NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Vera G. Sibirtseva

COMPUTER-BASED PROCESSING OF LITERARY WORKS AND STUDY OF LITERATURE

BASIC RESEARCH PROGRAM

WORKING PAPERS

SERIES: LINGUISTICS

WP BRP 07/LNG/2014

Vera G. Sibirtseva[1]

COMPUTER-BASED PROCESSING OF LITERARY WORKS AND STUDY OF LITERATURE[2]

Currently many software applications that enable text analysis are being created for different purposes (semantic reference tools, concordancers, sentiment analysis etc.), but are not being used by literary researchers. Computer software allows to facilitate the search of required information and to considerably save time. With such an approach to the field of linguistic and literary analysis – comparative analysis in particular – new opportunities and unexpected horizons are being offered.

This paper suggests a critical review of existing computer resources related to text processing and a consistent description of program applications, successfully tested on literary materials and used for text analysis at the Faculty of Humanities (HSE Branch in Nizhny Novgorod): linguistic annotated text corpora; collections of literary texts of one author; different computer tools such as AntConc concordancer, multifunctional text analyzer LEKTA, LF aligner for text alignment – those tools which allow a variety of loaded and analyzed text collections. Computer-based text analysis shall be practiced only with further literary description and interpretation. This comparison of data retrieved in the process of computer-based analysis with existing traditional researches may mark the dawn of a new stage of literary text analysis.

JEL Classification: Z19

Keywords: text analysis, linguistic corpora, concordancer, literary works, translation.

[1] National Research University Higher School of Economics. Department of Applied Linguistics and Intercultural Communication, associate professor; E-mail: vsibirtseva@hse.ru

[2] The author is grateful to Anna Odintsova, Lubov Violentova, Ekaterina Fomina, and Marina Tsvetkova (Higher School of Economics) for providing papers and data

Everything should be as simple as it is, but not simpler.

A.Einstein

Introduction

Literary studies related to text mining and interpretation are traditionally viewed as a field implying purely intuitive, creative, "human" approaches. Currently many software applications that enable text analysis are being created for different purposes (authorship verification, semantic reference tools, concordancers, sentiment analysis, etc.). However, the conservatism of research in literary studies remains significant. Software created for humanities, such as sociology, psychology, and management, are rarely applied to text mining.

The idea of "measuring harmony by arithmetic" certainly shall not be taken literally; a computer cannot replace a professional researcher. But in the process of text analysis there are always tasks related to the examination of a big volume of material or to a routine search of homogeneous information. In such cases computer software facilitates the search for the required information and to considerably save time. With such an approach to the field of linguistic and literary analysis, and comparative analysis in particular, new opportunities and unexpected horizons are being offered.

Here are the main reasons for the non-use of the existing tools in the field of literary studies: many software products lack a manual optimized for a user who is not familiar with software engineering; automatic text processing tools are unstructured from the point of view point of a literary researcher; the number of software application examples in the analysis of literary works is tiny; potential analysis algorithms for an automatically processed text are absent.

The present paper suggest a critical review of existing computer resources related to text processing, and a consistent description of program applications, successfully tested on literary materials and used for text analysis at the Faculty of Humanities (HSE Branch in Nizhny Novgorod).

Program resources that are suitable for literary texts represent two large groups: linguistic test corpora and supplementary computer software, which can be used in the process of literature analysis.

National linguistic corpora

Linguistic annotated text corpora became widely used in the field of text analysis. Work with the corpora is one of the leading linguistic research methods whereby many different problems can be solved. The two major projects based on the Russian National Corpus (RNC) are worth mentioning: "Problems of Russian Stylistics", under the supervision of A.I. Levinzon and Y.M. Kuvshinskaya, and "A Willful Meaning: An Attempt of Microhistorical Study of 19th-21st-century Lexis", supervised by M.A. Daniel and N.R. Dobrushina. The goal of the "Willful Meaning" project, implemented in 2010-2012, was to study and describe the changes in meaning and use of 30 Russian lexemes throughout the 19th-21st centuries. Eventually a series of popular-scientific features, each of which was dedicated to the meaning history of one specific word (obshchezhitie, klassnyĭ, ruchka, eres', rasstroennyĭ, etc.), was created. The continuing "Problems of Russian Stylistics" project is a number of mini-studies completed by different authors, where each study is dedicated to a certain problem from the field of stylistic variation of the contemporary Russian discourse.

