. Properties

The distribution of polysemies in individual poems is a very pretentious task. The poems are usually short, the basic POS have a small polysemy, the relations between them expressed by prepositions, conjunctions, etc. are strongly polysemic. It depends on the structure of language and on the extent of the dictionary used, and on the way a researcher understands the dictionary, how the distribution of polysemies is structured. In some cases one can use a simple decreasing function but not in all cases. In order to overcome this problem we shall use a "trick" which is legal in these cases. We pool three neighbouring classes taking polysemy 1,2,3 as the class x = 2; the classes 4,5,6 as the class 5, etc. For the individual poems we obtain the results presented in Table 1.

Since we perform a pooling, we conjecture that the rate of change of the frequency *y* can be relativized by *y*^{2}. We obtain a differential equation

$$\frac{y'}{y^2} = \frac{2(x-b)}{a * c^2}$$

in which the parameter *b* is applied by the writer and the parameters *c* and *a* represent the language and the reader's effort respectively. Solving the equation and reparametrizing it we obtain

$$y = \frac{a}{1+\left(\frac{x-b}{c}\right)^2}\,.$$

Applying this formula, called Lorentzian function, to all our data we obtain the results presented in Table 1.

Table 1
Pooled polysemies in some languages and fitting the Lorentzian function

Polysemy	English		Chinese		Russian	
(pooled)	f_x	Lorentzian	f_x	Lorentzian	f_x	Lorentzian
2	43	42.57	83	83.03	99	99.04
5	46	47.01	47	46.64	48	47.37
8	25	19.86	14	17.44	17	21.46

11	3	9.03	17	8.35	22	11.55
14	2	4.94	0	4.79	0	7.09
17	2	3.07	2	3.09	2	4.76
20	3	2.09			7	3.41
23	0	1.50			3	2.56
26	1	1.13			2	1.99
29					1	1.59
	a = 54.3723 b = 3.7129 c = 3.2527 R^2 = 0.9733		a = 85.0138 b = 2.4369 c = -2.8261 R^2 = 0.9785		a = 104.0953 b = 1.2192 c = 3.4553 R^2 = 0.9774	

Polysemy	Italian		Slovak		German	
(pooled)	f_x	Lorentzian	f_x	Lorentzian	f_x	Lorentzian
2	23	23.26	68	67.78	128	127.96
5	32	32.58	24	27.56	46	46.83
8	15	16.54	24	14.78	23	20.01
11	10	7.73	8	9.18	12	10.68
14	5	4.19	2	6.25	2	6.56
17	0	2.58	1	4.52	2	4.42
20	5	1.73				
23	1	1.24				
26	0	0.93				
29	0	0.72				
32	0	0.58				
35	1	0.47				
	a = 32.5557 b = 4.3473 c = 3.7121 R^2 = 0.9773		a = 1247.5266 b = -3.0451 c = 1.2092 R^2 = 0.9589		a = 150.9986 b = 0.8069 c = -2.8115 R^2 = 0.9967	

Evidently, the parameters are not characteristic for a language but rather for the given poem. This automatically means that in order to bring a general characterization, many poems must be analyzed. In any case, the Lorentzian function expresses adequately the situation.

4. Positional representation

Now considering each column of a poem separately, we obtain for the Slovak poem the means of columns as follows:

1	2	3	4	5	6
5,28	4,47	4,03	5,1	1,9	2,4

Since in the fourth, fifth and sixth column the number of words is smaller and in the fourth column the mean is too great, one may smooth the course averaging all values in x = 4,5,6 obtaining a curve with y = 3.8 for x = 4. Since this is a smooth course, one may fit to it the exponential function and obtain the results presented in Table 2. In some cases, the exponential function is not adequate. We add one parameter and obtain the Menzerath function defined as

$$y = ax^b\exp(-cx);$$

used frequently in linguistics. Its differential equation distinguishes from the exponential one only by the constant *c* representing the influence of language.

Table 2
Positional consensus means

	Slovak		English		Chinese	
Position	Mean	Exponential f.	Mean	Menzerath f.	Mean	Exponential f.
1	5.28	5.17	6.42	6.52	4.76	4.74
2	4.47	4.61	6.29	5.79	4.07	4.18
3	4.03	4.12	4.33	5.09	3.86	3.68
4	3.80	3.67	4.81	4.45	3.16	3.25
	a = 5.7948, b = 0.1139, R^2 = 0.9560		a = 7.5267, b = 0.0353 c = 0.1435, R^2 = 0.7091		a = 5.3796, b = 0.1262, R^2 = 0.9609	

	Russian		Italian		German	
Position	Mean	Menzerath f..	Mean	Menzerath f..	Mean	Menzerath f.
1	6.30	6.30	8.78	8.76	3.25	3.26
2	4.78	4.79	6.57	6.67	4.34	4.30
3	5.03	5.02	6.71	6.58	4.19	4.25
4	6.01	6.01	7.16	7.21	3.76	3.73
	a = 3.8102, b = -1.1214, c = -0.5027, R^2 = 0.9999		a = 6.1688, b =-0.9003 c = -0.3510 , R^2 =0.9901		a = 4.9467, b = 0.9977, c = 0.4162, R^2 = 0.9916	

Since the texts have small units (lines) and they can have different length a pooling technique can help to find an adequate model. Perhaps, in prosaic texts it will be necessary, too, because sentences have different number of words and the word order may be quite different from that in European languages. The above procedure is merely one of the possibilities.

