Glottometrics 42
2018

RAM-Verlag

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

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. 42 (2018), Lüdenscheid: RAM-Verlag, 2017. Erscheint unregelmäßig.
Diese elektronische Ressource ist im Internet unter der Adresse
https://www.ram-verlag.eu/journals-e-journals/glottometrics/ verfügbar (Open Access).
Bibliographische Deskription nach 42 (2018) **ISSN 2625-8226**

Contents

Glottometrics 42, 2018, 1-12

Quantitative Linguistics and R

Xu Yingying, Yu Yang, Fan Fengxiang[1]
School of Foreign languages
Dalian Maritime University

Abstract. R is a vectorized language with combined features of a high-level computer language and dedicated software package. It has a wide range of string manipulation and pattern matching capabilities, of which its regular expressions are particularly useful. It provides a full-range of easy-to-use math and statistic functions. In addition, there are also versatile plotting systems for data visualization. It is a powerful tool for quantitative linguistic computing.

Keywords: R, quantitative linguistics, string manipulation, mathematical operation, statistical function, visual presentation

1. Introduction

Empirical researchers in linguistics, in particular in quantitative linguistics, rely to a high degree on the acquisition of large amounts of appropriate data and, as a matter of course, on sometimes intricate computation (cf. Köhler, from *Preface to Quantitative Linguistics Computing with Perl* published by RAM-Verlag).

In a research article, Mark Pagel et al (2007) raised an interesting question: why, being speakers of the Indo-European tongues, Greeks say *ουάρ*, Germans *schwanz* and the French *queue* to describe what Britons call a *tail*, but all of them use a related form of *two* to describe the number after one? Using big data extracted from mega-corpora with Bayesian Markov chain Monte Carlo (MCMC) framework and statistical models such as word evolution likelihood and lexical replacement regression, they found that frequency affects lexical evolution, and that for words such as *die*, *night*, *tongue*, *two* etc to evolve into their cognates it would take at least 10,000 years while for words such as *dirty*, *gut*, *stab*, *turn*, etc this time span would be shortened by a factor of nine!

Another less otherworldly linguistic inquiry was made by Fan (2013) into the relationship between vocabulary size and text coverage. Employing a 100-million word corpus with complicated data extraction and processing, he discovered that for the same set of vocabulary its text coverage is a constant regardless of the length of the covered text, and that the relationship between vocabulary size and text coverage can be captured by the re-parametrized mathematical models of Altmann (1980), Tuldava (1995) and Köhler & Martináková-Rendeková (1998).

Both the above examples share the same quantified linguistic research paradigm: the use

[1] Fan Fengxiang, Univ. Dalian (China), E-Mail: Fanfengxiang@yahoo.com

of huge amounts of data, intensive string manipulation and number crunching, and the establishment of mathematic models, apart from sets of illuminating figures. Research of this nature enhances its descriptive and explanatory power and is gaining popularity in the linguistic circles. However, without the use of a computer language, research like the above is difficult or even impossible to carry out.

Of the impressive array of high-level computer languages, one is particularly suitable for such exacting tasks—the R language, which can single-handedly do data extraction and processing, mathematic modeling, statistic testing and data visualization all at one sitting so that researchers do not have to lug cartloads of raw data around from one computing tool to another.

R is a freeware initially written by Ross Ihaka and Robert Gentleman at the Department of Statistics at the University of Auckland in the early 1990's and, as we can see, the initials of its creators were used to name the language. After more than 20 years of development, R has become a full-fledged powerful computer language. Now the language is maintained by the R core team, of which the two creators are members. Its functionality and flexibility, as well as its popularity, are increasing by the day thanks to the R core team and those outside in different parts of the world contributing packages of various functionality to R, numbering more than 10,000 as of the time of writing. The latest version of R is R 3.50 released on April 23, 2018, which can be downloaded from https://www.r-project.org/. R is a stand-alone computer language with a console that has menus for file and package management, and program (script) editing and running, apart from many other functions, such as executing commands entered in it and displaying the results etc. Figure 1 is the R console:

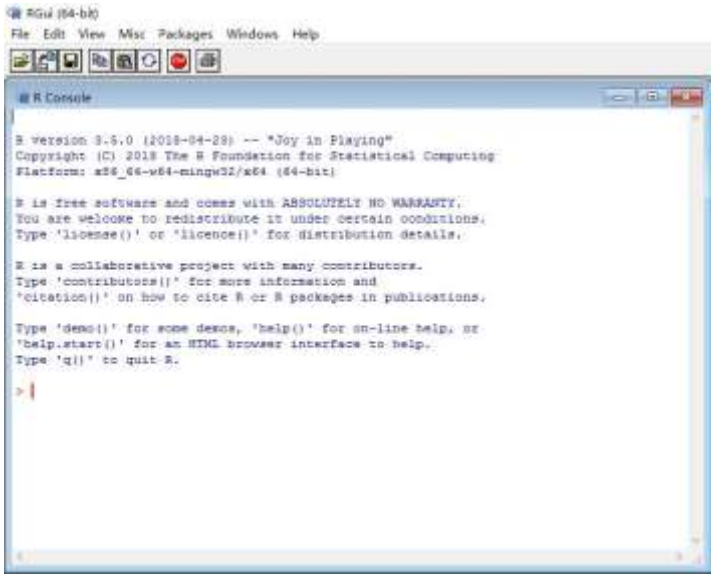

Figure 1. The R console.

However, there is a freeware IDE (Integrated Development Environment) for efficiently managing and running R called RStudio, which can be downloaded from www.rstudio.com. It has a more sophisticated console with four panels: R Source (for writing, editing and running R programs), R Console, Files (for displaying files, plots, packages etc), and Environment

(for displaying variables etc), each can be maximized or minimized and their positions can be freely arranged to suit the user's preference. Figure 2 is the RStudio's console:

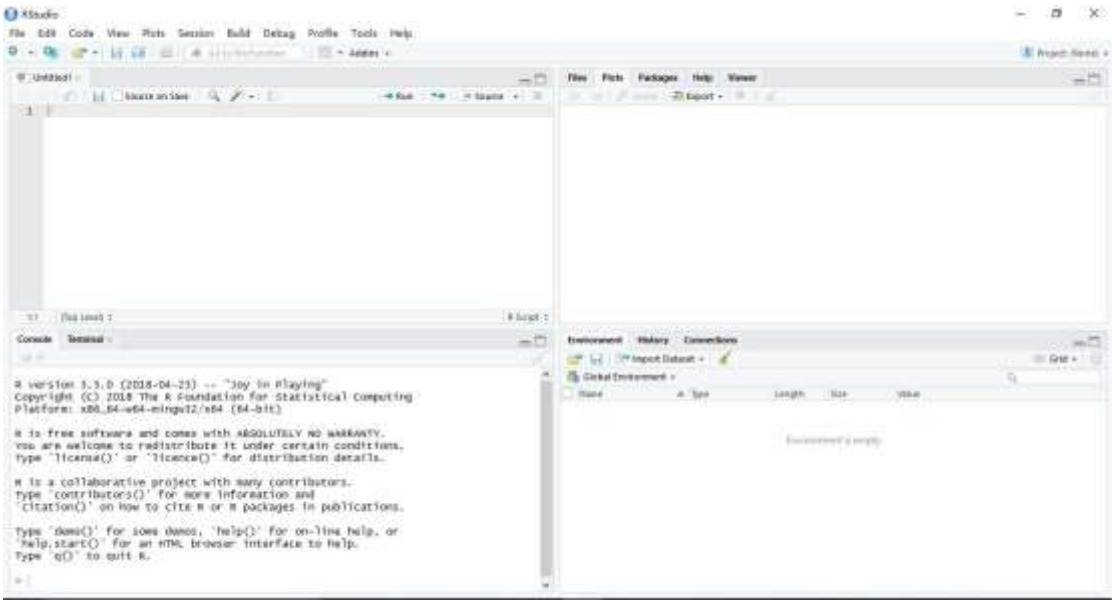

Figure 2. The RStudio console.

There are many good books on R for different readers, and two of which are *R in a Nutshell* by Joseph Adler published by O'Reilly Media, Inc, and *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani published by Springer. There are also some books on R dedicated to language research, one of which is *Analyzing Linguistic Data, a practical introduction to statistics* by R. H. Baayen published by Cambridge University; the title is a bit misleading but its contents are actually on "Analyzing Linguistic Data with R". The draft version of Baayen's book, as well as *An Introduction to Statistical Learning with Applications in R*, can be freely downloaded at http:// freecomputerbooks.com/langRBooks.html.

2. A vectorized language

R seems peculiar to people familiar with other computer languages. Firstly, R uses two symbols for assigning values to variables: <- and =. For example, *number <- 23*; *fruit* = "*apple, banana, peach*". In this respect, <- and = are the same. However, for choosing a setting within a function only = can be used, such as *formatC("WORDS", width = -25)*. By convention <- is generally used for assigning values to variables, although = can do the same thing. Secondly, it does not have the scalar variable but a set of different data structures such as the vector, the data frame, the list, the factor, the matrix, etc, apart from the common array. People who have already learned a computer language other than R may find R strange and would initially stumble over these seemingly similar data structures, with the scalar variables common in other languages missing. However, for those learning R as their first computer language this poses no problem at all but they would feel at a loss at the beginning when learning another computer language because of the absence of these handy data structures. R

is a vectoriszed language, that is, its basic variable is the vector, a data structure for an ordered set of data in which all the components of the data are numbered starting from 1 and can be indexed by their numbers. Assigning "Hello World!" to a variable *greetings* in the R console using the strange looking value assignment symbol<-:

```
greetings <- "Hello World!"
```

you have created a vector named *greetings* that has only one element. Type *greetings* in the console and press Enter, the following is displayed:

```
[1] "Hello World!"
```

[1] means *Hello World!* is the first element of the vector. Other R data structures have similar ways of indexing the data components stored in them. The vectorized feature of R may seem a bit cumbersome adding a number to everything stored in its variables, but actually it is a blessing in disguise. It is very efficient and easy for data manipulation. One simple example can show this convenience. To divide 237, 342, 812, 920, 99, 473, 110, 998, 34, 88, 12, 7887 respectively by 4740, one just puts them in a vector, say, *data*:

```
data <- c(237, 342, 812, 920, 99, 473, 110, 998, 34, 88, 12, 7887)
```

then enter data/4740 at the command line. The result is also a vector as shown below:

```
[1] 0.0500000 0.0721519 0.1713080 0.1940928 0.0208861 0.0997890
[7] 0.0232068 0.2105485 0.0071730 0.0185654 0.0025316 1.6639241
```

This vectorized computing avoids complex loops common in other languages performing similar routines, hence greatly reducing the length of an R program and makes it more readable; this feature is particularly welcome by students of linguistics learning R who generally are scared of program loops, traps easy for them to fall into but difficult to get out. To further illustrate the advantage of vectorization, take the computation of entropy and arc length as examples. The former is computed with (1) whilst the latter (2)

(1) $H = -\sum_{x \in X} p(x) \log_2 p(x)$ (Manning and Schutze,. 1999, p 61)

(2) $L = \sum_{r=1}^{V-1} \{[f(r) - f(r+1)]^2 + 1\}^{1/2}$ (Popescu et.al, 2013)

Suppose *data* contains word frequencies, and the total number of words is 237 + 342 + 812 + 920 + 99 + 473 + 110 + 998 + 34 + 88 + 12 + 7887 = 12012 (the sum function in R does this). An R script computing *H* using *data* is as follows:

```
data <- c(237, 342, 812, 920, 99, 473, 110, 998, 34, 88, 12, 7887)
wordprobility <- data /sum(data)
wordentropy <- wordprobility *log2(wordprobility)
H <- -sum(wordentropy)
```

The result is 1.889976. To compute *L* using *data*:

```
data <- c(237, 342, 812, 920, 99, 473, 110, 998, 34, 88, 12, 7887)
numberofwords <- length(data)
fr <- data[1:(numberofwords - 1)]
```

```
fr1 <- data[2:numberofwords]
arclength <- sum(((fr - fr1)**2 + 1)**(1 / 2))
```

The result is 12098.03. The two little programs are loop-free and straight forward. Using other languages, the scripts would be much longer and more involved because of the complex loops entailed, hence mind boggling.

3. A powerful string cruncher and math performer

One of the nightmares of quantitative linguists is to winnow down a sea of data to extract useful information for the research at hand; another may be arranging such data in different ways so as to find patterns or laws or both hidden in these data or to metricize them. With R such burdens can be greatly reduced. R has a set of regular expressions compatible with those of Perl plus many easy-to-use string manipulation functions. For example, in the following word list called *word* consisting of 14 words: *begin*, *begins*, *beginning*, *began*, *begun*, *beginner*, *beginners*, *begs*, *begged*, *begging*, *beggar*, *beggars*, *beggaring*, *beggared*, we want to return the variations of the word *begin* to their original form keeping other words unchanged. This is done with just one statement in R:

```
gsub("beg.n+e?r?s?i?n?g?", "begin", word)
```

The result is *begin*, *begin*, *begin*, *begin*, *begin*, *begin*, *begin*, *begs*, *begged*, *begging*, *beggar*, *beggars*, *beggaring*, *beggared*. Suppose we want to turn the following short text of 10 words into a trigram: *R is a vectorized computer language suitable for quantitative linguists*; the following R statement does it:

```
paste0(text[1:(10 - 2)], " ", text[2:(10 - 1)], " ", text[3:10], "\n")
```

The result is as follows:

```
R is a
is a vectorized
a vectorized computer
vectorized computer language
computer language suitable
language suitable for
suitable for quantitative
for quantitative linguist
```

This seems incredibly simple and easy; again, with other language it would be a mind bender.

R has many built-in functions for string manipulation. For example, if we want to find the union, intersection and difference between the following sets of data. set1: *a*, *b*, *c*, *d*, e; set2: *d*, *e*, *f*, *f*, *g*.

```
union(set1, set2)
intersect(set1, set2)
setdiff(set1, set2)
```

The results are shown below:

```
a b c d e f g
```

d e
a b c

It is not difficult for us to see the potential applications of such natural language like functions in processing textual data.

R has a full spectrum of mathematical and geometrical functions and operators for us to pick, ranging from those for elementary math operations to those for calculus and others. Look at the following two integration problems (Wheeler and Peeples, 1986, p591, p621):

$$\int_9^{12} \frac{1}{x^2\sqrt{225-x^2}}\,dx$$

$$\iint_R 3y^2e^{2x+5}\,dxdy$$

$R = \{(x, y) \mid 0 \le x \le 2,\ 1 \le y \le 3\}$

The first one can be solved with the *integrate* function in R, the second with the *dblquad* function of the *pracma* package for R:

```
compute <- function(x) 1/(x**2*sqrt(225-x**2))
integrate(compute,9, 12)
```

The result is 0.002592593.

```
compute <- function(x, y) 3*y**2*exp(2*x + 5)
dblquad(compute, 0, 2, 1, 3)
```

The result is 103410.7. The next example is to find the dot product of two matrices *A* and *B*:

$$A = \begin{bmatrix} 2 & 4 \\ 3 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

We can use the *%*%* operator for matrix inner multiplication:

```
A%*%B
```

The result is as follows:

```
     [,1] [,2]
[1,]   10   22
[2,]   13   29
```

4. All-inclusive statistical functionality

Perhaps one of the reasons that a computer-language-shy quantitative linguist may be lured into the strange world of R would be its statistical functionality. After all the hassle of data extraction, no sigh of relief can be heaved because of the loads of statistical work ahead: is the distribution Gaussian, Gamma, binomial or Poisson? What is the Mahalanobis distance between the two sets of data? Is the covariance homogeneous? What test to use, parametric or

non-parametric? What classification methods to use, logistic regression, K-means cluster, linear or quadratic discriminant analysis? What non-linear regression models to use to fit the data, Shenton-skees geometric distribution or the 1-displaced extended binomial distribution and so forth. R is a statistical all-rounder with functions and facilities covering practically all the areas of statistics. The following are some examples.

The Shapiro-Wilk Normality Test is used to check whether a set of data is normally distributed. R has a function for this test: *shapiro.test*. Suppose there is a set of data in *A*, and *A* = 0.17, 1.03, 0.23, 0.58, 1.19, 0.13, 0.05, 0.90, 0.67, 0.91, 2.12, 0.27, 0.53, 0.47, 0.43, 0.15, 0.10, 0.54, 3.67, 0.01. To test whether *A* is normally distributed:

```
shapiro.test(A)
```

The result is shown below:

```
Shapiro-Wilk normality test
data:  A
W = 0.71252, p-value = 5.563e-05
```

The *p* value is much smaller than 0.05, indicating A is not normally distributed.

Altmann (1980) proposed that the longer a linguistic construct the shorter its constituents, which is now known as the Menzerath-Altmann law. He mathematically presented this theory with the following differential equation:

$$\frac{y'}{y} = -c + \frac{b}{x} \tag{3}$$

which can be solved into the following:

$$y = ax^b e^{-cx} \tag{4}$$

where *y* is the (mean) size of the immediate constituents, *x* is the size of the construct, and *a*, *b* and *c* are parameters which seem to depend mainly on the level of the units under investigation (Köhler 2012, p. 147). Now we use (3) to fit the word length (in number of syllables) distribution of the 100000000-word BNC Corpus with R's *nls* (non-linear least squares) funtion. The word length data is as follows:

Length	Frequency
1	68372337
2	17994755
3	7422756
4	3053470
5	935713
6	151518
7	15639
8	2519
9	378
10	76
11	25
12	8

The following R statement does the fitting:

```
fit <- nls(frequency ~ a*length**b*exp(-c*length), start = list(a = 3, b = 1, c = 3), control
= list(maxiter = 500),    algorithm = "port")
```

The result is as follows: a = 114700000, b = −1.157, c = 0.5177, R^2 = 0.9997634.

Köhler (2014) used the mixed Poisson model to capture the relationship between compound length and the corresponding frequency. It is also called the two-Poisson in natural language processing (Manning and Schutze, 1999，p548). It is computed with the following:

$$P_{x=} \frac{\alpha e^{-\lambda}}{x!} + \frac{(1-\alpha)e^{-\mu}\mu^{x}}{x!}, x = 1, 2, 3, \ldots$$

Now we use it to fit the *NP* length and corresponding number of patterns in the written section of the ICE-GB Corpus (Wang, 2012). The data is as follows:

Length	Pattern	Length	Pattern
1	4	8	128
2	53	9	52
3	183	10	21
4	341	11	10
5	439	12	8
6	405	13	4
7	244	15	2

The R statement is as follows:

```
fit <- nls(y~pi*exp(-lambda1)*lambda1**x/factorial(x) + ((1-pi)*exp(-lambda2)
*lambda2**x)/factorial(x), start = list(pi = 6, lambda1 = 1, lambda2 = 0.1),
control = list(maxiter = 500), algorithm = "port")
```

The result is as follows: *pi* = 2149, *lambda1* = 5.491, *lambda2* = 0.02211, R^2 = 0.9140636.

5. Graphic master

The R base has graphic functions for scatter plot, line chart, pie chart, barplot, boxplot, and histogram. These functions are very simple and easy to use. Suppose we have three sets of data *A*, *B* and *C*. *A* = 10, 23, 30, 34, 37, 39, 41, 44, 47, 50, 55, 60, 63, 69, 71, 78, 82, 88, 94, 101; *B* = 24, 28, 29, 35, 45, 50, 61, 87, 90, 110, 121, 130, 133, 140, 155, 159, 167, 170, 178, 190. To make a scatter plot of *B* with *A* as the *x* axis:

```
plot (A, B)
```

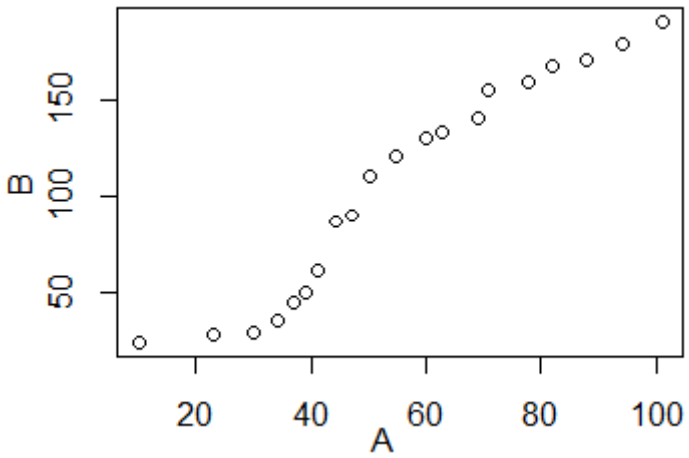

Figure 3. Scatter plot.

To produce a line chart with *A* as the *x* axis:

```
plot(A, B, type = "l")
```

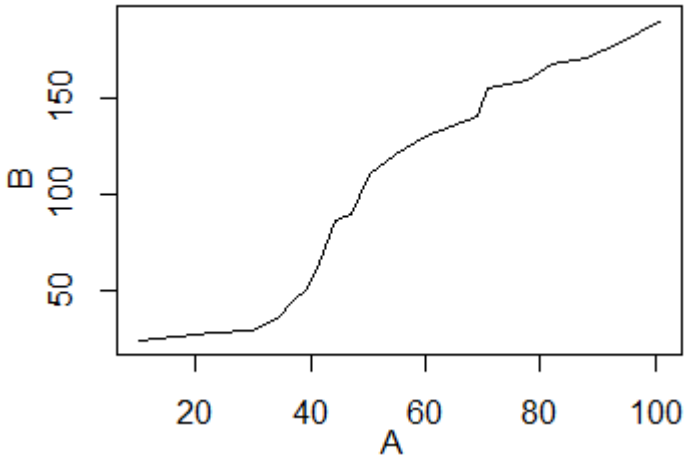

Figure 4. Line chart.

To produce a pie chart for A:

```
pie(A)
```

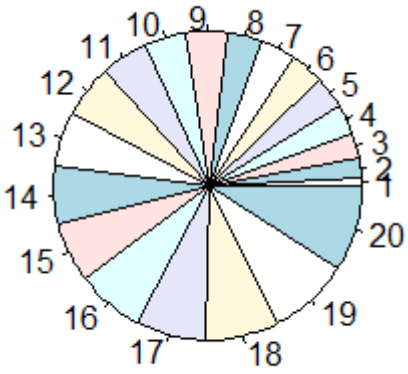

Figure 5. Pie chart.

To produce a barplot for A:

barplot(A)

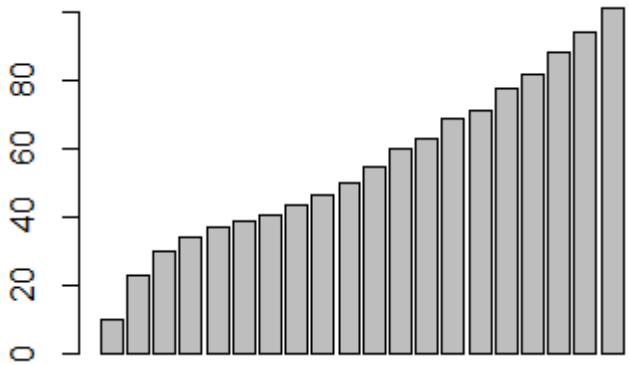

Figure 6. Barplot.

To produce a boxplot for *A* and *B*:

Boxplot (A, B)

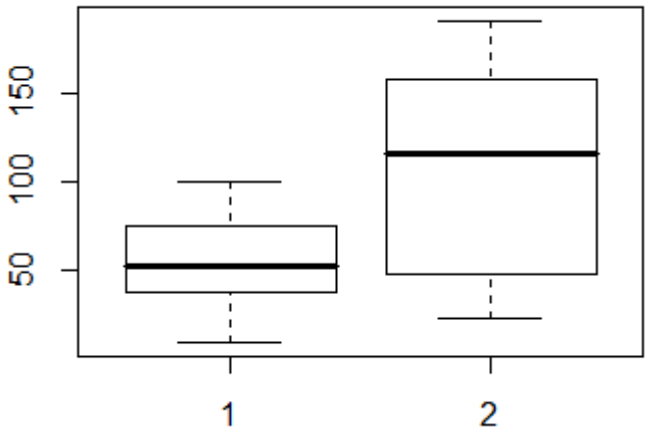

Figure 7. Boxplot.

To produce a histogram with number of bins set to 4:

```
hist(A, nclass = 4)
```

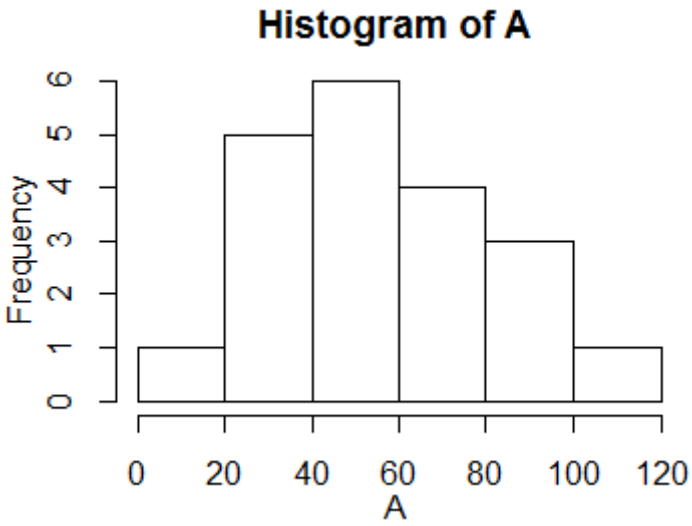

Figure 8. Histogram.

We used only the bare minimum of these functions. As a matter of fact, there are different settings for these functions for the selection of colour, line types (solid, dotted, etc), point size and so on.

There are two packages for more sophisticated graphics, the *lattice* package and *ggplot2*, and the latter is very popular and can produce a full-range of elegant, high quality graphics.

6. Concluding remarks

R is a powerful, flexible, all-purpose and user-friendly computing tool for quantified linguistic research. Its syntax is simple and straight forward. It is vectorized and has the combined features of a high-level language and a dedicated software package. Unlike many other commonly used computer languages, R is equipped with a console. Short R commands can be entered in it with results displayed instantly after pressing Enter. This makes it extremely handy since R codes do not have to be written in the form of a program and then run it to perform household tasks such as $889.7^{(2.13/210)}-334$.

Everything has two sides. R's speed for string processing is relatively low. For example, it takes about 2 seconds for Perl to tokenize, lemmatize and compute the vocabulary size and word frequency of a 2000000-word text, but it is about 5 seconds using R. However, considering the time and effort saved by not hopping around different computing tools for math modeling, statistics and graphics, this slowness is more adequately offset; in addition, what significant difference does it make between

2 and 5 seconds? Thanks to the contributed packages, R's string manipulation power and speed are increasing by leaps and bounds. It seems R was born for quantitative linguistics, and to quantitative linguists R is a handy and efficient tool facilitating their research.

References

Altmann, G. (1980). Prolegomena to Menzerath's Law. In: *Glottometrika 2, 1-10*. Bochum: Brockmeyer.

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

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

Köhler, R. (2012). *Quantitative Syntax Analysis*. Mouton de Gruyter, Berlin/Boston.

Köhler, R. (2014). Towards a theory of compounding. *Glottometrics* 28, 75-86.

Köhler, R., Martináková-Rendeková, Z. (1998). A systems theoretical approach to language and music. In G. Altmann & W. A. Koch (Eds.), *Systems. New Paradigms for the Human Sciences* (pp. 514–546). Berlin: Walter de Gruyter.

Manning, C., Schutze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Pagel, M., Atkinson, Q., Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history *Nature* 11: 717-721.

Popescu, I.-I., Zörnig, P., Almann, G. (2013). Arc length, vocabulary richness and text size. *Glottometircs*. 25, 43-53.

Tuldava, J. (1995). *Methods in Quantitative Linguistics*. Trier: Wissenschaftlicher Verlag Trier.

Wang, H. (2012), Length and complexity of NPs in Written English, *Glottometrics 24, 79-87.*

Wheeler, R.E., Peeples, W.D. (1986). *Modern Mathematics with Applications to Business and the Social Sciences*. Montery: Brooks/Cole Publishing.

Glottometrics 42, 2018, 13-20

On Word Length in German and Polish

Otto Rottmann

Abstract. Word length is one of the most examined properties of language. It can be measured in terms of syllable or morpheme numbers. In the present article we bring some new models.

Keywords: Word length, Zipf-Alekseev function Ord's criterion, repeat rate, German, Polish

Until recently, one used – for modelling word length – some kind of probability distribution. One used a family of discrete univariate distributions and modified them if it was necessary. Sometimes, there were some problems with testing, which was performed by means of the chi-square test requiring values at least of 5 in individual classes; otherwise, one was forced to pool some neighboring ones. Popescu, Best, and Altmann (2014) showed that any phenomenon of length in language can be modelled by the same function, namely by the Zipf-Alekseev function. Although this function is continuous, methodologically there is no difference. The discrete and the continuous do not exist in the real nature – they are our scientific concepts, just as the names of language entities. Until 2014, thousands of data have been evaluated, but it could be shown that one sole function can capture all phenomena of length in language. The authors, text types, levels, languages, etc., differ only in the parameters of the function.

For recapitulating some of the results, we take the data published by A. Kühner (2007), who applied the 1-displaced Hyperpoisson and the 1-displaced Poisson distributions with very good results. Unfortunately, within 20 poems by Goethe, 7 did not want to abide by the Hyperpoisson, and in many cases, the probability of the chi-square was too small.

