Internal Dynamics of Text: Parts of Speech Distribution in Verse[*]

Vadim Andreev[1]
vadim.andreev@ymail.com

Larisa Beliaeva[2]
lauranbel@gmail.com

[1]Smolensk State University, Smolensk
[2] Herzen State Pedagogical University of Russia,
Russian Federation

Abstract

The research is aimed at the study of the degree of regularity in the relationship between the frequencies of different parts of speech in a verse text, in particular between verbs and nouns. The data-base for the analysis includes 20 sonnets of famous Russian poets of the Silver Age of Russian poetry. The results demonstrate regularity in the distribution of parts of speech frequencies. The exponential function provides a good fit.

Keywords: *parts of speech, exponential function, static and dynamic description.*

1 Introduction

Parts of speech (PoS) are often used in research in the sphere of quantitative linguistics, stylometry and others [Best, 1994; Stamou, 2008]. Analysis of the frequencies of different PoS in a text and proportions between them allow to solve important problems in “linguistics of verse” which has been intensively developing in Russia in the 20th century [Gasparov, 2012].

Depending on the peculiarities of individual styles the frequency of PoS vary to some extent, and sometimes rather considerably [Čech, Altmann, 2013]. Nevertheless it is possible to raise a question of the possible limits of variation and if there are any tendencies of keeping certain proportions between PoS, if there is any general regularity in their frequencies common for all the speakers of the same language. In some studies the results obtained demonstrated the existence of certain order in PoS distribution in speech [Andreev, Popescu, Altmann, 2017]. The present research is aimed at exploring the possibility of such general tendencies in the distribution of parts of speech in verse texts written by authors, differing in style and creativity manner.

2 Data-base

Verse is characterized by a much bigger number of restrictions and rules in choosing words than prose whereas sonnets is a poetic genre with one of the most formalized structure and a big number of strict schemes.

The data-base includes 20 sonnets by famous Russian authors V. Brusov, K. Balmont, V. Ivanov and M. Voloshin, written by these poets in the first part of their creative activities. All these poets belong to the period known as the Silver Age of Russian poetry (the beginning of the 20th century) when much attention was paid exactly to this strictly structured genre of sonnet. Below the names (titles) of these sonnets are given with their text numbers in the data-base.

Valery Brusov

T1	Teny proshlogo
T2	K portretu R. D. Balmonta
T3	Zhenshchine
T4	Kleopatra
T5	Sonet o poete

Konstantin Balmont

T6	Bretan’
T7	Propovednikam
T8	Proklyatiye gluposty
T9	Razluka
T10	Put’ pravdy

Maksimilyan Voloshin

T11	“Starinnym zolotom i zhelchyu napital…”
T12	“Zdes’ byl svyaschenny les. Bozhestvenny gonets…”
T13	“Ravnina vod kolishitsa shiroko…”
T14	“Nad zibkoy ryab’u vod vsayet iz glubiny…”
T15	“Mare internum”

Vyacheslav Ivanov

T16	Na mig (“Den’ purpur tsarstvenny dayet…”)
T17	Polyet
T18	La superba
T19	La pineta
T20	Nostal’giya

3 Methods and feature set

The following parts of speech were counted in the sonnets: nouns (N), adjectives (A), verbs (V), adverbs (ADV), personal pronouns (PRNP), other types of pronouns which can be used in attributive function (PRNA), participles (PTL), adjectivized participles (PTLA), category of state words – adjectives, used as predicates in non personal sentences (STW).

After counting the PoS in the sonnets quantitative data were obtained, specifying their frequencies. Thus in the sonnet Zhenshchine by Brusov (3) the following numbers of PoS were obtained: 27 nouns, 15 personal pronouns, 9 verbs, 5 adjectives, 4 adjectivized participles, 3 participles and 2 adverbs. (Category of state words and pronouns-adjectives were not registered).

The frequencies of all morphological classes in the samples were ranked in decreasing order so that the most frequent PoS is ranked higher than all the others, the second frequency PoS receives Rank 2, etc. To fit the distribution of such ranked PoS frequencies the formula of the exponential function which was suggested for such purposes in [Andreev, Mistecky, Altmann, 2018] was used:

$$y = a * \exp^{(-b*x)} \ ,$$

where a and b are parameters.