5. The properties of motifs

If one has the sequence of polysemies, motifs can be produced mechanically Using the polysemies of the English poem, we obtain (without taking line boundaries into account) the following motifs:

[11], [2,4], [1,5],[3,5,9], [2,5], [3,8]. [4], [2,7]. [5], [2,3,6], [1,7], [3], [2,4], [3,8], [1,4,5,6], [1,4,4], [1,9,9,9], [5,7], [3,9], [5,8,9], [5,6], [3,9], [3,6,7], [1,5], [3,7,7], [4,6,8], [3,3,6,8],

[5,6], [2,5], [2], [1,14], [1,5], [4,21], [5], [1,3,5], [1,10], [2,3,19], [14,21], [5,7], [5,9], [2], [1,9], [3,5,10], [3,6,16], [5,5], [3], [1,3,3], [1,3], [2,4], [3], [1,1,3,6,27], [5,17], [3], [2,4], [3], [1,1,6], [2,5,5], [1,7], [6,7], [4,6,8], [3,3,8], [5,6], [2,5], [2].

We conjecture that the length of motifs – measured in terms of the size of motifs – follows some regularity. Computing the length we obtain the results presented in Table 3. There are merely 5 length and their distribution is quite regular. Again, the Lorentzian function can be used.

Table 3
Length of English motifs

Length	Frequency	Lorentzian
1	11	10.02
2	33	33.26
3	17	15.92
4	2	5.05
5	1	2.29
a = 35.2113, b = 2.1803, c = -0.7445, R^2 = 0.9808		

The results of computation for the other languages are presented in Table 4.

Table 4
Polysemy motif lengths

Length	Russian		Chinese		Slovak	
	Frequency	Lorentzian	Frequency	Lorentzian	Frequency	Lorentzian
1	21	20.88	15	14.11	14	13.20
2	39	39,08	23	23.30	25	25.64
3	18	17.83	16	13.29	17	15.19
4	8	7.17	7	5.85	3	6.40
5	2	3.63	2	3.04		
6	1	2.15	1	1.82		
	a = 39,3252, b = 1.9225, c = 0.9813, R^2 = 0.9954		a = 23.9179, b = 1.9647, c = 1.1576, R^2 = 0.9420		a = 25.7635, b = 2.0780, c = -1.1050, R^2 = 0.9361	

Length	German		Italian	
	Frequency	Lorentzian	Frequency	Lorentzian
1	21	20.27	9	8.86
2	42	42.49	25	25.05
3	26	24.58	8	7.44
4	7	9.92	1	2.52
5	6	4.91	1	1.21
6	1	2.86		
	a = 42.850, b = 2.1997, c = 1.0421, R^2 = 0.9867		a = 25.2707, b = 1.9355, c = 0.6876, R^2 = 0.9936	

As can be seen, the Lorentzian function excellently captures the motif length.

The **range** of a motif is given by the difference of the greatest and the smallest number in the motif. If there is only one number, then the range is equal to this number. Computing the ranges from the above number for English we obtain:

[11,2,4,6,3,5,4,5,5,4,6,3,2,5,5,3,8,2,6,4,1,6,4,4,4,4,5,1,3,2,13,4,17,5,4,9,17,7,2,4,2,8,7, 13,0,3,2,2,2,3,26,12,3,2,3,5,3,6,1,4,5,1,3,2].

Here the ranges have the values beginning with 0 up to 26 but some of them are very seldom (because of short texts). For English, we obtain the results presented in Table 5.

Table 5
Ranges of motifs in the given English poem

	English	
Range	Frequency	Lorentzian
0	1	3.55
1	6	5.67
2	11	9.24
3	12	13.00
4	12	12.10
5	9	7.98
6	5	4.88
7	2	3.11
8	2	2.10
9	1	1.50
11	1	0.86
12	1	0.69
17	2	0.28
26	1	0.10

	a = 13.3859, b = 3.3461,
	c = 2.01012, R^2 = 0.9333	

In some cases, especially if the texts are too short, one must pool the frequencies. In the next table we show the pooled values of ranges. The pooling is simply a method which helps us to overcome some irregularities. We repeat that models do not represent truth but a method whose results can further be developed. If one wants to have all computed number > 1, one may add 1 to the Lorentzian function. In the differential equation it means merely to consider $y^2 - 1$. The German data are at the lower boundary of significance but this holds merely for the given poem. Further research will show whether in German the curve has two peaks.