Although the usual frequency data are discrete distributions, one can model them using a continuous function with discrete values of the independent variable because mathematical models do not represent the truth, they are merely formal concepts which can be further formally processed. As has been show in several publications, one can use usually the Zipf-Alekseev function defined as

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

which is the solution of the differential equation given as

$$y`/(y-1) = (A+B \ln x)/Cx,$$

taking into account the fact that the relative rate of change of *y* depends on the previous value, *y*-1. Reparametrizing the result, we obtain formula (1). In what follows, we shall use the formula for fitting it to the same text type of the same author, namely the poems by J. W. von Goethe written in different years, as shown in Table 2 and applied by I. Kühner (2007). Kühner prepared a table of the examined texts and added to all the poems the years of creation, as can be seen in Table 1.

Table 1
20 texts by Goethe with years

Text Nr	Text and first appearance	Text Nr	Text and first appearance
1	Hochzeitlslied (1767)	11	Harzreise im Winter (1777)
2	An Schwager Kronos (1774)	12	Das Göttliche /1783)
3	Prometheus (1774)	13	Der Besuch (1788)
4	Smbolum (1815)	14	Wiederfinden (1815)
5	Willkomm und Abschied (1789)	15	Morgenklagen (1788)
6	Lauf der Welt (1825)	16	Philine (1795)
7	Vermächtnis (1829)	17	Magisches Netz (1803)
8	Der Becher (1776)	18	Adler und Taube (no year)
9	An Frau von Stein (1784)	19	An den Mond (1789)
10	Grenzen der Menscheit (1778)	20	Seefahrt /1776)

Table 2
Twenty poems by J.W.v. Goethe

Length	T1		T2		T3		T4	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	92	91.99	91	91.00	130	129.96	105	105.00
2	41	41.06	56	55.98	65	65.44	50	50.01
3	9	8.49	16	16.09	24	21.97	6	5.92
4	1	2.33	5	4.84	5	7.86	1	1.42
8	–	–	–	–	1	1.16	-	-
	a = 0.6785 b = -2.6878 c = 90.9949 R^2 = 0.9996		a = 0.8529 b = -2.2552 c = 90.0018 R^2 = 1.0000		a = 0.1141 b = -1.6987 c = 128.9598 R^2 = 0.9989		a = 1.8054 b = -4.1708 c = 103.9996 R^2 = 1.0000	

Length	T5		T6		T7		T8	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	116	116.01	75	75.03	146	145.89	75	75.00
2	54	53.88	46	45.73	72	73.11	58	57.99
3	8	9.17	10	11.72	30	25.45	16	16.00
4	6	2.15	7	3.34	5	9.40	4	4.43
5	–	–	–	–	1	4.09	3	1.81
	a = 1.0776 b = -3.1718 c = 115.0101 R^2 = 0.9980		a = 1.0367 b = -2.5444 c = 74.0303 R^2 = 0.9948		a = 0.0411 b = -1.5116 c =144.8929 R^2 = 0.9965		a = 1.4621 b = -2.6532 c = 74.0028 R^2 = 0.9996	

Length	T9		T10		T11		T12	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	167	167.06	67	67.00	181	181.04	92	92.02
2	102	101.50	47	47.03	110	109.59	69	68.82
3	25	27.86	12	11.83	30	32.29	17	18.07
4	13	7.65	3	3.23	14	9.46	7	4.74
5	–	–	1	1.48	2	3.41	–	–

	a = 0.8717 b = -2.3029 c = 166.0561 R^2 = 0.9976	a = 1.4041 b = -2.7756 c = 65.9958 R^2 = 0.9999	a = 0.7466 b = -2.1294 c = 180.0443 R^2 = 0.9988	a = 1.4545 b = -2.7018 c = 91.0243 R^2 = 0.9987

Length	T13		T14		T15		T16	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	176	176.02	151	151.02	214	214.02	95	95.00
2	128	127.83	56	55.71	132	131.79	50	49.99
3	27	28.19	19	20.26	23	24.79	11	11.04
4	9	5.98	10	8.50	10	4.76	3	2.90
	a = 1.6384 b = -3.0340 c = 175.0225 R^2 = 0.9995		a = -0.7493 b = -1.0187 c = 150.0197 R^2 = 0.9997		a = 1.5043 b = -3.1857 c = 213.0242 R^2 = 0.9989		a = 0.9334 b = -2.7028 c = 94.0005 R^2 = 1.0000	

Length	T17		T18		T19		T20	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	84	83.99	153	153.04	85	85.00	138	138
2	63	63.07	66	65.53	54	54.03	108	108
3	13	12.43	10	13.55	8	7.63	10	10
4	1	2.74	11	3.36	0	1.67	15	15
5	–	–	5	1.48	–	–	–	–
	a = 1.9496 b = -3.4173 c = 82.9904 R^2 = 0.9993		a = 0.5313 b = -2.5504 c = 152.0379 R^2 = 0.9948		a = 2.1528 b = -4.0630 c = 83.9965 R^2 = 0.9994		a = 2.9206 b = -4.7313 c = 137.3050 R^2 = 0.9860	

Needless to say, there are also other functions which could capture the given values satisfactorily – e.g., the Menzerathian function, but using it, one obtains very large *c*-values, while with the Zipf-Alekseev function *c* is almost identical with the frequency in $x = 1$, i.e. with the greatest frequency. In Text 19, the number of classes was only 3. Since we have three parameters, we added the fourth value as zero.

The above-presented results show that either the situation in German poetry was so that poetic texts were written under the influence of this subconscious law, or that Goethe had a special style. In order to state which of the hypotheses is correct, a number of other pomes by other German writers must be analyzed and compared.

Another set of data published in the same volume (cf. Rottmann 2007) concerns Polish texts selected by Borawski and Furdal (1980). There were 27 evaluated texts, list of which can be found in Rottmann (2007). The fitting of the Zipf-Alekseev function is presented in Table 3. As can be seen, several data are concave. Whereas Rottmann used the extended positive binomial, the Hyperpoisson and the positive Cohen-Poisson distributions, we shall try to use only one function and simplify the relation.

Table 3
Fitting the Zipf-Alekseev function to Polish texts

Length	T 1		T 2		T 3		T 4	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	153	150.18	275	271.50	349	348.80	173	172.31
2	148	160.54	259	275.23	238	239.40	162	165.20
3	116	89.88	183	149.05	93	88.62	87	80.64
4	40	45.38	69	73.29	31	31.37	39	35.86
5	6	23.07	11	36.34	2	11.94	6	16.36
6	1	12.25	1	18.75	–	–	1	8.00
	a = 1.0682 b = -1.4014 c = 149.1804 R^2 = 0.9472		a = 0.9914 b = -1.4018 c = 270.50 R^2 = 0.9679		a = 0.6690 b = -1.7512 c = 347.7994 R^2 = 0.9986		a = 1.0263 b = -1.5688 c = 171.3109 R^2 = 0.9925	

Length	T 5		T 6		T 7		T 8	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	171	169.48	92	91.16	101	100.08	196	194.18
2	177	183.52	130	132.59	91	95.80	190	198.50
3	114	100.32	76	71.08	63	51.55	123	104.01
4	45	49.09	33	31.62	21	25.71	45	49.22
5	20	24.17	7	13.96	5	13.15	9	23.58
6	4	12.45	1	6.57	–	–	3	11.88
7	1	6.84	–	–	–	–	–	–
	a = 1.1352 b = -1.4712 c = 168.4849 R^2 = 0.9896		a = 1.8699 b = -1.9108 c = 90.1639 R^2 = 0.9914		a = 0.8744 b = -1.3534 c = 99.0783 R^2 = 0.9658		a = 1.9650 b = -1.4904 c = 193.1761 R^2 = 0.9806	

Length	T 9		T 10		T 11		T 12	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	213	208.96	285	280.60	135	133.98	160	158.64
2	235	251.10	251	272.50	140	144.43	151	157.06
3	175	141.00	197	152.38	83	73.13	95	84.45
4	60	69.16	69	78.59	32	32.92	48	41.55
5	12	33.71	17	40.97	5	15.07	4	20.77
6	2	17.01	1	22.16	1	7.38	3	10.92
	a = 1.3370 b = -1.5448 c = 207.9591 R^2 = 0.9590		a = 0.8398 b = -1.2728 c = 279.5963 R^2 = 0.9528		a = 1.2476 b = -1.6424 c = 132-9819 R^2 = 0.9867		a = 0.9503 b = -1.3920 c = 157.6439 R^2 = 0.9780	
Length	T 13		T 14		T 15		T 16	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	241	240.37	135	132.91	166	163.97	105	104.75
2	246	248.93	132	140.93	197	205.28	109	109.95
3	133	126.34	99	83.39	125	102.53	57	55.97
4	57	57.40	47	44.95	28	43.61	30	25.59

5 6	17 –	26.36 –	13 3	24.37 13.70	4 –	18.56 –	4 –	11.98 –
	a = 1.1442 b = -1.5775 c = 239.3655 R^2 = 0.9968		a = 0.9631 b = -1.2666 c = 131.9106 R^2 = 0.9665		a = 1.6197 b = -1.8664 c = 162.9664 R^2 = 0.9640		a = 1.1792 b = -1.5996 c = 103.7540 R^2 = 0.9899	

Length	T 17		T 18		T 19		T 20	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	178	177.86	219	219.38	157	158.88	193	193.06
2	176	176.16	307	305.25	271	264.53	144	143.53
3	93	95.52	129	136.61	129	148.07	57	58.66
4	58	47.41	65	49.80	89	66.49	24	22.88
5	13	23.86	11	18.09	30	28.89	13	9.56
6	–	–	1	7.14	4	13.00	1	4.52
	a = 0.9372 b = -1.3722 c = 176.8649 R^2 = 0.9888		a = 2.0378 b = -2.2496 c = 218.3752 R^2 = 0.9949		a = 2.1130 b = -1.9821 c = 157.8817 R^2 = 0.9642		a = 0.7065 b = -1.6400 c = 192.0633 R^2 = 0.9991	

Length	T 21		T 22		T 23		T 24	
	Fr.	ZA	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	174	173.35	206	205.10	182	181.26	131	130.86
2	142	145.76	213	217.25	188	191.55	219	219.44
3	80	70.82	112	100.27	90	78.314	65	62.73
4	31	32.43	35	40.72	18	27.83	9	13.48
5	4	15.41	4	16.85	1	10.29	1	3.42
	–	–	–	–	1	4.35	–	–
	a = 0.7244 b = -1.4081 c = 172-3453 R^2 = 0.9888		a = 1.3476 b = -1.8238 c = 204.0950 R^2 = 0.9903		a = 1.5343 b = -2.0981 c = 180.2605 R^2 = 0.9911		a = 3.1901 b = -3.5199 c = 129.8595 R^2 = 0.9991	

Length	T 25		T 26		T 27	
	Fr.	ZA	Fr.	ZA	Fr.	ZA
1	99	96.57	149	147.71	130	129.53
2	142	149.29	163	168.68	192	193.37
3	102	88.50	97	81.72	94	91.43
4	43	43.36	26	34.45	38	34.97
5	9	20.74	1	14.73	4	13.34
6	1	10.29	–	–	2	5.58
	a = 1.8546 b = -1.7612 c = 95.5650 R^2 = 0.9712		a = 1.4520 b = -1.8167 c = 146.7059 R^2 = 0.9747		a = 2.1233 b = -2.2240 c = 128.5311 R^2 = 0.9959	

In any case, one can see that the proposed model yields excellent results.

The relation between the parameters *a* and *b* is not smooth. As can be seen in Figure 1, *b* decreases with increasing *a*, but the curve oscillates strongly, and we do not propose any hypothesis.

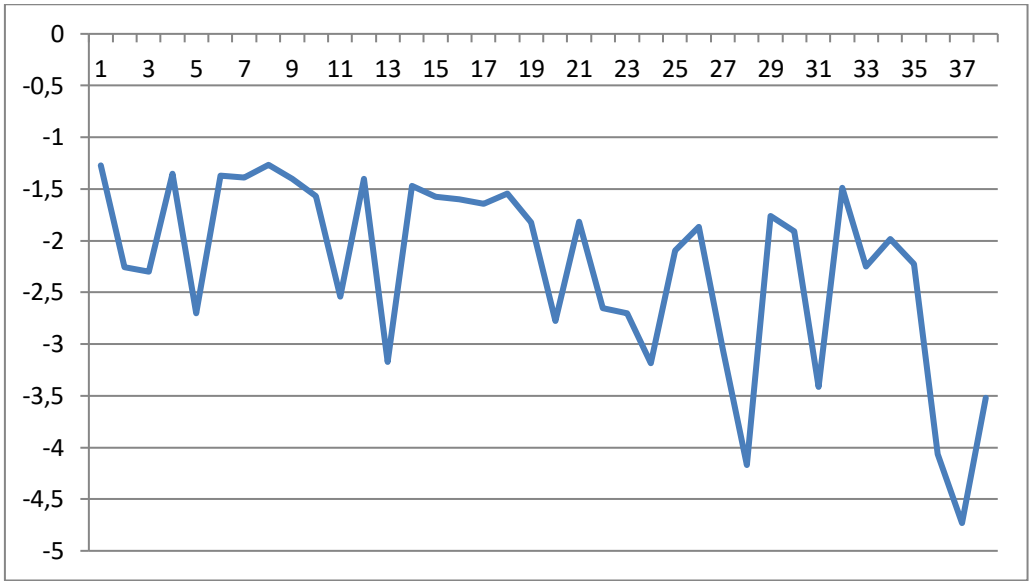

Figure 1. The relation between the parameters *a* and *b*

In order to state whether there was some development with Goethe, it is sufficient to compare an indicator of the given frequencies. In order to do it, we use here merely the Ord's criterion <I, S>, representing the moments of the distributions, namely

$$I = \frac{m_2}{m_1'},$$

and

$$S = \frac{m_3}{m_2},$$

where m_r are the central moments, and m_1' is the first moment. Ordering the data from Table 1 according to years, we obtain the results presented in Table 4.

Table 4
Ord's criterion for word length in poems by Goethe ordered according to years

Text	Year	I	S	Text	Year	I	S
T 18	None	0.5368	1.7823	T 13	1788	0.3034	0.8506
T 1	1767	0.2884	0.8703	T 15	1788	0.3405	0.9723
T 2	1774	0.3758	0.9231	T 19	1789	0.2427	0.5105
T 3	1774	0.4017	1.0398	T 5	1789	0.3600	1.2193
T 8	1776	0.4544	1.2726	T 16	1795	0.3322	0.9546
T 20	1776	0.3887	1.1128	T 17	1803	0.4834	0.9586
T 11	1777	0.4387	1.1606	T 4	1815	0.2507	0.9085

T 10	1778	0.3913	1.0826	T 14	1815	0.4357	1.2530
T 12	1783	0.3783	0.8834	T 6	1825	0.4185	1.0871
T 9	1784	0.4019	1.0264	T 7	1829	0.4030	1.0201

The individual poems can be characterized by <I, S>, but historically, they display a rather oscillating concave course. That means, either one needs to analyze all poems by Goethe, or tries to characterize the development by means of works of other German poems.

In Polish, we may test the hypotheses using Rottmann's data (2007).

Table 5
Ord`s criterion for Polish texts

Text	I	S	Text	I	S
T 1	0.5058	0.5850	T 15	0.4101	0.5091
T 2	0.5011	0.6635	T 16	0.5483	1.0349
T 3	0.4317	0.9016	T 17	0.5520	0.7882
T 4	0.5111	0.8411	T 18	0.4639	0.7897
T 5	0.6024	1.0748	T 19	0.5327	0.9679
T 6	0.4813	0.7086	T 20	0.5791	1.3065
T 7	0.5021	0.7053	T 21	0.4906	0.7736
T 8	0.5240	0.8622	T 22	0.4377	0.6586
T 9	0.4903	0.6518	T 23	0.3996	0.7005
T 10	0.5235	0.6943	T 24	0.2935	0.4595
T 11	0.4916	0.7608	T 25	0.4755	0.5737
T 12	0.5368	0.8209	T 26	0.4103	0.4891
T 13	0.5192	0.8286	T 27	0.4451	0.7619

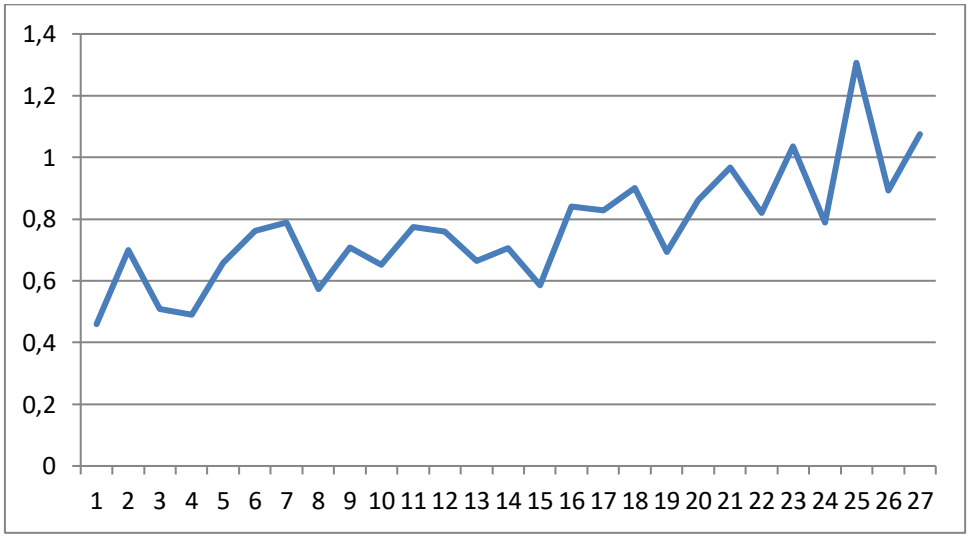

Figure 2. Relation between I and S in Polish

The relation between I and S is increasing, but the oscillation is too strong to yield a simple curve according to I.

It can further be shown that the Repeat rate of frequencies in Polish is constant. One obtains an oscillating horizontal straight line presented in Figure 3.

Table 6
Repeat rate in the Polish text

T 1	**T 2**	**T 3**	**T 4**	**T 5**	**T 6**	**T 7**	**T 8**	**T 9**	**T 10**
0.2806	0.2844	0.3699	0.2981	0.2686	0.2865	0.2902	0.2864	0.2778	0.2797
T 11	**T 12**	**T 13**	**T 14**	**T 15**	**T 16**	**T 17**	**T 18**	**T 19**	**T 20**
0.2908	0.2812	0.2903	0.2576	0.3062	0.3466	0.2789	0.2036	0.2820	0.3321
T 21	**T 22**	**T 23**	**T 24**	**T 25**	**T 26**	**T 27**			
0.3112	0.3127	0.3337	0.3844	0.2697	0.3096	0.3028			

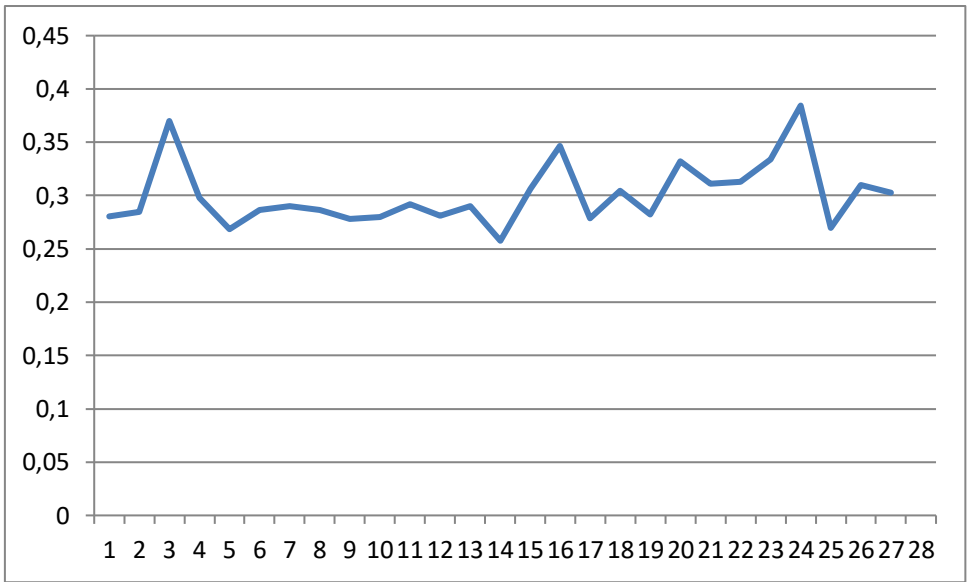

Figure 3. Repeat rate of frequencies in Polish

The present study yields data for comparison. It raises questions whether the distribution of lengths is identical in all languages, the Ord's criterion is similar in all languages, the repeat rate is a constant, etc. We conjecture that text type can have a strong influence on these functions or indicators, and hope that at least one language will be thoroughly analyzed.

References

Borawski, S., Furdal A. (1980). *Wybór tekstów do historii języka Polskiego*. Warszawa: Państwowe Wydawnictvo Naukowe.

Kühner, A. (2007). Wortlängenhäufigkeit in J. W. Goethe's Gedichten. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 361–370.* Berlin/New York: Mouton de Gruyter.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified Modeling of Length in Language*. Lüdenscheid: RAM-Verlag.

Rottmann, O. (2007). Wortlänge im Polnishen in diachroner Sicht. In: Grzybek, P., Köhler, R. (eds.), *Exact Methods in the Study of Language and Text: 597–604*. Berlin/New York: Mouton de Gruyter

Glottometrics 42, 2018, 21-38

Piotrowski Law in Sequences of Activity and Attributiveness: A Four-Language Survey

Michal Místecký[1], Sergej Andreev[2], Gabriel Altmann

Abstract. The present study investigates the possibility of applying Piotrowski Law, a general mathematical principle of historical change in language, on two types of binary sequences – the activity strings, which consist of adjectives and verbs, and adnominal strings, comprising attributes and genitives. The fit is tested on various samples (sonnets, long poems, newspaper texts) written in Czech, Russian, English, and French. In the interpretation part, some ways of further research are sketched, including general employments in the theory of linguistic laws and parameter analyses. As to the latter, the study emphasises their use in testing the hypotheses on texts and in their classifications.

Keywords: Piotrowski Law; sequences; activity; attributiveness; parameters; testing

1. Theory

The same as many phenomena in nature, developments of linguistic properties can be modelled by mathematical functions, which are able both to take into account the situations in the past, and predict possible trends for the future. For instance, the workings of Menzerath Law, which links units of individual levels of language (cf. Altmann 1980), confirms Zipf's assumption of the principle of minimal effort being a powerful organizing feature in using language. The first two laws of Zipf's (cf. Zipf 1949), on the other hand, are used to prove workability of linguistic units – including these that are not easy to be interpreted from the qualitative research viewpoint, such as motifs or n-grams.

As to language change, which has been a much studied issue in modern language science (Labov 1994; Crystal 2000; Labov 2001), Piotrovskaja and Piotrovskij (1978) and, later, Altmann et al. (1983) and Altmann (1983) derived a law which is able to capture several forms of the development. The intuition of historical linguists, who had presupposed that the curve should be S-shaped, was turned into the mathematical function of the form –

$$y = \frac{c}{1 + a * e^{-bx}}, \tag{1}$$

which is supposed to seize the evolution of a phenomenon in a language. The curve, which is exemplified in Figure 1, makes a good sense: first, there is a rise in the usage of a new form (the decreasing denominator), which breaks at a certain point (x = 0); from this place on, the increase gets slower, following, finally, a trend of a linear-like constant. At this stage, the change has taken root in the language, and its distribution remains stable.

[1] Michal Místecký, University of Ostrava; e-mail: MMistecky@seznam.cz.
[2] Sergey Andreev, University of Smolensk; e-mail: smol.an@mail.ru

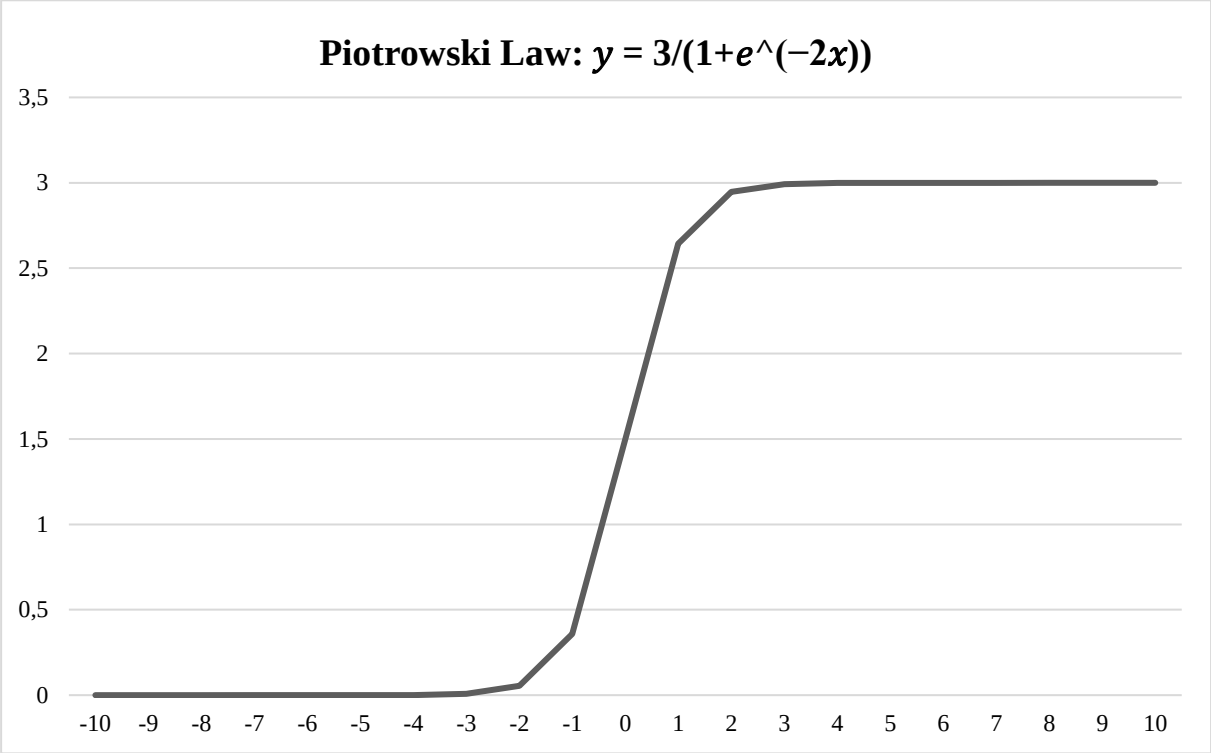

Fig. 1. An example of a Piotrowski-law curve, with the parameters a = 1, b = 2, and c = 3.

In the present paper, there will be an endeavour to extend the scope of application of what has become to be called Piotrowski Law over two types of textual sequences which are going to be presented in Section 2. If it were proved that these vectors abide by the same function, it would mean that there may exist a unifying mechanism lying beneath all possible developments in language; however, before multiple types of sequences are tested, it is too early to make such assumptions.

2. Examined Sequences

In recent years, the interest of linguists in various sequential measurements has been on the rise (Popescu et al. 2010; Zörnig 2015); a lot of attention was paid to the properties, but their open-scale research is complicated by the absence of a data-processing software. That is why in the paper, use will be made of the accumulated investigations carried out in sonnets, with the focus being paid on activity and attributiveness strings.

As to the former, stretches contain adjectives (A) and verbs (V). For example, the sentence

I have fallen in love with a beautiful girl in a black dress I met at the university ball

is scripted as the adjective-verb vector of

V–A–A–V–A.

Concerning the latter, strings comprise attributes (T) and genitives (G). By attributes, adjectives modifying nouns are understood (“beautiful girl”); the status of genitives is clear in synthetic languages, but in English, where the case system is virtually non-existent, *of*-phrases are usually taken as their equivalents. Respecting this method, one parses the sentence

At the ball of my secondary school, I met a sad girl of extraordinary beauty

into the sequence

T–G–T–T–G,

corresponding to, respectively, "secondary" (T), "school" (G), "sad" (T), "extraordinary (T), and "beauty (G).[3]

The stretches may undergo multiple analyses, such as measuring distances, run decomposition, or various motif clusterings; here, the research will concentrate on the development projection, which it will try to match with the workings of Piotrowski Law.

3. Analysis

The application of Piotrowski Law in the sequences will be tested on a corpus the details of which are given in Table 1. Its G–T section comprises ten Czech and ten Russian sonnets; the A–V structures are covered by thirty English and thirty French sonnets, which are accompanied by ten Czech poems by Jiří Wolker, and a selection of the same number of Czech newspaper crime stories (from the internet server of *Novinky* – "News"). The texts by Jiří Wolker, which are mostly long, were chosen as counterparts to short sonnets. The total of the texts is 100.

Table 1
Overview of the studied corpus

Author	Poem
K. H. Mácha	4
	6
J. Kvapil	Duše
	Iniciála
	Koflík
	Portrét
Jiří Karásek ze Lvovic	Kalný západ
	Noční sonet
	Narkózy
V, Nezval	měšťák
E. Baratynskij	My p'jom v ljubvi otravu sladkuju (…)
N. Jazykov	K. K. Janish
	Na prazdnik vash prines ja dva priveta (…)
A. Pushkin	Madona
V. Benediktov	Cvetok
V. Brjusov	Egipetskij rab

[3] In English, there are basically two contradictory ways of treating the genitive positions (G) – they may be taken either as starting with the "of" preposition, or with the noun as the central object. In the former case, the sequence would be inverted, yielding G [of] – T [secondary] – T [sad] – G [of] – T [extraordinary]; however, as it seems to make more sense to take into account nouns as carriers of the grammatical case, the latter parsing was given priority in the paper. Moreover, it is not of much importance, as the T–G sequences studied in the article are taken from synthetic languages only.

