Mathematics of Harmony, Generative Grammars and Quantitative Linguistics

Gregory Martynenko

Preface

During last years I stay in close contact with a well-known Ukrainian and Canadian scientist **Aleksey Shakhov**, who is the founder and the leader of the International Club «The Golden Section».

When he learnt that I am going to make a trip to Graz, he told me that in 1990th in Graz two international conferences on Golden Section and Fibonacci numbers took place. His presentations on these conferences had great success. Now he has the warmest memories about Graz. He sends his best greetings to all of us from Canada. He asked me to find the monument to Johannes Kepler, who made a significant contribution to the Theory of Golden Section and Fibonacci numbers.

Aleksey STAKHOV and Johannes KEPLER in GRAZ

Johannes KEPLER
(1571-1630)

Johannes KEPLER observed that the ratio of consecutive Fibonacci numbers converges. He wrote that

"as 5 is to 8 so is 8 to 13, practically, and as 8 is to 13, so is 13 to 21 almost",

and concluded that the limit approaches the golden ratio, which equals.

$$\lim_{n\to\infty} \frac{F(n+1)}{F(n)} = \varphi, \quad \varphi = \frac{1+\sqrt{5}}{2} \approx 1.6180339887\ldots$$

For example, the initial values 19 and 31 generate the sequence 19, 31, 50, 81, 131, 212, 343, 555 ... etc. The ratio of consecutive terms in this sequence shows the same convergence towards the golden ratio.

Fibonacci numbers

are the nucleus of

Mathematics of Harmony

Mathematics of Harmony

Mathematics of harmony is based on:

- the theory of Golden Section,
- Fibonacci numbers,
- theory of proportions and progression,
- iterated radicals, and continued fractions,
- symmetry theory,
- number theory,
- combinatorial analysis,
- etc.

Application in philological studies

Ideas of *mathematics of harmony* become actively embedded in philological studies:

- stylometrics,
- poetics,
- prosody studies,
- semiotics of mathematical languages,
- text and speech dynamics,
- etc.

A demonstration of using Mathematics of Harmony techniques in linguistics:

Investigation of Real and Potential Lexical Richness in Text Corpus

Introduction

In lexical statistics and statistic lexicography various **indices** are used to measure **lexical richness** (diversity) of text vocabulary. The necessity to use these indices is caused in particular by the fact that the size of frequency words lists strongly depends on sample size of text or corpus. Therefore, it is valid to compare frequency words lists using standard methods only in case of equal text samples.

Because of that fact it is important to find those parameters, which do not depend on sample size.

For that purpose various analytic dependencies **«frequency list - text»** are built and indices of lexical richness are further built taking into account these dependencies.

The investigation was made on material of frequency lists of prose texts by **Anton Chekhov, Leonid Andreev** and **Alexander Kuprin**. These frequency lists are representative enough: the first two were made on samples of 200 (two hundred thousend) of words and the third - on that of about 300 (three hundred thousand) of words.

Anton Chekhov (1860-1904)

Leonid Andreev (1871-1919)

Alexander Kuprin (1870-1938)

- The dependency «text sample size - vocabulary size» was analyzed using special methodology, which was based on least square technique with considerable modifications caused by specific of research material.

- Corpus of short stories by each author was presented as a sequence of stories randomly attached one to another. Frequency list was calculated each time with attachment of each new story. It was found that the size of vocabulary for each author increase with fading rate. However, it's difficult to say

  1) **Whether vocabulary size tend to some upper limit or not?** and

  2) **What is the analytic form of these tendency?**

Approximation of Dependency «Text Size - Vocabulary Size»

For approximation of empiric dependencies between text size and vocabulary size a number of theoretical functions of both asymptotic and not-asymptotic growth were used.

For not-asymptotic dependency were used four elemental functions:

$y = ax + b$	linear function
$y = ax^b$	power function
$y = ae^{bx}$	exponential function
$y = a(\ln x)^b$	logarithmic function

What concerns three other elementary functions, it is possible to convert them into linear functions taking the logarithm. In the result the equations have the following form:

$\ln y = \ln a + b \ln x$	power function
$\ln y = \ln a + bx$	exponential function
$\ln y = \ln a + b \ln \ln x$	logarithmic function

The list of asymptotic-growth functions which we used in our research was considerably longer. There are three groups of these functions. The first group contain difference- fractional functions:

$y = k - \frac{a}{x^b}$	power function
$y = k - ke^{-ax^b}$	exponential-power function (Weibull function)
$y = k - \frac{a}{(\ln x)^b}$	logarithmic-power function