If some PoS class was not found in the sample it was omitted (no zero classes were used).

Consider, for example, the above-mentioned sonnet (3). PoS counting in it brought about the following numbers, represented in Table 1. The first column of this table shows rank numbers, the second represents the PoS classes, the third is their observed frequencies in the sample, the forth column shows theoretically expected frequencies which should be according to the formula. Besides, at the bottom of the table the values of a and b parameters and of the determination coefficient R^2 are presented. The coefficient of determination is a measure of goodness of fit and provides information on whether a statistical model fitted to empirical data is successful. R^2 ranges between 0 and 1. When $R^2 > 0.8$ the model fits well.

Table 1. Fitting PoS distribution in T3

Rank	PoS	Observed frequencies	Theoretical frequencies
1	N	27	26.56
2	PRNP	15	15.73
3	V	9	9.31
4	A	5	5.51
5	PRNA	4	3.27
6	PTL	3	1.93
7	ADV	2	1.15
a=44.848, b=0.524 R^2= 0.9928			

In our case the value of R^2 is over 0.99 which implies a very good fit. In Figure 1 this is demonstrated graphically.

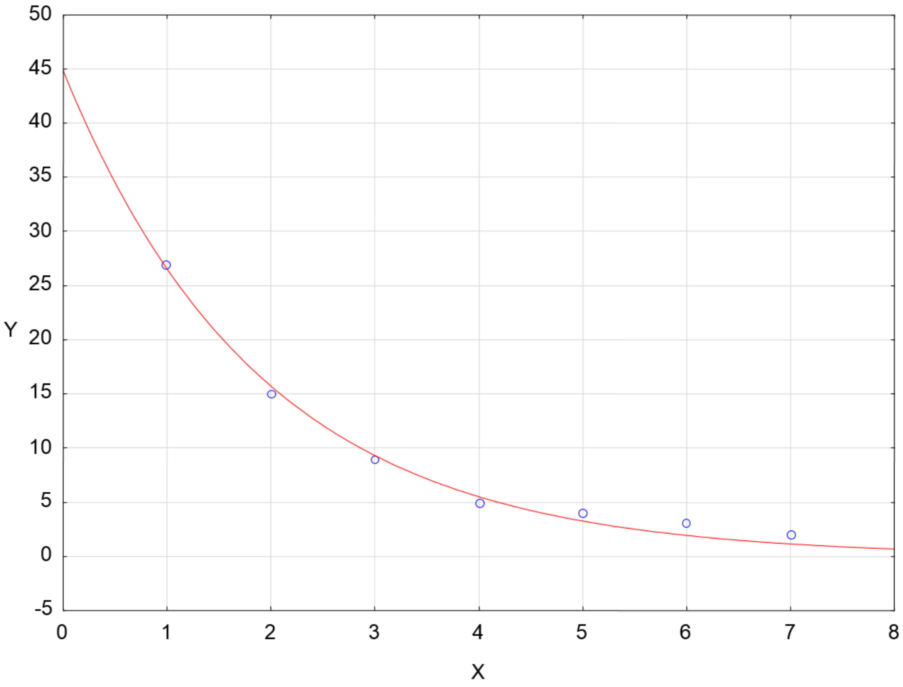

Figure 1: Observed and theoretically expected frequencies

Along the x-axis we have ranks and along the y-axis—the values of the observed (empirical) and theoretically expected frequencies. As shown in Figure 1 the observed frequencies (OBSF) are very close to those expected on the curve (EXPF).

For all the sonnets the results of fitting were obtained, they are shown in Table 2.

All the values of the determination coefficient R^2 are very high. Thus even the lowest R^2 value for fitting the distribu-tion of PoS in 4 $(R^2 = 0.8897)$ and in 7 $(R^2 = 0.8924)$ should be considered as a proof of a good fit. Since the sonnets were written by different poets whose style and manner of writing as well as the topics were different, it should be recognized that the distribution of PoS does not depend on such individual matters, but displays some kind of regularity.