A. Blok	Ne ty l' v moih mechtah-pevuchaja-proshla (…)
M. Voloshin	Venok sonetov. Sonet 1
N. Gumilev	Popugaj
K. Bal'mont	Marlo
W. Wordsworth	Beloved Vale!
	Calm Is All Nature (…)
	Glasmore Lake
	How Sweet It Is (…)
	London, 1802
	Nuns Fret Not (…)
	Scorn Not the Sonnet (…)
	The World Is Too Much with Us (…)
	Those Words Were Uttered (…)
	With How Sad Steps (…)
S. T. Coleridge	To Southey
	To Priestley
	To Pitt
	To Mrs Siddons
	To Kosciusko
	To Godwin
	To Fayette
	To Burke
	To Bowles
	To Erskine
D. G. Rossetti	A Venetian Pastoral
	Heart’s Haven
	Life-in-Love
	Mary Magdalene
	Silent Noon
	The Heart of the Night
	The Choice, I
	The Choice, II
	Through Death to Love
	Without Her
Ch. Baudelaire	Bohémiens en voyage
	Correspondances
	L’Ennemi
	L’Idéal
	La beauté
	La masque
	La Muse malade
	La Muse vénale
	La vie antérieure
	Le guignon
	Le mauvais moine

	Parfum exotique
T. Corbière	1 Sonnet
	À l'éternel madame
	Bonsoir
	Déclin
	Duel aux camélias
	Féminin singulier
	Fleur d'art
	Litanie
	Pauvre garçon
	Pudentiane
	Sonnet à sir Bob
	Sonnet de nuit
S. Mallarmé	Angoisse
	Dame sans trop d'ardeur (…)
	Le pitre châtié
	Le sonneur
	O si chère de loin (…)
	Placet futile
	Quand l'ombre menaça (…)
	Remémoration d'amis belges
	Renouveau
	Salut
	Sonnet
	Tristesse d'été
J. Wolker	Moře
	Balada o očích topičových
	Dům v noci
	Fotografie
	Muž
	Pohřeb
	Sloky
	Jaro
	Oči
	Kázání na hoře
Novinky	1
	2
	3
	4
	5
	6
	7
	8
	9
	10

The procedure of the investigation will be exemplified upon William Wordsworth's sonnet *Glasmore Lake*. The poem manifests the A–V string of

A–V–A–A–A–A–A–V–A–A–V–A–A–A–A–V–A–A–V–V–A–V–A–A–A–A–A–V,

which can be captured in the form

$$V = f(A).$$

Practically, it means that one always counts the number of verbs preceding an adjective – thus, as the sequence starts with an adjective, there is no verb before it, the first coordinate of the function being [1;0]. Next, the second adjective is preceded by one verb, this yielding the second coordinate – [2;1]. What follows is another, third adjective, so the number of verbs remains the same, the third coordinate equalling [3;1]. All the coordinates of the function are listed in Table 2.

Table 2
Coordinates of the A–V sequence function in Wordsworth's *Glasmore Lake*

x (A)	**y (V)**
1	0
2	1
3	1
4	1
5	1
6	1
7	2
8	2
9	3
10	3
11	3
12	3
13	4
14	4
15	6
16	7
17	7
18	7
19	7
20	7
21	8

In order to cover all the verbs, an "artificial" adjective is added at the end of the sequence.[4] Now, the distribution of the data is fitted by the Piotrowski Law function; in the present case, the nearest match is represented by the formula

[4] The G–T sequences are treated in the same way; the function they form is $T = f(G)$, as the projected property is attributiveness.

$$y = \frac{9.20}{1 + 24.51 * e^{-0.24x}},$$

which covers the data with the precision of $R^2 = 0.96$. The projection is visualised in Figure 2; the real and approximated figures are compared in Table 3.

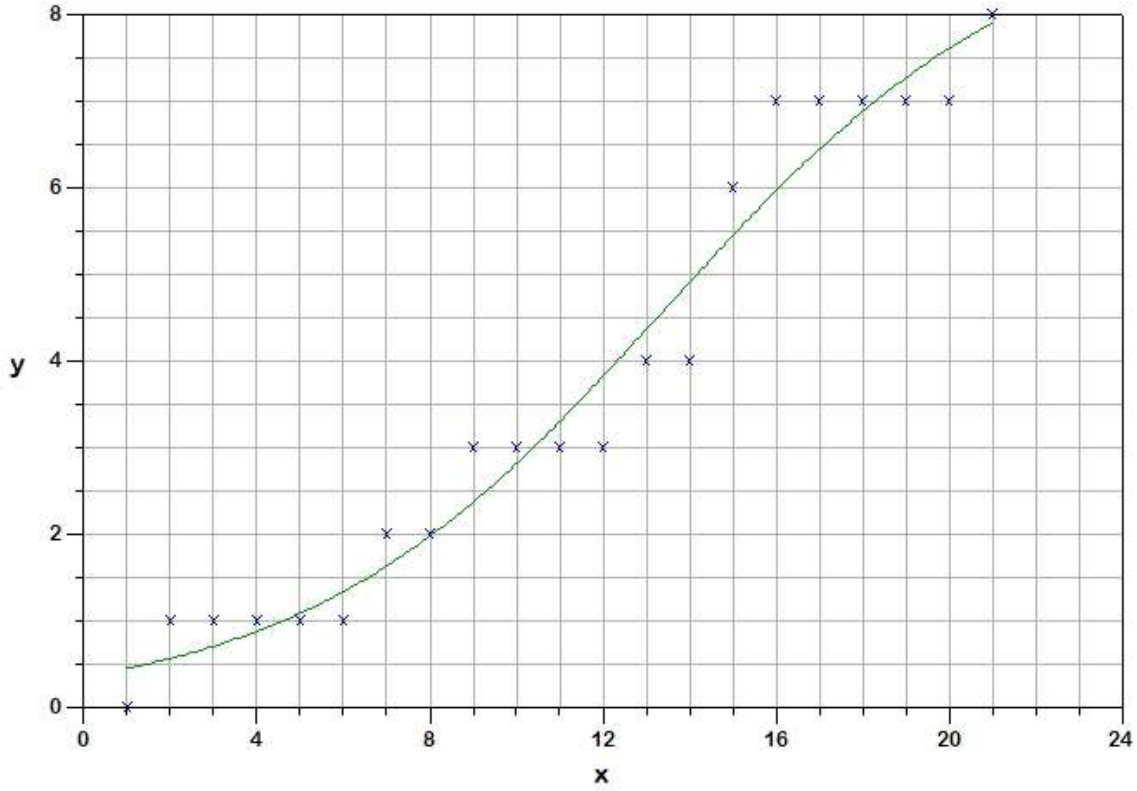

Fig. 2. Wordsworth's *Glasmore Lake*: Piotrowski-Law fit

Table 3
Comparison of the real and calculated figures of the A–V sequence function in Wordsworth's *Glasmore Lake*

x (A)	y (V)	ŷ (Piotrowski Law)
1	0	0.45
2	1	0.57
3	1	0.71
4	1	0.88
5	1	1.09
6	1	1.34
7	2	1.64
8	2	1.98
9	3	2.37
10	3	2.82
11	3	3.30
12	3	3.82
13	4	4.36

14	4	4.91
15	6	5.45
16	7	5.96
17	7	6.44
18	7	6.88
19	7	7.27
20	7	7.61
21	8	7.90

The results of the analysis will be exemplified upon three excerpts – Table 4 lists the outputs in the G–T sequences of five Czech sonnets, and in the A–V stretches of five French poems and five Czech newspaper texts. Other results are presented in the form of the averages of the determination coefficients (R^2), capturing the fit with which the Piotrowski-Law function matches the given data (see Table 5).

Table 4
Piotrowski-Law function fit in the selected poems

Author	**Poem**	**Sequence**	**G / A**	**T / V**	**T / V (Estimate)**	**Parameters**
K. H. Mácha	4	T, G, T, T, T, G, T, T, T, T, T, G, T, T, T, T, T	1	1	1.19	a = 52.33 b = 1.36 c = 17.21 R^2 = 0.99
			2	4	3.87	
			3	9	9.13	
			4	14	14.03	
	6	T, T, G, T, G, T, T, T, T, T, G, T, T, T, T, T, T	1	2	1.49	a = 63.61 b = 0.91 c = 39.69 R^2 = 0.99
			2	3	3.51	
			3	8	7.71	
			4	15	14.87	
J. Kvapil	Duše	T, T, G, T, T, T, G, G, T, T, G, T, T, G, G, T, T, T, G, T, T	1	2	3.06	a = 8.90 b = 0.33 c = 22.61 R^2 = 0.96
			2	5	4.04	
			3	5	5.25	
			4	7	6.69	
			5	9	8.35	
			6	9	10.14	
			7	12	12.00	
			8	14	13.83	
	Iniciála	T, T, T, G, T, T, T, G, T, T, G, T, G	1	3	3.02	a = 7.61 b = 1.29 c = 9.35 R^2 = 0.99
			2	6	5.93	
			3	8	8.07	
			4	9	8.96	
	Koflík	T, T, G, T, T, G, T, G, T, G, T, G, G, T, T, G, T, T, T, T, G	1	2	2.75	a = 10281810.6 b = 0.22 c = 22709479.2
			2	4	3.43	
			3	5	4.27	
			4	6	5.32	
			5	7	6.64	

			6	7	8.27	$R^2 = 0.92$
			7	9	10.30	
			8	14	12.84	
Ch. Baudelaire	Bohémiens en voyage	A, A, V, A, V, A, A, A, V, A, A, V, A, A, A, A, V, V, V, V, V, V, A, A, A	1	0	0.66	a = 96.14 b = 0.22 c = 51.98 $R^2 = 0.90$
			2	0	0.81	
			3	1	1.00	
			4	2	1.22	
			5	2	1.50	
			6	2	1.84	
			7	3	2.25	
			8	3	2.75	
			9	4	3.35	
			10	4	4.07	
			11	4	4.93	
			12	4	5.95	
			13	10	7.15	
			14	10	8.55	
			15	10	10.15	
	Correspondances	A, V, V, A, V, V, A, A, V, A, A, A, V, A, A, A, A, A, A, A, V	1	0	1.03	a = 10.78 b = 0.80 c = 6.01 $R^2 = 0.92$
			2	2	1.89	
			3	4	3.04	
			4	4	4.18	
			5	5	5.02	
			6	5	5.52	
			7	5	5.78	
			8	6	5.90	
			9	6	5.96	
			10	6	5.99	
			11	6	6.00	
			12	6	6.01	
			13	6	6.01	
			14	6	6.01	
			15	7	6.01	
	L’Ennemi	A, A, V, V, A, V, V, V, V, A, V, A, V, A, V, V, A, A, V, V, A, V, V, V, V	1	0	1.15	a = 21.95 b = 0.50 c = 16.40 $R^2 = 0.93$
			2	0	1.81	
			3	2	2.78	
			4	6	4.13	
			5	7	5.85	
			6	8	7.84	
			7	10	9.86	
			8	10	11.70	
			9	12	13.18	
			10	16	14.29	
	L’Idéal	A, A, A, V, V, V, V, V,	1	0	0.00	a= 7414131.17 b = 4.15
			2	0	0.00	

		A, V, A, V, A, A, A, A, V, A, A	3	0	0.24	c = 7.16 R^2 = 0.97
			4	5	4.91	
			5	6	7.11	
			6	7	7.16	
			7	7	7.16	
			8	7	7.16	
			9	7	7.16	
			10	8	7.16	
			11	8	7.16	
	La beauté	A, V, V, V, A, A, V, A, V, V, V, V, V, A, V, V, A, V, A, V, A, A, V, A, A, A	1	0	0.71	a = 42.07 b = 0.82 c = 13.91 R^2 = 0.98
			2	3	1.52	
			3	3	3.03	
			4	4	5.39	
			5	9	8.20	
			6	11	10.64	
			7	12	12.25	
			8	13	13.13	
			9	13	13.55	
			10	14	13.75	
			11	14	13.84	
			12	14	13.88	
Novinky	1	V, A, A, V, V, A, A, V, V, A, V, A, A, V, V, A, A, A, A, V, A, A, A, A, A, V, V, V, A, A, V, A, V, A, A, A, V, V	1	1	2.73	a = 6.61 b = 0.14 c = 18.40 R^2 = 0.93
			2	1	3.07	
			3	3	3.44	
			4	3	3.85	
			5	5	4.30	
			6	6	4.77	
			7	6	5.29	
			8	8	5.83	
			9	8	6.40	
			10	8	7.00	
			11	8	7.61	
			12	9	8.24	
			13	9	8.88	
			14	9	9.53	
			15	9	10.17	
			16	9	10.80	
			17	12	11.42	
			18	12	12.01	
			19	13	12.58	
			20	14	13.12	
			21	14	13.64	
			22	14	14.11	
			23	16	14.56	
	2	A, V, A, A, A, V, A, A,	1	0	0.97	a = 13.83 b = 0.22
			2	1	1.19	

		V, A, V, A, V, A, A, V, A, A, A, V, A, A, V, A, A, V, A, A, V, A, V, A, A	3	1	1.44	c = 11.74 R^2 = 0.97
			4	1	1.74	
			5	2	2.10	
			6	2	2.50	
			7	3	2.96	
			8	4	3.47	
			9	5	4.04	
			10	5	4.64	
			11	6	5.27	
			12	6	5.91	
			13	6	6.55	
			14	7	7.18	
			15	7	7.77	
			16	8	8.33	
			17	8	8.84	
			18	9	9.29	
			19	9	9.69	
			20	10	10.04	
			21	11	10.33	
			22	11	10.58	
	3	V, A, V, V, V, V, A, A, A, V, V, A, V, V, V, V, V	1	1	2.23	a = 483848.94 b = 0.33 c = 775089.36 R^2 = 0.87
			2	5	3.10	
			3	5	4.31	
			4	5	6.00	
			5	7	8.34	
			6	12	11.60	
	4	V, A, A, A, A, V, V, A, V, A, V, V, A, A, V, V, V, A, A, A, A, A, A, V, V, V, V, A, V, V, V	1	1	1.27	a = 15.46 b = 0.26 c = 16.37 R^2 = 0.90
			2	1	1.61	
			3	1	2.02	
			4	1	2.53	
			5	3	3.14	
			6	4	3.85	
			7	6	4.67	
			8	6	5.58	
			9	9	6.58	
			10	9	7.62	
			11	9	8.68	
			12	9	9.73	
			13	9	10.72	
			14	9	11.64	
			15	13	12.47	
			16	16	13.19	
	5	V, A, V, A, V, A, V, V, A, A, V, A, A, V, V, V,	1	1	1.64	a = 8.83 b = 0.38 c = 11.54 R^2 = 0.96
			2	2	2.25	
			3	3	3.02	
			4	5	3.94	

		A, A, A, V, A, V	5	5	4.97	
			6	6	6.06	
			7	6	7.13	
			8	9	8.11	
			9	9	8.95	
			10	9	9.64	
			11	10	10.17	
			12	11	0.56	

Table 5
Intervals of the determination coefficients in the studied text groups

Text Group	Sequence Type	R^2 Average
Czech sonnets	G–T	0.95
Russian sonnets	G–T	0.88
English sonnets	A–V	0.94
French sonnets	A–V	0.94
Jiří Wolker	A–V	0.97
Czech newspaper	A–V	0.94

Needless to say, the development can be expressed by other functions as well; this is especially recommendable if the parameters are too great. For example, in *Koflík* by Kvapil, the parameters *a* and *c* are enormous and cannot be well interpreted. We simplify the Piotrowski function and take

$$y = \frac{1}{a * e^{-bx}}, \tag{2}$$

which is, as a matter of fact, a special case of "Piotrowski". Here, we assume that the relative rate of change of the dependent variable is simply constant, i.e. –

$$\frac{dy}{y} = b * dx\ . \tag{3}$$

For Kvapil's data, we obtain $R^2 = 0.9272$, and the parameters are a = 0.4528, b = 0.2200; for short texts – like sonnets –, the approach thus seems adequate. The results are presented in Table 6.

Table 6
Comparison of the real and calculated figures of the G–T sequence function in Kvapil's *Koflík*

x (G)	y (T)	ŷ (Simplified Piotrowski)
1	2	2.75
2	4	3.43
3	5	4.27
4	6	5.33

5	7	6.64
6	7	8.27
7	9	10.3
8	14	12.84

However, it is obvious that even in case of the general function (see Formula 1), the figures of the determination coefficient are generally very high, indicating the fact that Piotrowski Law can be in operation in the development of the studied properties. Given $R^2 > 0.8$ is the satisfactory figure, there are only three poems escaping the trend; whereas Wordsworth's *Those Words Were Uttered (...)* [$R^2 = 0.75$], and Karásek's *Noční sonet* ("A Night Sonnet"; $R^2 = 0.79$) may be considered on the verge of acceptance, Gumilev's *Popugaj* ("A Parrot"; $R^2 = 0.36$), whose G–T sequence reads

T–T–G–G–G–G–G–G–G–T–G,

manifests almost no evolution in the studied property (attributiveness). The fact that Piotrowski Law cannot handle such a case may thus mean that it is the "counter-property" that develops, as the function for the inverted, T–G sequence matches the data with the 100-percent precision. The excessive use of genitives may be attributed to the Acmeist, Classicism-infused tendency of Gumilev's production, which yearned to discover the beautiful in science-like clarity (Wachtel 2004).

4. Applications

The finding that Piotrowski Law is able to account for two types of sequences in various languages may significantly contribute to multifarious domains of linguistic research. First, it provides the grounds for a unified theory of sequences, as the present results can be, *mutatis mutandis*, extended over other types of language vectors. Second – and given the previous –, it may be used as a test whether a particular sequence behaves according to the same rules as most of them; this way, it imitates the employment of the Zipfian power-law function for linguistic units. And last but not least, since individual Piotrowski-Law functions differ in parameters, the study of them may provide a deep insight into the structure of poems, being thus of use for literary and stylistic scholarships. If a sufficient number of texts is analysed, a picture of the language type can also be obtained. Translations of texts – not poetic ones – could show the most evident differences.

The pertinence of the last point may be developed further. Parameter *a* determines the turning point of the change, where the convex part of the function turns into the concave one – linguistically, it means the stage of the rise in a particular feature is substituted by the period when its increase slows down. If *a* is a low number, it means that the change starts in the initial phase of the development, whereas if it is high, it marks a late onset of the rise of the given phenomenon (in the present case, it concerns numbers of verbs or attributes). Parameter *b*, on the other hand, points at the degree of briskness of the change – the higher the figure becomes, the more abrupt the increase in the given sequence is. And last but not least, parameter *c* focuses on the overall range of the evolution; however, in short sequences, its interpretational value may be discredited.

The aforementioned reflections are illustrated upon the examples in Figure 2, which presents the projections of two Czech newspaper samples (*Novinky1* and *Novinky5*), and two poems by Baudelaire (*Correspondances* and *L'Idéal*). As can be seen in Table 6, each text

displays a different set of parameters – for instance, *L'Idéal*'s change is the steepest, as it possesses the highest figure of parameter *b*; on the other hand, the text of *Novinky1* tends to have a linear-like development of activity, scoring also top as to the overall measure of the change (parameter *c*). It is an open question whether newspaper texts tend to a moderate development of activity, whereas poetic ones incline to unpredicted metamorphoses of tone; these issues should be dealt with within the domains of genre studies, stylistics, or literary history.

Table 7
Piotrowski-Law functions of the selected samples

Text	Function
Novinky1	$y = \frac{18.4}{1 + 6.61 * e^{-0.14x}}$
Novinky5	$y = \frac{11.54}{1 + 8.83 * e^{-0.38x}}$
Correspondances	$y = \frac{6.01}{1 + 10.78 * e^{-0.8x}}$
L'Idéal	$y = \frac{7.16}{1 + 7414131.17 * e^{-4.15x}}$

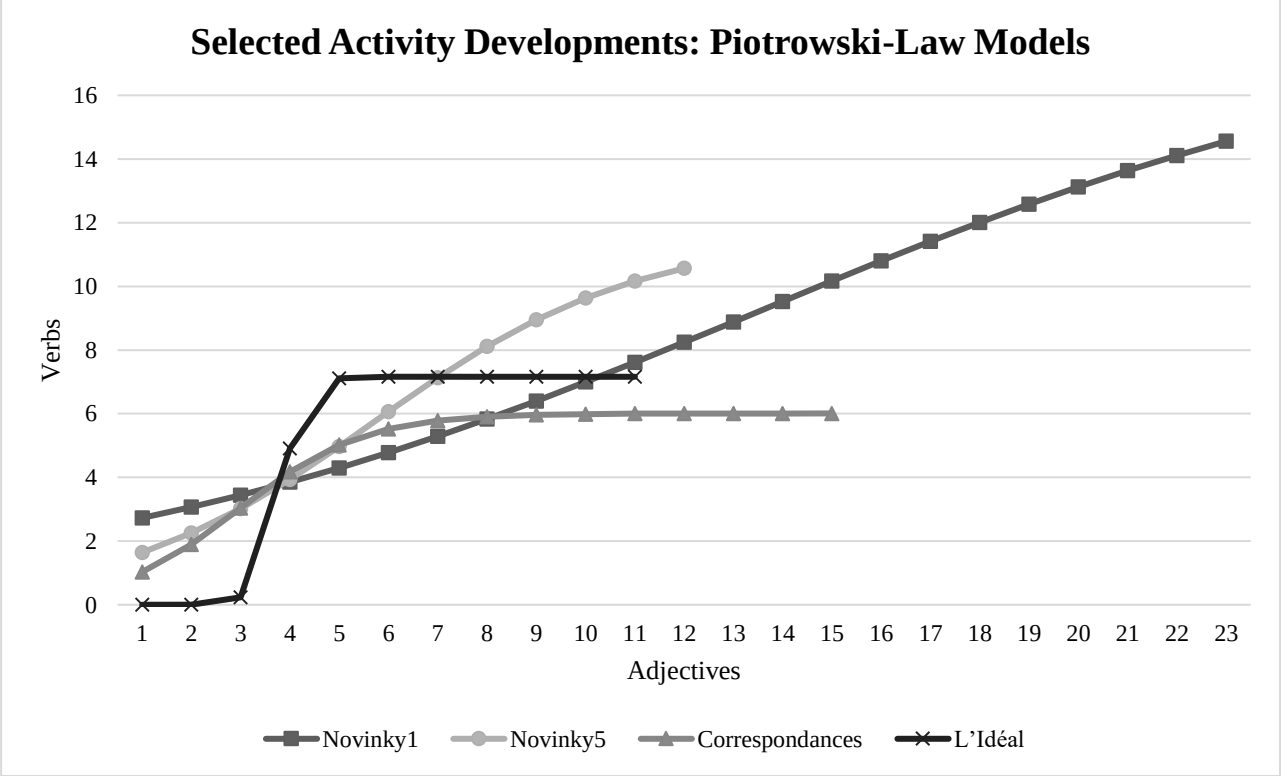

Fig. 3. Fitted activity developments in the selected samples

The parameters can be visualised in multiple manners, including scatterplots of various configurations; for instance, Figure 4 shows the confrontation of parameters *b* and *c* in the selected samples. It indicates that, for instance, *Novinky1* manifests an elevated level of the change, which proceeds very slowly; this is contrasted by the situation in *L'Idéal*, the development of which is very abrupt, though the level of the change does not match the one in the former. More generally, texts that score low in both parameters tend to possess a low number of verbs, which are interspersed in the long chains of adjectives (e.g., A–A–A–V–A–A–A–V), whilst those with high figures in both demonstrate a lot of verbs crowded in one particular phase of the development (e.g., A–A–A–V–V–V–V–V–A–A–A).

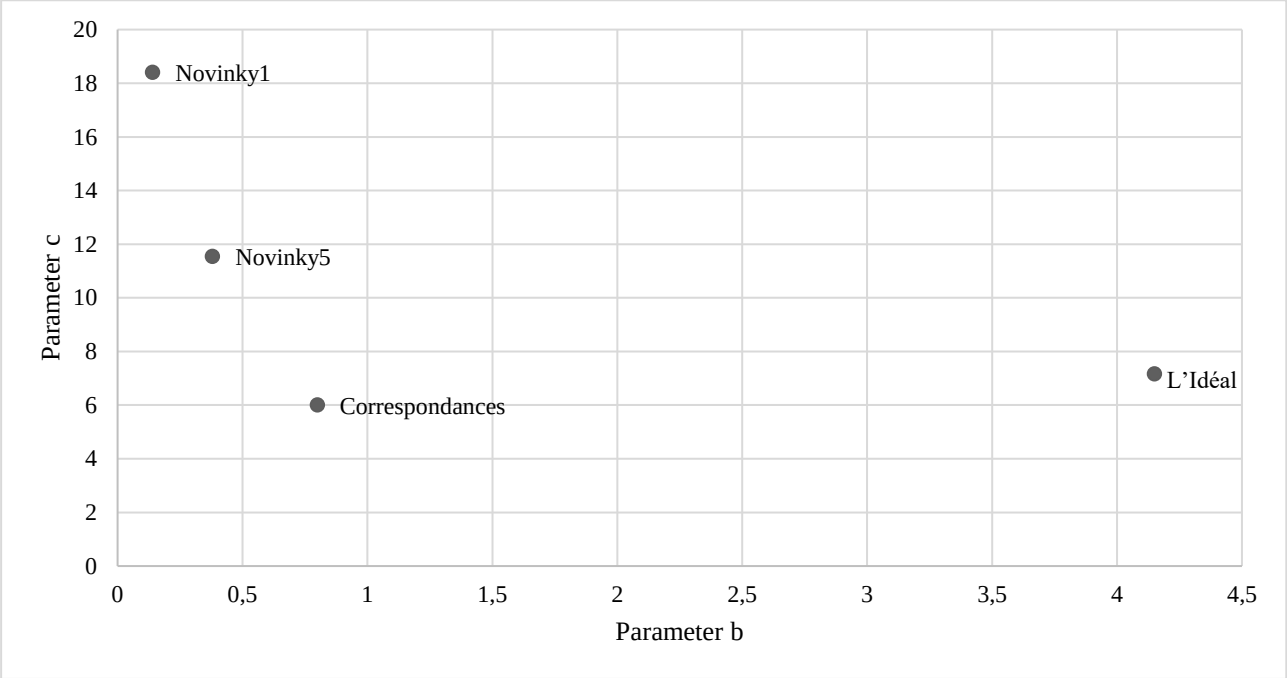

Fig. 4. Scatterplot of the Piotrowski-Law function parameters in the selected texts

The second, general question is whether the parameters differ as to the type of the language studied; in this research, statistical tests can be used when various groups of samples are compared. For the sake of illustration, *b* parameters of the Czech and Russian G–T sequences were contrasted on the basis of Mann–Whitney–Wilcoxon test (MWW). As the test is nonparametric, it is suitable for the data one does not know the distribution of.[5] The result is 1.81, which corresponds to the p-value of 0.07; according to the chosen level ($\alpha = 0.05$; MWW = 1.96), the difference is thus not significant. Given the respective averages

$$b_{Cze} = 1.68\,;$$
$$b_{Rus} = 0.55\,,$$

it is to be concluded that the Czech G–T developments are insignificantly steeper than the Russian ones.

Some more investigations are listed in Table 8; they may be utilized as subjects of separated studies.

[5] More information on the calculation of the MWW test is to be found on the Wikipedia page https://en.wikipedia.org/wiki/Mann–Whitney_U_test.

Table 8
Testing the significance of differences in the parameters *b* of the selected samples

Texts	Sequences	Averages	MWW Test Value	p-Value	Result
W. Wordsworth	A–V	0.2917	0.9294	0.3527	not significant
S. T. Coleridge	A–V	0.2198			
English sonnets	A–V	0.2753	3.8785	0.0001	significant
French sonnets	A–V	1.1310			
Ch. Baudelaire	A–V	0.7792	1.1712	0.2415	not significant
T. Corbière	A–V	2.1229			
Czech sonnets	G–T	1.6790	3.4878	0.0005	significant
Jiří Wolker	A–V	0.2155			

5. Conclusions

The research has shown that within the scope of the studied samples, the sequences of genitives and attributes, and of adjectives and verbs abide by Piotrowski Law, as 97 out of 100 texts manifest more than the 80-percent fit. This may entail that all sequences in language behave in the same way, and that the functioning of the law can thus possibly be universal.

Besides theoretical investigations, the present paper wants to pioneer the idea that the parameters of the function can also be of practical employment in stylometric analyses, providing an unmatchable tool for capturing the development of a particular phenomenon in a text. Last but not least, it is probable that comparing various types of sequences on the basis of statistical testing will uncover interior workings of language, and delimit the characteristics of many a linguistic feature.

To sum it up, the following questions arise:

(1) Is the principle similar in all languages?
(2) Does it hold for any type of language properties?
(3) Can one characterize texts or text types by some of the parameters of the function?
(4) What is the relation of the parameters to other properties of the text like – i.e., does the development depend on its position in the (self-regulation) system?