Large perspectives for the text study were offered when analyzing parallel linguistic corpora. In the field of comparative studies, parallel linguistic corpora represent a peculiar kind of comparative text bank. They allow for studying the adequacy of language material conveyance and styles of the writer and the translator. Parallel text fragments are arranged in the same sequence in both parts of the corpus (original text and translation), thus facilitating the search of the reference data: dates, proper names, inquit markers, etc. Linguistic lacunae – translation gaps – may also provide interesting material for analysis. Detecting lacunae in different translations can help to evaluate the quality of the translation. The parallel RNC subcorpus allows for using precise quantitative data on the frequency of different grammatical structures, as well as considering synonymic vocabulary in the context:

ru-de

- *Prileteli ptichki,—zlobno progovoril Artem.—Ėkh, i kuter'ma nachnetsia, ediat ego mukhi!—I voshel v dom. [N. A. Ostrovskiĭ. Kak zakalialas' stal' (ch. 1) (1930-1934)]*

"Ja, ja, jetzt fliegen die Vögelchen wieder ein", flüsterte Artjom erbittert, "das kann jetzt heiter hergehen, hol's der Teufel!" Und er ging ins Haus. [Nikolai Ostrowski. Wie der Stahl gehärtet wurde (erster Teil) (1936-1977)]

de-ru

Der Rote! keuchte sie. Sie wog zweihundertvierzig Pfund und zitterte vor Lachen wie ein Berg von Gelee im Erdbeben. [Erich Maria Remarque. Der schwarze Obelisk (1956)]

- A, ryzhiĭ! — progovorila ona, zadykhaias'. Khoziaĭka vesila sto dvadtsat' kilo, i vse ee telo khodilo khodunom ot khokhota, slovno gora zhele vo vremia zemletriaseniia. [Ėrikh Mariia Remark. Chërnyĭ obelisk (V. Stanevich, 1961)]

de-ru

"Dann weinte sie plötzlich wieder und sagte:" Ich kann mich hier nicht mehr blicken lassen. [Heinrich Böll. Ansichten eines Clowns (1963)]

No potom vdrug opiat' zaplakala i progovorila:—Kak ia teper' pokazhus' liudiam na glaza? [Genrikh Bëll'. Glazami klouna (R. Raĭt-Kovaleva, 1964)]

A study of contextual translations of the Russian verb "*progovorit*'" (say, utter) shows the difference in perception of this verb in the German and Russian languages. A frequency analysis of performative verbs on the basis of the parallel subcorpus allows for comparing their usage in different languages as well.

One of the possible text analysis types is a mini-research: a search for the word *"nakonets"* as parenthesis or an adverbial, for example. Punctuation marking of parentheses in the Russian language is different from the one in English, German, or Polish, which is why this type of work does not only enriche the vocabulary, but also develops contextual and punctuation literacy:

ru-pl

Kogda zhe slukhi ob ėtom doshli, nakonets, i do muzha, Anna, ne zadumyvaias', pokinula ego dom i pereekhala k svoemu liubovniku. [V. IA. Briusov. Cherez piatnadtsat' let (1909)].

Kiedy zaś pogłoski o tym dotarły wreszcie i do męża, Anna, niewiele myśląc, porzuciła jego dom i przeniosła się do swego kochanka. [Walery Briusow. Po piętnastu latach].

pl-ru

Udawałam szalone nim zainteresowanie wyłącznie w tym celu, żeby mąż się wreszcie zaniepokoił. [Joanna Chmielewska. Wszyscy jesteśmy podejrzani (1966)].

IA pritvorialas' bezumno zainteresovannoĭ im s edinstvennoĭ tsel'iu – chtoby muzh nakonets zabespokoilsia. [Ioanna Khmelevskaia. My vse pod podozreniem].