From the point of view of how description takes place in sonnets one can group different PoS into two classes. One of these classes actualizes a static vision of the poetic world [Naumann, Popescu, Altmann, 2012]. In this case the author depicts the world attributing to the themes, expressed by nouns, some features which are viewed as more or less permanent qualities. This is achieved by using such PoS as A, PRNA, PTLA, STW. The other class, on the contrary, gives a description which can be called dynamic, because the features ascribed to the themes in the sonnet are represented as a process or action. This class includes V and PTL. Further on we shall analyze how these two classes of PoS interrelate with one another [Martynenko, 2004].

Table 2: Fitting PoS distribution in 20 sonnets

T1			T2			T3			T4		
PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF
N	33	31.96	N	18	17.82	N	27	26.56	N	29	26.80
A	12	15.51	V	10	11.28	PRNP	15	15.73	PRNP	10	15.15
V	10	7.53	PRNP	9	7.15	V	9	9.31	A	8	8.56
PRNP	3	3.65	PRNA	5	4.53	A	5	5.51	PRNA	7	4.84
ADV	3	1.77	A	2	2.87	PRNA	4	3.27	V	7	2.73
PTLA	3	0.86	PTLA	1	1.82	PTL	3	1.93	ADV	2	1.54
PRNA	3	0.42	PTL	1	1.15	ADV	2	1.15	PTL	2	0.87
a=65.866, b=0.723			a=28.130, b=0.457			a=44.848, b=0.571			a=47.4355, b=0.571		
R^2=0.9551			R^2=0.9709			R^2=0.9928			R^2=0.8897		

T5			T6			T7			T8		
PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF
N	24	24.32	N	27	26.81	N	28	26.30	N	32	31.00
PRNP	16	16.64	A	13	13.81	PRNP	8	13.01	PRNP	11	14.24
V	14	11.38	V	8	7.11	A	7	6.43	V	7	6.55
PRNA	7	7.78	PRNP	4	3.67	V	7	3.18	A	6	3.01
A	5	5.32	ADV	1	1.89	ADV	3	1.57	PRNA	3	1.38
ADV	4	3.64	PTLA	1	0.97	STW	3	0.78	ADVR	2	0.64
PTL	1	2.49	PRNA	1	0.50	PRNA	2	0.38			
STW	1	1.70									
a=35.567, b=0.381			a=52.045, b=0.663			a=53.183, b=0.704			a=67.442, b=0.777		
R^2=0.9768			R^2=0.9952			R^2=0.8924			R^2=0.9596		

T9			T10			T11			T12		
PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF
N	20	19.34	N	31	29.94	N	40	39.60	N	31	30.09
PRNP	11	12.82	V	13	15.24	A	14	14.47	A	14	16.85
V	10	8.50	A	6	7.76	V	5	5.29	V	11	9.44
A	5	5.63	ADV	6	3.95	ADV	4	1.93	ADV	6	5.29
ADV	4	3.73	PRNP	4	2.01	PTL	4	0.71	PRNA	4	2.96
PTLA	2	2.48	PRNA	4	1.02	PRNA	2	0.26	PRNP	1	1.66
PRNA	2	1.64	PTL	2	0.52	PRNP	1	0.03	PTL	1	0.93
PTL	2	1.09	PTLA	1	0.27						
a=29.174, b=0.411			a=58.805, b=0.675			a=108,379, b=1,007			a=83.473, b=0.580		
R^2=0.9729			R^2=0.9572			R^2=0.9817			R^2=0.9800		

T13			T14			T15			T16		
PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF
N	36	35.37	N	33	33.50	N	36	34.45	N	34	33.26
V	12	14.98	A	18	16.26	A	14	18.62	A	15	16.37
A	10	6.35	V	7	7.90	V	12	10.06	V	9	8.05
ADV	2	2.69	PTLA	3	3.83	PRNP	7	5.44	PRNP	5	3.96
PTL	2	1.14	PRNP	1	1.86	PRNA	4	2.94	PTL	3	1.95
			PRNA	1	0.90	ADV	2	1.59	ADV	1	0.47
			PTLP	1	0.44	PTLA	2	0.86	PTLA	1	0.23
a=83.473, b=0.859			a=68.997, b=0.723			a=68,745, b=0,615			a=67.608, b=0.709		
R^2=0.9694			R^2=0.9934			R^2=0,9623			R^2=0.9887		