The second group contains variants of fractional-exponential function:

$y = \frac{k}{e^{\frac{a}{x^b}}}$	exponential-power function
$y = \frac{k}{e^{\frac{a}{e^{bx}}}}$	double exponential function
$y = \frac{k}{e^{\frac{a}{(\ln x)^b}}}$	exponential-logarithmic function

The third group form logistic functions:

$y = \frac{k}{1 + \frac{a}{x^b}}$	power logistic function (delayed logistic function)
$y = \frac{k}{1 + \frac{a}{e^{bx}}}$	exponential logistic function
$y = \frac{k}{1 + \frac{a}{(\ln x)^b}}$	logarithmic logistic function

Each of nine functions of asymptotic growth by means of taking the single or repeated logarithm may by also transform into linear dependency. Linear variants of these functions have the following forms:

The first group:

$\ln(k-y) = \ln a - b \ln x$	power function
$\ln\ln\frac{k}{k-y} = \ln a - b \ln x$	exponential-power function (Weibull function)
$\ln(k-y) = \ln a - b \ln\ln x$	logarithmic-power function

The second group:

$\ln\ln\frac{k}{y} = \ln a + b\ln x$	exponential-power function
$\ln\ln\frac{k}{y} = \ln a - bx$	double exponential function
$\ln\ln\frac{k}{y} = \ln a - b\ln\ln x$	exponential-logarithmic function

The third group:

$\ln\frac{k}{k-y} = \ln a - b\ln x$	power logistic function (delayed logistic function)
$\ln\frac{k}{k-y} = \ln a - bx$	exponential logistic function
$\ln\frac{k}{k-y} = \ln a - b\ln\ln x$	logarithmic logistic function

A system of normal equations for linear dependency is very simple, so it is not a problem to solve it. However, in our case on of the parameters (asymptote) is included into dependent parameter. This fact does not allow to use the least square technique as it is.

Method of approximation is described in the book *G. Martynenko «Fundamentals of Stylometrics»*.

The results of approximation :

The best similarity with empiric data was obtained by asymptotic-growth functions, and the **Weibull function** was the best:

$$y = k - ke^{-ax^b}$$

Its linear variant has the following form:

$$\ln \ln \frac{k}{k-y} = \ln a - b \ln x \, .$$

Empiric dependency of vocabulary size from sample size for Anton Chekhov

Number of stories	Number of word forms	Empiric number of lexemes	Theoretic number of lexemes
10	4250	1593	1517
20	8735	2600	2603
30	13304	3488	3528
40	19413	4550	4588
50	24938	5344	5424
60	32340	6188	6408
70	38343	7006	7114
80	52139	8316	8502
90	59940	9103	9171
100	74670	10221	10257
110	88823	11355	11128
120	105084	12121	11966
130	127428	13360	12897
140	167076	14103	14105
150	198066	14610	14776
FORECAST			
189	250000		15554
379	500000		16838
757	1000000		17088
1514	2000000		17100

Empiric dependency of vocabulary size from sample size for Leonid Andreev

Number of stories	Number of word forms	Empiric number of lexemes	Theoretic number of lexemes
5	16255	3976	3964
10	40244	6504	6575
15	74789	9324	9173
20	115802	11296	11498
25	133444	12420	12348
30	175281	14205	14118
35	198592	14942	14988
FORECAST			
44	250000		16689
88	500000		22482
176	1000000		28839
352	2 млн		34777