Table 6
Ranges of motifs in the given poems (pooled)

	Slovak		Chinese		German	
Range	Frequency	Lorentzian	Frequency	Lorentzian	Frequency	Lorentzian
1	18	17.50	11	11.94	27	24.61
4	17	18.23	23	22.15	23	28.03
7	14	11.68	17	18.32	28	19.46
10	6	6.70	11	8.93	8	11.40
13	3	4.06	2	2.70	2	6.91
16	1	2.66			2	4.50

	a = 19.3088, b = 2.7096 c = 5.3116, R^2 = 0.9582		a = 23.2870, b = 4.9112, c = -4.0134, R^2 = 0.9667		a = 28.5921, b = 3.2199, c = -5.5201, R^2 = 0.8092
	Russian		Italiian		
Range	Frequency	Lorentzian+1	Frequency	Lorentzian	
1	17	16.89	6	6.18	
4	26	25.89	14	13.92	
7	18	19.13	7	7.24	
10	15	10.27	3	3.26	
13	1	6.04	4	2.08	
16	3	4.06	3	1.62	
19	4	3.03	0	1.40	
22	2	2.44	1	1.28	
25	2	2.06	0	1.21	
28	1	1.82	1	1.16	
	a = 25.0101, b = 4.3089, c = 4.3668, R^2 = 0.9313		a = 13.0362, b = 4.2479, c = 2.6370, R^2 = 0.9439		

In the same way the mean of quantitative motifs could be computed. The motif [1,5,6] has the mean 12/3 = 4 but the above results are sufficient. The determination coefficient is in all cases very high. In any case, one should consider longer texts.

6. POS-polysemy

As a matter of fact, parts of speech differ in polysemy. Usually prepositions and conjunctions are more polysemic than e.g. nouns. However, in strongly agglutinating languages where prepositions are mostly missing, e.g. in Hungarian, and are replaced either by affixes or adverbs, the situation may differ. This property can be used also in typology but up to now nobody tried to express it quantitatively.

Let us consider the English poem by Poe and replace the words by parts-of-speech abbreviations. We obtain using the abbreviations: N = noun, V = verbs, A = adjective, Pn = pronoun, Pr = preposition, Art = article, I = interjection, Av = adverb, C = conjunction, Pa = particle (e.g. “not”)

Table 3
POS-polysemies in the English text

Take this kiss upon the brow!	V A N Pr Art N
19 7 9 1 8 3	19 7 9 1 8 3
And, in parting from you now,	C Pr N Pr Pn Av
6 33 3 5 3 12	6 33 3 5 3 12
Thus much let me avow—	Av A V Pn V
4 6 9 5 2	4 6 9 5 2
You are not wrong, who deem	Pn V Av A Pn V
3 10 1 14 3 3	3 10 1 14 3 3
That my days have been a dream;	C A N V V Art N
15 5 11 17 10 33 14	15 5 11 17 10 33 14

Yet if hope has flown away	C C N V V Pa
7 5 7 17 22 14	7 5 7 17 22 14
In a night, or in a day,	Pr Art N C Pr Art N
33 33 12 3 33 33 11	33 33 12 3 33 33 11
In a vision, or in none,	Pr Art N C Pr Pn
33 33 8 3 33 5	33 33 8 3 33 5
Is it therefore the less gone?	V Pn Av Art Av V
10 8 1 8 9 31	10 8 1 8 9 31
All that we see or seem	Pn C Pn V C V
13 15 6 14 3 3	13 15 6 14 3 3
Is but a dream within a dream.	V C Art N Pr Art N
10 13 33 14 4 33 14	10 13 33 14 4 33 14
I stand amid the roar	Pn V Pr Art N
11 33 1 8 11	11 33 1 8 11
Of a surf-tormented shore,	Pr Art A N
21 33 1 6	21 33 1 6
And I hold within my hand	C Pn V Pr A N
6 11 27 4 5 24	6 11 27 4 5 24
Grains of the golden sand—	N C Art N N+
19 21 8 7 9	19 21 8 7 9
How few! yet how they creep	Av A C Av Pn V
13 5 7 13 3 12	13 5 7 13 3 12
Through my fingers to the deep,	Pr A N Pr Art N
21 5 12 21 8 18	21 5 12 21 8 18
While I weep--while I weep!	C Pn V C Pn V
6 11 8 6 11 8	6 11 8 6 11 8
O God! can I not grasp	I N V Pn Av V
15 4 13 11 1 9	15 4 13 11 1 9
Them with a tighter clasp?	Pn Pr Art A N
7 28 33 20 6	7 28 33 20 6
O God! can I not save	I N V Pn Av V
15 4 13 11 1 15	15 4 13 11 1 15
One from the pitiless wave?	Pn Pr Art A N
16 5 8 1 14	16 5 8 1 14
Is all that we see or seem	V A C Pn V C V
10 13 15 6 14 3 3	10 13 15 6 14 3 3
But a dream within a dream?	C Art N Pr Art N
13 33 14 4 33 14	13 33 14 4 33 14

Now, for each POS we compute the mean polysemy and rank the individual averages. If one orders the POS according to their frequency in text, one obtains a quite different ranking.We apply to this regularity the exponential function representing satisfactorily the trend. The results are presented in Table 4.

Table 4
Ranking of POS polysemies in the English text

POS	Rank	Frequency	Sum	Average	Exponential
Art	1	16	378	25.63	23.05
Pr	2	16	280	17.50	19.70
I	3	2	30	15.00	16.84
Pa	4	1	14	14.00	14.39
V	5	26	342	13.15	12.30
N	6	25	268	10.72	10.52
C	7	17	147	8.65	8.99
Pn	8	18	144	8.00	7.68
A	9	11	82	7.45	6.57
Av	10	9	55	6.11	5.61
	a = 26.9632, b = 0.1569, R^2 = 0.9451				

The other languages are presented in further tables. It must be remarked that the ordering according to simple frequency would yield a different function.

