Statistical Taxonomization of Literary Corpus

Gregory Martynenko

St. Petersburg State University
Department of Mathematical, Computational and Applied Linguistics
Universitetskaya nab. 11, 199034, St. Petersburg, Russia
gymart@ts4306.spb.edu

Abstract. The paper proposes the research methodology aimed to study corpora of literary texts based on the synthesis of stylistic, semiotic, classification and statistical conceptions. The research has been carried out on the material of Computer Anthology of Russian Stories of the XX century, which is currently created at the Department of Mathematical Linguistics of St. Petersburg State University. The paper presents detailed illustrations of two typical tasks: 1) statistical taxonomization of fiction texts based on their syntactic similarity, and 2) reconstruction of semantics of the world-perception of writers by means of lexical investigations.

1 Systems Approach and Study of Fiction

Systems approach was announced already in 1929 to be one of the fundamentals of general poetics when studying large literary/fiction systems (Tynianov, Y.N. [1]). However, even nowadays the systems approach is not consistently observed both in studying the total creative heritage of some particular writer and in investigations aiming at research of highest possible comprehension of works, belonging to some concrete literary period. This paper launches an attempt to fill this deficiency at least partially and offers a methodology based on conjunction of system-statistical concepts with semiotic understanding (referential semantics viewpoint). Our main aim is to reconstruct individual and collective fiction "world-comprehension" characterizing writers of some particular literary epoch. This goal may be achieved by means of processing and analyzing the exterior and superficial text parameters related to its deep-laid semiotic characteristics on semantic and syntactic levels.

2 Research Material

The research has been carried out mainly on the material of Computer Anthology of Russian Stories of the XX century, which is currently created at the Department of Mathematical Linguistics of St. Petersburg State University. The selection of story as a literary genre for such kind of research is justified by the fact, that it is widely distributed in any literary epoch and allows to involve in comparative research more numerous quantities of authors. Moreover, the story represents the main object of the narrative studies, and in the process of literary evolution it is the story that fulfills the “prospecting” function: as compared with other prosaic genres (e.g. novel), it better responds to any shifts of the public mood and social values.

Computer Anthology of Russian Stories represents the full-text database, now consisting of about 2000 stories. The corpus is subdivided into synchronic groups (sub-Anthologies) according to traditional scheme of dividing Russian literature into periods. Sub-anthologies include selected works of the maximum number of writers, active in correspondent literary period. For most prominent writers (Chekhov, Bunin, Kuprin, Gorky, Sologub, Platonov, Bulgakov, Zoschenko and others) individual author’s anthologies are being made. The system of frequency dictionaries is further built for the whole Anthology, for concrete literary periods and for works of individual writers. Then each dictionary is exposed to structurization based on the system of parameters [2].

Examination of poetical texts was carried out on material of Computer Anthology of Russian Sonnets. Belonging to the poetic class of firm forms both in “plan of expression” and “plan of content”, sonnet is extremely interesting primarily because of its semantic concentration and may be rated as a peculiar semantic “bomb”. Besides sonnet we studied some other poetic genres, in particular – text corpus of several famous poets of the Silver Age.

3 Research Methodology

Research Methodology is based on the synthesis of stylistic [3], semiotic [4], theoretically classificational [5] and statistical conceptions [6]. Principles of linguistic data processing are determined by the following facts: 1) data ascription to the concrete linguistic level, 2) type of the reflected objects liable to classification, 3) genre of text, 4) current typology of fiction texts, 5) striving for minimization of diagnostic parameters, 6) necessity of taking into account for diagnostic tasks most general linguistic and semiotic mechanisms, and some other factors.

Let us consider two rather typical tasks.

4 Task I

We have the following input objects: 1) the totality of text works belonging to different writers of some particular literary period (in demo variant it consists of 100 stories), and 2) a number of diagnostic parameters, related to syntactic level. Though these parameters relate to surface structures, they all correlate to some extent with deep, cryptic parameters and even with conceptions of general type. For example, such outwardly “primitive” extensive parameters as “length of sentence”, “length of paragraph”, “extent of word-combination”, etc., which reflect the measure of text partition, correlate to cognitions of analytic and synthetic prose. Such parameter as comparative share of one or other notional word in the sentence correspond to dynamic and static style measure, and to degree of its discriptiveness or narrativeness. The measure of distance of syntactically related parts of a sentence in combination with variants of positional word class realization (e.g., in initial or final position of a sentence) and above mentioned extensive parameters point to ornamentalness or decorative neutrality, whereas parameters like the number of homogeneous groups in a sentence, the length of enumeration series and the height of hierarchic tree of dependences testify about the degree of associative and hierarchic coherency of text chains, etc. [6].