Due to the context, the corpora of parallel texts encourage more precise definition of a word and choice of a synonym out of range. The found concordances extend the notion of lexical system arrangement of the compared languages, of the context influence on the choice of this or that word, and build up more certain ideas of the language as a whole and of the language of literary texts in particular.

As we can see, when addressing to parallel corpora, an opportunity of literary studies of not only polysemy and synonymy in different languages, but also a stylistic exact transfer of connotations in texts of different cultures becomes available.

Systematized collections of literary texts

The national linguistic corpora is a multidimensional global resource for analyzing language events in the art of declamation at large. At that, an analysis of works of only one author on the basis of national corpora is unpractical just because the representativeness of these resources hinders the deep study of authorship works.

Along with the national corpora, there are resources containing collections of literary texts of one author, or different translation variants of one text, systematized and annotated in accordance with particular goals. Concordance vocabulary of journalism of F.M. Dostoevsky, parallel corpus of "The Tale of Igor's Campaign" translations, and the Russian-French poetry corpus of the first third of the 21st century may be referred to as examples.

The concordance vocabulary of journalism of F.M. Dostoevsky enables one to receive a use of context and frequency characteristics of any word or word form, as well as their position in the text in the complete set of writer's woks in 30 volumes. This search is possible through all the journalistic texts of Dostoevsky, including his Diary. Materials containing the concordance dictionary, together with the dictionary of the writer's literary texts, created by the Russian Language Institute of the Russian Academy of Sciences, give a complete idea of frequency characteristics of the main corpus of Dostoevsky's texts

The Russian-French poetry corpus of the first third of the 21st century is created for the purpose of different studies related to detecting a specific character of translations of certain authors, and to the study of the specific character of their poetical and translation idiolect.

A parallel corpus of translations of "The Tale of Igor's Campaign" is another resource for translation comparison. The capacity of this translation corpus of one specific text is quite big: the user can compare the existing translations, analyze language events in concurrent versions, and receive a complete picture of the perception of "The Tale of Igor's Campaign" in the contemporary world.

Since these resources are dedicated to literary texts, it would be logical to assume that their application for the purpose of literary analysis should become a frequent practice. In particular, the perspective of studies of an author's individual style, or diverse prosodic study, seem interesting: the study of rhyme, stylistic clichés, and so forth. All the above-mentioned

programs have a user-friendly interface, which is quite important for a literary researcher. But use of these resources in linguistic and literary studies is considerably limited: all of them are dedicated to non-flexible literature segments. By means of the Dostoevsky corpus, one cannot interpret Gogol and Bulgakov, the same as a collection of translations of "The Tale of Igor's Campaign" will not help to study the translations of "Alice In Wonderland". But the perspective and relevance of such collections is obvious – literary researchers receive everything necessary for studying literary works with a complete toolkit: a system of search and citation, bibliographies, dictionaries, and thesaurus.

Natural language processing tools

The natural language processing tool is a special product that to a certain extent can be used in analyzing literary texts. Some research literature contains a consolidated description of tools, enabling them to perform linguistic text processing. All existing tools and programs can be divided into several groups:

1. Programs for morphological and syntactical text analysis.
2. Programs for statistical text analysis.
3. Linguistic technologies and systems.

Some of these tools may be interesting for literary researchers.

The TextAnalyst SDK contains programmatic components for realizing a complex of functions of automatic analysis for Russian and English texts. The capabilities of the library will help to you to abstract the texts, form hypertext databases, execute information retrieval in the texts, and so forth (http://www.analyst.ru/index.php?lang=eng&dir= content/products/). Interesting functions of this resource are described in the annotation to the program:

- content text analysis with automatic build-up of a subject tree with hyperlinks – identification of semantic structure of the text in the form of hierarchy of topics and subtopics;

- semantic search considering concealed semantic links between the requested words and text words;

- automatic text abstraction – build-up of its semantic portrait in the terms of the most informative phrases.

Russian Context Optimizer (http://www.rco.ru). RCO demonstrates another approach to the text. It solves such problems as creation of a substantial text portrait, extraction of named objects, links and facts, and analysis of the text tone. The above-mentioned resources are used in

humanities, especially in sociology. They are commercial, which limits their experimental using in the field of literary text analysis.