References

Altmann, Gabriel (1980). Prolegomena to Menzerath's law. *Glottometrika*. 2, 1–10.

Altmann, Gabriel (1983). Das Piotrowski-Gesetz und seine Verallgemeinerungen. In: Best, Karl-Heinz, Kohlhase, Jürgen (eds.). *Exakte Sprachwandelforschung.* Göttingen: Herodot, 54–90.

Altmann, Gabriel – von Buttlar, Haro – Rott, Walter – Strauß, Udo (1983). A law of change in language. In: Brainerd, B. (ed.). *Historical Linguistics*. Bochum: Brockmeyer, 104–115.

Crystal, David (2000). *Language Death.* Cambridge: CUP.

Labov, William (1994). *Principles of Linguistic Change. Volume I – Internal Factors.* Hoboken: Blackwell.

Labov, William (2001). *Principles of Linguistic Change. Volume II – External Factors.* Hoboken: Blackwell.

Mann–Whitney U Test. Available at: https://en.wikipedia.org/wiki/Mann–Whitney_U_test.
Piotrovskaja, Anna A. – Piotrovskij, Rajmund G. (1974). Matematičeskie modeli v diachronii i tekstoobrazovanii. In: *Statistika reči i avtomatičeskij analiz teksta.* Leningrad: Nauka, 361–400.
Wachtel, Michael (2004). *The Cambridge Introduction to Russian Poetry.* Cambridge: CUP.
Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley Press.
Zörnig, Peter – Stachowski, Kamil – Popescu, Ioan-Iovitz – Miyangah, Tayebeh Mosavi – Mohanty, Panchanan – Kelih, Emmerich – Chen, Ruina – Altmann, Gabriel (2015). *Descriptiveness, Activity and Nominality in Formalized Text Sequences.* Lüdenscheid: RAM-Verlag.

Sources

Baudelaire, Charles. Available at: <https://fr.wikisource.org/wiki/Les_Fleurs_du_mal>.
Coleridge, Samuel Taylor. Available at: <https://en.wikipedia.org/wiki/Sonnets_on_ Eminent_Characters>.
Corbière, Tristan. Available at: <https://fr.wikisource.org/wiki/Les_Amours_jaunes>.
Karásek, Jiří ze Lvovic. Available at: <http://sonety.blog.cz/0810/jirikarasek-ze-lvovic-vsechny-sonety-ze-sb- zazdena-okna-1894>.
Kvapil, Jaroslav. Available at: <http://sonety.blog.cz/0906/jaroslav-kvapil-1868-1950-intermezzo-sonety-7-sonetu-ze-sbirky-ruzovy-ker-1890>.
Mácha, Karel Hynek. Available at: <http://sonety.blog.cz/0707/karel-hynek-macha-vsechny-sonety-znelky-vcetne-uvodniho-sonetu-z-maje-v-cestine-a-v-anglickem-prekladu-jamese-naughtona>.
Mallarmé, Stéphane. Available at: <https://fr.wikisource.org/wiki/Poésies_(Mallarmé,_1914,_8e_éd.)>.
Nezval, Vítězslav. Available at: <http://sonety.blog.cz/1205/vitezslav-nezval-vsech-16-sonetu-ze-sb-skleneny-havelok-1932>.
Novinky1. Available at: <https://www.novinky.cz/krimi/467193-v-plzni-mela-utocit-kyselinou-prostitutka-tri-lide-skoncili-v-nemocnici.html>.
Novinky10. Available at: <https://www.novinky.cz/krimi/467147-zlodej-odnesl-z-domu-na-teplicku-sest-milionu.html>.
Novinky2. Available at:
< https://m.novinky.cz/articleDetails?aId=467200&sznu=ZHSZqngw13ZMnN4I>.
Novinky3. Available at: <https://www.novinky.cz/krimi/467187-na-rychnovsku-horel-byvaly-veprin-skoda-temer-milion.html>.
Novinky4. Available at: <https://www.novinky.cz/krimi/467173-zena-lhala-ze-je-tehotna-a-chtela-po-muzi-dalsi-a-dalsi-penize.html>.
Novinky5. Available at:
https://m.novinky.cz/articleDetails?sznu=aC9tLITsGStYlAgF&sId=1&aId=467169&cId=0&mId=998>.
Novinky6. Available at: <https://www.novinky.cz/krimi/467158-ridic-horiciho-auta-v-jablonci-nadychal-pres-ctyri-promile.html>.
Novinky7. Available at:
<https://m.novinky.cz/articleDetails?aId=467159&sznu=lMIFWNCwkbNSPeEt>.
Novinky8. Available at: <https://www.novinky.cz/krimi/467196-zlodej-v-dolnim-dvoristi-vyhodil-bankomat-do-povetri.html>.
Novinky9. Available at: <https://www.novinky.cz/krimi/467161-byk-na-jatkach-na-jihlavsku-zautocil-na-zamestnance-a-tezce-ho-zranil.html>.

Rossetti, Dante Gabriel. Available at: <https://www.poemhunter.com/dante–gabriel–rossetti/poems/>.

Sovalin, Vladimir Sergeyevich – Velikanova, L. O. (eds.) [1986]. *Russkij sonet XVIII – natchalo XX veka*. Moscow: Moskovslij rabotchij.

Wolker, Jiří. Available at: <http://www.sosoom-karvina.cz/media/e-knihovna/wolker-tezka-hodina.pdf>.

Wordsworth, William. Available at: <www.sonnets.org/wordsworth.htm>.

Glottometrics 42, 2018, 39-45

Polysemy of some Parts of Speech

Emmerich Kelih[1], Sergey Andreev[2], Gabriel Altmann

Abstract. The article analyzes the distribution of polysemy with nouns, verbs and adjectives in German, Russian, Slovak, Italian and Hungarian. An adequate model has been found and the problems of data collection, computation and modeling were discussed.

Keywords: Polysemy, parts of speech, homogeneity, German, Hungarian, Italian, Russian, Slovak,

In the present article we use the results attained by Levickij, Drebet and Kiiko (1999) who analyzed the German DUDEN-dictionary (1989) and for texts they used Ortmann's (1975) dictionary in order to state the polysemy of verbs, nouns and adjectives in German. The Russian data have been won from the *Malyj akademicheskij slovar* (1999) considering the letter "t", the Slovak data from *Krátky slovník slovenského jazyka* (1989), considering only letter "b", and for Hungarian we used the *Értelmezö szótár* (2007), considering only letter "p". The Italian data, letter "a" have been won from *Grande Dizionario Italiano*.

Polysemy can be measured in different ways but it can be shown that the number of nouns/verbs/adjectives having x = 1,2,3,… meanings abides by a regular function. The Ukrainian authors applied the Poisson distribution and the mixed geometric distribution, both having two parameters, with good results, however, without substantiating them. Modeling of distributions has been usually performed by applying a distribution, i.e. a normalized function. Since we know that mathematical modeling does not represent truth but merely a capturing of our conceptual picture of the reality by a mathematical form which can be further processed and evaluated, we shall use here a simple function. Each word, whether created or borrowed, has at its birth one meaning but according to the domain of its meaning it can soon become polysemic. However, polysemy is effort reduction for the speaker, but increases of effort for the hearer. Hence the field of polysemy must be restructured. The hearer prefers words with one meaning and if the polysemy increases, he tries to reduce it, e.g. when he himself is the speaker. The ideal control of this phenomenon is given in such a way that the relative rate of change of polysemy is constant and negative, i.e. (cf. Kelih, Altmann 2015)

(1) $dy/y = -b*dx,$

whose solution yields the exponential function $y = a*exp(-b*x)$. Here, the parameter b is the control instrument applied both by speaker and the hearer. Now, since one computes only words having at least one meaning and omits all polysemies which are not represented in the dictionary, one may add -1 to y in the left-hand side of formula (1) and obtains

(2) $y = 1 + a*exp(-b*x)$.

[1] Emmerich Kelih, Universität Wien, E-Mail_emmerich.kelih@univie.ac.at
[2] Sergey Andreev. Smolensk State University, 214000 Smolensk Przhevalskij str. 4, Russia. E-mail: smol.an@mail.ru

This formula has been used also for modeling diversification (cf. Altmann 2018). It has the advantage of being simple, well substantiated by Köhler's (2005) synergetic linguistics and easy to compute. The parameter *a* depends on the most frequent polysemy, i.e. on the frequency of polysemy x = 1 and may be quite different both for different parts-of-speech and for different languages.

The fit of (1) to the results of counting by Levickij, Drebet and Kiiko (1999) are presented in Table 1.

Table 1
Fit of the formula (1) to the polysemy of German verbs

Polysemy	Number of verbs	Exponential fit
1	1385	1395.54
2	721	690.37
3	341	341.78
4	145	169.46
5	62	84.28
6	37	42.17
7	24	21.35
8	9	11.06
9	5	5.97
10	6	3.46
11	5	2.22
12	5	1.60
13	2	1.30
14	2	1.15
15	5	1.07
16	2	1.04
17	1	1.02
18	1	1.01
20	2	1.00
22	2	1.00
23	1	1.00
24	1	1.00
26	1	1.00
a = 2821.0487, b = 0.7045, R^2 = 0.9990		

The same procedure is applied to the polysemy of nouns presented in Table 2

Table 2
Fit of the formula (1) to the polysemy of German nouns

Polysemy	Number of nouns	Exponential fit
1	1571	1595.55
2	913	821.34
3	342	423.03
4	214	218.12
5	108	112.70
6	51	58.47

7	28	30.56
8	20	16.21
9	9	8.82
10	8	5.03
11	6	3.07
12	2	2.07
13	2	1.56
14	1	1.28
15	2	1.15
16	1	1.07
a = 3099.4518, b = 0.6646, R^2 = 0.9944		

The results of fitting for adjectives are presented in Table 3.

Table 3
Fit of the formula (1) to the polysemy of German adjectives

Polysemy	Number of adjectives	Exponential fit
1	213	214.56
2	125	119.42
3	62	66.66
4	36	37.41
5	23	21.19
6	12	12.19
7	6	7.21
8	6	4.44
9	2	2.91
10	1	2.08
21	1	1.00
a = 385.1566, b = 0.5897, R^2 = 0.9985		

As can be seen, the determination coefficient is in all cases greater than 0.99. The smaller the number of classes, the smaller is the parameter *b* but this observation must be tested in various languages.

As soon as one begins to perform a classification of the field, problems arise. All classifications in linguistics are some conceptual constructs set up differently by different researchers. The problem is that the whole field is decomposed differently but quite legally. Some researchers go in detail and obtain more classes, other ones begin with a small number of classes and subdivide each class separately. The question of truth cannot be asked because both language and its analysis are our creations. Nevertheless, one should set up a criterion whose violation is, unfortunately, no sign of falseness of the analysis or classification but rather a sign of forgetting the boundary conditions. If in a classification there is no quantification, then the only criterion is the rank-frequency order proposed already by G.K. Zipf.

This way of distinguishing POS does not function in all languages but in Slavic languages it is quite evident. For Russian data we obtain the results concerning polysemy as presented in Table 4.

Table 4
Polysemy of some Russian parts of speech (letter "t")

Polysemy	Nouns	Exp	Adjectives	Exp.	Verbs	Exp.
1	1025	1023.88	511	511.63	146	146.95
2	248	255.99	155	150.75	79	76.05
3	75	64.57	38	44.92	38	39.59
4	31	16.85	13	13.88	21	20.84
5	6	4.95	5	4.78	9	11.20
6	6	1.98	2	2.11	7	6.25
7	3	1.25	2	1.32	1	3.70
8	2	1.06	2	1.10	3	2.39
9	1	1.02	-	-	-	-
10	-	-	-	-	-	-
11	1	1.00	-	-	-	-
12			1	1.00	-	-
13			-	-	-	-
14			1	1.00	-	-
15					1	1.01
22					1	1.00
	a = 4103.1876 b = 0.7199 R^2 = 0.9996		a = 1741.2182 b = 0.8152 R^2 = 0.9997		a = 283.8235 b = 1.5035 R^2 = 0.9998	

The high value of the determination coefficient shows that the tendency is very similar to that in German.

For the Slovak data we used the letter "b" and obtained the results presented in Table 5.

Table 5
Polysemy of some Slovak parts of speech (letter "b")

Polysemy	Nouns	Exp	Adjectives	Exp.	Verbs	Exp
1	507	506.67	272	271.95	103	102.70
2	98	101.26	30	30.84	33	34.48
3	29	20.88	8	4.29	13	12.03
4	5	4.94	2	1.36	6	4.63
5	3	1.78	2	1.04	5	2.20
6	-	-			1	1.39
7	-	-			2	1.13
8	-	-			-	-
9	-	-			-	-
10	1	1.00			-	-
11					1	1.00
	a = 2550.3127 b = 0.6180 R^2 = 0.9996		a = 2460.2309 b = 0.4533 R^2 = 0.9997		a = 8558.8704 b = 0.9002 R^2 = 0.9984	

For Italian, using the *Grande Dizionario Italiano* available on the Internet we obtained the data for letter "a" presented in Table 6. In Italian, some words may be at the same time verbs,

nouns and adjectives, or nouns and adjectives, etc., a circumstance that has been taken into account. The fitting of noun polysemy using the exponential function was not sufficient. Adding one parameter to the differential equation we obtained the Zipf-Alekseev function $y = 1 + c*x^{(a+b*\ln(x))}$ yielding even an exact parameter *c* equal to the frequency of x = 1. In our data, this was the only exception.

Table 6
Polysemy of Italian parts of speech (letter "a")

Polysemy	Nouns	ZI-AL	Adjectives	Exp.	Verbs	Exp
1	1755	1755.00	784	783.38	394	388.77
2	418	417.79	166	171.15	140	163.83
3	131	133.86	46	38.00	95	69.37
4	61	52.56	21	9.05	31	29.71
5	20	23.90	2	2.75	17	13.06
6	9	12.22	2	1.38	9	6.06
7	4	6.92	-	-	-	-
8	4	4.32	-	-	2	1.89
9	3	2.95	1	1.00	-	-
10	-	-			1	1.16
11	-	-			1	1.07
12	1	1.49				
	a = -1.6023 b = -0.6794 c = 1753.9980 R^2 = 1.0000		a = 3597.5205 b = 0.6555 R^2 = 0.9995		a = 923.4463 b = 1.1525 R^2 = 0.9903	

For Hungarian, we used the *Értelmezö szótar* (2007) and analyzed the letter "p". The results are presented in Table 7.

Table 7
Polysemy of Hungarian parts of speech (letter "p")

Polysemy	Nouns	Exp	Adjectives	Exp.	Verbs	Exp
1	198	204.14	25	28.62	47	50.52
2	125	111.53	26	17.20	39	28.73
3	70	61.14	10	10.50	12	16.53
4	15	33.72	2	6.57	7	9.70
5	14	18.81	2	4.27	4	5.87
6	3	10.69	1	2.92	1	3.73
7	-	-	1	2.12		
9	-	-	1	1.39		
12	1	1.25				
	a = 373.3479 b = 1.6431 R^2 = 0.9787		a = 47.0923 b = 1.8738 R^2 = 0.8538		a = 88.4286 b = 1.7247 R^2 = 0.9186	

The number of Hungarian adjectives deviates slightly from the usual decreasing trend but the fitting is satisfactory. Analysis of further letters would surely improve the fitting.

From the above results one can derive some consequences: (1) In the given languages, all polysemies except for the nouns in Italian abide by the exponential law. The determination coefficient is in all cases very high and corroborates the model. (2) In the examined languages – though we analyzed for Russian, Slovak, Italian and Hungarian only one letter in the dictionary, the number of nouns is the greatest. (3) With the exception of Russian verbs, the parameter *b* is always smaller than 1.0. In Russian, this may be caused by the relatively small size of the sample (using only letter "t"). The situation in the Italian polysemy of nouns is different. (4) This kind of research should be performed in many languages – using monolingual dictionaries - in order to see whether the law is general enough. (6) The parameter *a* does not play any role, .but further research could reveal its behavior. (7) Polysemy and its frequency are part of the Köhlerian self-regulating cycle but for corroborating this regularity, one must test it in many languages. (8) What is the state of all the other parts of speech?

The exponential function we used here is a strong simplification. Still better results can be attained using the Zipf-Alekseev or the Menzerathian functions both having three parameters. The Zipf-Alekseev function yields one parameter which is almost equal to the frequency of x = 1. We were forced to use only in one case. But the future will bring, perhaps, also boundary conditions depending on the language type or text type. In any case, the research is not finished (cf. Kelih, Altmann 2015).

We performed this investigation because the German data stayed at our disposal. But in general, the distinction of lexical items in classical word classes is not always adequate. Some traces can be found in the syntax or in the morphology, but seldom in the dictionary. It is always possible to distinguish a noun, a verb and an adjective but in many cases we look at the data with Indo-European eyes. Consider the problems appearing in the Polynesian languages or any other analytic language. Is our way of specifying word classes adequate? If we can manage to delimit even one class, the study of its polysemy should be performed in order to see whether the model is adequate. If not, then either another model must be sought or the use of Latin grammar in other languages should be reformed.

A further problem is the question whether the meaning of a word in a compound should be considered. A word may be the head or a semantic modification of another word and it depends on the researcher which way he chooses. As can be seen, the number of problems will rather increase. At last, one must find the relation of the polysemy to other properties. But it will depend on the way of stating polysemy and parts of speech.

Since noun is in language and in philosophy our primary concept of reality, it can be expected that the number of nouns in the dictionary will be greater than that of the Aristotelian predicates of the first degree, namely adjectives and verbs. It we compare the above results, we find the numbers for the languages analyzed presented in Table 8. It should be remarked that if in the dictionary a word belonged to two or three classes at the same time, the polysemy has been computed separately.

Table 8
Numbers of nouns, adjectives and verbs in the given data

Language	**Nouns**	**Adjectives**	**Verbs**
German	3278	487	2765
Russian	1398	730	306
Slovak	643	314	164
Hungarian	426	68	110
Italian	2406	1022	690

As can be seen, the relation of the number of verbs and adjectives is not constant. This is caused, perhaps, by the fact that in four languages only one letter has been evaluated. That means, preliminarily, that one cannot conjecture hypotheses about the state of predicates of the first level. Further, this relation must be studied also in *texts* of the given languages in which some words may be repeated. In texts, the polysemy is strongly reduced – this is the aim of the text – but one can compute the extent of reduction.

Even if the numbers of individual POS is not equal – it depends on the extent of the dictionary – one may ask whether the two predicates of the first order are distributed homogeneously. Though the number of adjectives and verbs may be different, there may be a similarity/homogeneity in the distribution of polysemy. The answer can be obtained simply by performing the chi-square text for homogeneity of two samples and compute the appropriate formula. It can be shown that for the given cases there is homogeneity between adjectives and verbs only in Hungarian but since we analyzed only one letter the results is not definitive.

The situation in the dictionary may differ from that in texts. In the dictionary one considers all respective words, in a text one cares only for those that occur in it. Hence the investigation in texts may furnish quite different results. If in texts one considers each word and its polysemy in the dictionary, one obtains a strongly oscillating curve. If one considers only the three main parts of speech and at last, computes the distribution, one could obtain a quite characteristic function for an author, for the text type, for the language, for the development of texts, etc.

References

Altmann, G. (2018). *Unified modeling of diversification in language*. Lüdenscheid: RAM-Verlag.

Bánki, J., Bíró, Á, Szirmai, D. (eds.) (2007). *Értelmezö szótár.* Budapest: Tinta Könyvkiadó.

Duden (1989). *Deutsches Universalwörterbuch* 2nd ed. Mannheim/Vienna/Zürich: Duden-Verlag.

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.

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, nound and adjectives in the German language. *Journal of Quantitative Linguistics 6(2), 172-187*

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

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

Glottometrics 42, 2018, 46-55

Adnominal Valency Motifs in Sonnets

Sergey Andreev[1]

Abstract. Adnominal characteristics form an important element of syntactic structure of a sentence, being important both for its formal and semantic organization. One of possible approaches to the study of attributive relations consists in analyzing nouns from the point of view of their adnominal collocability. The number of attributes modifying nouns reflects their adnominal valencies, strings of which form the basis for the quantitative analysis of text structure. This article is devoted to capturing – in Russian sonnets – the distribution of adnominal valency sequences, modelled into motifs, by mathematical functions. The exponential function plus 1 demonstrates a very good fit.

Key words: *adnominal valency, motif, exponential function, sonnets*

The analysis of adnominal text structure can be conducted in two ways: the study of the distribution of adnominals, and the analysis of the distribution of nouns' adnominal valencies. In the first case, the study of adnominals is focused on the attributes which are used to modify nouns – on their type, number, often their position regarding the noun which they modify, the type of syntactic connections with the modified noun (coordination, government, and adjournment), etc.

Another approach is to study attributive relations in the direction from the head words – nouns modified by attributives: their semantic, derivational characteristics, syntactic position, etc. One of the most obvious features to be taken into account is the number of links which are observed between the given noun and its attributes – its adnominal valency (AV). The same as the valency of verbs in syntax or of derivational stems reflecting their collocability with affixes in word-building, this adnominal valency has a range of variance $AV \geq 0$.

In this study, we used the second approach.

The data-base includes 46 sonnets written by prominent Russian authors during the period of two and a half centuries (see Appendix). These sonnets can be divided into 4 subgroups according to the time of creation:

Period 1 – 18th century (7 sonnets);

Period 2 – 19th century (18 sonnets);

Period 3 – two first decades of the 20th century (16 sonnets);

Period 4 – the end of the 20th–21st century (5 sonnets).

The list of adnominals includes the following types: adjectives, participles, determiners (possessive, demonstrative, qualifying pronouns), infinitives, adverbial modifiers, nouns in the genitive, dative, instrumental cases, nouns in prepositional constructions, adjectival and participial constructions, appositions, attributive clauses.

To demonstrate the approach to annotating the adnominal valencies, let us take the sonnet by V. Brusov "Yegipetskyj Rab" /Egyptian Slave/ (T.28).

Я жалкий раб царя. С восхода до заката,
Среди других рабов, свершаю тяжкий труд,
И хлеба кус гнилой - единственная плата
За слёзы и за пот, за тысячи минут.

[1] Sergej Andreev, Smolensk State University, 214000 Przhevalskij str. 4, Smolensk, Russia. Email: smol.an@mail.ru.

Когда порой душа отчаяньем объята,
Над сгорбленной спиной свистит жестокий кнут,
И каждый новый день товарища иль брата
В могилу общую крюками волокут.
Я жалкий раб царя, и жребий мой безвестен;
Как утренняя тень, исчезну без следа,
Меня с лица земли века сотрут, как плесень;
Но не исчезнет след упорного труда,
И вечность простоит, близ озера Мерида,
Гробница царская, святая пирамида.

The sonnet has 25 adnominals and 29 nouns (N). Adnominals include 12 adjectives (A): *жалкий раб*, *тяжкий труд*, *кус гнилой*, *единственная плата*, *жестокий прут*, *новый день*, *могилу общую*, *жалкий раб*, *утренняя тень*, *упорного труда*, *гробница царская*, *святая пирамида*, 1 adjectivized participle (A-PT): *сгорбленной спиной*, 1 apposition (AP): *озеро Мерида*, 5 nouns in genitive case (G): *раб царя* (twice), *хлеба кус*, *лица земли*, след *труда*, 3 nouns in different cases with a preposition (PR): *плата за слезы*, *плата за пот*, *плата за тысячи*, 2 qualitative pronouns (DETQ): *других рабов*, *каждый день*, and 1 possessive pronoun (DETS): *жребий мой*.

Using tags, it is possible to describe the adnominal structure of the sonnet as follows:

A, N, G, N, N, DETQ, N, A, N, G, N, A, A, N, PR, PR, PR, N, N, A-PT, N, A, N, DETQ, A, N, N, N, N, A, N, A, N, G, N, DETS, A, N, N, N, G, N, N, N, A, N, G, N, N, APX, N, A, A, N.

The first noun *раб* (*slave*) is modified by two adnominals – the adjective *жалкий* (*miserable*), which precedes the noun, and – as it is the noun in the genitive case – *царя* (*of the tsar*) in postposition (V = 2). The next two nouns (*восхода*, *заката*) do not have any modifiers at all, having thus zero valencies. The forth noun is modified by a determiner – the qualitative pronoun *другой* (thus, its valency is 1), the fifth noun – by an adjective *тяжкий* (valency 1), the sixth noun has valency 2 (modified by a noun in the genitive case, and an adjective), the seventh one possesses valency 4 (one adjective and 3 nouns in prepositional cases), etc. Inserting the values of the valencies into the adnominal structures of the sonnet, one gets the following sequence, in which adnominal valency of each noun is expressed in brackets:

A, N(2), G, N(0), N(0), DETQ, N(1), A, N(1), G, N(2), A, A, N(4), PRR, PRR, PRR, N(0), N(0), PT, N(1), A, N(1), DETQ, A, N(2), N(0), N(0), N(1), A, N(0), A, N(2),GR, N(1), DETS, A, N(1), N(0), N(1),G, N(0), N(0), N(1), A, N(1), G, N(0), N(1), AP, N(1), A, A, N(1).

The adnominals being omitted, the sequence becomes: 2, 0, 0, 1, 1, 2, 4, 0, 0, 1, 1, 2, 0, 0, 1, 0, 2, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1. Thus, text 28 is characterized by 29 nouns with the following valencies: 0 (11 nouns), 1 (13), 2 (4), 4 (1).

Table 1 presents the data of the adnominal valencies in all 46 texts.

Table 1
Frequencies of noun adnominal valencies

Valency	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13
0	2	13	17	12	2	12	18	25	11	14	13	4	0
1	9	7	8	4	10	4	6	6	12	10	8	9	11

2	3	1	1	3	1	–	–	–	1	0	2	4	1
3	–	–	–	–	–	–	–	–	–	0	–	–	–
4	–	–	–	–	–	–	–	–	–	1	–	–	–

Valency	T14	T15	T16	T17	T18	T19	T20	T21	T22	T23	T24	T25	T26
0	0	2	8	3	8	9	5	13	12	11	7	12	9
1	17	14	12	19	13	14	17	8	16	16	10	11	4
2	3	6	2	–	5	–	1	1	1	1	3	–	–
3	0	3	–	–	1	–	–	–	–	–	–	–	–
4	1	–	–	–	–	–	–	–	–	–	–	–	–

Valency	T27	T28	T29	T30	T31	T32	T33	T34	T35	T36	T37	T38	T39
0	7	11	7	22	14	4	6	16	8	11	8	10	11
1	8	13	11	7	5	18	17	7	3	12	6	11	13
2	1	4	4	–	5	4	1	1	–	2	4	4	–
3	1	0	–	–	–	–	1	–	–	–	–	–	–
4	–	1	–	–	–	–	–	–	–	–	–	–	–

Valency	T40	T41	T42	T43	T44	T45	T46
0	7	8	7	17	11	14	11
1	8	13	5	3	11	7	6
2	3	5	2	–	1	2	2
3	1	–	–	–	1	–	1
4	–	–	–	–	–	–	–

Since the numbers in many sonnets are too small, we added them. The sum of valencies in all the sonnets (x = 0, 1, 2, 3, 4) can be captured by Lorentzian function (Wimmer, Altmann 2005) with the proportion of variance explained (= determination coefficient) being 99.71% (Table 2).

Table 2
Fitting the Lorentzian function to the adnominal valencies in all sonnets

Adnominal valency	Frequency	Lorentzian function
0	452	451.89
1	460	460.32
2	87	76.19
3	9	28.51
4	3	14.70
a = 1221.9384, b = 0.5037, c = -0.3858, R^2 = 0.9971		

The study of adnominal distribution can be further conducted in two ways – using the frequencies of adnominal valencies directly, or by transferring them into a number of sequences, organized according to certain rules. In the first case, we obtain the type of distribution of the adnominal valencies directly, on the surface level; in the second case, the distribution is studied at a higher level, reflecting less obvious general tendencies.

In this study, the second approach was chosen. It is based on the formation of a certain type of sequences – motifs, discovered by R. Köhler (2008, 2015) and defined as non-decreasing sequences of numbers.

Using this rule for the above-mentioned sonnet by Brusov *Yegipetskyj Rab*, one gets the following motifs, placed in brackets: [2], [0, 0, 1, 1, 2, 4], [0, 0, 1, 1, 2], [0, 0, 1], [0, 2], [1, 1], [0, 1], [0, 0, 1, 1], [0, 1, 1, 1].

Motifs possess their own properties and can be analyzed from different angles; one is free to measure the number of their elements, distances between the same types in the text, their structure, etc. (Cech, Vincze, Altmann 2016; Köhler, Naumann 2016; Sanada 2010; Liu, Fang 2016; Wang 2016). One can suggest that in our case, a possible feature for motif typology is the total valency of all its elements, received by adding all valencies in a motif. According to it, the motifs of *Yegipetskyj Rab* can be classified as follows: the first motif, which consists of one element only, has valency 2, the total valency of the second motif is 8, to the third motif, total valency 4 is ascribed, etc. As a result, every sonnet is described by a string of motifs of different types which are characterized by their total adnominal valency (TAV).

The results of such classification of the motifs in all the sonnets are presented in Table 3.