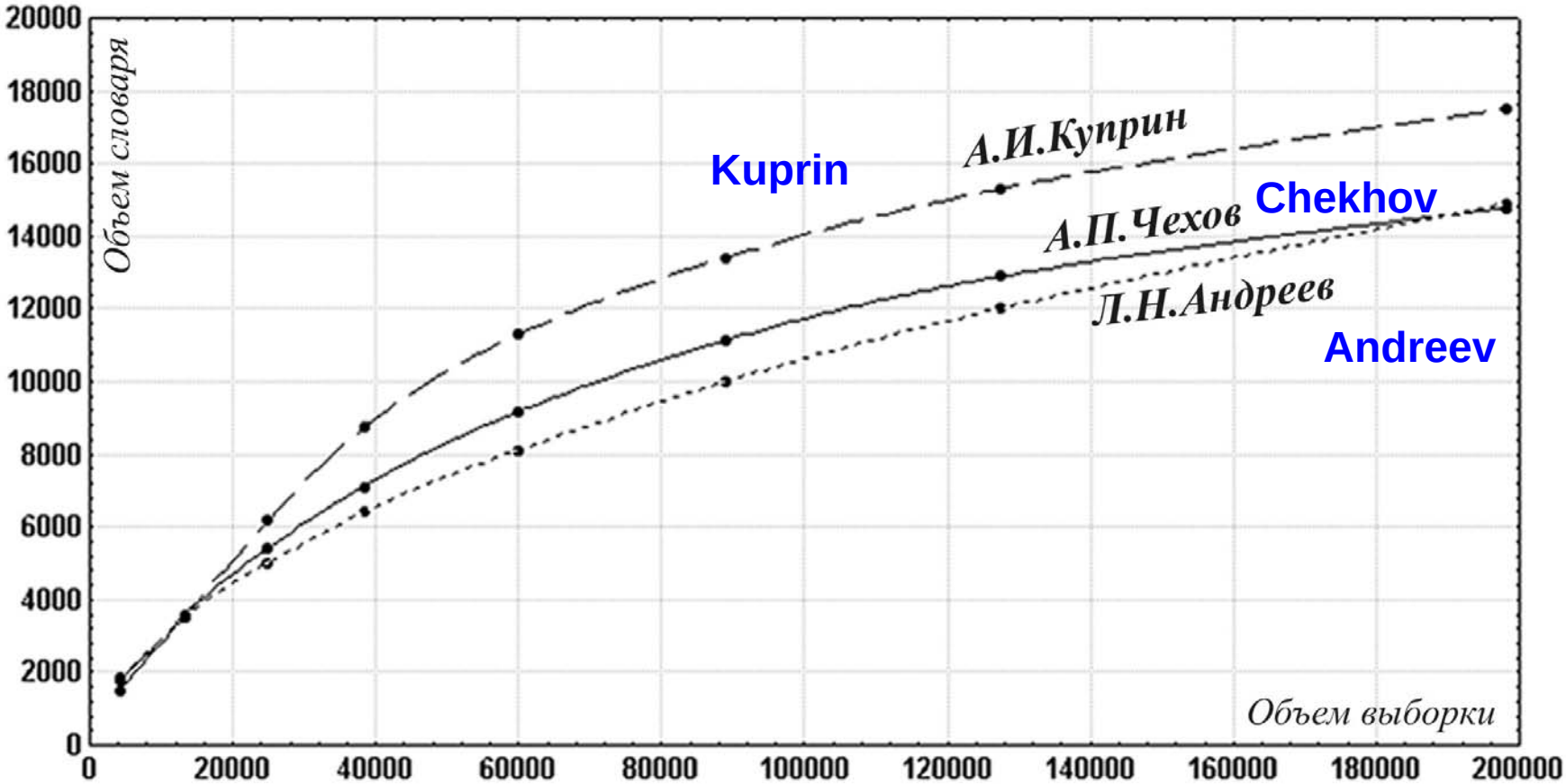
Empiric dependency of vocabulary size from sample size for Alexander Kuprin

Number of stories	Number of word forms	Empiric number of lexemes	Theoretic number of lexemes
5	18340	4952	4844
10	42825	8408	8305
15	51244	9004	9269
20	62779	10222	10471
25	79192	11883	11998
30	91851	13099	13060
35	102680	13070	13903
40	129278	15830	15763
45	14391	16670	16651
50	156834	17416	17442
55	171239	18274	18239
60	183538	18981	18881
65	208193	20184	20075
70	225079	20959	20830
80	268150	22549	22560
90	310372	24103	24032
FORECAST			
101	350000		25249
145	500000		28819
290	1000000		34789
580	2000000		38303

Constant coefficients of Weibull function in dependency «text size - vocabulary size» for Chekhov, Andreev, Kuprin

	Chekhov	Andreev	Kuprin
k	17100	42172	39612
a	$2{,}18 \cdot 10^{-4}$	$3{,}04 \cdot 10^{-4}$	$2{,}52 \cdot 10^{-4}$
b	0,798	0,596	0,65