T17			T18			T19			T20		
PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF	PoS	OBSF	EXPF
N	47	45.87	N	41	41.04	N	34	33.75	N	35	35.18
A	13	18.02	A	18	17.68	A	17	17.52	A	16	14.58
V	11	7.08	V	7	7.61	V	9	9.10	PRNP	3	6.04
PRNP	6	2.78	ADV	3	3.28	PTLA	4	4.72	PRNA	3	2.51
ADV	2	1.09	PTL	2	1.41	PRNA	4	2.45	V	3	1.04
PRNA	2	0.43	PRNP	1	0.61	PRNP	2	1.27	PTLA	2	0.43
PTL	1	0.17	PTLA	1	0.26	ADV	1	0.66	ADV	1	0.18
			PRNA	1	0.11						
a=116.719, b=0.934			a=95.270, b=0.842			a=65.002, b=0.655			a=84.879, b=0.881		
R^2=0.9645			R^2=0.9983			R^2=0.9954			R^2=0.9804		

Replacing the PoS of two classes by the name of the class to which they belong: S for the static description and D for the dynamic one, and omitting all other PoS, one obtains sequences which characterize the level of homogeneity of description.

Let us consider, as example, T3 again. After marking up the two above-mentioned classes we get the following sequence:

$$D-D-S-S-S-S-S-D-D-D-D-D-S-D-S-S-D-S-D-D-D,$$

This sequence consists of a number of strings formed by repeated elements which further on will be called "runs" [Andreev, Mistecky, Altmann, 2018, p. 50–52]. Here the following runs can be singled out:

$$[D-D]-[S-S-S-S-S]-[D-D-D-D-D]-[S]-[D]-[S-S]-[D]-[S]-[D-D-D].$$

Little number of runs, including big chains of similar members, indicates intensified monotony of description, big number of runs with few elements in them, on the contrary, suggests something like variability in depicting poetic world. In this example the total number of elements in all the runs equals 21, the number of runs equels 9. Thus it follows that the index of the homogeneity of description is $I_{total} = 21/9 = 2.33$. Measuring homogeneity of static and dynamic descriptions separately we get the following:

(1) static $I_c = 2,25(9/4)$;

(2) dynamic $I_v = 2,4(12/5)$.

Table 3: Indices of homogeneity description in all 20 sonnets

Text	Runs index I_c	Runs index I_v	I_{total}
T1	2.00	1.13	1.59
T2	1.71	2.17	1.92
T3	2.25	2.40	2.33
T4	3.75	1.80	2.67
T5	1.71	2.50	2.08
T6	2.50	1.33	1.92
T7	2.25	1.60	1.89
T8	2.25	1.75	2.00
T9	2.25	2.40	2.33
T10	1.38	1.67	1.53
T11	2.29	1.29	1.79
T12	3.33	1.71	2.46
T13	1.25	1.56	1.41
T14	3.67	1.33	2.50
T15	2.22	1.50	1.88
T16	1.45	1.20	1.33
T17	3.00	2.40	2.70
T18	3.33	1.50	2.42
T19	2.78	1.13	2.00
T20	10.50	1.50	6.00

Table 3 contains indices of homogeneity of different types of description in all 20 sonnets. With the exception of 20 in the samples homogeneity of description demonstrates a comparatively limited range