Our main hypothesis implies that despite of abundance of individual literary/fiction systems, each literary period is characterized by mutual "over-individual" syntactic organization, which is peculiar to the whole period. Complementary hypothesis implies that in frames of each period's integrity there exist some grouping of writers which are drawn towards one or another literary school, fiction trend or stylistic leader.

Semiotic meaning of this task signifies that each text taken separately and the totality of all texts are exponents of a complex sign, in which the world of heroes, conditions, ideas, believes and delusions, typical to the literary period in concern and therefore for each writer living in this period, serve the role of significant. Neither insignificant writer nor lofty genius can avoid all-penetrating influence of the "the spirit of the age" [7]. The role of referent of this complex text sign plays the world of structurized fiction meanings, correlating to the sense of syntactic variables, which form the area of parameters and their meanings. These deep implications of fiction are reconstructed with the aid of special procedure of multidimensional analysis, which was proposed by the author of this article [6]. The relations between these implications constitute *syntactics* of sygnificants. What concerns pragmatics of the given fiction period, it is being formed as a result of integrity interaction of historical, social and political, fiction and literary and some other actual conditions, typical for the given chronological cut. We may mention here: 1) the diverse literary associations (e.g., "Sreda" ("Wednesday") headed by N. Teleshov, "Bashnia" ("Tower") by Viacheslav Ivanov, Religiously Philosophic Society by D. Merezhkovsky), 2) printing editions (e.g., "Vesy" ("Libra"), "Golden Fleece", "Znanije" ("Knowledge"), "Shipovnik" ("Dogrose"), "World of Arts" and others), around which the groups of writers were formed, 3) literary manifestos of diverse literary schools belonging to different art trends (symbolists, acmeists, futurists, etc.), and 4) all the totality of visible and invisible contacts of all the writers of the concrete epoch and their groupings.

We shall not discuss here the structure of classificational algorithm (its detailed description may be found in [6]). I would like merely to present the final result of this algorithm functioning, depictured in the form of graph, in the knots of which we may see the names of writers, heading classificational groups, each of which is congeneric from stylistic point of view (see Figure 1).

The structure of one syntactic class (that of the writer P. Boborykin) is presented on the figure 2.

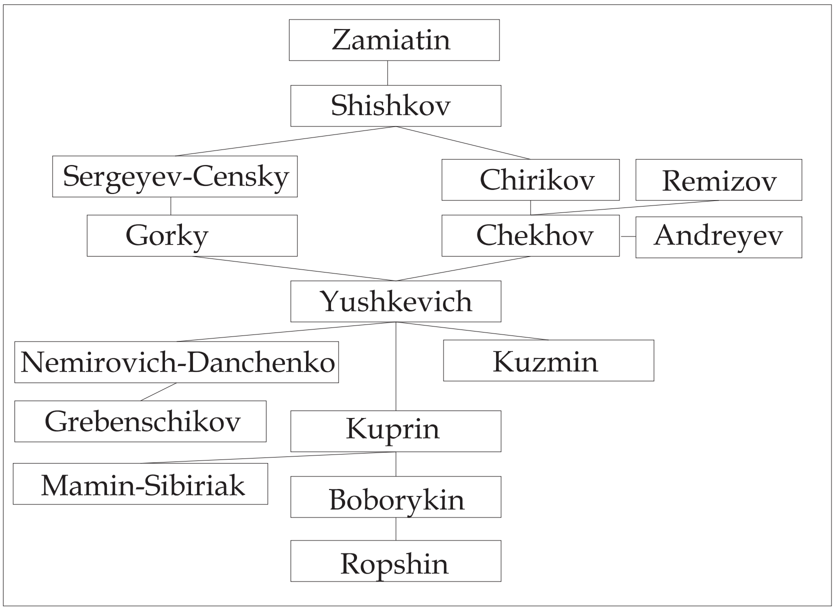

Fig. 1. System of syntactic classes of Russian prose writers in the beginning of XX century