Table 5
Ranking of POS polysemies in the Chinese text

POS	Rank	Frequency	Sum	Average	Exponential
Num	1	6	60	10.00	9.29
A	2	9	59	6.56	7.77
Pr	3	10	64	6.40	6.49
U	4	20	110	5.50	5.42
V	5	36	188	5,37	4.53
Cl	6	4	15	3.75	3.79
Av	7	12	50	3.33	3.16
N	8	44	99	2.25	2.64
Pn	9	12	24	2.00	2.21
C	10	1	2	2.00	1.85
	a =11.1234 , b = 0.7962, R^2 = 0.9504				

In Chinese, three new classes have been added, namely U - auxiliary word, Num - numeric word, and Cl - classifier word; in Slovak: Part = particle.

Table 6
Ranking of POS polysemies in the Russian text

POS	Rank	Frequency	Sum	Average	Exponential
Pr	1	15	266	17.73	15.99
C	2	22	214	9.73	11.27
Pa	3	14	81	5.79	7.94
A	4	15	71	4.73	5.59
V	5	33	141	4.27	3.94
Pn	6	36	149	4.14	2.78

N	7	50	187	3.74	1.96
Av	8	12	35	2.92	1.38
Part	9	4	7	1.75	0.97
	a = 22.6962, b = 0.3501, R^2 = 0.9018				

Table 7
Ranking of POS polysemies in the German text

POS	Rank	Frequency	Sum	Average	Exponential
C	1	11	82	7.45	8.25
Art	2	15	110	7.33	6.96
V	3	41	264	6.44	5.87
Pr	4	10	55	5.50	4.95
Av	5	21	96	4.57	4.18
A	6	15	52	3.46	3.52
N	7	53	131	2.47	2.97
Pn	8	57	98	1,72	2.51
	a = 9.7765, b = 0.1700, R^2 = 0.9270				

Table 8
Ranking of POS polysemies in the Slovak text

POS	Rank	Frequency	Sum	Average	Exponential
Pn	1	20	78	7.65	7.32
C	2	10	53	5.30	6.08
A	3	20	99	4.95	5.05
Pr	4	11	78	4.36	4.20
V	5	25	108	4.32	3.49
N	6	44	148	3.36	2.90
Part	7	1	2	2.00	2.41
Av	8	1	1	2.00	2.00
Num	9	1	2	1.00	1.66
	a = 8,8133, b = 0.1855, R^2 = 0.9322=				

Table 9
Ranking of POS polysemies in the Italian text

POS	Rank	Frequency	Sum	Average	Exponential
Pr	1	10	159	15.90	15.78
C	2	7	62	8.86	9.29
N	3	20	133	6.65	6.76
V	4	23	151	6.57	5.77
Pa	5	2	12	6.00	5.39
A	6	10	54	5.40	5.24
Art	7	6	31	5.17	5.18
Av	8	7	35	5.00	5.16
Pa	9	12	50	4.17	5.15
	a = 5.1455, b = 27.3128, c = 1.0601, R^2 = 0.9781				

For the Italian text we used the exponential function with an added parameter, i.e. y = a+b*exp(-x/c). Needless to say, the number of texts should be multiplied in order to derive some general trends. In any case we see that texts contain a regularity which is created on the basis of the nature of the given language and on the intuitive use of some regularity. For the present, we attained satisfactory fittings using a very simple function. Automatically the question arises, how must be the structure of a language if it deviates from this pattern. Automatically the question arises, how must be the structure of a language if it deviates from this pattern.

7. Conclusions

Mathematical models are not text-inherent, they are our creations. The data must not be falsed but one can study in what form they correspond to the model, that means, one may search for data transformation. Above, we introduced a pooling by three classes because the frequencies were not sufficiently distinguished. Further analyses will show which technique is appropriate. One can also choose a model and ask under which condition the data can be fitted by it. In any case one should take longer texts.

We tried to model the problems without probability distributions, used quite simple models and obtained good results.

The comparison of languages is slightly problematic because the parts of speech do not correspond to the classical (Latin) view and both their number and their basis differs. For example, in some languages there are numeratives for which one can use both ususal words and special words having no ther meaning. But mutatis mutandis one could perform a ranking test. Nevertheless, mathematics yields a good possibility to find a common background.

Here merely a restricted number of data (languages, texts) has been used and some special problems were analyzed in order to show that polysemy has its regularities which are worth of further research.

It is already known that polysemy is associated with length of the word, the age of the word, the synthetism of language, the polytexty of the word, but the relationship to other properties from the Köhlerian contro cycle have not been stufied up to now.

References

Dante, A. *A ciascun'alma presa e gentil core.*

Dictionary Editing Office of Institute of Linguistics at Chinese Academy of Social Sciences. (2005). *Contemporary Chinese Dictionary (5th Edition)*. Beijing: The Commercial Press.

Editors of the American Heritage Dictionaries. (2015). *The American Heritage Dictionary of the English Language (Fifth Edition)*. Boston: Houghton Mifflin Harcourt. (available at: https://ahdictionary.com)

Gabrielli, A. (2018). *Grande Dizionario Italiano*, available at**:** http://www.grandidizionari.it/Dizionario_Italiano.aspx

Kačala, J, (ed.) (1989). *Krátky slovník slovenského jazyka*. Bratislava: Veda.

Kempcke, G. et al. (eds.) (1984). *Handwörterbuch der deutschen Gegenwartssprache*. Berlin: Akademie-Verlag.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 760-774.* Berlin: de Gruyter.