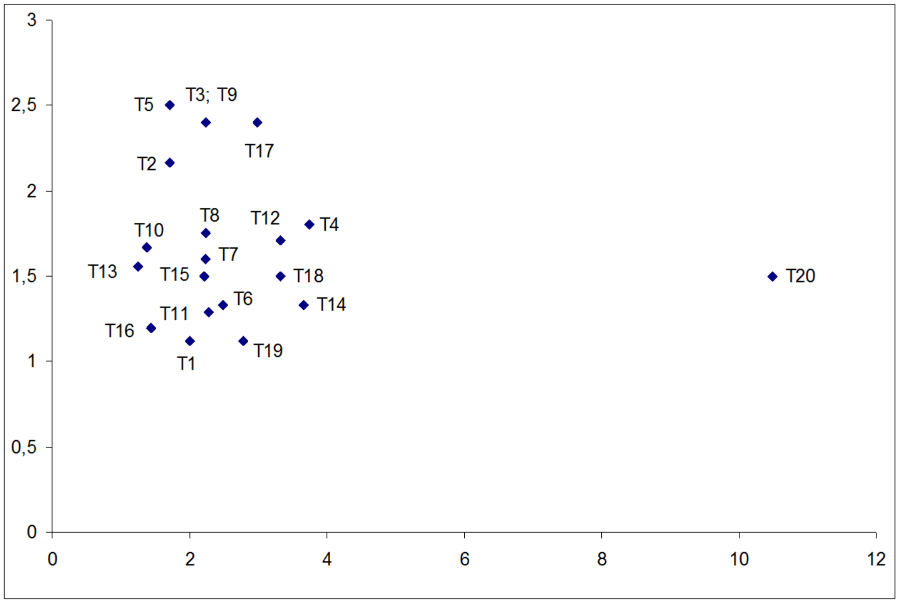

Figure 2: Scatterplot of texts arrangement based on indices I_c and I_v

of variability. It should be noted that sonnets are rather brief limited to 14 lines, and nevertheless in a number of cases the differences are apparent. This can be shown graphically. In Figure 2 the scatterplot demonstrates the relations of two indices—I_c and I_v in 20 sonnets. On the horizontal axis the values of I_c are set, the y-axis sets the values of I_v in the sonnets.

4 Conclusion & Discussion

The scatterplot shows that priority should be given to y-axis coordinate. X-axis does not provide a basis for classification, except for 1 outlier (T20) all the other texts form a rather dense group. On the other hand, index I_v, showing the dynamic homogeneity, divides the sonnets into 2 groups. The first one consists of 2, 3, 5, 9 17. It should be noted that 2 and 9 completely overlap and are marked by a common dot. All the other sonnets form another group. At the level of $I_v = 1.5$ it is split into two subgroups of equal number of texts. The homogeneity index of dynamics in description $I_v = 1.5$ is observed in three texts (15, 18, 20) thus forming a basis for the splitting of the whole group. Above and below this borderline there are 6 texts in each subgroup.

On the whole it is possible to conclude that this research demonstrated some order in all PoS distribution in sonnets and the relations between words which depict static and dynamic description of the poetic world.

References

[Best, 1994] Best K.-H. (1994). Word class frequencies in contemporary German short prose texts // Journal of Quantitative Linguistics. 1994. Vol. 1. P. 144–147.

[Stamou, 2008] Stamou C. (2008) Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating // Literary Linguistic Computing, 23(2). P 181–199.

[Gasparov, 2012] Gasparov M. L. (2012). Exact methods of grammar analysis in verse. In.: M.L Gasparov. Selected works. V. 4. Moscow: Languages of Slave culture, 2012. P. 23–35. (In Russian) = Tochniye metody analiza grammatiki v styhe // M.L. Gasparov. Izbranniye trudy. .4. 2012. S. 23–35.

[Čech, Altmann, 2013] Čech R., Altmann G. (2013) Descriptivity in Slovak lyrics. Glottotheory. 2013. Vol. 4 (1). P. 92–104.

[Andreev, Popescu, Altmann, 2017] Andreev, S., Popescu, I.-I., Altmann, G. (2017). Skinner's hypothesis applied to Russian adnominals. In: Glottometrics 36. RAM-Verlag. P. 32–69.

[Andreev, Mistecky, Altmann, 2018] Andreev S., Mistecky M., Altmann G (2018). Studies in quantitative linguistics - 29. Lûdenscheid: RAM-Verlag, 2018. – 130 p.

[Naumann, Popescu, Altmann, 2012] Naumann S., Popescu I.-I., Altmann G. (2012). Aspects of nominal style // Glottometrics. 2012. V. 23. P. 23–55.

[Martynenko, 2004] Martynenko G. Ya. (2004) Rhythmic-semantic dynamics of the Russian classical sonnet. Saint-Petersburg: SPb University, 2004. – 30 p. (In Russian) = Ritmiko-smyslovaya dynamika russkogo klassicheskogo soneta. SPb: SPbGU, 2004. – 30 s.