Vocabulary size in dependency of sample size

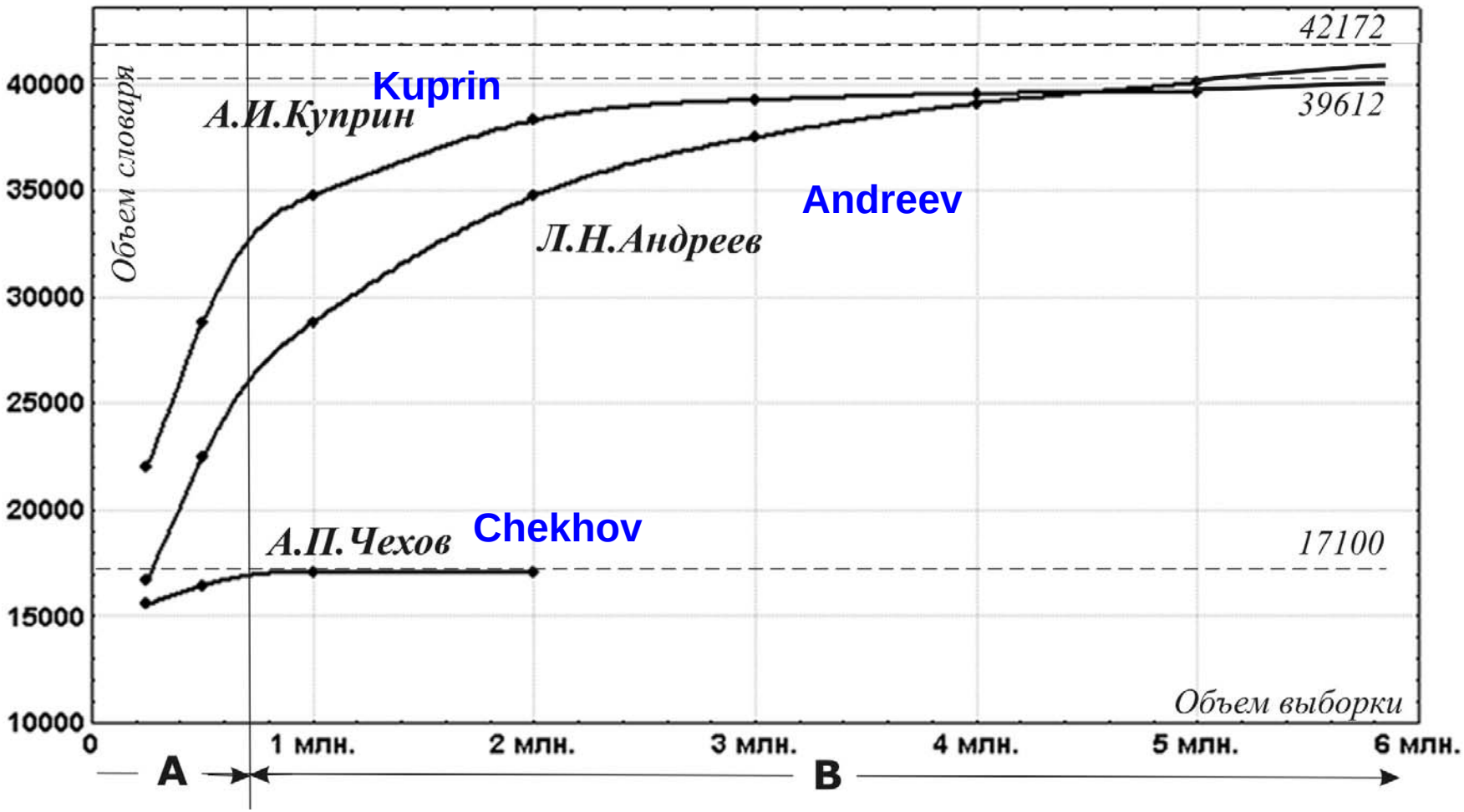


Extrapolation And Forecast of Vocabulary Size

Having determined theoretical parameters of distributions in concern we may forecast values of parameters for sample sizes considerably exceeding the analyzed size

and moreover – the size exceeding real bounds all literary works by some writer.

Forecast of Vocabulary Size



Golden section of asymptotic levels for vocabulary size

Many literary critics and philologists agree in opinion that **Leonid Andreev** was extraordinary in his style. Our prognostic results confirmed that he was an extraordinary writer even in size of his vocabulary. In this aspect he was the direct opposite of **Anton Chekhov**, which had a comparatively poor vocabulary because of his inclination to extraordinary generalization.

The **ideal short** story for Chekhov was the following

«*They loved one another*».

That' all!

We may suppose that **other writers** in the beginning of the 20-th century are **located between these two stylistic poles**.

Last night contribution

to the Golden section theory

The difference between potential vocabulary of Chekhov (17100 words) and Andreev (42172) which equals to **25072** is very near to the geometrical mean of these values, which **26854**.

Let Chekhov' vocabulary size be *a*,

and Andreev's vocabulary size be *c*.

Conclusion

If Chekhov's vocabulary size equals to 1,000

then middle (virtual) Russian prose writer equals to $1,618 = \varphi$

Andreev' vocabulary size equals to $2,618 = \varphi^2$

This a pure Harmony!

The Golden section theory is going to the top of Beauty step by step!