Köhler R. (2015). Linguistic Motifs. In: Mikros, G.K., Mačutek. J. (eds.) *Sequences in Language and Text: .89-108.* Berlin: de Gruyter 2015.

Kelih, E., Altmann, G. (2015). A continuous model for polysemy. *Glottometrics 31, 13-37.*

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook*: *760-774*. Berlin/New York: de Gruyter.

Levickij, V.V., Drebet, V.V., Kiiko, S.V. (1999). Some quantitative characteristics of polysemy of verbs, nouns and adjectives in the German language. *Journal of Quantitative Linguistics 6(2), 172-187*

Ortmann, W.D. (1975). *Hochfrequente deutsche Wortformen.* München: Goethe Institut.

Parker, P.M. (ed.) (2005). *Collected Works of Poe (Volume V)*. San Diego: ICON Group International (Poem: *A dream within a dream)*

Slovar' Russkogo Jazyka v chetyreh tomakh. (1984), 4-e izdanie, stereotipnoe 1999. Moskva: Izdatel'stvo «Russkij Jazyk»..

Skinner, B.F. (1957/1992). *Verbal behavior.* Acton: Copley Publishing Group

Sládkovič, A. (1976). *Poézia*. Edited by C. Kraus. Bratislava: Tatran. (Poem: *Dcérka a mat`*)

Xu, Z. (2016). *Zhimo's Poems*. Beijing: Writers' Press. (Poem: *Zai bie kangqiao*)

Zörnig, P. et al. (2016). *Positional Occurrences in Texts: Weighted Consensus Strings*. Lüdenscheid: RAM-Verlag.

History

In Remembrance of Fengxiang Fan, 1950–2018

A Pioneer of Quantitative Linguistics in China

Professor Fengxiang Fan died suddenly on August 19th, 2018, which is a shocking news to the entire quantitative linguistics community. People have since mourned for the loss of their dedicated and talented colleague, for he holds a prominent position in the field of quantitative linguistics in China. In remembrance of him, the following is a short statement of his life, findings, and ideas.

Fan was born on Dec 23th, 1950, in the Heilongjiang Province of China. Graduating from the University of Heilongjiang, where he majored in English, he pursued his graduate studies at Dalian Marine College (currently Dalian Maritime University) and University of Leeds, the UK, in the 1980s. Much of his career was intertwined with Dalian Maritime University, where he served as a professor from the 1980s until retirement in 2015. He was the vital force in the construction and development of the School of Foreign Languages at the university. The Department of Foreign Languages (currently the School of Foreign Languages) was established by him and his colleagues in 1998, and he served as the first dean of the department. From 2000 to 2004, he was also employed as a professor by the Faculty of Arts and Humanities of the University of Macau.

He was one of the pioneers of Chinese quantitative linguistics. He wrote several monographs and numerous journal articles, and edited several volumes. In 2008, he became one of the editors of the *Glottometrics* journal and of the *Studies in Quantitative Linguistics* book series. He co-edited *Problems in Quantitative Linguistics 1* with Strauss and Altmann (Strauss et al. 2008), the first monograph of *Studies in Quantitative Linguistics*. The editors collected recent problems and questions in the quantitative-linguistics field and recommended practical procedures for the implementation of theoretical hypotheses. This book can be used as a reference book for researchers in the field.

Above all, Fan was an expert in effective harnessing of computer programming languages for linguistic research (Fan 1995, 2005, 2010b, 2010c). His two books *Data Processing and Management for Quantitative Linguistics with Foxpro* (Fan 2010b) and *Quantitative Linguistic Computing with Perl* (Fan 2010c) give systematic descriptions and instructions on addressing common quantitative, computational, and corpus linguistic issues with two computer programming languages, Microsoft Visual FoxPro and Perl. The detailed programs were written for specific applications, such as lemmatization, creation of a word frequency spectrum, word length in syllables, etc., with specially devised exercises attached at the end of each chapter. With the clear purpose and accurate codes of each programme, novices can directly extract them for their own research. Two volumes were highly complemented (Hollósy 2011, Lei 2012, Feng 2015); as for what Hollósy (2011: 480) reviewed, Fan not only is an outstanding scientist, but also, he “has an excellent pedagogical gift, too". They were recommended as excellent textbooks for students and researchers in this field.

Moreover, he made important contributions to the application of quantitative methods to linguistic issues. One of his research interests lies in the characteristics of textual vocabulary of English – to name a few, the inter-textual vocabulary growth (Fan 2006a), the dynamic inter-textual type-token relationship (2006b), or the probability of textual vocabulary in the English language (Fan et al. 2016). Fan et al. (2016) studied the relationship between text size and the probability of textual vocabulary. The study reveals an interesting finding that as the text expands continually, instead of monotonically decreasing, the probability of the original textual vocabulary quickly reaches a point from which on it stabilizes, despite further expansion of the text. This indicates that a relatively small language sample can basically reflect the collective quantitative characteristics of textual vocabulary. It would be of interest to extend their finding further, as to whether this probability distribution patterns of textual vocabulary exist in other languages, and whether it is one of universals in human languages.