Table 3
Adnominal valency motifs in Russian sonnets

Period	Text	Total adnominal valencies in motifs										
		0	1	2	3	4	5	6	7	8	9	10
1	Text 1	–	1	1	–	3	–	–	–	–	–	–
1	Text 2	1	3	1	–	1	–	–	–	–	–	–
1	Text 3	–	5	1	1	–	–	–	–	–	–	–
1	Text 4	1	2	2	–	1	–	–	–	–	–	–
1	Text 5	–	1	–	1	–	–	–	–	1	–	–
1	Text 6	–	2	1	–	–	–	–	–	–	–	–
1	Text 7	1	4	1	–	–	–	–	–	–	–	–
2	Text 8	1	1	1	1	–	–	–	–	–	–	–
2	Text 9	1	3	1	3	–	–	–	–	–	–	–
2	Text 10	1	5	1	1	1	–	–	–	–	–	–
2	Text 11	1	3	3	1	–	–	–	–	–	–	–
2	Text 12	1	2	–	1	–	1	–	1	–	–	–
2	Text 13	–	–	–	1	–	–	–	–	–	–	1
2	Text 14	–	–	–	1	1	1	1	–	–	1	–
2	Text 15	–	1	1	–	3	1	–	–	–	1	–
2	Text 16	–	4	1	1	–	–	–	1	–	–	–
2	Text 17	–	–	–	–	1	–	1	–	–	1	–
2	Text 18	–	1	1	1	2	1	–	1	–	–	–
2	Text 19	–	1	3	–	–	–	–	1	–	–	–
2	Text 20	–	–	2	–	–	–	–	1	–	1	–
2	Text 21	1	3	2	1	–	–	–	–	–	–	–

2	Text 22	–	3	2	2	–	1	–	–	–	–	–
2	Text 23	–	3	3	1	–	–	1	–	–	–	–
2	Text 24	–	1	1	1	1	–	1	–	–	–	–
2	Text 25	–	6	–	–	–	1	–	–	–	–	–
3	Text 26	1	2	1	–	–	–	–	–	–	–	–
3	Text 27	–	1	–	1	1	1	–	–	–	–	–
3	Text 28	–	2	4	1	1	–	–	–	1	–	–
3	Text 29	1	2	2	3	1	–	–	–	–	–	–
3	Text 30	1	7	–	–	–	–	–	–	–	–	–
3	Text 31	–	5	3	–	1	–	–	–	–	–	–
3	Text 32	–	1	1	1	1	–	–	1	–	1	–
3	Text 33	–	–	2	1	–	–	–	1	1	–	–
3	Text 34	1	2	–	1	1	–	–	–	–	–	–
3	Text 35	1	3	–	–	–	–	–	–	–	–	–
3	Text 36	1	–	–	2	1	–	1	–	–	–	–
3	Text 37	1	2	2	1	2	–	–	–	–	–	–
3	Text 38	–	2	2	1	–	2	–	–	–	–	–
3	Text 39	1	1	2	–	–	–	–	–	1	–	–
3	Text 40	1	1	2	4	–	–	–	–	–	–	–
3	Text 41	–	4	2	2	1	1	–	–	–	–	–
4	Text 42	–	2	–	1	1	–	–	–	–	–	–
4	Text 43	1	3	–	–	–	–	–	–	–	–	–
4	Text 44	1	2	4	–	–	–	1	–	–	–	–
4	Text 45	1	2	1	1	1	–	–	–	–	–	–
4	Text 46	–	1	1	–	–	2	–	–	–	–	–
Total		21	100	58	38	26	12	6	7	4	5	1

Adding the motifs of the same type, we receive the frequencies which were ranked. After ordering the ranks, we obtained a sequence of rank frequencies whose ranking trend was caught very satisfactorily by the exponential function with added 1. The exponential function with added 1 (Andreev, Popescu, Altmann 2017, 34–35) is defined as:

$$f_x = 1 + a * \exp^{-bx}$$

where a and b are parameters, $x \geq 1$. Parameter b shows the decrease of the function.

The results of such counts and fitting of the function are represented in Table 4.

Table 4
Fitting the exponential function with added 1 to the ranking
of total adnominal valency motifs in 46 Russian sonnets

Total adnominal valency	Frequency	Exponential ft. +1
1	100	97.52
2	58	62.25
3	38	39.87
4	26	25.67
5	21	16.65
6	12	10.93
7	7	7.30
8	6	5.00
9	5	3.54
10	4	2.61
11	1	2.02
a = 5.5794, b = 0.0185, R^2 = 0.9941		

As mentioned above, all the sonnets may be divided into four groups depending on the period when they were written. In Tables 5–8, the ordered frequencies of the TAV motifs and those predicted by the exponential function plus 1 are listed.

Table 5
Fitting the exponential function with added 1 to the ranking of total adnominal valency motifs in Russian sonnets of the 18th century

Adnominal valency	Number	Exponential ft. +1
1	18	17.70
2	7	8.22
3	5	4.12
4	3	2.35
5	2	1.58
6	1	1.25
a = 4.5734, b = 0.0884, R^2 = 0.9846		

Table 6
Fitting the exponential function with added 1 to the ranking of total adnominal valency motifs in Russian sonnets of the 19th century

Adnominal valency	Number	Exponential ft. +1
1	37	36.13
2	22	23.51
3	16	15.42
4	9	10.24
5	6	6.92

6	6	4.80
7	5	3.43
8	4	2.56
9	4	2.00
10	1	1.64
a = 3.1688, b = 0.0287, R^2 = 0.9854		

Table 7
Fitting the exponential function with added 1 to the ranking
of total adnominal valency motifs in Russian sonnets of the 20th century

Length	Number	Exponential ft. +1
1	35	35.29
2	23	23.58
3	18	15.86
4	10	10.78
5	9	7.44
6	4	5.24
7	3	3.79
8	2	2.84
9	1	2.21
10	1	1.80
a = 2.6647, b = 0.0243, R^2 = 0.9889		

Table 8
Fitting the exponential function with added 1 to the ranking of total adnominal valency motifs
in Russian sonnets of the modern period (second part of the 20th – 21st century)

Length	Number	Exponential ft. +1
1	10	10.04
2	6	5.70
3	3	3.45
4	2	2.27
5	2	1.66
6	2	1.34
7	1	1.18
a = 1.5479, b =0.0577, R^2 = 0.9847		

As can be seen in Tables 5–8, the exponential function + 1 is very good for capturing the distribution with more than the 98-percent fit.

The values of *b*-parameter are shown graphically in Fig.1.

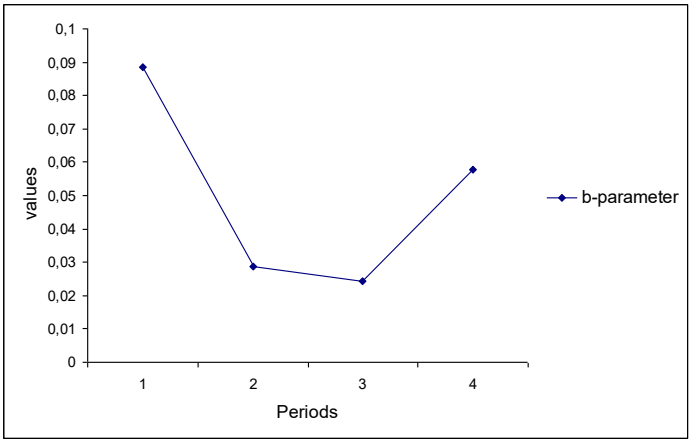

Fig. 1. Values of the parameter *b*

The 18th century is the time when sonnets as a genre appeared in Russian literature. This period – as well as the modern one – are characterized by a higher value of *b*-parameter, and thus by a stronger decrease of the function.

The sonnets of the second and the third periods are of special interest because these periods reflect a considerable rise of public interest in poetry in general and to the genre of sonnet (especially in the third period) when poetry was more popular than prose.

Their comparison reveals a high degree of similarity of both observed and predicted frequencies (Tables 6 and 7) as well as of the function parameter *b* (Fig. 1). This was not to be expected, because during 100 years, which separate these two periods, the genre of sonnets in Russia has been intensively developing. This included the changes of metric system (iambic hexameter or pentameter was often replaced by iambic tetrameter, trimeter, and even by trochee and other meters), considerable changes in the rhyming system, and some others. Yet, nevertheless, the rank sequences demonstrate unexpectedly high similarity.

The very good fitting of the motif ranking in these two (precision of $R^2 = 0.99$) and, generally, in all four periods shows that the distribution of motifs, based on total adnominal valencies, has been kept within a limited range, and thus may be supposed to be a certain constant in at least this genre of poetry. Further research, of course, is needed to confirm or reject this preliminary conclusion and show whether this constant exists in other genres of poetry and in prose.

It should be noted that motifs of this type proved to be very useful means of classifying strings of valencies in sonnets and can be used as the basis for comparison of other genres of poetry, especially with less strict formal rules of organization. The exponential function plus 1 was found to be very good at capturing such TAV motifs distribution.

References

Andreev, S., Popescu, I-I., Altmann, G. (2017). On Russian adnominals. *Glottometrics* 35, 64–83.

Čech, R., Vincze, V., Altmann, G. (2016). On motifs and verb valency. In: Liu, H., and Liang, J. (eds.), *Motifs in language and text*. De Gruyter, Mouton, 13–36.

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory* 1(1), 115–119.

Köhler, R. (2015). Linguistic motifs. In: Mikros, G.K., Mačutek, J. (eds.). *Sequences in Language and Text*, Berlin/Boston: de Gruyter Mouton, 89–108.

Köhler, R., Naumann, S. (2016). Syntactic text characterisation using linguistic S-motifs. *Glottometrics* 34, 1–8.

Liu H., Fang Yu (2016). Quantitative Aspects of Hierarchical Motifs. In: Emmerich Kelih, Róisín Knight, Ján Mačutek, Andrew Wilson (eds.), *Issues in Quantitative Linguistics*. 4. Dedicated to Reinhard Köhler on the occasion of his 65th birthday, 9–26.

Sanada, H. (2010). Distribution of motifs in Japanese texts. In: Grzybek, P., Kelih, E., Mačutek, J, (eds.), *Text and Language. Structures, functions, interrelations, quantitative perspectives*: Wien: Praesens, 183–194.

Wang Y. (2016). Quantitative Genre Analysis Using Linguistic Motifs. In: Liu, H., and Liang, J. (eds.), *Motifs in language and text.* De Gruyter, Mouton, 165–180.

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

Appendix

List of sonnets

Number	Author	Title	Period
Text 1	V. Trediakovskij	Sonet	I
Text 2	V. Trediakovskij	Sonet iz seja grecheskija rechi	I
Text 3	M. Heraskov	Sonet i jepitafija	I
Text 4	M. Heraskov	“Kol' budu v zhizni ja nakazan nishhetoju…”	I
Text 5	A. Rzhevskij	Sonet, zakljuchajushhij v sebe tri mysli:	I
Text 6	A. Rzhevskij	Sonet, tri raznye sistemy zakljuchajushhij	I
Text 7	I. Dmitriev	Sonet	I
Text 8	V. Zhukovskij	Sonet	II
Text 9	A. Del'vig	N. M. Jazykovu	II
Text 10	A. Del'vig	Vdohnovenie	II
Text 11	A. Del'vig	“Ja plyl odin s prekrasnoju v gondole...”	II
Text 12	E. Baratynskij	“My p'jom v ljubvi otravu sladkuju...”	II
Text 13	E. Baratynskij	“Hotja ty malyj molodoj... ”	II
Text 14	N. Jazykov	K. K. Janish	II
Text 15	N. Jazykov	“Na prazdnik vash prines ja dva priveta...”	II
Text 16	A. Pushkin	Sonet	II
Text 17	A. Pushkin	Pojetu	II
Text 18	A. Pushkin	Madona	II
Text 19	V. Benediktov	Priroda	II
Text 20	V. Benediktov	Kometa	II
Text 21	V. Benediktov	Vulkan	II
Text 22	V. Benediktov	Groza	II

Text 23	V. Benediktov	Cvetok	II
Text 24	V. Benediktov	“Krasavica, kak rajskoe viden'e...”	II
Text 25	V. Benediktov	“Kogda vdali ot suety vsemirnoj...”	II
Text 26	F. Sologub	Sonet	III
Text 27	V. Brjusov	Sonet	III
Text 28	V. Brjusov	Egipetskij rab	III
Text 29	A. Blok	“Ne ty l' v moih mechtah, pevuchaja, proshla…”	III
Text 30	V. Ivanov	Pritcha o devah	III
Text 31	V. Ivanov	Hramina chuda	III
Text 32	M. Voloshin	Venok sonetov. Sonet 1	III
Text 33	M. Voloshin	Venok sonetov. Sonet 2	III
Text 34	I. Severjanin	Sonet	III
Text 35	A. Belyj	Prosti	III
Text 36	N. Gumilev	Popugaj	III
Text 37	N. Gumilev	Roza	III
Text 38	S. Esenin	Moej carevne	III
Text 39	K. Bal'mont	Mikel' Andzhelo	III
Text 40	K. Bal'mont	Leonardo da vinchi	III
Text 41	K. Bal'mont	Marlo	III
Text 42	S. Gorodeckij	Mudrost'	IV
Text 43	I. Sel'vinskij	Sonet	IV
Text 44	V. Prokoshin	Deti RA	IV
Text 45	T. Averina	“Ochnjosh'sja– pogruzhjon po grud' v boloto…”	IV
Text 46	N. Beljaeva	“Schitaju vnov' chasy do nashej vstrechi…”	IV

Glottometrics 42, 2018, 56-74

A Study of Russian Adnominals

Sergey Andreev[1]

Abstract. In the present article, the development of adnominal types in Russian will be studied. On the selected texts from 1965 to 2008, the use of adnominals in terms of frequencies and the development of motif types are studied.

Keywords: Russian, adnominals, motifs, length, Menzerathian function, Zipf-Alekseev function, evolution

1. Introduction

The study of adnominal modifiers already has a rich history and seems to continue developing. Their classification and use have been described by several authors – e.g., Best, Boschtan (2010), Givón (2001), Gunkel, Zifonun (2009), Halliday (2004), Rijkhoff (2004); their symmetry, weight, distribution, complexity, cohesion, scaling, motifs, similarity, distances, and aggregation have been analyzed (Köhler, Altmann 2014; Altmann 2015; Andreev, Popescu, Altmann 2017, Andreev, Fan, Altmann 2018), and the study of adnominals seems to obtain ever more weight in the recent time. The fact that adnominals are linguistically defined entities gives them the same status as some other syntactic-semantic units of the sentence have – e.g., parts-of-speech, parts of the sentences, etc. Describing the text, one can characterize it also by operating with adnominals, which can be transformed also into motifs – one can study Frumkina chains, etc.

In the present study, we merely take some texts in a historical sequence and study whether something changes. The tests can be performed traditionally, but, perhaps, there are some trends which have not yet been observed. The texts have been selected by chance, but if there is some tendency, it must be observable in any texts. For a thorough investigation, one would be forced purposefully to select texts belonging to the same text type, the sample of texts should be large enough, etc., but not even in statistics can one learn what a sufficiently large sample means.

The abbreviations are as follows (cf. Andreev, Popescu, Altmann 2017: 78–79):

A – adjective (*Бледное* лицо – Pale face; Человек *спокойный* – *Man calm).
ADV – adverb (Комната *наверху* – Room upstairs; *Назад* козырьком – *With the backwards peak).
AO – adjective in an elliptical construction (У меня есть один красный карандаш и один *синий*. – *I have one red pencil and one blue).
AP – apposition (Его *костюм, галстук, рубашка* – вся одежда была абсолютно новой – His suit, tie, shirt – all clothes were brand new; Незнакомец, *мужчина среднего возраста*, подошел ко мне – The stranger, a middle-aged man, came up to me).
APAJ – type of apposition based on adjoinment type of connection with the headword, i.e. its syntactic links with the head word are not based on either agreement, or

[1] Sergej Andreev, Smolensk State University, 214000 Przhevalskij str. 4, Smolensk, Russia. Email: smol.an@mail.ru.

government (Гостиница «*Байкал*»; слово «*привет*» – The hotel Baikal; the word 'hello').

APX – type of apposition expressed by a proper name which agrees in number, case and gender with the appositive (Хирург *Иванов*, капитан *Смоллетт* – Surgeon Ivanov, Captain Smollett).

AY – adjectival phrase (*Бледное от волнения* лицо – *Pale from anxiety* face; Лицо, *бледное от волнения* – Face pale from anxiety).

CN – compound word with attributive relations of two stems, one of which is a modifier (*Страдальцы*-мальки – Sufferers-fries; Спортсмен-*чемпион* – sportsman-champion).

DAT – dative case (Письмо *другу* – Letter to a friend).

DETF – demonstrative pronoun (*Этот* дом – This house; Книга *эта – моя*. – *Book this is mine).

DETH – indefinite pronoun (*Какие-то* книги – Some books; Книги *какие*-то – *Books some).

DETN – negative pronoun (*Никакой* ошибки – No mistake; Знакомств *никаких* не желаю – *Acquaintances any I do not want).

DETQ – qualifying pronoun (*Все* книги – All the books; Книги *все* – Books all).

DETS – possessive pronoun (*Его* друг – His friend; Книги *мои* здесь – *Books mine are here).

DETV – relative pronoun (Я спросил, *какая* книга пропала – I asked which book was missing; Интересно, экономия *какая* будет – It is interesting what will happen to economy).

DETW – interrogative pronoun (*Какая* книга пропала? – Which book is missing?; А машина *какая* там была? – *And car which was there?)

G – genitive case (*Отца* брат – *Of the father brother; Книга *брата* – Book of the brother).

I – infinitive (*Поехать* желание было, *собирать* вещи желания не было – *To go there was a wish, to pack things – there was no wish; Желание *узнать* – Wish to learn).

INSTR – instrumental case (Восхищение *книгой* – Fascination with the book).

PR – prepositional noun (*на плече* чехол – On the shoulder a cover; Книга *для детей* – Book for children).

PT – participle (*Разбитый* стакан – Broken glass; Чудеса *невиданные* – Miracles unseen).

PTY – participial construction (*Разбитый* на куски стакан – Broken to pieces glass; Книга, *потерянная несколько дней назад* – Book lost a few days ago).

RC – subordinate clause (Это тот человек, *который может нам помочь* – This is the man who can help us; Вот план, *что делать дальше* – Here is a plan what to do next; Это – место, *где мы встретились* – This is the place where we met).

Here we omit the positioning (left and right) of the adnominal because we are interested only in its existence.

1. Frequency

First, let us consider the frequencies of individual adnominals. We conjecture that their ranking yields one of the usual functions, namely either the power function, the Zipf-Mandelbrot function or the Menzerathian function, etc., all used in Andreev, Popescu, Altmann (2017). We shall try as first also the simple exponential function because its

differential equation is the simplest of all used in linguistics. To each function we add 1 because there are no smaller frequencies. For the texts attached at the end of the article, we obtain the results presented in Table 1. The rank-order of adnominals is not the same in all texts. Hence, one can study each type of adnominal separately and set up a historical sequence of rank-order of individual adnominals.

Table 1
Frequencies of adnominals

Adnom	Text								
New	T 1 1965	T 2 1965	T 3 1965	T 4 1965	T 5 1969	T 6 1992	T 7 1993	T 8 1996	T 9 1999
A	39	79	74	38	79	84	84	87	56
ADV	–	–	1	2	–	–	–	–	–
AO	2	2	8	–	7	–	1	1	–
AP	3	5	2	6	1	17	5	5	3
APAJ	1	2	8	–	10	3	3	5	–
APX	–	–	1	–	–	–	–	3	–
AY	–	1	1	3	–	–	–	1	–
CN	–	3	2	1	1	1	2	–	2
DAT	–	–	–	2	–	–	1	1	–
DETF	4	11	15	6	8	10	4	12	6
DETH	–	4	–	4	–	–	–	2	1
DETN	–	–	–	–	–	–	–	–	–
DETQ	3	12	10	5	4	2	4	12	1
DETS	10	38	15	22	13	42	18	33	8
DETV	–	1	1	–	1	2	1	5	–
DETW	–	–	–	–	–	–	–	1	–
G	26	65	51	41	44	17	68	63	41
I	1	1	–	–	–	–	–	–	–
INSTR	–	–	1	–	2	–	–	–	–
PR	5	22	23	19	7	9	20	13	14
PT	5	19	17	19	16	11	9	10	13
PTY	8	14	6	14	14	3	9	12	12
RC	3	22	4	22	3	2	12	27	7
Sum	110	301	240	185	210	203	241	293	164

Adnom	Text							
New	T 10 2000	T 11 2002	T 12 2003	T 13 2007	T 14 2008	T 15 2008	T 16 2010	T 17 2011
A	82	61	117	29	320	53	125	31
ADV	3	–	3	–	–	–	1	1
AO	–	–	1	–	6	–	–	–
AP	26	3	8	6	11	2	20	3
APAJ	6	–	–	–	1	–	3	–
APX	–	–	12	–	1	1	1	–
AY	3	2	2	–	6	–	5	–

CN	4	1	2	–	6	–	4	1
DAT	2	2	–	–	–	–	–	–
DETF	4	14	6	1	26	4	15	3
DETH	1	1	–	1	3	1	–	1
DETN	–	–	–	–	–	–	–	–
DETQ	–	25	14	2	21	2	18	4
DETS	34	27	59	8	41	15	38	6
DETV	–	–	–	1	3	1	–	–
DETW	–	–	–	–	–	–	–	–
G	48	10	47	1	56	14	92	30
I	–	–	–	–	–	1	3	1
INSTR	–	–	1	–	1	–	1	–
PR	17	7	14	4	43	1	35	9
PT	25	2	7	2	21	8	34	11
PTY	9	4	7	–	20	–	25	5
RC	5	3	10	4	26	8	32	9
Sum	269	164	310	75	613	111	452	115

The development can be investigated either by comparing the proportions/ percentages of individual adnominal classes, or by observing their ranking. In Table 2, we present the percentages, that is, we multiply the proportion by 100.

Table 2
Percentages of adnominals

Adnom	Text							
	T 1 1965	T 2 1965	T 3 1965	T 4 1965	T 5 1969	T 6 1992	T 7 1993	T 8 1996
A	35.45	26.25	30.83	20.54	37.62	41.38	34.85	29.69
ADV	0	0	0.42	1.08	0	0	0	0
AO	1.82	0.66	3.33	0	3.33	0	0.41	0.34
AP	2.73	1.66	0.83	3.24	0.48	8.37	2.07	1.71
APAJ	0.91	0.66	3.33	0	4.76	1.48	1.24	1.71
APX	0	0	0.42	0	0	0	0	1.02
AY	0	0.33	0.42	1.62	0	0	0	0.34
CN	0	1.00	0.83	0.54	0.48	0.49	0.83	0
DAT	0	0	0	1.08	0	0	0.41	0.34
DETF	3.64	3.65	6.25	3.24	3.81	4.93	1.66	4.10
DETH	0	1.33	0	2.16	0	0	0	0.68
DETN	0	0	0	0	0	0	0	0
DETQ	2.73	3.99	4.17	2.70	1.90	0.99	1.66	4.10
DETS	9.09	12.62	6.25	11.89	6.19	20.69	7.47	11.26
DETV	0	0.33	0.42	0	0.48	0.99	0.41	1.71
DETW	0	0	0	0	0	0	0	0.34
G	23.64	21.59	21.25	22.16	20.95	8.37	28.22	21.50
I	0.91	0.33	0	0	0	0	0	0
INSTR	0	0	0.42	0	0.95	0	0	0
PR	4.55	7.31	9.58	10.27	3.33	4.43	8.30	4.44
PT	4.55	6.31	7.08	10.27	7.62	5.42	3.73	3.41

PTY	7.27	4.65	2.5	7.57	6.67	1.48	3.73	4.10
RC	2.73	7.31	1.6	11.89	1.43	0.99	4.98	9.22

Adnom	Text								
	T 9 1999	T 10 2000	T 11 2002	T 12 2003	T 13 2007	T 14 2008	T 15 2008	T 16 2010	17 2011
A	34.15	30.48	39.61	37.74	49,15	52.20	47.75	27.65	26.96
ADV	0	1.12	0	0.97	0	0	0	0.22	0.87
AO	0	0	0	0.32	0	0.98	0	0	0
AP	1.83	9.67	1.95	2.58	10,17	1.79	1.80	4.42	2.61
APAJ	0	2.23	0	0	0	0.16	0	0.66	0
APX	0	0	0	3.87	0	0.16	0.90	0.22	0
AY	0	1.12	1.30	0.65	0	0.98	0	1.11	0
CN	1.22	1.49	0.65	0.65	0	0.98	0	0.88	0.87
DAT	0	0.74	1.30	0	0	0	0	0	0
DETF	3.66	1.49	9.09	1.94	1,69	4.24	3.60	3.32	2.61
DETH	0.61	0.37	0.65	0	1,69	0.49	0.90	0	0.87
DETN	0	0	0	0	0	0	0	0	0
DETQ	0.61	0	16.23	4.52	3,39	3.43	1.80	3.98	3.48
DETS	4.88	12.64	17.53	19.03	13,56	6.69	13.51	8.41	5.22
DETV	0	0	0	0	1,69	0.49	0.90	0	0
DETW	0	0	0	0	0	0	0	0	0
G	25.00	17.84	6.49	15.16	1,69	9.14	12.61	20.35	26.09
I	0	0	0	0	0	0	0.90	0.66	0.87
INSTR	0	0	0	0.32	0	0.16	0	0.22	0
PR	8.54	6.32	4.55	4.52	6,78	7.01	0.90	7.74	7.83
PT	7.93	9.29	1.30	2.26	3,39	3.43	7.21	7.52	9.57
PTY	7.32	3.35	2.60	2.26	0	3.26	0	5.53	4.35
RC	4.27	1.86	1.95	3.23	6,78	4.24	7.21	7.08	7.83

Now, our problem is to study which of the adnominal types displays a smooth development. The smooth development means only one maximum, or a clear trend in the last part of the sequence. For the first three texts, which were written in 1965, one may consider the mean percentage computed from the original data. For example, the adnominal *A* would have the 1965 mean (74 + 39 + 79 + 38)/(240 + 110 + 301 + 185)*100 = 27.51 (and not the mean of the four percentages).

In order to obtain a clearer picture, we present the percentages as shown above for individual years in Table 3.

Table 3
Average proportions of adnominals

Adnom	Text												
	1965	1969	1992	1993	1996	1999	2000	2002	2003	2007	2008	2010	2011
A	27.51	37.62	41.38	34.85	29.69	34.15	30.48	39.61	7.74	49.15	51.52	27.65	26.96
ADV	0.35	0	0	0	0	0	1.12	0	0.97	0	0.00	0.22	0.87
AO	1.44	3.33	0	0.41	0.34	0	0	0	0.32	0	0.83	0	0
AP	1.91	0.48	8.37	2.07	1.71	1.83	9.67	1.95	2.58	10.17	1.80	4.42	2.61
APAJ	1.32	4.76	1.48	1.24	1.71	0	2.23	0	0	0	0.14	0.66	0
APX	0.12	0	0	0	1.02	0	0	0	3.87	0	0.28	0.22	0

AY	0.60	0	0	0	0.34	0	1.12	1.30	0.65	0	0.83	1.11	0
CN	0.72	0.48	0.49	0.83	0	1.22	1.49	0.65	0.65	0	0.83	0.88	0.87
DAT	0.24	0	0	0.41	0.34	0	0.74	1.30	0	0	0.00	0	0
DETF	4.31	3.81	4.93	1.66	4.10	3.66	1.49	9.09	1.94	1.69	4.14	3.32	2.61
DETH	0.96	0	0	0	0.68	0.61	0.37	0.65	0	1.69	0.55	0	0.87
DETN	0.00	0	0	0	0	0	0	0	0	0	0.00	0	0
DETQ	3.59	1.90	0.99	1.66	4.10	0.61	0	16.23	4.52	3.39	3.18	3.98	3.48
DETS	10.17	6.19	20.69	7.47	11.26	4.88	12.64	17.53	19.03	13.56	7.73	8.41	5.22
DETV	0.24	0.48	0.99	0.41	1.71	0	0	0	0	1.69	0.55	0	0
DETW	0.00	0	0	9	0.34	0	0	0	0	0	0.00	0	0
G	21.89	20.95	8.37	28.22	21.50	25.00	17.84	6.49	15.16	1.69	9.67	20.35	26.09
I	0.24	0	0	0	0	0	0	0	0	0	0.14	0.66	0.87
INSTR	0.12	0.95	0	0	0	0	0	0	0.32	0	0.14	0.22	0
PR	8.25	3.33	4.43	8.30	4.44	8.54	6.32	4.55	4.52	6.78	6.08	7.74	7.83
PT	7.18	7.62	5.42	3.73	3.41	7.93	9.29	1.30	2.26	3.39	4.01	7.52	9.57
PTY	5.02	6.67	1.48	3.73	4.10	7.32	3.35	2.60	2.26	0	2.76	5.53	4.35
RC	6.10	1.43	0.99	4.98	9.22	4.27	1.86	1.95	3.23	6.78	4.70	7.08	7.83

If one considers the development of the percentages in the course of years, one may only focus on one adnominal, namely the *I* (infinitive), which displays a non-zero percentage at the beginning (in 1965), then disappears completely and begins to increase in 2008 again. That means, either we have too few data, or there is no regular development, and the authors have different styles.

The same result follows if we consider the number of different adnominals in individual texts. The sequence is clearly oscillating, telling us that the use of adnominals is a problem of personal style.