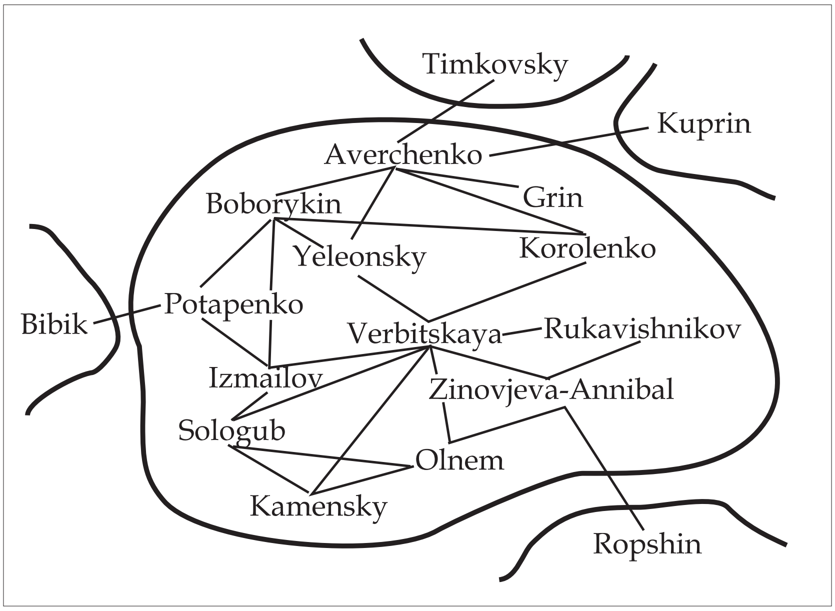

Fig. 2. Fiction class headed by writer P. Boborykin

Semiotic and statistic interpretation

The obtained fiction classification does not provoke any embarrassments. Thus, most "populated" is the class of S. Yushkevich. This class is characterized by the highly balanced syntax, without any stylistic extremes. Here we find writers, which were called by Y. Tynianov as "literary commonness". Group of A. Chehkov borders with this centre class. The writers of this group gravitate towards satiation of a sentence by conjunctive constructions with the minimum breaks between syntactically related components. This result does not seem to be an accidental one, and may be explained chiefly by the fact that prose of Chehkov has become the fiction standard, which had substantial influence on ways of forming of XX century literary schools [8]. Moreover, some most revealing syntactic "inventions" (e.g., over normative supercharging of conjunction groups with conjunction "*and*") had been adopted by many prose writers so directly that sometimes the pages of their works look like the exact copies of Chehkov ones. In the whole, texts written by prose writers of this class may be considered as belonging to associatively objectivistic type of prose.

Towards the end of given literary period Chekhov tradition was exposed to considerable transformations by writers of new generation, the style of which is characterized by partial replacement of smooth associative paratactic narration either by extreme syntactic primitivism (everyday life description with the use of brief, enumerating, short-distance phrases) or by "neurasthenic" syntax of the unusually short, highly parted and decorated prose (like in classes of E. Zamiatin, A. Remizov, V. Ropshin). In the works of these authors a threshold of the new post-revolutionary literary period was formed.

5 Task II

Our second task is connected with reconstruction of semantics of the world-perception of an individual writer or some group of writers by means of investigations on the lexical level. Let us consider systematization of verbal themes in the Russian poetry of the beginning of XX century. The main instrument for such kind of research is a number of frequency dictionaries, which are being structured by means of system of symptomatic parameters, mirroring lexical diversity of the dictionary, degree of its concentration and dispersion, possible variants of its stratification into separate zones, etc. Deep semantic structure of the world-perception may be presented in the form of literary themes, which may be further put in order according to degree of their activity in the works of one or several writers. Some of the literary themes constitute thematic core, which may be interpreted as a concept of poetical world-vision, whereas its covering along with the core itself may be interpreted like significant of this vision [9].

Let us demonstrate it on the example:

The important feature of Russian symbolists works (in particular in that of the older generation) was the theme of "endless, infinite, eternal, and perpetual" (ecstatic) intensity, excessiveness and extremeness of emotional experiences. This theme found its personification in correspondent text constructions, some of which had got broad propagation. Among such construction we can mention the so-called "endless" epithet — Russian adjective with the prefix "*bez-*" ("*bes-*") (similar to English postfix "*-less*"), depriving peculiarity of its quality distinctness.

Here is an example of the forcing usage of "endless" epithet:

V bezglazoj seroj mgle bezmernost', bezyzmernost',
Bezotnositel'nost', pustynia dnej bez vekh,
Beskrajnost' skatnaja, brodiachikh snov nevernost',
Otsutstvie putej, khotia b veduschikh v grekh.