Another concept that intrigues Fan is textual vocabulary coverage (Fan 2006c, 2008a, 2013). It refers to the proportion of words of a text or a collection of texts covered by a given set of vocabulary. His study (Fan 2013) investigates the relationship between text length, vocabulary size, and text coverage. Results showed that text coverage is affected only by vocabulary size, not by text length, and the relationship between text coverage and vocabulary size can be captured by mathematical models. These findings are important for EFL/ESL research and teaching, helping educators to estimate text coverage better and determine the quantity of textual input to improve reading comprehension for the learner.

A further important contribution is his investigation on quantitative features of low-frequency word classes and especially hapax legomena (Fan 2010a, Fan et al. 2014a, 2014b). For instance, Fan (2010a) examined a baffling phenomenon that the ratio between text length and the ratio of hapaxlegomena to the vocabulary size (HVR) seems to be always about 50%. Through a computer simulation, the study revealed that the HVR is not constant at all; instead, it follows a U-shaped pattern and approaches the horizontal asymptote 1 as the text length approaches the infinity. The study on hapax legomena has implications on linguistic typology, authorship identification, and the degree of analytic features in a language, etc.

As a teacher at a maritime university, he examined quantitative lexical characteristics of maritime engineering English (Fan 2006c, 2008c), which may shed new lights on language teaching in the major of the field.

He also endeavoured to investigate other linguistic issues, i.e., the written English change (Fan 2007, 2012a), English compounds (Fan and Altmann 2007a, 2007b), the writing system of the English language (Altmann and Fan 2008, Fan and Altmann 2008a), meaning diversification in English (Fan and Altmann 2008b, Fan et al. 2008), word length distributions (Fan 2008b, Fan et al. 2010), word frequency spectrum (Fan 2012b), and so on (Fan 2012c, 2016, Fan et al. 2013). His empirical studies display not only high statistical and methodological standards, including ingeniously self-compiled computer programs and properly employed mathematical models, but also insightful and creative views concerning quantitative linguistics. It is qualified to regard him as an excellent quantitative linguist and an important part in the development of quantitative linguistics.

Despite his retirement, he never lost his zest for academia, though. He continued to serve as one of the editors of *Glottometrics* and occasionally reviewed linguistic journals. He also had an interest in the popular R software, which led to an article he had published recently in *Glottometrics,* combining R and quantitative linguistics (Xu et al. 2018). It is convincing that his contribution on R package would be another remarkable "stepping stone" (Fan, 2010c: 2) for linguists without any programming background – had it not been for his unexpected death.

As a teacher, he not only imparted the knowledge to students, but also readily resolved their doubts. He mentored about 60 graduates, according to CNKI[1] from 2000 to 2015. Most of their theses are based on corpus linguistics and quantitative linguistics, the scope of which ranges widely from business English, maritime English, Chinese EFL learner's writing, stylometry, to literary translation, etc.[2] More importantly, he provided advice and warm support for countless students and young scholars – and for whoever came for his help. His personal traits, i.e., selflessness, kindness, modesty, generosity, are undoubtedly as important as his academic accomplishments.

Although Fan is gone, all of his writings and thoughts are rich and enduring legacies left for the community. As long as these live, they give life to him. They will constantly inspire people to follow up his academic path and complete his unfinished work.

Yaqin Wang and Haitao Liu
Department of Linguistics, Zhejiang University, China (mail: htliu@163.com)

References

Altmann, G., & Fan, F. (Eds.). (2008). *Analyses of Script: Properties of Characters and Writing Systems*. Berlin: Mouton de Gruyter.

Fan, F. (1995). Application of SNOBOL4 to English teaching and research. *Journal of Dalian Maritime University (Social Sciences Edition),* 21(4), 25-29. (in Chinese).

Fan, F. (2005). Quantitative Linguistic Computing with Foxpro. In: Gabriel Altmann, Viktor Levickij, and Valentina Perebyinis (eds.). *Problems of Quantitative Linguistics: A Collection of Papers*. Chernivtsi: Ruta.

Fan, F. (2006a). A Corpus-based empirical study on inter-textual vocabulary growth. *Journal of Quantitative Linguistics*, 13(1), 111–127.

Fan, F. (2006b). Models for dynamic inter-textual type-token relationship. *Glottometrics*, 12, 1–10.

Fan, F. (2006c). Quantitative lexical description of marine engineering English. *Journal of Dalian Maritime University (Social Sciences Edition)*, 5(3), 161–164. (In Chinese.)

Fan, F. (2007). A corpus based quantitative study on the change of TTR, word length and sentence length of the English. In: Peter Grzybek and Reinhard Kohler (eds.). *Exact Methods in the Study of Language and Text*. Berlin: Mouton de Gruyter.

Fan, F. (2008a). A corpus-based study on random textual vocabulary coverage. *Corpus*

[1] It is a key national information construction project established in 1996 and now has built China Integrated Knowledge Resources System, including journals, doctoral dissertations, masters' theses, proceedings, newspapers, yearbooks, statistical yearbooks, e-books, patents, standards, and so on. The website is http://www.cnki.net/.

[2] To obtain a better understanding of the research of those graduates, the author listed the information of some of masters' theses in the appendix.

Linguistics and Linguistic Theory, 4(1), 1–17.

Fan, F. (2008b). An empirical study on syllabic word length and compounding propensity in technical English. In: Gabriel Altmann, Iryna Zadorozhna and Yuliya Matskulyak (eds.). *Problems of General, Germanic and Slavic Linguistics*. Chernivtsi: Ruta.