We can, of course, compare the last century with the present one and test the difference of percentages. We may simply treat the percentages in Table 2 as usual characteristic values, and take their mean. For example, the adnominal *A* has been realized in the texts of the 20th century as (30.83 + 35.45 + 26.25 + 20.54 + 37.62 + 41.38 + 34.85 + 29.69 + 34.15)/9 = 32.3067, while the mean in the 21st century is 38.9425. Computing the common variance, which can be found in any text-book, we may perform the t-test defined with 15 DF, yielding in case of adnominal *A*: u(A) = 1.66, which is not significant. The significance level is 2.13. Having performed all tests for individual adnominals, we obtain the results presented in Table 4.

Table 4
Means of percentages and the t-tests for difference between means
of the 20th and 21st centuries

Adnominal	Mean p (20)	Mean p (21)	u
A	32.31	38.94	1.66
ADV	0.17	0.40	1.09
AO	1.10	0.16	1.85
AP	2.55	4.37	1.27
APAJ	1.57	0.38	1.93
APX	0.16	0.64	1.05
AY	0.30	0.64	1.30
CN	0.60	0.69	0.41
DAT	0.20	0.26	0.25
DETF	3.88	3.50	0.41
DETH	0.53	0.62	0.27
DETQ	2.54	4.60	1.21

DETS	10.04	10.95	0.42
DETV	0.89	0.38	0.98
DETW	0.04	0.00	0.94
G	21.41	13.67	2.38*
I	0.14	0.30	0.93
INSTR	0.07	0.09	0.22
PR	6.75	5.71	0.87
PT	6.26	5.50	0.57
PTY	5.03	2.67	2.30*
RC	4.94	5.02	0.05

There are merely two types of adnominals, namely *G* and *PTY* displaying a significant decrease of use. We hope that the number of texts is sufficient, but in any case, one should continue this research.

2. Motif Development

The change in motifs can be scrutinized in various ways. First, one must define them and then take a special property whose change can be observed. We define motifs in Köhler's spirit (Köhler, Naumann 2008; Köhler 2015). Let us consider the first text T1 by Rozhdestvenskij (1965) and subdivide it in adnominal motifs as shown below:

T1

{[APAJ],[APAJ,DETF,G,A],[G],[A],[A],[A,G,DETF],[A,AO,ADV],[A,G,APAJ,CN,DETS],[G],[G,APAJ],[APAJ,A],[A,PT],[A,PR],[PR,RC],[A,PR],[A,G,DETS,DETF],[A],[G,A],[A,DETS],[A,G],[A],[G,DETQ,AP,A],[AP],[DETQ,A],[A],[A,PR],[PR,PT,G],[PT],[G,RC],[RC,A,PR,DETF,PT],[A],[A,RC,G,PR],[A],[PR,G,PT],[G,A],[G,PR],[G],[G,DETS,PTY,DETQ,CN,PR,AY],[G],[PR,DETS],[DETS],[DETS,G],[G],[G,DETS,PTY],[DETS,A],[G,A],[A],[G,DETS],[G,A,PT,PR],[A],[A,DETS,G,DETF],[A,PTY,DETQ],[A],[A],[A,G,DETF,DETQ],[DETQ,PR],[PR],[PR,A],[PR],[A,PTY,PT],[PT],[A],[A],[A],[A,PTY,DETF,PR],[A,PR,G],[A],[G,DETQ],[A],[G,DETF,A],[DETF,DETS],[DETF,DETQ,G,A],[A,PR],[DETF,PR],[A,DETS,PT],[A],[A],[A],[A],[A,AO],[AO,PT],[A],[A],[A],[A,G],[G],[G],[G,PT,DETQ,A],[A],[G],[G,DETF],[G,PR,PT,A],[DETF,DETQ,AO],[AO,PT],[AO,APAJ,A,G,PTY],[A],[G],[G],[G],[G],[G],[G,APAJ,DETF,A],[A],[DETF,DETV,INSTR,APX,DETS,A,G],[G],[DETS,G,A],[A,PR],[A],[A,PT],[PT],[PT,PR,G,A],[G],[A,AO],[AO],[A],[A,APAJ],[A,G,PT]}

Since motifs are justified linguistic entities, we may study various properties of theirs, as mentioned above. Here, we shall consider merely the length, i.e. the number of adnominals in individual motifs. In this way, we obtain a distribution whose properties can be used in the sequel. The longer the motifs are, the more variegated the style of the texts is. Surely, poetic texts will be more variegated than receipts or scientific texts. Now, the distribution of the motifs in T1 according to length is presented in Table 4. Since we have to do with length, we automatically apply the Zipf-Alekseev function. The results of

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

are presented in Table 5. Here, we use the function (not the distribution).

Table 4
Distribution of adnominal motif lengths in T3 (1965)

Length	Frequency	Zipf-Alekseev ft.
1	52	52.09
2	34	33.10
3	13	16.36
4	13	8.40
5	3	4.72
7	2	2.09
a = 0.0527, b = -1.0437, c = 51.0946, R^2 = 0.9812		

The mean of the distribution is m_1 = 2.0513; that means, the mean number of adnominals in a motif is ca 2. This does not take into account the length of sentences, which may play a role in the setting up of dependencies in text. Sentence lengths could be studied in terms of numbers of clauses, but we shall omit this perspective here. (Sentences are marked in the Appendix with "/".)

The results of the other texts are presented in Table 5.

Table 5
Motif lengths in individual texts

L	T 1 1965		T 2 1965		T 4 1965		T 5 1969		T 6 1992		T 7 1993	
	f	Z-A	f	Z-A	f	Z-A	f	Z-A	f	Z-A	f	Z-A
1	21	20.69	56	55.92	25	25.14	53	52.88	44	44.03	57	56.82
2	14	16.10	39	38.84	25	24.27	33	33.95	32	31.70	36	37.33
3	15	9.74	18	21.96	12	13.73	18	14.87	13	14.57	21	17.62
4	2	5.98	23	12.66	8	7.40	4	6.74	10	6.77	8	8.52
5	–	–	3	7.73	6	4.24	3	3.48	2	3.53	1	4.55
6	–	–	1	5.03	2	2.68	1	2.13	1	2.17	2	2.76
7	–	–	–	–	–	–	–	–	–	–	–	–
8	1	1.71	–	–	–	–	–	–	–	–	–	–
	a = 0.2260 b = -0.8784 c = 19.6914 R^2 = 0.8411		a = 0.0427 b = -0.8372 c = 54.9227 R^2 = 0.9285		a = 0.8512 b = -1.3047 c = 24.1393 R^2 = 0.9845		a=0.2790 b=-1.3472 c = 51.8820 R^2=0.9909		a = 0.4763 b = -1.3894 c = 43.0264 R^2 = 0.9888		a = 0.2067 b = -1.1919 c = 55.8186 R^2 = 0.9891	

Length	T 8 1996		T 9 1999		T 10 2000		T 11 2002		T 12 2003	
	f	Z-A	f	Z-A	f	Z-A	f	Z-A	f	Z-A
1	52	52.07	30	30.00	42	42.29	45	44.82	74	73.78
2	48	47.63	26	25.94	47	45.58	22	23.72	46	47.51
3	20	21.06	12	12.67	15	20.00	16	11.31	26	22.93
4	9	8.74	8	6.04	14	7.94	4	5.90	13	11.33
5	6	4.02	2	3.22	4	3.52	1	3.48	1	6.08
6	2	2.23	1	2.01	2	1.95	1	2.33	2	3.63

7	1	1.52	–	–	–	–	–	–	–	–
	a = 1.0981 b = -1.7736 c = 51.0691 R^2 = 0.9980		a = 0.8264 b = -1.5062 c = 28.9966 R^2 = 0.9909		a = 1.5078 b = -2.0157 c = 41.2857 R^2 = 0.9654		a = -0.3150 b = -0.9124 c =43.8173 R^2 = 0.9750		a = 0.1158 b = -1.0993 c = 72.7849 R^2 = 0.9894	

Length	T 13 2007		T 14 2008		T 15 2008		T 16 2010		T 17 2011	
	f	Z-A	f	Z-A		Z-A	f	Z-A	f	Z-A
1	15	14.95	209	208.95	35	34.96	78	76.97	17	16.86
2	15	15.25	101	101.54	15	15.53	45	51.45	13	13.85
3	7	5.96	40	38.11	10	8.43	40	30.51	10	8.13
4	1	2.45	14	15.17	4	5.27	22	18.66	4	4.81
5	1	1.43	5	6.80	–	–	9	11.99	3	3.08
6	–	–	–	–	–	–	2	8.10	1	2.17
7	–	–	–	–	–	–	1	4.26	–	–
8	–	–	–	–	–	–	1	3.29	–	–
	a = 1.6918 b = -2.3970 c = 13.9523 R^2 = 0.9825		a = -0.1592 b = -1.2831 c = 207.9484 R^2 = 0.9997		a = -0.9537 b = -0.3912 c = 33.9563 R^2 = 0.9920		a = -0.1283 b = -0.6667 c = 75.9690 R^2 = 0.9617		a = 0.4227 b = -1.0467 c = 15.8554 R^2 = 0.9686	

In order to compare the texts, it is sufficient if one considers the parameter *a*, or *b*. As can be seen, the development is oscillating, i.e. no tendency can be observed.

If one compares the individual centuries, one obtains the results presented in Table 6. As can be seen, the Zipf-Alekseev formula holds true in both cases; the parameters *b* are almost identical, but the parameter *a* changed drastically.

Since parameter *b* represents rather the effort of the hearer, we have to do here with the change of motif construction in the texts caused either by the writers, or by the overall change in writing.

Table 6
Motif lengths in 20th and 21st centuries in Russian texts

Ranks	20th century		21st century	
	f	Z-A	f	Z-A
1	391	390.56	515	513.85
2	289	290.20	304	312.19
3	145	150.39	164	147.72
4	98	76.21	76	71.06
5	30	40.00	24	36.10
6	18	22.03	8	19.49
7	1	2.79	1	11.20
8	1	7.85	1	6.85
	a = 0.3269 b = -1.0916 c = 389.5562 R^2 = 0.9946		a = -0.0055 b = -1.0319 c = 512.8516 R^2 = 0.9968	

It would be very interesting to study a similar evolution in other languages. Comparing the two centuries by the usual chi-square test, we obtain an inhomogeneity of $\chi^2 = 12.15$. Since we pooled the classes 6, 7, and 8, we obtain 5 degrees of freedom, hence a difference at the 0.025 level. The main differences are caused by lengths 1, 4, and 6 in the 20th century.

3. Sentence in Terms of Motif Numbers

Sentence length can be measured in many different ways. Since one adheres to immediate components, the "correct" method is counting the number of clauses, as has been shown in many cases. However, one can also count the number of adnominals in a sentence and consider the result as "sentence containing *x* adnominals". From a specific point of view, one may consider this number as sentence length in a slightly figurative sense. In the Appendix, one finds each sentence marked with a slant line "/". Counting the number of adnominals in a sentence, one may obtain somewhat problematic results because in poetic texts, verses and sentences are sometimes mixed units. The authors may place commas instead of full stops at the end of stanzas, so several stanzas may form one sentence. This is, so to speak, a poetical vision. Such a case can be found in text T 16, where the author replaced several sentence marks by commas. Nevertheless, one can perform the computation. Starting from the conjecture that short sentences (i.e., containing few adnominals) are more frequent than those containing many, we adhere, again, to the exponential function with added 1, and omit all "lengths" having frequency 0. The results are presented in Table 7.

Table 7
Sentence "length" in terms of adnominals

"Length"	T 1		T 2		T 3		T 4	
	f_x	Exp	f_x	Exp	f_x	Exp	f_x	Exp
1	11	12.48	25	36.94	93	92.52	47	46.67
2	10	8.88	26	24.31	19	22.61	19	20.30
3	10	6.41	22	16.12	11	6.10	10	9.15
4	4	4.71	7	10.81	9	2.20	5	4.44
5	1	3.55	6	7.36	4	1.28	3	2.46
6	–	–	3	5.12	2	1.07	2	1.62
7	1	2.20	3	3.68	1	1.12	2	1.26
8	1	1.82	–	–	–	–	–	–
9	–	–	1	2.13	–	–	1	1.04
10	–	–	1	1.73	–	–	–	–
12	1	1.18	1	1.30	–	–	–	–
20	–	–	1	1.01	–	–	–	–
	a = 16.7214 b = 0.3763 R^2 = 0.8310		a = 55.4186 b = 0.4330 R^2 = 0.9562		a = 387.6516 b = 1.4435 R^2 = 0.9858		a = 108.0.935 b = 0.8615 R^2 = 0.9974	

	T 5		T 6		T 7		T 8	
	f_x	Exp	f_x	Exp	f_x	Exp	f_x	Exp
1	58	58.60	35	38.05	65	65.18	30	29.90
2	30	27.38	30	23.93	26	25.03	18	20.32
3	10	13.08	18	15.19	9	10.00	18	13.91
4	7	6.53	4	9.78	3	4.37	9	9.63

5	4	3.53	4	6.43	4	2.26	8	6.77
6	1	2.16	–	–	–	–	2	4.85
7	–	–	–	–	1	1.18	5	3.58
8	1	1.24	1	2.29	1	1.07	1	2.72
9	–	–	–	–	–	–	1	1.77
10	–	–	1	1.49	1	1.01	1	1.51
11	–	–	–	–	–	–	1	–
12	–	–	–	–	1	1.00	–	–
17	–	–	–	–	–	–	1	1.05
19	–	–	–	–	–	–	1	1.02
28	–	–	–	–	1	1.00	–	–
	a = 125.7523 b = 0.7809 R^2 = 0.9931		a = 59.9902 b = 0.4800 R^2 = 0.9236		a = 171.4187 b = 0.9824 R^2 = 0.9982		a = 43.2267 b = 0.4027 R^2 = 0.9614	

	T 9		T 10		T 11		T 12	
	f_x	Exp	f_x	Exp	f_x	Exp	f_x	Exp
1	11	11.56	54	56.24	54	52.99	29	26.03
2	7	8.26	38	32.44	18	22.95	11	16.45
3	11	5.99	18	18.90	16	10.19	8	10.54
4	3	4.43	10	11.19	5	4.86	9	6.89
5	3	3.36	5	6.80	–	–	9	4.63
6	1	2.62	1	4.30	1	1.68	4	3.24
7	1	2.11	2	2.88	–	–	6	2.38
8	–	–	–	–	–	–	1	1.85
9	1	1.52	–	–	–	–	–	–
10	–	–	–	–	–	–	1	1.33
11	–	–	–	–	–	–	1	1.20
12	–	–	–	–	–	–	1	1.12
13	–	–	–	–	–	–	1	1.08
17	–	–	–	–	–	–	1	1.01
18	–	–	–	–	–	–	1	1.01
20	1	1.01	–	–	–	–	–	–
34	1	1.00	–	–	–	–	–	–
	a = 15.3714 b = 0.3750 R^2 = 0.7836		a = 97.0331 b = 0.5634 R^2 = 0.9786		a = 123.6728 b = 0.8666 R^2 = 0.9665		a = 40.5550 b = 0.4825 R^2 = 0.8900	

	T 13		T 14		T 15		T 16	
	f_x	Exp	f_x	Exp	f_x	Exp	f_x	Exp
1	12	12.72	103	101.37	13	14.27	11	11.43
2	10	8.64	53	57.35	13	9.85	6	10.26
3	6	5.98	34	32.63	4	6.90	11	0.22
4	5	4.25	18	18.76	8	4.93	6	8.29
5	1	3.12	17	10.97	3	3.62	15	7.47
6	–	–	5	6.60	1	2.75	10	6.74
7	–	–	3	4.14	1	2.16	9	6.09
8	–	–	5	2.76	–	–	2	5.51
9	–	–	2	1.99	–	–	4	5.01
10	–	–	1	1.56	–	–	2	4.55

12	–	–	1	1.18	–	–	3	3.80
13	–	–	–	–	–	–	3	3.48
14	–	–	1	1.06	–	–	–	–
15	–	–	–	–	–	–	1	2.95
	a = 17.9805 b = 0.4279 R^2 = 0.9009		a = 178.7801 b = 0.5773 R^2 = 0.9933		a = 19.9037 b = 0.4055 R^2 = 0.7928		a = 11.7659 b = 0.1197 R^2 = 0.4542	

	T 17	
	f_x	Exp
1	19	18.89
2	10	10.61
3	7	6.16
4	4	3.77
5	2	2.49
6	1	1.80
7	2	1.43
9	1	1.12
	a = 33.3096 b = 0.6217 R^2 = 0.9913	

Although there is an exception in T 16, we accept preliminarily the proposed model. Further examinations, also in other languages, will bring more security. Historically, we obtain a mean *b* for the 20th century as 6.1353/9 = 0.6817, and for the 21st century 3.9449/7 = 0.5636. It would be premature to draw any consequences until one did not analyze at least 50 texts.

References

Altmann, G. (2015). *Problems in Quantitative Linguistics, Vol. 5*. Lüdenscheid: RAM-Verlag.

Andreev, S**., Fan, F., Altmann, G.** (2018). Adnominal aggregation. *Glottometrics* 40, 63–76.

Andreev, S., **Popescu, I.-I., Altmann, G.** (2017). Some properties of adnominals in Russian texts. *Glottometrics* 38. 77–106.

Best, K.-H., Boschtan, A. (2010). Diversification of simple attributes in German. *Glottotheory* 3(2), 5–9.

Givón, T. (2001). *Syntax 1, 2.* Amsterdam/Philadelphia: Benjamins.

Gunkel, L., Zifonun, G. (2009). Classifying modifiers in common names. *Word Structure* 2(2), 205–218.

Halliday, M. A. K. (2004). *Introduction to functional grammar*. London: Hodder Arnold.

Köhler, R. (2015). Linguistic motifs. In: Mikros, G., Mačutek, J. (eds.), *Sequences in Language and Text: 89–108.*Berlin/Boston: de Gruyter.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T- segments. In: Preisach, B., Schmidt-Thieme, D. (eds.), *Data analysis, machine learning and applications. Proceedings of the Jahrestagung der Deutschen Gesellschaft für Klassifikation 2007 in Freiburg: 627–646.* Berlin/Heidelberg: Springer.

Köhler, R., Altmann, G. (2014). *Problems in Quantitative Linguistics, Vol 4*. Lüdenscheid: RAM-Verlag.
Rijkhoff, J. (2004). *The noun phrase*. Oxford: Oxford University Press.

Appendix

T 1
Evg. Evtushenko *Pushkinskij pereval* (Pushkin pass) /Евг. Евтушенко «Пушкинский перевал»/, 1965. Words 709, adnominals 110, sentences 40.

A, I/, G/, G, AP/, A, DETS/, A, A, G/, A, A, A/, A/, A, G/, A/, DETS, PT, PR, A/, PR/, A/, PTY, A/, A, A/, PTY, PTY, G, A, G, PT, G, G/, PTY, A, A/, A, G, PT, G, DETQ, DETS, DETQ, DETS, PTY, PT, G, RC/, DETF/, AP/, A, A/, RC, DETQ, A, G/, A, A, G/, A, A, G/, A, G/, AP, APAJ, G/, A, G, A, DETS, PTY/, A/, A, A, G/, G, G, G, PR, G, PT, G/, A, A, DETS/, DETF/, G, DETS/, A, PTY, PTY, A/, DETS, PR, PR/, DETS, DETF, AO, AO/, A/, G, DETS, A/, DETF/, G, RC/, A/

T 2
E. Evtushenko *Bratskaja GES* (Bratskaja hydroelectric power plant) /E. Евтушенко «Братская ГЭС»/, 1965. Words 2026, adnominals 301, sentences 106.

PR/, RC, A, G, RC/, G, DETS, A/, RC/, G/, G/, AO, A, A/, DETS, DETS, PT, DETS, PT/, DETS, A, DETS, G, PT, G, RC, PTY, DETS, G, AP, A, G, DETS, PT, G, DETS, A, G, G/, DETS, G/, A, DETS, DETQ/, A, PT, A/, G, G, G, G, G, DETS/, PR, PR, PR, PR, DETQ, PR, RC/, A, PR, CN/, AP, DETS, RC, A, DETS, A, A/, DETS/, PT, G/, DETS/, DETQ, G/, I, G, RC/, A, DETQ, G/, A, G/, G/, PR, PR, A, DETS, G, G/, DETH, PT, G, A, G/, DETQ, DETQ/, A/, DETS/, A, A, PR, PR, DETS/, DETF, RC/, DETQ, G, G/, PT, PTY, G/, A/, DETS, A, G, PR, G, A, A/, DETQ/, A, G/, PT, DETS/, DETS, DETQ/, DETH, PTY, DETS, A/, PTY, A, G/, PT, A, PR/, PR, AP, DETF, DETQ, DETF, DETF, PR, DETF, PT, G/, A, APAJ/, DETF, A, A/, A, A, DETF/, DETH, RC, RC, AY, DETS/, G, DETF, PT, A, DETS, A, PT, G, G, A, PTY, G/, DETS, A, DETS, A/, A, A, PT, DETS/, DETS, G, G/, A, G/, G, G/, G, G/, DETV, DETS/, A, DETS/, DETS, A, G/, A, PTY/, A/, G/, RC, RC, A, PTY/, RC, PR, PTY, RC, RC, RC, RC, A, DETF/, PTY/, PT/, A, A, RC, PTY, A, A/, A, G, G, PTY/, PT, DETS, G/, G, PR/, A, G/, CN, G, G, G/, G, PTY/, G/, G/, PT, G/, DETS/, A/, A/, DETQ/, A/, AP/, A, A, A/, DETS/, G, A, A/, PR/, DETF, DETH, A/, A, G/, A, PT, PT/, DETS, A, AO, PR, PTY/, PR, A/, A/, A/, A/, A/, A/, CN, DETS/, DETF, A/, A, DETS/, PR/, RC, A, RC, RC, A/, A, PTY, PR/, DETQ/, AP, RC/, G, G, G/, A/, A/, APAJ/

T 3
R. Rozhdestvenskij *Pojema o razlichnyh tochkah zrenija* (Poem about different points of view) /P. Рождественский «Поэма о различных точках зрения»/, 1965. Words 2205, adnominals 240, sentences 141.

APAJ/, APAJ/, DETF/, G, A, G/, A, A, A/, G/, DETF/, A/, AO, ADV, A, G/, APAJ/, CN/, DETS/, G/, G/, APAJ, APAJ, A/, A, PT, A/, PR/, PR/, RC, A/, PR/, A/, G/, DETS/, DETF, A, G, A, A/, DETS/, A/, G/, A, G/, DETQ, AP, A, AP/, DETQ, A/, A, A, PR, PR/, PT, G, PT, G/, RC/, RC/, A, PR/, DETF, PT/, A, A/, RC/, G/, PR, A, PR, G/, PT, G, A, G/, PR/, G, G/, DETS/, PTY/, DETQ, CN, PR, AY, G, PR/, DETS/, DETS/, DETS, G, G, G/, DETS/, PTY/, DETS/, A/, G, A, A/, G, DETS, G/, A/, PT/, PR/, A/, A/, DETS/, G/, DETF, A, PTY/, DETQ,

A, A/, A, G/, DETF/, DETQ, DETQ, PR, PR, PR/, A/, PR/, A, PTY/, PT/, PT/, A, A/, A/, A, PTY, DETF, PR/, A/, PR/, G, A, G/, DETQ/, A, G/, DETF/, A/, DETF/, DETS, DETF/, DETQ/, G/, A/, A/, PR/, DETF, PR/, A/, DETS, PT, A, A, A/, A/, A/, AO/, AO/, PT/, A, A/, A/, A, G, G, G, G/, PT/, DETQ/, A/, A/, G/, G/, DETF/, G/, PR/, PT/, A, DETF, DETQ, AO, AO, PT, AO/, APAJ/, A, G/, PTY/, A, G/, G, G, G, G, G, APAJ/, DETF/, A/, A/, DETF, DETV/, INSTR/, APX/, DETS/, A/, G/, G, DETS/, G, A, A/, PR/, A/, A, PT, PT, PT/, PR/, G/, A/, G, A/, AO, AO/, A/, A/, APAJ/, A/, G/, PT/

T 4
I. Brodsky *Feliks* (Felix) /И. Бродский «Феликс»/, 1965. Words 1401, adnominals 185, sentences 89.

G, PR/, DETS/, DETS/, DETF/, A, G/, DETS/, G, G/, A, PT/, DETQ, G, A/, DETS, DETH, A, G, RC/, A, DETS, G/, A/, A/, A, PT/, DETS, PTY/, PT/, DETF, PR, PT/, ADV/, AP/, DETF/, G, G/, A/, RC/, DETH, PT/, A, A, G, G/, A/, PT/, G, G, PR, G, PTY, A, G/, PR/, A/, ADV/, DETS/, A/, DETQ/, DETQ, AP, AP, DETQ, AP, RC, PTY, G, A/, DETQ, AY, G/, PT, PTY, A/, DETS, A, PR/, PTY/, AP, DETF, CN, DETS, A, DETF, AY/, G/, A, A/, G/, DETS/, DETS/, PR/, G/, G/, DETH, A, DAT/, PR/, PR/, DETS, PR/, PR/, G/, G, DETF, A, PTY, RC/, G, AY/, DETS, A/, DETH, A/, RC/, PT, DETS, DETS, PR, G/, DETS, A, DETS, G/, A, G/, PTY, G/, G, DETS, G, G/, G/, PR, DETS/, DAT, PR/, DETS, PR/, PTY/, PR/, PR, PT, G/, PR, PT, G, G, A, G/, PT/, A/, RC, RC/, RC/, DETS, AP, A/, A/, DETS/, A/, G, PR, PT/, PTY/, A/, G/, A/, A/, A, A, PR, G/, PTY, PT, PT, PTY, PTY, PTY/, G, PTY, A, G/

T 5
R. Rozhdestvenskij *Posvjashhenie* (Dedication) /Р. Рождественский «Посвящение»/, 1969. Words 1265, adnominals 210, sentences 111.

A/, A/, DETF/, PT, A/, A/, A, G, A/, AO/, AO/, DETQ, DETS/, A, G/, G/, A/, A/, A/, A/, AO/, AO, AO/, A/, G/, G, G/, A, APAJ, DETQ, PR/, DETS, DETS/, PTY/, A, G, PT/, A, A, A, G, PTY, A, APAJ, APAJ/, PT, PTY, G, G, A, G/, G/, A/, A, PR/, A, G/, A/, DETF, DETF/, A, A/, DETF, PT/, G, G/, G, A/, PT/, A/, AO/, AO/, PT/, RC/, A/, PTY, A, A/, DETS/, DETS/, A, G/, A, A/, A, CN, INSTR/, A/, PT, G, G, A, G/, PT, A, A, PTY, PTY/, PT, A, G, G/, DETS/, DETS/, DETF/, A/, G, A/, DETF/, A/, A/, A/, A, G/, G, G/, G/, A, A, G, A, A/, PTY, PTY, PTY/, PTY, PTY, RC/, DETQ, PT/, A, A/, DETF, DETV/, A/, A, G/, A/, A, A/, A/, DETS/, A/, A, A, DETS, PT/, A/, A, DETS/, A/, A/, PT/, DETF/, PT, A, G, G/, DETS/, DETS/, G, A, A, G/, PR, APAJ/, PR, APAJ/, A, RC/, APAJ, APAJ, APAJ, APAJ/, A/, G, A, G/, G, PTY/, PR, A, PT, A, PT/, A/, A, A/, DETS, INSTR/, G/, G, G, APAJ/, A/, PR/, A/, G/, G, PR, PTY/, G, PTY, A/, A, DETQ, AP, G/, PT/, G/

T 6
R. Dyshalenkova *Begu po cementu* (I am running on the cement) /Р. Дышаленкова «Бегу по цементу»/, 1992. Words 1571, adnominals 203, sentences 93.

DETQ, AP, PR/, A/, A/, DETF, AP, AP/, DETF, DETS, A, A, DETS/, A/, DETS, PT/, G/, DETS/, DETS, A/, AP, AP/, PR, DETS/, A, PT/, G/, A/, CN, DETS, A, AP/, PR, APAJ, A/, A, DETS, DETV/, G, A, A/, DETS, A/, A/, DETS, AP, G/, DETS, A/, AP, A, AP, A/, DETS, A/, G, A, G, A, G, A, G, A/, PR/, A/, DETS, PT/, RC/, DETF, AP/, AP, AP, PR/, AP, A, G/, A, DETF, A/, A, A, DETS/, A/, A, PR, A, DETS/, A/, DETS, PTY, A, DETS, DETF, PT, DETF, A, PT, G/, PT, DETF, A, A, PR/, PT/, A, A, A/, RC/, DETS, PT/, A, A/, DETS, A, A/,

DETS, G/, A/, DETS/, A, G/, G/, PR/, DETS/, DETF/, DETS, A, A, DETS, DETS/, A, A/, DETS, A, AP, G/, DETS, DETS/, DETS/, DETV, A, G/, A, A, DETS/, AP, DETS/, G, A/, DETS, A/, A, APAJ/, APAJ/, A/, DETS/, PTY, A/, A, A/, A/, DETQ, PT/, PTY/, DETS, A/, PT, DETS, A, A, AP/, A, PT/, DETF, A/, A, A, A/, AP, A/, A/, A/, A, DETS, A/, DETS, A/, A/, G, DETS/, DETS/, A/, A, A/, PR/, DETS/, A, DETF, A/, DETS/, DETS, A, A/

T 7
A. Voznesensky *Rossija voskrese* (Russia resurrect) /А. Вознесенский «Россия воскресе»/, 1993. Words 3696, adnominals 241, sentences 112.