K.Balmont, "Unsteadyness"

Along with word-formative model using "*bez-*", we find here the other variants of "deprivation": lexically semantic ("*otsutstvie*" (meaning "absence; lack") in the role of predicate, "*pustynia*" (desert) — object), one more word-formative variant using particle "*ne-*" (English "*in-*", "*un-*") — "*nevernost'*" (infidelity) and syntax variant with the preposition "*bez*" (without) — "*bez vekh*".

The totality of words with prefix "*bez-*" may be characterized from the outside as a number of realizations of one concrete stylistic model, and from the inside — from the point of view of composition and activity of micro-themes, peculiar to the world-perception of the concrete poet, literary school or literary trend.

We had made four samples, each containing approximately 300 words with the prefix "*bez-*", from rhymes and poems by F. Sologub, V. Briusov, K. Balmont, and A. Blok. Individual and general frequency dictionaries were built for these selections.

What concerns the quality composition of the words with prefix "*bez-*", in works of all authors we may distinguish two groups of themes. The first one includes lexemes, which characterize "*beskonechnost'*" (infinity, endlessness), "*bezdonnost'*" (fathomless), "bezpredel'nost"' (limitless) of the universe. In combination with words reproducing Tiutchev's idea of "*vsemirnoe molchanije*" (world-wide silence) — "*bezmolvnyj*" (speechless, mute), "*bezglasnyj*" (mute, dumb), "*bezzvuchnyj*" (soundless; noiseless), "*bezglagol'nyj*" ("soundless; silent"). At the same time outer world is being pictured like "*besposchadnyj*", "*bezzhalostnyj*" (pitiless, merciless), "*bezdushnyj*" (heartless, callous). Inner life of the poet is also portrayed in dismal mood: "*bezradostnyj*" (joyless, dismal, cheerless), "*beznadiozhnyj*" (hopeless, despairing), "*bezzhiznennyj*" (lifeless, spiritless) world; poet is "*bessil'nyj*" (powerless) and "*bezzaschtnyj*" (defenseless) in his "*beznadiozhnyj*" (hopeless), "*bescel'nyj*" (aimless; pointless), and "*bezvol'nyj*" (weak-willed) attempts to resist this hostile world, because of that he should be "*bespristrastnyj*" (impartial, unprejudiced) and "*bezuchastnyj*" (apathetic; indifferent).

The listed thematic constituents are presented in works of four poets in different proportions. F. Sologub and K. Balmont occupy here diametrically opposed positions: somber and rigorous poetry of Sologub stands against light and "feminine" that of Balmont.

Evidently, the proposed interpretation cannot be considered to be accomplished, because sphere of analyzing stylistic techniques should be considerably wider. Thus, examination of one or two word-forming modes in addition to described above "endless" epithet noticeably enriches semantic interpretation. The following example analyses the list of micro-themes of K. Balmont's world-perception, obtained in the result of semantic and statistical investigation of abstractly characteristic nouns with postfix "*-ost'*" (similar to English "*-ness*"), which were found in his verse. Herebelow we present concise list of these micro-themes:

– Harmonious "*celostnost'*" (integrity), "*strojnost'*" (orderliness, harmony), "*zakonchennost'*" (completeness, perfection) and "*mudrost'*" (wisdom) of the universe;
– Its "*pyshnost'*" (splendour), "*nariadnost'*" (smartness; elegance), "*jarkost'*" (brilliance) and "*radostnost'*" (gladness, joyfulness);
– Infinite "*mnogorazlichnost'*" (multideversity) of existence forms of the objective reality, its "*slozhnost'*" (complexity), "*zaputannost'*" (intricacy), "*mnogostoronnost'*" (variety, multiaspectiveness) and because of that its "*zagadochnost'*" (inscrutability, mystique), "*tainstvennost'*" (mystery), "*volshebnost'*" (magic, witchery) and "*nepostizhimost'*" (incomprehensibility, inscrutability);
– "*kontrastnost*"' (contrastness), "*poliarnost'*" (polarity), "*razdel'nost'*" (separateness) of the world phenomena with their simultaneous "*slitnost'*" (unity), "*jedinost'*" (uniformity), "*cel'nost'*" (wholeness) ("general uniformity in multidiversities" and "multidiversity in general uniformity");
– "*nechetkost'*" (illegibility), "*tekuchest'*" (fluctuation, instability), "*perelivchatost'*" (iridescence) and "*mnogozyblemost'*" (vast instability and vagueness) of borders between contrast and even opposite phenomena, their mutual "*dopolnitel'nost'*" (subsidiariness) and "*zerkal'nost'*" (mirrorness). At the same time "*slijannost'*" (confluence) dominates over "*razlichnost'*" (distinction) ("*chetvertichno razlichnoje — stokratno jedinoje*");
– "*krugovratnost'*" (the continuous round of events), "*povtornost'*" (repeatedness), "*mnogoizmenchivost'*" (extreme changeability and unsteadiness) of the objective reality, its "*mernost'*" (regularity), "*razmernost'*" (dimensionness), "*ritmichnost'*" (rhythm), "*volnistost'*" (corrugation, waviness), "*zmejnost'*" ("snakeness"), "*pevuchest'*" (melodiousness) and even "*svirel'nost'*" ("pipeness");
– "*zvezdnost'*" ("starness"), "*vznesennost'*" ("ascensionness"), "*vozvyshennost'*" (loftiness, elevation), "*vershinnost'*" (apicalness) of poet's attitude toward reality;
– his "*nenasytnost'*" (insatiability), "*alchnost'*" (greediness), "*isstuplionnost'*" (ecstasy, franticness), "*derzost'*" (daring, audacity) right up to "*raznuzdannost'*" (unruliness, licentiousness) and "*vnestydnost'*" (shamelessness) in "*beskonechnost' stremlenija*" (infinity of aspiration) for "*beskonechnost' dostizhenija*" (infinity of achievements).