Rhymes is a program that easily picks up a rhyme, synonym, or an epithet, shows the word's meaning, its pronunciation, and examples of its usage. It is a universal dictionary tool that will be helpful to anyone who writes texts in Russian. (http://rifmovnik.ru/).

However, the authors of the mentioned reviews do not go beyond a short summary of program functions and do not demonstrate the capacity of it. All the tools to a greater or lesser degree ascend from a list of tools in the Russian Virtual Library, which has not been updated since 2006. As a result, many links indicated in the table do not work. A number of the available programs became a part of paid content.

It is worth mentioning that within the context of the present work, programs for linguistic text analysis are not the same as programs that help to make a linguistic or literary analysis of a literary text. Thus, this paper leaves behind such software products as tools for phonetic and phonologic text analysis, dictionary builders, tools of automatic indexing, and terminology extraction.

As we can see, the application of software programs for analysis of literary texts is hindered by the following factors: unreliable information on the operating status of one text analysis resource or another; the necessity to buy a product; and the absence of a comprehensible non-professional user manual.

Based on the example of student and professor research papers, we will demonstrate the possibilities to analyze and then interpret literary texts by means of modern concordancers, content analyzers, and other programs. Let us consider such programs as AntConc concordance, multifunctional text analyzer LEKTA, LF aligner for text alignment – in other words, those programs that allow a varying of loaded and analyzed text collections, and that are almost never used by literary researchers.

LF Aligner

In the process of studying different translations of one and the same text, one can track the stylistic changes in the vocabulary and detect peculiarities of translation strategy. An electronic text collection is made for such purposes. Literary texts as a rule contain a large number of gaps and cases of asymmetrical translation, thus giving abundant material for analysis

and interpretation. Besides the non-observance of authorial paragraphing and sentence breakdown, in different languages (even in chronologically different editions of one text) various means of text visualization – such as graphic design of direct speech and dialogues – may be used.

A free program, LF Aligner can be used to study non-occurrence in translations and corresponding distortions of meaning (due to different reasons). It allows for comparing several translations done by different authors, as well as translations in different directions (from and to the language). For educational purposes, a collection of parallel translations of "All Red" by I. Khmelevskaya was created.

LF Aligner, used for aligning parallel texts (which, in fact, relates to supplementary tools of translation) and statistic calculation in linguistic studies, can be used as a supplementary tool for the analysis of translation due to several reasons. First of all, visual representations of aligned texts are convenient for analysis. Secondly, the program capacity is not limited to two texts; four or more texts in different languages can be added if necessary. And finally, the program interface is philologist-friendly, simple commands do not require specific programming skills, and the whole process of text processing takes just several minutes. For faultless operation of LF Aligner, only the Latin alphabet should be used both in the storage locations and in the names of the files themselves. The result is convenient for visual evaluation of alignment quality, since it presents texts in the form of a table consisting of several columns: with original sentences, three translation variants, and a column containing the metadata of text sources. LF Aligner offers different types of text breakdown (two types of alignment): with sentence segmentation or without it. In a separate window the program shows information on the number of segments in the original text and in its translations. A revision and correction of segmented text is done by means of a built-in editor.

In the process of program mastering, several translations of fragments of "All Red" by I. Khmelevskaya were compared: those by V. Selivanova, M. Krongauz, and O. Kuznetsova (for the course of foreign language study according to the method of Ilya Frank). The analyzed fragments include the one at the beginning where the name of Allerod town is being discussed, and the dialogues with the participation of police officer Mister Muldgaard.

The first fragment illustrates the choice of different translation strategies. The translation of M. Krongauz is the most neat one, different details are omitted that does not impact the overall plot development, but also does not convey the original resemblance to girl talk and chatter about everything. The translation of O. Kuznetsova, made for reading in accordance with the method of Ilya Frank, is the most accurate one. In this case the translator deliberately chose

to closely follow the letter of the original, and with this the ironic style of Ioanna Khmelevskaya was lost. The translation of V. Selivanova is recognized as the best one: irony and humor, which boost the author's love for the details, were successfully conveyed by the interpreter.