A/, A, AP/, A, DETF, DETQ, RC, RC, A, G/, A/, A, PR/, PR, PR, A, G, RC, RC, G, PR, CN, A/, G, G/, PR/, RC, G, PR, A, PTY/, PT, A/, PR, RC, A/, RC/, RC, DETS, PR, A, G/, DETQ/, G/, A/, PR/, G, A/, DETF/, APAJ/, G/, DETS, CN/, DETF/, AP/, PT/, A/, G, A, A, A, A, A, PR, A, G, DAT, G, A, A, G, G, PTY, RC, DETS, A, DETS, APAJ, G, G, G, PR, G, A, A/, G/, G, G, A/, A/, A/, A/, A/, A, DETS, G/, G, G, G/, G/, DETS/, PR/, DETS, PTY/, G, A, RC, A/, DETS/, A/, AP/, A, PTY, G/, A/, A/, A, G/, PR, A/, PT, G, PTY, A, AO, A, DETS, A/, A/, G, A/, DETS, G/, A, DETS/, G, G, A/, G, RC, A, A, G/, G/, G, G/, DETS/, A/, A, G/, G, AP, DETS, G, DETS, PTY, A, A, RC, G, A, A/, G/, G, A/, A, A/, G/, A, PR/, A, PT, A, G/, G/, A/, PT/, G, A/, A/, PR/, G/, G/, A/, A/, A, G/, A, A/, A, G, A, G/, PR/, DETS/, DETS, PT/, PT/, A/, A/, DETF/, G/, G/, G/, A, AP, G/, PTY, PTY/, G/, A/, DETQ, PR, PR/, G/, PT/, PR, G/, A/, G, A/, DETS, A/, G, G/, A/, DETQ/, DETV/, G/, A/, A/, APAJ, A, A/, PR/, A, PTY, PT, DETS, G/, G/

T 8
Svetlana Kekova *Po obe storony imeni* (On both sides of the name), /С. Кекова «По обе стороны имени»/, 1996. Words 1767, adnominals 293, sentences 88.

G, A, A, AP, AP, DAT, PTY/, DETQ/, DETS, G/, DETH, A, G/, DETH, A/, RC/, G, A, PR, A, PR, DETQ, PR/, DETS/, A/, DETV, DETF, DETV/, G, DETF, RC, DETS, PR/, G, DETS, DETS/, DETS, PT, RC/, DETS/, A, G, A, A, PTY, PT, PT, A, AO/, A, A/, DETF, G, G, DETF, G, G, DETF, A, G, RC/, A, PT, G, RC, RC, DETV/, DETF, G, DETF, G, A/, DETQ/, DETS, A, PR, DETQ, A/, APX, RC, G/, A, APAJ, DETS, A/, DETS/, G, DETS/, A, DETS, A, PTY/, A, PR, PR, A/, RC, RC, A, RC, RC, RC, DETS, RC, RC, RC, DETQ, RC, DETS, DETS, RC, A, A/, G/, A, A, G/, PT, G, DETS, RC/, PTY/, G/, DETS, DETS/, A, PT, PR, AP, G/, DETS/, G/, A, A/, G/, G/, DETF, A, DETS, A/, G, A/, RC, RC/, DETS, APAJ, APX/, G, G, A, AP/, DETF, A, G/, G/, DETS, APX/, APAJ, RC/, PTY, APAJ, DETQ, RC, A, A, G, PTY/, G/, G, A/, AP/, DETF, G, RC, RC, G, A, A, DETS, A, A, G, PT, RC, G, A, G, DETF, G, DETS/, A, G/, DETV, G, A/, DETS, A, G, A, DETS, A/, DETQ, G, DETQ, A, G/, DETQ, DETS, DETS, A, A, G, RC/, G, DETQ/, PR, PT, A, G, A/, DETS, A, PTY/, G/, A, G/, A, AY/, DETQ, A, G/, DETW/, PTY/, G, A, DETS/, A, APAJ, PTY, A, A/, A, A, G/, A/, G/, G, G, PR, G, A, G, PTY/, A, G/, DETF, A, PT/, A, G, A, G/, A, A, A, A, A, PR, A, PR, DETQ, RC, DETS/, A, A, A, PTY/, A, A, DETV, A/, A, A, G/, PR/, G/, A/, PT/, DETS, G, PTY/, DETS/, A/

T 9
A. Parshhikov *Neft'* (Oil) /А. Парщиков «Нефть»/, 1999–2003. Words 869, adnominals 164, sentences 40.

G, G, PT, AP/, G, PTY/, AP, RC, A/, PT/, A, PR, AP/, CN, G, A/, A/, A, PTY/, DETS, A, RC/, DETF, A, PR, PR/, PT, A, A/, G, G, PR/, DETF, A, G, PTY, RC/, G, G, A/, A, A, G,

DETS, A/, A/, G/, G, G, G, G, A/, G, DETS, A, DETS, PT, PR, PT, A, DETS, A, PT, A, A, G, PTY, A, DETS, A, G, A, PTY, DETS, PT, A, A, PR, A, G, PTY, RC, G, G, A, A/, G, A, A/, PT/, DETS, A, G, PT/, PR, A, PT/, CN/, PTY/, G/, A/, A/, PR, G, G/, A, G/, G/, A, PR, G/, A, A/, DETQ, G, A, RC, PR, A, G/, DETF, A/, G, A, G, DETF, G, C:\Users\User\AppData\Local\Microsoft\Windows\G, PT, A, G/, A, DETF, A, A, A, G/, A, PTY, RC, A, DETF, PTY, A, A, A, PR, A, PR, PR, PR, RC, PT, A, G, PTY, PTY/, G, G/, DETH, PT, A, PTY, G/

T 10
A. Voznesensky *ru Poem* (ru Poem) /A. Вознесенский "ru Поэма"/, 2000. Words 1578, adnominals 269, sentences 128.

A/, G/, G, AP/, DETS/, DETS/, DETS/, AP, A, AP, DETS/, PT, ADV, DETS, DETS/, PT, G/, A/, PR, PR/, DETS, G, DETS, G/, A/, A, G, A, AP/, A/, PR, A, A/, PT, PT, DETS, A, DETS, G, PR/, G, AP, AP/, G/, ADV/, AP, AP/, CN/, DETS, A, A, G/, PT, G, G/, DETS, DETS, AP/, G, A/, PT/, A, G/, AP, DAT/, G/, DETS, AP/, A/, PR, G/, G/, PTY, DETS, A/, ADV, A, G/, A, G, A, DETS, A, PR/, PR/, PT, PT/, AY, A/, DETS/, A, A/, A/, AP, PT/, G/, G/, PR/, A/, DETS, AP/, CN/, DETS, G, G/, PR/, A, APAJ/, A, A, APAJ/, A, PT/, PT, G/, A, G/, DETS, G, A/, AY, PT, DETS, PT, DETS/, A/, PR, APAJ/, A, G, PTY/, G, A, PR, PR/, A, A, A/, DETH/, APAJ/, PT/, AP/, PT, PR/, DETS/, RC/, RC/, PR/, PT/, A/, DETS/, A/, A, A/, PR/, A, G/, A, A, G/, G/, PTY, PT/, PT, DETS/, G, AP, G, PT, G/, DAT, AP/, APAJ/, DETS/, A, G, A/, A/, A/, PT, A/, AP/, AP, DETF, A, AP, A/, G, G, A/, G, A, PT/, G, AP/, G, DETS, G, G/, DETS, RC/, A/, A, A, PTY/, CN, PT, A, PR/, A, DETS/, AY, RC/, A, PTY/, PTY, DETF, A, A, PTY/, A, DETF, A, A, A, A, PR/, PT, AP, PTY, APAJ/, A, AP, A, CN, G/, RC/, A, G, A/, A, G/, DETF/, A/, A, AP/, AP, PT/, A, DETS/, A, G/, A/, A/, A/, A, DETS/, DETS/, DETS/, A/, PTY, PT, AP/, DETS, G/, A, A, G, AP/

T 11
F. Gimberg *Andrej Ivanovich vozvrashhaetsja domoj* (Andrej Ivanovich comes home) / Ф. Гримберг "Андрей Иванович возвращается домой"/, 2002. Words 1942, adnominals 164, sentences 94.

DETF, A/, PR/, A/, DETF, A, A/, A, PR/, DAT, DAT, AP/, AY/, A/, AY/, A, A, A, A/, DETQ, DETQ, DETQ, A/, A, A/, DETQ, A/, A, DETQ, A/, A, A, DETH/, DETS, DETS, DETS/, DETQ, DETQ, DETQ/, DETS, DETS, DETS/, DETQ/, DETQ/, DETQ/, DETN/, A/, DETS, DETQ, DETS/, DETQ/, DETN/, DETS/, DETS/, DETS/, RC/, DETQ, DETS/, DETF/, A/, A/, DETQ/, A/, A/, A/, A, A, A, A/, DETS, A, DETS/, A/, DETQ/, DETS/, DETS/, A, PT, DETS/, A, A/, DETQ, A, A, A, PR, PR/, G, A, A/, A/, A/, A, A/, G, G, G/, DETF/, DETF/, A, DETF, DETQ, PR/, DETF/, G/, A/, A/, G, A/, G/, DETQ/, PTY/, A/, DETQ, DETS/, PTY, PTY/, DETS, CN/, G/, A/, PR/, PTY/, A/, A, PT, DETS/, A/, DETF, RC, PR, DETQ/, DETF/, A/, DETS/, DETQ, A/, DETF/, AP/, DETF, RC/, AP/, DETQ, DETF, A/, DETQ, DETF, A/, DETS, G/, DETS, G/, DETS, DETS/, A, A/, DETF, A, A/, A/, A/, DETQ/, DETS, DETS/

T 12
F. Grimberg *Versija krysolova* (Pied Piper's version) /Ф. Гримберг «Версия крысолова»/, 2003. Words 1826, adnominals 310, sentences 82.

A, A/, A, PT, A, A, PTY, A, APX/, A, A, G, A/, A, PTY, G/, A, G, A/, DETF, RC, A, A, A, A, A, A, DETQ, A, A, G/, PT, A, A, PT/, A, G, DETS, G/, G/, DETQ/, A, A, A, RC, DETQ/, AY, PR, DETS, A/, DETS/, A/, DETS, A, A/, DETQ, A, A, A, A, G, A/, PT, AO, G, A, A, A,

G/, A/, A, A, PR/, DETS, AP, AP, AP/, A, A, A, G, G, AP/, DETQ/, A/, G, PTY, A, DETQ, A/, G, G, RC, RC, RC/, A, G, A, DETS, A, A, PT/, DETS, DETQ/, DETQ/, A, DETS/, DETS, APX, A, A, AP, A/, CN, A, A, DETS, G, DETF, A, RC/, DETS/, DETQ, G/, DETS, DETQ/, A, DETS/, A, ADV, INSTR, A, A/, A, A, G/, A, AY/, DETQ/, A, G/, DETF/, PTY, A, A, A, PT, A, A/, A/, A/, A/, APX/, G/, G, A, G, DETS/, DETS, A, DETS, APX, A, PR, G, DETS, A, A/, G, DETF, G, DETS, PR/, DETS/, DETQ, DETS/, A, DETS/, DETF/, DETQ/, DETQ/, DETS, APX, PR, PTY, DETS/, DETS, APX, AP, A, A, G/, A, APX, DETF/, DETS, A, DETS, DETS, A/, DETS/, DETS/, APX/, DETS/, DETS/, DETS/, DETS/, A, A/, DETS, APX, DETS, DETS, DETS, AP, G, G, G, G, A, DETS, A, APX, A, DETS, DETS/, DETS, DETS, G, RC, G, G/, DETS, A, APX, DETS/, DETS, A, A, G/, A, PTY, A, G, AP, G, G/, DETS, A, A, G, A, A, PR, A, G, G, G/, DETS, A, PR, CN, RC/, DETS, DETS, G/, DETS, A, PR/, DETS, PT, PR, G/, DETS, A, PR, DETS, G, PR, A, DETS, DETS, G, APX, A, PR/, DETS, A, A, A, A/, A/, A, PR, A, G, A, A, RC, DETS, RC, DETS, A, A, DETS, ADV, ADV, PTY, A, PR/

T 13
V. Emelin *Pechen'* (Liver) /В. Емелин «Печень»), 2007. Words 557, adnominals 75, sentences 34.

A, A, A/, A, A, A/, DETS, A, DETH, DETS/, A, AP/, A/, RC/, DETS, A/, DETV/, AP/, G/, G, DETS/, G, G/, RC/, RC/, DETS, DETS/, DETQ, DETQ, G/, A, A/, RC/, A/, G, G/, A/, G/, PR/, A, PR, A/, AP, A/, A, G, G/, G, A, G/, A, G, A, A/, A, G, A, AP/, DETF, A/, G, G, AP, PT/, A, G, DETS, DETS/, A, PT/, AP, PR, A, A, PR/

T 14
M. Stepanova *Proza Ivana Sidorova* (The prose of Ivan Sidorov) /М. Степанова «Проза Ивана Сидорова»/, 2008. Words 5049, adnominals 613, sentences 243.

A, A, A, A, PR, A, A, G, PTY, A, PR, CN/, DETN, PTY/, G/, A, PR/, AP, A, G, AP, A/, PR/, A/, PTY/, A/, A, RC/, A, A, RC, A/, G/, A, A/, PT, A, A, G, DETS, G/, A, A/, G, AY/, A/, A, A/, DETS, G/, A/, A, A, A, A, A/, A, PTY/, A, A, A, G/, A, RC/, A/, PT, A, A/, RC, A, A/, PT/, A/, DETV, DETV/, A, A, A, G, G, RC, DETQ, A/, A, A, A/, RC/, DETF/, DETF/, RC, DETF, PR, G/, A, A, PR, A, A/, A, A, G, DETS/, A, CN, A, DETS/, A, DETS/, DETS/, A, DETH/, G, A, G, G, DETF, A, A, A, A, G/, DETF/, G/, A/, A/, A/, RC/, A/, RC, A/, A, A/, DETQ/, A, A, A/, DETF, RC, RC/, G, RC, DETQ, DETQ/, A, A, PR/, PR, G/, G/, A, A, A, A/, DETS/, AY/, A/, G, G, AY/, AY/, DETH, A/, PR, A/, AY/, PT/, A, A/, AO/, A/, PT, A/, A, A/, DETF/, A/, AP, A, G, A, A, A, DETS, A/, CN/, A/, PR/, DETS/, DETV/, DETS/, CN/, A/, A, A, A, DETQ, A, RC/, RC/, A/, A, A, A, PR, PR/, CN/, A, PTY, A, A, DETQ, A, A/, DETF, A, G, A, A/, G/, DETF/, A/, A, A, A, A, A/, DETS, PR, DETS/, A, DETQ/, DETS, A, G, DETF, G/, DETS, AP, DETQ/, A, A, DETF/, A, A, A, A, A/, A, PR, DETS/, A, DETQ/, DETS, A, G, DETF, G/, DETS, AP, DETQ/, DETS, A, A/, DETF, A, RC, RC/, DETF, DETF, RC/, DETF/, PT, A, A, DETF/, DETF, DETF, RC/, PTY/, A, PTY/, A/, PR, A, PR, A, PR, PR, PT, PTY/, DETF, A/, DETF, A, A/, A/, A/, DETS, A/, DETS/, AP/, A/, G/, A, A, G/, A, A, A, PR/, DETQ, G, PTY, A, PR/, G, AY, G, G, A/, A/, A, A, DETQ/, DETQ/, DETQ/, A, DETS, A, DETS/, A, A, DETQ, G, A, A/, A, A, A/, DETF/, A/, A, DETF/, DETS/, AP/, RC/, A, PT, PTY, PR, A, A, A, PR, A/, A, DETQ, DETS/, A/, A/, DETQ/, DETS, A, A, A, DETS/, DETS, A/, AP/, G, DETF/, AP/, A/, PT/, PTY, A/, CN/, A/, A, A, A, A, A, PR, PT, PR/, A/, A/, A/, A, DETS/, DETS/, A, RC, A/, A/, DETQ/, A, PR/, A, A/, A, PT, PT/, A, A, A, PR, PR, A, PR/, PR/, A, PR/, A, DETF/, A, A, A, A/, A/, A/, A/, A, A, A, DETS, A/, A, A/, A, PT/, A, A/, A, A, PR, G/, PT/, A, PR/, A, RC/, A, A, G, A, PR, A, A, A/, A, A, A/, A,

DETS/, DETH, DETS, A, A/, RC, RC, A/, A, PR, INSTR, PT, DETS, A, RC/, PT, G/, A, A, PTY, DETS, A/, APX, A, PR/, A, RC, PTY, A/, DETQ, G/, A, A/, DETQ/, RC, PR, PR/, A/, A, A, A, PT, A, A/, A/, A/, A, A/, A/, PT, A, A, A/, AO, AO/, A/, PR/, AO, A/, PTY, PTY, A/, A/, A, A, PT, A, A/, AP, A, A/, A/, A/, A, A/, A, DETS/, A, A/, A/, G, PTY, A/, G, PR, PR, G/, DETS, DETQ, A, A, PR/, A, G, AP/, A, G, G, G, A, A, A, PTY, A/, A, G/, PR, G, A, A, G, G/, DETF, A, PTY, APAJ, PTY/, DETS, PR, A/, A, G/, G/, A/, AO/, DETS/, DETS/, A, A/, AO/, A, A, A/, A, A, A, A, G, A, A, PTY, A, A, A, A, A, DETS/, PT, A, DETS, PT/, A, A, A/, A, A, A/, A/

T 15
Vsevolod Emelin *Pojema truby* (The poem of the pipe) /«Поэма трубы»/, 2008. Words 869, adnominals 111, sentences 43.

RC, RC, A, DETQ/, A/, DETS, G, DETS, A, G, G/, A, A/, A/, PT, A/, DETS, A, A, AP, AP/, DETS, A, DETV, RC, A, A, PT/, DETF, A, DETS, DETS/, A, A/, A, DETS/, DETS, A/, PT, DETS/, A/, A, A/, G, A/, DETQ/, DETS, DETS/, A/, A/, RC, A, A/, A, G, RC/, RC, A, A, G, PT/, DETH, A, A, PT/, A, A, A, PT/, A, G/, A/, A, G, DETF/, A/, DETF, G, G/, A, A, A, A/, A/, A/, DETS, DETF, A, I/, G/, A/, A, PT/, PT, A, G, APX/, PR, A/, A, G, RC, DETS, A/, A, A, RC, G/, DETS, DETS/, A/

T 16
A. Kalinina *Peterburggo* /A. Калинина «Петербургго»/, 2010. Words 1925, adnominals 452, sentences 85.

DETQ, RC, G, RC, A, DETS, G/, PR, A, CN/, A/, PR/, DETS, DETF, RC, DETS, A/, PT, PR, PT/, A, PR, PTY, G, DETS, A, RC, DETQ, A, PT, DETS, A/, DETS, A, PT/, G/, A, A, PT, PR, G, A, PT, PR, A/, PT, RC, PT, G, PR/, RC, G, DETQ, DETF, RC, PTY, G, A, PR/, PT, PR, PT, PR, RC, PT, RC, A, RC, G/, AY, DETQ, G, G, G, A, G, G, A, PTY/, A, G, PT, PTY, PT, DETS/, A, CN, AP, PTY, A, DETS/, A, A, A, PR, G, A, PR/, A, G, PTY, G, A, INSTR/, G, G, PTY, RC, A, PR, G, G/, DETS, A, PTY, CN, A, A, A/, DETS, G, G, A, G, PR, G/, DETS, A, A, A, A/, G, A, AP, G, G/, A, A, PR, A, A, G/, AP, G, G/, A, DETS, AY/, A, A, AP, A, A, RC, DETF, AY, PT/, A, DETF, DETF/, A, G, A, A, G, PR, A/, A, A/, G/, PR, PR, A, DETS, PR, PT, A/, RC, A, AP, DETS, A/, DETQ, G, G, G, G, G, G, G, DETQ, G, DETQ, G/, PT, A, PR, DETQ/, G, G, A/, PTY, PT/, PT, A, G, PR, DETS/, PT, A, A, G/, DETQ, DETF, A, A, G/, A, G, PR, G, PTY/, DETS, DETS, G, RC, DETS, PR, G, G, I, A, PR, A, G/, CN, G, PR, A, PR, A, G, AY, PR/, AP, AP, AP, DETQ, RC, DETF/, DETS, A, G, PR, RC/, DETQ, DETS, A, PTY, PTY/, DETF, A, G, A, DETS, A/, A, A, RC/, APAJ, A/, A, A/, A, G, A, A, PR, A, PR, PT, PT, PT, G, PTY, I/, I, A, A, G, A, A, A/, DETS, G, DETQ, G/, PT, G, G, A/, G/, DETQ, G, G/, DETQ, A, RC, PR, PT, G, RC, PT, RC, DETS, A, A, A, DETS, A/, DETS, A, RC/, PTY, PTY, PR/, DETS, A, PT/, PT, G, G, A, RC/, APAJ, PT, G, PT/, A, DETF/, RC/, G/, DETF, A, PT, A, A, DETS/, A, PTY, AP, AP, AP, AP, AP/, AP, DETF, RC, G, G, DETQ/, A/, A, PR, G, RC, A, G, A, G/, A, A/, RC/, DETS, DETQ, PR, G, PT/, APX, AP, A, G, DETS/, DETS, A, DETQ, PT, PTY/, DETS, A, DETS, AP, DETS, PT, G, G, PTY/, DETS, AP, A, AP, G, RC/, DETF, A, G, AP, AY, G, A, G, G, DETF, RC, A, RC/, A, AP, ADV, A/, G, DETS, PTY, PTY, DETS, A/, A/, PTY, DETF, PTY, RC, DETQ, A, DETF, RC, PTY, PTY, PT, RC/, A/, DETS, PR, APAJ, G, A, DETS, PTY/, A, A, G, A, G/

T 17
E. Mihajlichenko *Grecheskie strasti* (Greek passions) /E. Михайличенко «Греческие страсти»/, 2011. Words 968, adnominals 115, sentences 46.

RC, A, G, PT, AP, PT, G/, G/, PT, PT, PT, G, PTY, A/, RC, A, DETF, DETQ, A/, PR/, A, A, DETS/, DETH/, DETQ, DETF, A, G, A, A, PR/, G, DETQ/, A/, RC, DETS, G/, DETS, PTY, A/, A, PT, G, RC/, DETS/, A/, PTY, PT, AP/, A, A/, G, G/, A, PT, G/, A/, A/, DETS/, G/, A, G, PT, PR/, A, G/, G, G/, I/, A, A, G, PT, G/, G, A, G, PR, PR, PR, PR, A, PR/, G/, A/, G, ADV/, RC/, A/, DETF, G/, G, A, G, CN/, RC/, A/, PT/, PTY, RC/, G, PTY, A/, AP, G, RC, DETQ/, DETS, A/, A, G, G/, PR, RC/, G/

Glottometrics 42, 2018, 75-85

The Nature and Hierarchy of Belza-Chains

Gabriel Altmann

Abstract. Belza-chains are uninterrupted sequences of any type of linguistic entities. Their study shows the inertia of some kind of entities. Their length may be captured by usual functions.

Keywords: *Belza chain, phonemes, syllables, words, herbs, parts of speech, sentences, motifs, weighting, combining, exponential function*

Introduction

Belza-chains represent a means enabling us to study the inertia of texts. Based on the Skinner hypothesis, high inertia is a sign of spontaneity, semantic concentration, few pauses in writing, few additional changes of the text by the author or the editors. However, Belza-chains made up of special units – e.g., scientific terms – need not be a sign of spontaneity, just the contrary. Hence, one must search for various kinds of inertia. Originally, it concerned only identical words (cf. Belza 1971, Skorochod'ko 1981), but a text does not consist of words only. Beginning from the lowest level, one must/may consider also *phonemes* (sounds), which are basic especially in poetry (e.g., rhyme, assonance, alliteration, selected vowel sequences in Old Javanese poetry, etc.). If a given phoneme occurs in all higher entities, then one can consider rather subsequent pairs, triads, etc. of phonemes or even, say, the vowels of phonetic words. *Syllables* have no meaning, but are basic for rhythm and rhyme. Even in hexameter, there are inertial structures and changing structures of dactyls and spondees. One may speak about rhythmic inertia here. *Morphemes* and their types represent rather the morphological type of language. If they are classified in one way or another – e.g., as bearers of various grammatical categories, as semantic and grammatical morphemes, as belonging to different or no word classes, prefixes, infixes, suffixes, free and bound morphemes, etc. –, one obtains various chains. They may refer to words, to phrases, to clauses, etc. *Words*, which are the true bearers of meaning, may be classified in an enormous number of ways, as can be seen in any grammar. At the same time, each word can be quantified according to various properties – e.g., length, polysemy, complexity, left and right adnominality of different kinds, verbs according to their valency or semantic classes, meaning classes, etc. For higher entities – such as phrases, clauses, or sentences –, there are both different aspects and different schools working with them.

Further, since the properties of entities can be quantified (which is a higher level of research), the values can be ordered in sequences, in distributions, in motifs, and the results can, again, be classified and have their own properties.

Inertias can appear at any level, and one of them is represented by Belza-chains. They are constructed in the following way: one defines a frame entity – say a sentence, verse, strophe, paragraph, or Frumkina passage –, and studies the repetition of a given property in subsequent units. The number of subsequent frame entities in which the property occurs represents the length of the chain. The various chains need not be separated: a chain can begin (and end) in another chain. Computing the lengths of chains, one obtains a distribution/function whose properties may be used as a characteristic of the inertia. Many short chains testify to small inertia in the given sense; hence, some properties of the distribution may be used, e.g. mean, steepness, the lambda indicator (Popescu, Čech, Altmann 2011; Popescu, Altmann

2014), Ord's criterion (Ord 1972), etc., for characterizing it. It is very important to compare the texts – not only by ordering, but by a test. That means, one either compares the distributions/functions themselves as wholes, or some of their properties.

Now, inertia may be a property of a text, of a writer, of a text type, even of a language (according to the level of the hierarchy), of the author's age or education, of the development of a language, etc. There is a number of possible interpretations and specifications of parameters of the resulting functions.

The lengths of Belza-chains, if written as a sequence, may be examined for runs; the sequence can be segmented into Köhlerian motifs; one can measure the distance between chains of the same length, the lengths constituting themselves a distribution; etc. Evidently, chains are a special discipline lying higher than the usual linguistic levels. It cannot be performed on the langue level, it is restricted to texts.

In the following, we shall provide some examples concerning various levels and units.

Phonemic Level

At the lowest level, one may take into account all phonemes, or only a special class, say vowels or a class of consonants. Below, we present the vocalic structure of the German poem *Der Erlkönig* by J. W. v. Goethe, as has been analyzed be K.-H. Best and presented in Altmann, V., Altmann, G. (2008: 59):

1. e: ai e o: ä: u a u i
2. e i e: a: e i ai e i
3 e: a e: a: e o: i e: a
4. e: a i: i e e e i: a:

5. ai o: a i u: o: a ai e i o
6. i: a: e u: e: e o: i i
7. e: e e ö: i i o: u ai
8. ai o: e i ai e: e ai

9. u: i: e i o e: i i: e a
10. a: ö: e i: e i: i i i: a
11. a u e u: e i a e: a a
12. ai e u e a a ü e e a

13. ai a: e ai a e u ö: e u: i
14. a e e ö: i i: ai e e i
15. ai u: i ai e u: i ai i
16. i ü e e e oi e e: i

17. i ai e a: e u: i i: e:
18. ai e ö e o e i a e ö:
19. ai e ö e ü: e e: e i e ai
20. u i: e u a e u i e i ai

21. ai a: e ai a: e u i: u: i
22. e ö: i o: e a ü: e o
23. ai o: ai o: i e: e e au
24. e ai i: a e ai e o: au

25. i i: e i i ai ai e ö: e
26. u i u: i i i o: au i e
27. ai a: e ai a: e e a e: i
28. e ö: i a i: ai ai e a:

29. e: a: e au e e: ai e e i
30. e: e i a e a e e e i
31. e ai e: o: i ü: e u o:
32. i ai e a e a i a: o:

Here, we find the following vowels:

[a, a:, e, e:, i, i:, o, o:, u, u:, ä:, ö, ö:, ü, ü:, ai, au, oi]

Considering the verse as the frame unit (possible elements of chains we can find the following chains of vowels (cf. Table 1).

Table 1
Lengths of Belza-chains of vowels in verses of *Der Erlkönig*

a	1,3,6,3,1,2,1,1
a:	3,1,1,1,1,1,3,1
e	32
e:	4,4,1,2,1,1,1,3
i	32
i:	1,1,2,1,1,2,1
o	1,1,1,1
o:	1,1,4,1,2,1,2
u	1,1,3,2,1,1
u:	2,1,1,1,1,1,1
ä:	1
ö	2
ö:	1,1,2,1,1,1,1
ü	1,1
ü:	1,1,1
ai	2,1,2,4,5,3,3,2
au	2,1,1
oi	1

In poetry, where one has short framing units (verses), one expects rather longer vocalic chains. It depends also on the character of language, the character of poetry, the author, the way of defining the phonemes, etc. A characteristic feature is, in any case, the mean of chain lengths. Ordering the chains according to length, we obtain the results presented in Table 2.