Fan, F. (2008c). Inter-textual vocabulary repetition of marine engineering English. *Journal of Dalian Maritime University (Social Sciences Edition)*, 7(2), 128–132. (In Chinese)

Fan, F. (2010a). An asymptotic model for the English hapax/vocabulary ratio. *Computational Linguistics*, 36(4), 631–637.

Fan, F. (2010b). *Data Processing and Management for Quantitative Linguistics with Foxpro*. Lüdenscheid: RAM-Verlag.

Fan, F. (2010c). *Quantitative Linguistic Computing with Perl*. Lüdenscheid: RAM-Verlag.

Fan, F. (2012a). A quantitative study on the lexical change of American English. *Journal of Quantitative Linguistics*, 19(3), 171–180.

Fan, F. (2012b). A study on word frequency spectra. In: Gabriel Altmann, Peter Grzybek, Sven Naumann and Relja Vulanović (eds.): *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens Verlag, p. 39–48.

Fan, F. (2012c). Review of Text and Language. *Journal of Quantitative Linguistics*, 19(2), 162–170.

Fan, F. (2013). Text length, vocabulary size and text coverage constancy. *Journal of Quantitative Linguistics*, 20(4), 288–300.

Fan, F. (2016). A study on segmental TTR, word Length and sentence Length. In: Emmerich Kelih, Róisín Knight, Ján Mačutek, Andrew Wilson (eds.). In: *Issues in Quantitative Linguistics 4*. Lüdenscheid: RAM-Verlag, p. 183–195.

Fan, F., & Altmann, G. (2007a). Measuring the cohesion of compounds. In: Volodymir Kaliuscenko, Reinhard Köhler and Viktor Levickij (eds.). *Problems of Typological and Quantitative Lexicology: A Collection of Papers*. Chernivtsi: Ruta, p. 190–209.

Fan, F., & Altmann, G. (2007b). Some properties of English compounds, In: Volodymir Kaliuščenko, Reinhard Köhler and Viktor Levickij (eds.). *Problems of Typological and Quantitative Lexicology: A Collection of Papers*. Chernivtsi: Ruta, p. 170–189.

Fan, F., & Altmann, G. (2008a). Graphemic representation of English phonemes. In: Gabriel Altmann and Fan Fengxiang (eds.). *Analyses of Script: Properties of Characters and Writing Systems*, Berlin: Mouton de Gruyter, p. 25–59.

Fan, F., & Altmann, G. (2008b). On meaning diversification in English. *Glottometrics*, 17, 66–78.

Fan, F., Grzybek, P. & Altmann, G. (2010). Dynamics of word length in sentence. *Glottometrics*, 20, 70–109.

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

Fan, F., Yu, Y., & Wang, H. (2013). Subjectival position and syntactic complexity in English sentences. In: Reinhard Köhler and Gabriel Altmann (eds.). *Issues in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag, p. 137–149.

Fan, F., Wang, Y, & Gao, Z. (2014a). Some macro-quantitative features of low-frequency word classes. *Glottometrics*, 28, 1–12.

Fan, F., Zhou, P., & Su H. (2014b). The use of the POR in macro-lexical analyses. In: Gabriel

Altmann, Radek Čech, Ján Mačutek and Ludmila Uhlířová (eds.). *Empirical Approaches to Text and Language Analysis*. Lüdenscheid: RAM-Verlag, p. 60–68.

Fan, F., Yu, Y., & Wang, Y. (2016). The Probability Distribution of Textual Vocabulary in the English Language. *Journal of Quantitative Linguistics*, 23(1), 49–70.

Feng, H. (2015). Review of Quantitative Linguistic Computing with Perl. *Australian Journal of Linguistics,* 35(2), 195–196.

Hollósy, B. (2011). Review of Data Processing and Management for Quantitative Linguistics with Foxpro. *Literary and Linguistic Computing*, 26(4), 1, 479–481.

Lei, L. (2012). Review of Quantitative Linguistic Computing with Perl. *Literary and Linguistic Computing*, 27(2), 227–230.

Strauss, U., Fan, F., & Altmann, G. (Eds.). (2008). *Problems in Quantitative Linguistics 1*. Lüdenscheid: RAM-Verlag.

Xu, Y., Yang, Y., & Fan, F. (2018). Quantitative Linguistics and R. *Glottometrics*, 42, 1–12.

Appendix

Bibliography of masters' theses supervised by Fengxiang Fan

Cao, K. (2008). *An empirical study on mathematical models for vocabulary growth* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Deng, Y. (2004). *Collocation Patterns of delexical verbs in Chinese EFL learners' writing* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Dong, H. (2012). *Analysis of modal verb transformation in the translation of Great Expectations* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

He, J. (2003). *Quantitative stylistic analysis of conversation in modern English novels* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