Quantitative analysis of stylometric techniques may help not only to interpret poetic world-perception (both individual and general), but the degree of its correspondence to the system of scientific conceptions of the period in concern and the following epochs. Many micro-themes of the Russian poetry of the beginning of XX century anticipated the serial of scientific discoveries and hypothesizes: principle of complementarity by N. Bohr, uncertainty principle by V.Geizenberg, principle of the unity of diversity and diversity of unity by L.Von Bertalanffy, hypothesis on rhythm of substance and consciousness dialectics by P.K. Anokhin, etc. This fact ones more confirms justice of the subtle remark made by Russian linguist A.A. Potebnia: "In some, even not accessory, but rather essential aspects, the artist comprehends the nature better than botanist, zoologist or anatomist" [10].

References

1. Tynianov, Y.N.: Archaisty i novatory (Archaists and Innovators). Leningrad. Priboj (1929)
2. Martynenko, G.Y., Sherstinova, T.Y.: Statistical Parametrization of Text Corpora. Proceedings of the Workshop on Text, Speech and Dialog (TSD'2000). September 13-16, 2000. Brno, Czech (2000)
3. Arutiunova, N.D.: O sintaksicheskikh tipax khudozhestvennoj prosy (About Syntax Types of Russian Prose). In: Obscheje I romanskoje jazykoznanije (General and Romance Linguistics). Moscow: Nauka. (1977) 188–199
4. Chebanov, S.V., Martynenko, G.Y.: Semiotika opisatel'nyx tekstov (typologicheskij aspect) (Semiotics of the Descriptive Texts (Typology Aspect)). St. Petersburg: Edition of St. Petersburg State University (1999)
5. Chebanov, S.V.: Logicheskije osnovanija lingvisticheskoj tipologii (prilozhenije teorii klassifikacii v linguistike). Logical Fundamentals of Linguistic Typology (Linguistic Application of Theory of Classification). Vilnius: VLANI (1996)
6. Martynenko, G.Y.: Osnovy stilemetrii (Fundamentals of Stylometrics). VLeningrad: Edition of Leningrad State University (1988)
7. Balmont, K.D.: Predislovije k polnomu sobraniju sochinenij Shelly (Preface to Complete Set of Works by Shelly). In: Shelly. Complete Set of Works in translation by K.D. Bal'mont. St. Petersburg: Edition of Association "Znanije" (1904)
8. Usmanov, L.D.: Khudozhestvennyje iskanija v russkoj proze konca XIX veka (Artistic Seekings in the Russian Prose of the End of XIX Century). Tashkent: FAN. (1975)
9. Zhirmunsky, V.M.: Valerij Briusov i nasledije Pushkina. Opyt sravnitel'no-sopostavitel'nogo issledovanija (Valerij Briusov and Pushkin's Heritage. Essay on Comparative Research). In: Zhirmunsky V.M. Teorija literatury. Poetika. Stilistika (Theory of Literature. Poetics. Stylistics). Leningrad: Nauka (1977) 15–56.
10. Potebnia, A.A.: Estetika i poetika (Aesthetics and Poetics). Moscow: Iskusstvo (1976)