The mean length of vowel chains in *Erlkönig* can easily be computed as $m_1 = 2.3614$. Although two special vowels occur in each verse and their chain length is 32, the overall mean is rather low. This means that the author tried to avoid vocalic monotony – i.e., the vocalic inertia is small. The fact that two vowels (/i/ and /e/) occur in each verse is rather a property of German.

Now, since we have to do with length, we automatically may model the data by the Zipf-Alekseev function (cf. Popescu, Best, Altmann 2014). However, one always prefers functions having fewer parameters; hence, we apply the exponential function defined as

$$y = 1 + a * e^{-bx}, \tag{1}$$

and obtain the results presented in Table 2.

Table 2
Vocalic Belza-chain length (exponential function + 1) in *Erlkönig*

Chain lengths	Frequency	Computed
1	53	52.70
2	14	15.95
3	8	5.32
4	4	2.25
5	1	1.36
6	1	1.10
32	2	1.00
a = 178.8023, b = 1.2408, R^2 = 0.9928		

We use the function (not the distribution), and since we omit all zero frequencies, we add 1 to the exponential function. The fitting is optimal, as shown by the determination coefficient. The adding of 1 may also be interpreted on the basis of the differential equation leading to (1).

One obtains longer chains if one considers long vowels as representatives of the parallel short vowel – e.g., /a:/ is considered /a/. One obtains shorter chains if one considers, say, pairs of successive vowels. The definitions can, evidently, be variegated. In many languages, there are problems with the interpretation of diphthongs.

For prose texts, one must consider the sentence as a chain unit. However, in that case, many vowels will display a unique, or a very long chain. It depends on the type of language, too.

Another way of investigating the chains of individual phonemes is taking into account the number of chains formed by individual phonemes – for example, in the poem above, there are 4 vowels forming 8 chains (/a/, /a:/, /e:/, /ai/); there are 4 vowels forming 7 chains; there is 1 vowel forming 6 chains; there is no vowel forming 5 chains, etc. For such a short text, the course of numbers is very irregular, and cannot be modeled preliminarily. At the same time, the number of vowels is too small to display a smooth function.

For the sake of comparison, we analyze the poem concerning consonantal chains of verses, and obtain the results presented in Table 3.

Table 3
Lengths of Belza-chains of consonants in verses of *Der Erlkönig*

p	1,1
t	22,9
k	2,2,1,2,1,1,1,1,1
b	1,1,2,1,3,2
d	3,3,5,7,3,5
g	3,2,1,1,1,1,5,2
f	1,1,2,2,1,1,1,1,1
s	1,3,2,2,1,4,5,1
ʃ	1,2,2,1,1,2,1
h	2,2,1,2
x	1,3,5,6,3,2
v	1,3,1,1,3,1,1,1,1,1
z	2,3,1,1,3,1,1,2,1,2,1
l	2,7,6,1,3,1,1
r	19,2,9
m	4,9,3,3,1,6
n	32
ŋ	1

Summing up the individual lengths, we obtain the results presented in Table 4. Again, we fit the exponential function (with added 1) and obtain excellent results.

Table 4
Consonantal chain length (exponential function + 1)

Chain lengths	Frequency	Computed
1	48	47.93
2	23	23.78
3	15	12.05
4	2	6.36
5	5	3.60
6	3	2.26
7	2	1.61
9	3	1.14
19	1	1.00
22	1	1.00
32	1	1.00
$a = 96.6901$, $b = 0.7229$, $R^2 = 0.9838$		

The mean of the consonantal chain length is $m_1 = 2.9231$ – that is, greater than that for vowels, but any tests and conjectures would be premature. A number of texts in many languages should be examined.

Long phonetic chains have a merely euphonic background, and are used rather in poetry. However, there are languages having only five vowels and eight consonants; hence, long chains are a necessity here. Nevertheless, this is also a method for comparing languages typologically. One can take the same (translated) text and evaluate the results.

In general, the smaller the parameter *b* of the exponential function is, the stronger is the phonic inertia of the texts. Comparing the results for the text above, we see that consonantal inertia of Erlkönig is greater than the vocalic inertia. Nonetheless, this is probably caused by the type of language.

Formula (1) expresses the conjecture that the relative rate of change of chain lengths is constant – i.e., one can write

$$\frac{dy}{y-1} = -b\,dx \tag{2}$$

That is, the parameter *b* signals the strength of inertia. The greater *b* is, the more short chains are present – i.e., the inertia is smaller. Parameter *a* depends on the size of the text.

Syllable

Syllables can be classified as open and closed, as stressed or not stressed, as initial, central, or final in the word; they can be left as they are (in their qualitative forms) forming as many forms as there are in language. In the last case, they clearly depend on the words they constitute. Again, the framing chain element can be a sentence or a verse, but one may consider also clauses and phrases. It can be conjectured that strongly synthetic languages have smaller repetitions of syllables than analytic ones. If one examines short frames (e.g., verses), it can happen that there are very short chains only. On the other hand, if one sets up few classes, consisting of, say, 2–3 categories, the chains may be very long. Thus, the first problem is to find an appropriate classification of syllables. In any case, it may be a matter of style how syllables are repeated. A quite strong influence on syllable chains is exerted by rhymes if they have the form AABBCC…, but not by ABAB CDCD… However, in strongly synthetic languages, not even the rhymes need contain the same syllables. Further, in some languages, there may be many assimilations which must be taken into account.

Since the syllable is the bearer of accent, in many languages, the rhythm of poetry is based upon it. Well-known are the classical meters, but even in modern poetry, one can look at the verses from this point of view. Here, the frame unit is the verse, and a Belza-chain is formed by verses having the same sequence of "heavy" and "light" feet. Nevertheless, even this is difficult to decipher because in poetry the stress may be shifted. Here, longer chains mean rhythmic monotony, regularity; long chains support Skinner's hypothesis. The chains may give the poem a special character which may also distinguish poems written in classical meters.

Let us consider a short Slovak poem *K Nitre* by A. Sládkovič, yielding the following syllabic chain lengths with a verse as the frame unit: [2,3,2,2,2,2,2,2,2,2,2,2,2] – that means, a poem without a strong syllabic inertia. In one case only, a syllable is repeated in three verses, the rest are rather random coincidences.

However, even prosaic texts can be analyzed as to the syllables. Taking again the sentence as the frame unit, each sentence consists of syllables, which can also be written phonemically. Preliminarily, one cannot conjecture that there will be a very small number of very short chains because sentences may use the same syllable in various words. The counting of chain length may proceed in the same way as that of sounds: each syllable is considered, and for each the length of the chain (= a number of subsequent sentences containing it), it must be computed separately. Another variant of counting is omitting the chains of length 1 because there are too many syllables not repeated in the next verse. A third possibility is to consider length 1 only if there is no chain in the verse – i.e., for the whole verse. The results must be interpreted in the given sense. A further possibility is to numerate the sentences, set

up a vector consisting of the relevant syllables and the numbers of sentences containing it, and measure the distance between the given chains. For example, let us have the syllable *BA* and write the order number of sentences containing it – e.g., *BA*: 1,2,3,7,9,10,13,14,15. Here, we have the three chains: 1–3, 9–10, and 13–15. Number 7 is a one-member chain. Now, the distances are the numbers of steps necessary to come from one chain to the next – hence, 4 steps from the first chain (1–3) to sentence 7; 2 steps from sentence 7 to chain 9–10; and 3 steps from chain 9–10 to chain 13–15. One can search for the function expressing the length of chain, or for another one expressing the distances between chains; or, finally, one can compute the distances between sentences containing the given syllable. The domain is very rich, and we can make conjectures only after various languages and text types have been analyzed. In all cases, one should compare the computation with the original Skinner hypothesis.

The length of chains – i.e., an indicator – presents the degree of phonic inertia of the text. This will be different even for the same translated text.

Rhythmic chains can be found in poetry. One can consider the verse as a frame unit, and either one constructs a sequence consisting of feet abbreviations (D = dactyl, S = spondee, I = iamb, T = trochee, etc.) and considers a chain the sequence of rhythmically equal verses, or one marks merely the number of non-accented syllables between two accented ones – e.g., *-*--*---*-- is a sequence (1,2,3,2). Then, one may compare the subsequent verses and note the equalities. As a matter of fact, a quite new domain opens for the text-phoneticians.

Morpheme

Morphemes are usually classified in some way, e.g., grammatical vs. lexical, affixal vs. independent, basic vs. specifying (e.g., in compounds or derived words), part-of-speech forming vs. only specifying – for example, the German personal affixes are POS-forming (*-en, -e, -st, -et*), rendering the noun into the category of verbs (e.g., Bild vs. bildet), whereas some prefixes are merely specifying (e.g., arbeiten vs. verarbeiten). In German, some prefixes may be detached and put behind the word. In this position, they are considered adverbs. The problem of finding a Belza-chaining specific for the given text is very complex and very broad. The "same" text (e.g., a translation) can have quite different morphological chains, and one cannot know whether this is caused by the translator, by the language, or by our classification. The simplest way is to consider morphemes from the semantic point of view – e.g., the morphemes *I*, *me*, *my*, *we*, *us*, *our* all contain the same meaning + another meaning belonging to another class. The question also is whether zero morphemes should be taken into account, but if we also treat grammatical categories, then it must be done in any case. In strongly synthetic languages, one may find texts in which almost all sentences contain the same category expressed by variants of a morpheme, or by a zero morpheme. Evidently, this aspect is very complex and one does not know which procedure would be successful.

This enumeration is sufficient to show that not all aspects are reasonable in all languages. Some languages do not express some grammatical categories (e.g., Hungarian has no gender); hence, a comparison is impossible. Some languages do not have affixes, etc. Thus, the morphemic level is a very complex domain for forming Belza-chains.

Word and Hrebs

The classical Belza-chains have been created for words with all their forms, i.e. for lemmas. However, there are still "higher" units, namely hrebs, encompassing all words in which some morphemes are associated with other words – e.g., German personal morphemes such as *-e, -st, -t, -en*, or cases in the singular or the plural of nouns, etc. For example, the German *ich*

zeige represents two hrebs: {*ich, (zeig)-e*} and {*zeig(-e)*} It can be conjectured that considering hrebs yields longer chains than considering merely words. The same word may belong to different hrebs – for example, possessive pronouns or relative pronouns; some possessive morphemes may ascribe the noun to different persons/things. In some cases – e.g., if the same affix concerns a different object –, one must decide whether one ascribes the verb to a certain person – for instance, the difference between "I work" and "he works" is evident, but there is none between "I/you/we/they work". Does "work" belong to a special herb, or to none? If there are no personal verbal affixes in the language, then the verb does not belong to any nominal herb – this is the situation in the Austronesian languages. In Hungarian, even adverbs may contain personal endings – e.g., in Hungarian, there is "elött, elöttünk, elötte, elötted, elöttetek, …" (*in front of/before, in front of us, in front of him, in front of you* /sg. and pl./). Measuring inertia, one must use rather hrebs than words because one wants to distinguish between the homonyms, take into account the references, synonyms, etc.

Consider the 32 verses of *Der Erlkönig* by J. W. v. Goethe. The first word, "wer" (*who*), belongs to the hreb concerning "Vater" (*father*), and one finds directly "Vater" or references to him in verses (1, 2, 3, 4, 5, 6, 8, 13, 15, 21, 23, 27, 28, 29, 30, 31, 32); hence, there are 7 chains (1–6, 8, 13, 15, 21, 23, 27–32). The second word, "reitet", occurs in the first and the 29th verse, forming two chains. The third word, "so", occurs three times (verses 1, 5, and 26), but in two fully different meanings; hence, one can speak about words forming chains of length 1. As can be seen, the hreb analysis is rather complex, and even in short texts, it causes many problems. Hrebs were introduced by L. Hřebíček (e.g., 1997) under the name "sentence aggregates", and later on were renamed as hrebs by other linguists.

Hrebs have many properties (cf. Ziegler, Altmann 2002) which were not yet examined intensively. One can mention, e.g., hreb size, distances between elements of hrebs, all text properties (entropy, repeat rate, etc.), possible scaling of hreb elements from various points of view (e.g., morphological one), diffuseness, etc.

Parts of speech

Since sentences usually contain several parts of speech, it is not very productive to work with them directly. If each sentence contains a verb, one would always obtain a chain whose length is identical with the number of sentences in the text. But if we consider one special POS and its properties, the results may be interesting. A good example is the valency of verbs – i.e., one does not care for sentence frames, but considers merely the sequence of verb valencies. In such a case, one may obtain both runs, and chains of valencies. Runs are automatically chains – i.e., one cannot use the well-known theory of runs because something is added. Another case is that of adnominals (cf. Andreev, Popescu, Altmann 2017); however, they do not consist necessarily of the same parts of speech.

Parts of speech may be classified into motifs (see below), and one can construct hrebs containing the same sequence of parts of speech. Here, not a unique POS is important, but the identical sequence of POS. A simple text will have quite other properties than, for example, a poetic text. Up to now, no examinations in this direction have been performed, but one must try to enter this domain, too.

Considering only a special class of POS, one can find in each of them subclasses that can be quantified and measured. For example, considering the verbs, one can quantify their modality in various ways. In German, modality is expressed by separate words which can be scaled. We obtain, e.g., the sequence and a possible scaling as:

Müssen, sollen, tun, können, dürfen, mögen, wollen ,
1 2 3 4 5 6 7

where "tun" represents a neutral modality, simply the verb itself. In other languages, the situation may be different, and the researcher can choose his/her own scaling. In a stage play, each sentence has a degree of modality, but this should not be mixed up with speech acts.

Sentences

In sentences, there is a sheer infinite number of research possibilities. They can be equally structured, they can express a special speech acts – which can be productively studied especially in stage plays –, they may express a relation to the same object (human, animal, thing), they may express the same psycholinguistic relation, etc. This view is similar to the original Belza-chains, in which a word was decisive. Nevertheless, chains are built not only of words; one can consider also clauses. In poetry, there are three basic entities: the line, the strophe, and the sentence; hence, one can analyze a poetic text in at least three ways. One can say that there are at least as many clause/sentence Belza-chains as there are sentence definitions. The Belza-chaining could, perhaps, reduce the number of definitions, namely eliminating those ones that do not contribute to a deeper insight into the text – e.g., those that have no relation to other properties in the Köhlerian control cycle. In stage plays, each sentence is a speech act having a fully different classification. The speech acts can be symbolized by abbreviations, or scaled in different ways.

Motifs

Motifs, introduced by R. Köhler (2008, 2015, cf. also Köhler, Naumann 2008), may be either qualitative, or quantitative. Qualitative motifs contain symbols representing some entities. Any linguistic entities can be utilized for setting up motifs. A new motif must not contain two symbols of the preceding motif, or the same symbol twice, but a complete text may form one unique chain. That means, this way of setting up Belza-chains does not hold. Qualitative motifs cannot form Belza-chains because they are defined by Köhler in a special way. One searches for a law controlling the creation of chains, but up to now, there has been no idea of what the background properties correlated with this type of entities can be.

Quantitative motifs are constructed as entities containing only non-decreasing numbers. They can be constructed on the basis of word length, number of elements in front of and behind the verb, sentence lengths, degrees of properties, etc. A Belza-chain can be constructed if the subsequent motifs are identical – i.e., they contain the same numbers in the same order. However, this is rather an exception in prosaic texts; one can expect them mostly in poetry. Quantitative motifs can be constructed in various ways, e.g., as non-increasing sequences. One can search for their Belza chains not only concerning their identical membership, but also concerning their other properties – e.g., length, ranges, or means.

Weighting

Weighting of chains is a very complex procedure representing the view of the analyst. Some chains or their members may be weighted according to numerous properties. For example, a vocalic chain may obtain the weight of the given vowel, which may depend on its place in the mouth, on the opening of the mouth; syllables can be weighted according to their length (= number of consonants in them), placing within the word; morphemes can be weighted according to their grammatical features, placing within the word, number of categories to which they may belong, their length; for words, there are a number of possibilities considering various properties, beginning with phonetics up to the lexical level, borrowing status, dialectal status, etc. The higher the level, the more possibilities there are. If we choose

a special level, then forming of hierarchies are possible. That means, the study of inertia of text is a discipline whose departments have not been even mentioned up to now.

Further, if one describes the text in terms of chain weights, one obtains new Köhlerian motifs, a new distribution or function based on an approach deduced from the general theory; one obtains some relations to other linguistic phenomena; etc. That is, this way is infinite. In any case, one should not assume that one discovers the truth, but one investigates some aspects of our human invention concerning language, and at the same time concerning our views of language.

Above, we captured the inertia and length of chains by means of the exponential function which is very simple, but it does not mean that it holds true in all possible cases. Even within the class of the same entities, there may be differences not only in parameters, but also in the substantiation, derivation, and relationship to other properties. It may be conjectured that texts, especially poetic ones, will contain their own inertia, length types, etc. We merely hope that one will publish more results of examinations.

Combining

The above-mentioned levels are in no case an exhaustive description. The number of aspects can be multiplied by their possibilities of combination. Thus, for example, the types of vowels may be combined with the types of syllable in which they occur. In the poem by Goethe, we find the long /e:/ in the first four lines in the words (1) *wer*; (2) *der*; (3) *er*, *den*, *den*; (4) *er*, *er*. Here, we have the chains: (1) Ce:C; (2) Ce:C; (3) e:C, Ce:C, Ce:C; (4) e:C, e:C. That means, there is one chain of length 3 (Ce:C; 1–3), and another chain (e:C) of length 2. The same element in the same frame is counted only once. The situation changed, we could also include the environment of vowels. Of course, the more properties are taken simultaneously into account, the shorter the chains will be. Our a-posteriori question is not "which is better", or "which is more true?", but "which of the combinations can be derived from the background theory?" In this way, one obtains a great number of possible descriptions. The elaboration of possible combinations is an eternal work for a team of linguists.

It is to be noted that Belza chains represent a search for commonalities of any sort. The concentration/inertia of a text may be performed in many ways, and whatever entity is examined, there are always several ways to see the textual reality. In some cases, one must deviate from the classical view of the text, but the capturing of reality leads frequently – even in physics – to fundamental change of our worldview. As a matter of fact, Belza-chains represent another view of runs in text and, as can be seen, yield different results – they are more linguistically substantiated than the runs.

A more general view of texts is represented by "hrebs", whose construction does not depend on the continuity of the units analyzed. Again, one can consider any type of units. The difference to the usual frequency counts lies in the fact that, e.g., for morphemes, one considers all morphemes with the same meaning as members of the same hreb. These units need not be morphs (i.e., forms of the same morpheme), but must have the same meaning – e.g., in English *I* and *me*, in German *du* (mach)-*st*, in Slovak *my* (vidí)-*me*, in Italian *noi* (parl)-*iamo*, where the italic parts represent morphemes having the same meaning, but not belonging to the same morpheme formally.

References

Altmann, V., Altmann, G. (2008). *Anleitung zu quantitativen Textanalysen*. Lüdenscheid: RAM-Verlag.

Andreev, S., Popescu, I.-I., Altmann, G. (2017). Some properties of adnominals in Russian texts. *Glottometrics* 38, 77–106.

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

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

Köhler, R. (2008). Sequences of linguistic quantities. Report on a new unit of investigation. *Glottotheory* 1(1), 115–119.

Köhler, R. (2015). Linguistic motifs. In: Mikros, G.K., Mačutek, J. (Eds.) (2015). *Sequences in Language and Text: 89–108.* Berlin/Boston: de Gruyter Mouton.

Köhler, R., Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In: Preisach, Ch., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (eds.), *Data Analysis, Machine Learing and Applications: 635–646*. Berlin/ Heidelberg: Springer.

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

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

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

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

Glottometrics 42, 2018, 86-90

Word Length with G. Herdan

To the memory of G. Herdan who died 16. 11. 1968

Karl-Heinz Best
Gabriel Altmann

The problem of word length is as old as quantitative linguistics itself. G. Herdan published many data but did not care for modeling, he rather computed the proportions, mean length and the entropy which was very frequently used at his times, that is, he characterized the distribution using an indicator. Nevertheless, he evaluated the length in terms of syllable numbers while many authors after him used for English the number of letters which is, especially for English, a nonsense. What more, there were authors who computed the English sentence length in terms of letter numbers. Today, we know that syllables and morphemes are the immediate constituents of words and the only possibility to measure the length of a linguistic unit is the counting of its immediate constituents. For example, sentence length must be measured in terms of clause numbers – not in word numbers.

Herdan's life and activity have been described in Best, Altmann (2007), he is mentioned many times on the Internet because he was lawyer, linguists, sinologist and mathematician and wrote five books on language (cf. also Best 2015).

In order to exploit the data collected by Herdan (1966: 284 ff.) for English, German and Russian, we try to find a model capturing all. We do not use a distribution but rather a function, a method which is, from the methodological point of view, of the same value because one can use the results formally. We know that languages differ in the use of word length, e.g. in agglutinating languages, they are longer than in isolating languages but this difference can be stated comparing the parameters of the function.

Here we use the Menzerathian function, $y = ax^b\exp(-cx)$ whose differential equation

$$y'/y = -c + b/x$$

says that the *relative rate of change* of frequency (y'/y) depends on the language constant *c* modified by the function representing the requirements of the speaker (b) and is equilibrated by the simple function of the hearer (b/x). Applying this function to Herdan's data, we obtain the results presented in Table 1. Unfortunately, in several cases, Herdan did not give the exact name of the text analyzed. The alternative derivation of the law of length presented in Popescu, Best, Altmann (2014: 5) is $y'/y = a/x + b*\ln(x)/x$ yielded after reparametrization the Zipf-Alekseev function. Concerning word length, the Zipf-Alekseev function yielded in some cases unsatisfactory results (e.g. in Inuktitut and Vogul, cf. Meyer 1997, Kahl 2002) but there are, most probably, some boundary condition which must be taken into account.

Table 1
Fitting the Menzerathian function to G. Herdan's results

	Shakespeare Henry IV, Pt. 2, Prose		Shakespeare Henry IV, Pt. 2, Verse		Bacon Essays	
Length	Frequ.	Menz.	Frequ.	Menz.	Frequ.	Menz.
1	10965	10965.04	9076	9075.85	4640	4638.77
2	2177	2176.02	1918	1920.83	1000	1020.97
3	430	436.23	476	462.89	420	367.10
4	99	87.82	108	117.64	167	161.37
5	23	17.72	4	30.78	41	79.20
6	2	3.58			3	41.67
7	2	0.72				
	a = 53919.7106 b = -0.0352 c = 1.5928 R^2 = 1.00		a= 31362.6474 b = -0.4514 c = 1.2400 R^2 = 1.00		a = 6459.3503 b = -1.7062 c = 0.3311 R^2 = 0.9996	

	Gray, Poems		Gray, Letters		Johnson	
Length	Frequ	Menz	Frequ	Menz	Frequ	Menz
1	1769	1769.02	3987	3986.68	1268	1267.50
2	585	584.67	831	836.24	423	428.99
3	70	73.43 6.21	281	269.93	195	179.60
4	23	0.42	121	103.95	77	82.03
5	1		15	44.10	29	30.30
6			2	19.89	-	-
7					8	9.78
	a = 55090.1914 b = 3.3633 c = 3.4384 R^2 = 0.9999		a =6727.1175 b = -1.4984 c = 0.5232 R^2 = 0.9999		a = 2243.3654 b = -0.7393 c = 0.5709 R^2 = 0.9996	

	Gibbons, 2 samples		Carlyle, French revolution		Carlyle Past and Present	
Length	Frequ	Menz	Frequ	Menz	Frequ	Menz
1	1860	1857.65	713	712.51	692	691.95
2	614	640.69	180	187.20	193	193.50
3	342	279.80	92	74.00	73	72.74
4	123	134.60	32	34.57	34	30.86
5	27	68.27	8	17.70	12	13.99
6	-	-	2	9.60	3	6.62
7	4	19.22				
	a = 3049.6613 b = -0.8206 c = 0.4957 R^2 = 0.9973		a = 1013.7738 b = -2.4196 c = 0.3526 R^2 = 0.9986		a = 1213.1428 b = -1.0283 c = 0.5615 R^2 = 0.9999	

	Carlyle, Heros and hero worship		Macaulay Clive		Macaulay, Hastings and Milton	
Length	Frequ	Menz	Frequ	Menz	Frequ	Menz
1	811	810.78	1260	1258.72	723	722.27
2	208	211.21	375	391.11	169	179.40
3	88	80.52	201	161.41	95	74.48
4	39	35.87	67	74.82	42	38.16
5	5	17.41	14	36.96	17	21.94
6			5	19.01	2	13.57
7					1	8.82
	a = 1243. 6351 b = -1.3234 c = 0.4278 R^2 = 0.9995		a = 2044.4798 b = -0.9865 c = 0.4850 R^2 = 0.9977		a = 843.4224 b = -1.7857 c = 0.1551 R^2 = 0.9981	

	Shaw Dramatic criticism		Goethe Wilhelm Meister		Lichtenberg	
Length	Frequ	Menz	Frequ	Menz	Frequ	Menz
1	584	583.86	587	587.16	539	538.76
2	184	185.88	410	408.97	317	318.52
3	72	65.78	146	149.93	136	132.64
4	20	24.31	49	42.26	49	47.85
5	5	9.20	8	10.32	7	15.96
	a = 1421.0639 b = -0.3680 c = 0.8895 R^2 = 0.9997		a = 3957.5143 b = 2.2310 c = 1.9081 R^2 = 0.9997		a = 2120.1859 b = 1.2182 c = 1.3700 R^2 = 0.9995	

	Pushkin		Turgeneff		Tolstoy, Autobiography	
Length	Frequ	Menz	Frequ	Menz	Frequ	Menz
1	290	286.08	334	326.86	393	389.51
2	307	321.78	293	319.91	369	382.30
3	204	170.09	241	205.27	253	233.40
4	39	65.99	123	110.80	120	117.33
5	7	21.62	19	54.40	37	53.02
6	1	6.36	6	25.16	10	22.40
7					5	9.03
	a = 1569.0044 b = 2.6250 c = 1.7019 R^2 = 0.9768		a = 923.6976 b = 1.4677 c = 1.0388 R^2 = 0.9614		a = 1245.7019 b = 2.6503 c = 1.1626 R^2 = 0.9941	

Nevertheless, the function captures the data excellently, as can be seen considering the determination coefficient R^2 which is in all cases greater than 0.96.

One expects that the parameters *b* and c are in some way associated. If one sets $c = f(b)$, one obtains an increasing but strongly oscillating function. This may be caused by the

mixing of languages. Eliminating the Russian data, we obtain a simple exponential function given as

$$c = 0.9193*\exp(0.3852b)$$

yielding the determination coefficient $R^2 = 0.9047$. The results are presented in Table 2.

The fact that in the Russian data the parameter dependence deviates from the English-German relations may be caused by typological differences, or, by Herdan's way of counting which was not documented in his book.

Table 2
Dependence of parameter c on parameter b

Parameter b	Parameter c	Computed exponential
-2.4198	0.3526	0.361964
-1.7857	0.1551	0.462101
-1.7062	0.3341	0.476470
-1.4984	0.5232	0.516173
-1.3234	0.4278	0.552165
-1.0283	0.5615	0.618633
-0.9865	0.4850	0.628673
-0.8206	0.4857	0.670157
-0.7393	0.5709	0.691474
-0.4514	1.2400	0.772566
-0.3680	0.8893	0.797786
-0.0352	1.5928	0.906894
1.2172	1.3700	1.469113
2.2310	1.9081	2.170909
3.3633	3.4384	3.357765

The given relation *c = f(b)* is a step into a more abstract dimension and may be examined for different texts in different languages. It will surely be different for Finno-Ugric or Austronesian languages but one could exploit all computations performed up to now. The fact that this would be an enormous work should not hinder young scientists to continue performing the Herdanian heritage.

If we characterize the sequences using Ord's criterion <I,S> we may easily see that in the three given languages, the greater is I, the greater is S. However, the position of the lines is quite different. This could be one of the possible characterizations of languages.

The number of analyses of word length has been performed in dozens of languages and thousands of texts. Usually, a probability distribution has been applied. Here, we have tried to simplify and to unify the computation and the derivation. Checking the model in other languages we obtained always positive results using the Menzerathian function.

In the last years, all lengths in language were modeled uniformly, namely applying the Zipf-Alekseev function (cf. Popescu, Best, Altmann (2014) but most probably it depends on the way of measurement which function should be used. For example, in some Slavic languages one can consider some prepositions as proclitics of the following word and omit fully the length zero. In any case, it has been shown that all lengths abide by a law; the present contribution shows that boundary conditions can be taken into account directly when modeling.

References

Best, K.-H. (2009). Herdan, Gustav. In: Stammerjohann, Harro (Ed.): *Lexicon Grammaticorum. A Bio-Bibliographical Companion to the History of Linguistics. Volume II: A – K, 41 ff.*. Tübingen: Niemeyer

Best, K.H., Altmann, G. (2007). Gustav Herdan (1897-1968). *Glottometrics 15, 92-96.*

Best, K.-H. (2015). *Studien zur Geschichte der Quantitativen Linguistik: 68-73.* Lüdenscheid: RAM-Verlag.

Herdan G. (1966). *The Advanced Theory of Language as Choice and Chance*. Berlin/Heidelberg: Springer-Verlag.

Kahl, S. (2002). Wortlängenverteilungen in wogulischen Texten. *Göttinger Beiträge zur Sprachwissenschaft 7, 51-63.*

Meyer, P. (1997). Word-length distribution in Inuktitut narratives. Empirical and theoretical findings. *Journal of Quantitative Linguistics 4(1-3), 143-155.*

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

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.