He, L. (2007). *A corpus-based study on the lexical change of the English language* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Huang, J. (2010). *A Corpus-based Study on the Lexical Characteristics of Commercial English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Ji, C. (2012). *A study on the structures of NPs in Written English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Li, J. (2006). *Inter-textual vocabulary growth patterns for marine engineering English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Liu, J. (2008). *A functionalist study on the translation of international maritime conventions* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Liu, Y. (2006). *Inter-textual lexical repetition in marine engineering English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Qi, X. (2009). *A corpus-based study on noun-phrase types and their syntactic functions* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Song, Y. (2006). *The distribution of hapax legomena in maritime engineering English (MEE)* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Sun, Y. (2003). *A computational stylistic analysis of Gone with the Wind and Scarlett* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Wang, F. (2002). *Lexical features of persuasive public speaking English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Wang, H. (2013). *The distribution of sentence-initial and sentence-final phrase length in written English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Wang, L. (2009). *A corpus-based study on the translation of Chinese political documents at the sentence and word levels* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Wang, Y. (2015). *A study on the distribution of dependency distances in different domains of written English in the BNC* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Xiao, Z. (2000). *Quantitative analysis of temporal clauses in T4* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Xu, Y. (2010). *The distribution of word families in college English textbooks* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Xue, H. (2007). *A computational stylistic analysis of Dan Brown's four novels* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Yan, S. (2001). *Statistical Analysis of Lexical Density in Corpora* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Yu, H. (2009). *Distribution of tenses in Chinese students' compositions* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Yu, Y. (2012). *A corpus-based descriptive study on the translation of high-frequency adjectives in international maritime conventions* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Zhang, X. (2007). *A descriptive study on two English versions of Hong Lou Meng* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Zhou, P. (2014). *A study on the subjectival position and the syntactic complexity in spoken English* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Zou, Y. (2002). *Textual cohesion and coherence of news script* (Master's thesis). Retrieved from CNKI's China Masters' Theses Full-text Database.

Other linguistic publications of RAM-Verlag:

Studies in Quantitative Linguistics

Up to now, the following volumes appeared:

1. U. Strauss, F. Fan, G. Altmann, *Problems in Quantitative Linguistics 1*. 2008, VIII + 134 pp.
2. V. Altmann, G. Altmann, *Anleitung zu quantitativen Textanalysen. Methoden und Anwendungen*. 2008, IV+193 pp.
3. I.-I. Popescu, J. Mačutek, G. Altmann, *Aspects of word frequencies*. 2009, IV +198 pp.
4. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics 2*. 2009, VII + 142 pp.
5. R. Köhler (ed.), *Issues in Quantitative Linguistics*. 2009, VI + 205 pp.
6. A. Tuzzi, I.-I. Popescu, G. Altmann, *Quantitative aspects of Italian texts*. 2010, IV+161 pp.
7. F. Fan, Y. Deng, *Quantitative linguistic computing with Perl*. 2010, VIII + 205 pp.
8. I.-I. Popescu et al., *Vectors and codes of text*. 2010, III + 162 pp.
9. F. Fan, *Data processing and management for quantitative linguistics with Foxpro*. 2010, V + 233 pp.
10. I.-I. Popescu, R. Čech, G. Altmann, *The lambda-structure of texts*. 2011, II + 181 pp
11. E. Kelih et al. (eds.), *Issues in Quantitative Linguistics Vol. 2*. 2011, IV + 188 pp.
12. R. Čech, G. Altmann, *Problems in Quantitative linguistics 3*. 2011, VI + 168 pp.
13. R. Köhler, G. Altmann (eds.), *Issues in Quantitative Linguistics Vol 3*. 2013, IV + 403 pp.
14. R. Köhler, G. Altmann, *Problems in Quantitative Linguistics Vol. 4*. 2014, VI + 148 pp.
15. K.-H. Best, E. Kelih (Hrsg.), *Entlehnungen und Fremdwörter: Quantitative Aspekte*. 2014, IV + 163 pp.
16. I.-I. Popescu, K.-H. Best, G. Altmann, *Unified modeling of length in language*. 2014. III + 123 pp.
17. G. Altmann, R. Čech, J. Mačutek, L. Uhlířová (eds.), *Empirical approaches to text and language analysis*. 2014, IV + 230 pp.
18. M. Kubát, V. Matlach, R. Čech, *QUITA. Quantitative Index Text Analyzer*. 2014, IV + 106 pp.
19. K.-H. Best (Hrsg.), *Studies zur Geschichte der Quantitativen Linguistik. Band 1*. 2015, III + 159 pp.
20. P. Zörnig et al., *Descriptiveness, activity and nominality in formalized text sequences*. 2015, IV+120 pp.
21. G. Altmann, *Problems in Quantitative Linguistics Vol. 5*. 2015, III+146 pp.
22. P. Zörnig et al. *Positional occurrences in texts: Weighted Consensus Strings*. 2016. II+179 pp.

23. E. Kelih, E. Knight, J. Mačutek, A. Wilson (eds.), *Issues in Quantitative Linguistics Vol 4.* 2016, 287 pp.

24. J. Léon, S. Loiseau (eds). *History of Quantitative Linguistics in France*. 2016, 232 pp.

25. K.-H. Best, O. Rottmann, *Quantitative Linguistics, an Invitation*. 2017, V+171 pp.

26.M. Lupea, M. Rukk, I.-I. Popescu, G. Altmann, *Some Properties of Rhyme.* 2017, VI+125 pp.

27. G. Altmann, *Unified Modeling of Diversification in Language*. 2018, VIII+119 pp.

28. E. Kelih, G. Altmann, *Problems in Quantitative Linguistics, Vol. 6.* 2018, IX+118 pp.

29. S. Andreev, M. Místecký, G. Altmann, *Sonnets: Quantitative Inquiries.* 2018, 129 pp.