The fragments of dialogues with participation of Mister Muldgaard are interesting for the ways of conveying a particular Polish language of the Danish policeman. In these dialogues, grammatical meanings of the Polish language not existing in Russian are the most challenging to convey. The example hereafter shows the conveyance of a specific plural form in the Polish language, which is used to denote plurality, excluding males:

— Co robiły one? — spytał pan Muldgaard. — Dlaczego tylko my? — zaprotestowała Zosia z oburzeniem i pretensją, w przekonaniu, iż pytanie odnosi się wyłącznie do kobiet.

— Vo onozhe vremia, - sformuliroval sledovatel' svoĭ sleduiushchiĭ vopros, - shto ona tvorikhu? - Kto "ona"? I pochemu on sprashivaet tol'ko o zhenshchinakh? – vozmutilas' Zosia [V.Selivanova].

— Kto i chto delali osoby? - Pochemu tol'ko my? - voznegodovala Zosia, schitaia, chto vopros otnositsia k zhenshchinam [M.Krongauz].

In the study translation of Olga Kuznetsova, clarifications are required:

— Chto delali oni (one - oni, po otnosheniiu tol'ko k litsam zhenskogo pola)? - sprosil pan Mul'gor. - Pochemu tol'ko my? - zaprotestovala Zosia s vozmushcheniem i obidoĭ, ubezhdennaia, chto vopros otnositsia tol'ko k zhenshchinam.

Certainly the capacity of LF Aligner is much wider than the visualization of the compared fragments, but demonstrativeness in the process of comparison can considerably facilitate the perception of several texts and save time on searching for required fragments in each text taken by itself.

LEKTA

The LEKTA content analyzer, developed at Nizhny Novgorod State University, was also used in studies dealing with the analysis of literary texts with the application of computer means. Content analysis is a method of detection and evaluation of specific characteristics of texts and other information carriers (videotapes, TV and radio programs, interviews, answers to open questions, etc.), which implies assigning certain semantic units of content and form of information in accordance with the goals of the research. This method is widely used in sociology, consisting in the "cross" processing of texts with further interpretation of results, and allows for identifying concealed informational content and form.

Content analysis can be used in linguistics and literary studies. The practicability of this approach lies in the fact that "content analysis uses purely linguistic information on text characteristics, and tries to identify its semantic peculiarities. The essence of content analysis is – on the basis of the external (quantitative) characteristics of the text at the level of words and word combinations – to make a credible conjecture about its content plane, and, as a consequence, to draw conclusions about cognition peculiarities of the author – his intentions, attitudes, wishes, value system, etc." Usage of content analysis in text interpretation is limited: text analysis is formalized to a great extent, quantitative registration of language units is done, giving basis to drawing assumptions on correspondence of a text fragment to the semantics of units included in it. Texts with a large number of metaphors, implications, language games, and a sophisticated plot cannot be credibly interpreted by means of content analysis. Factual and folklore prose, as well as authorial fairytales, are the most suitable.

It should not go unmentioned that fact analysis is considerably difficult for an unprepared researcher-philologist. He has to learn how to work in LEKTA by himself or with the help of a specialist. Studies mentioned in this paper were done by master's students of the Computer Linguistics Branch after being lectured on operating a program for content analysis.

Multiple translations of "How the Leopard Got His Spots" by R. Kipling were chosen for a literary text study by means of LEKTA.

Kipling's fairytales are notable for the simplicity of style and language, an expression plane directly corresponds to the content plane, and the linguistic characteristics of the text show its high affinity to news items and informant answers in different surveys.

Such literary texts allow successful application of content analysis using a comparison of terms or actions in the text. In Kipling's case, these are the main moments of the plot and characteristics of the heroes. LEKTA is used to find key words and phrases that form cross points of text meanings. The final goals of content analysis predetermine emphasis of the comparative aspect, since with its help text communities and differences can be clearly demonstrated. Factor analysis (a comparison of collocated words) helps to draw more reasonable conclusions about semantic aspect of the text that are both surface and implied.

The keywords in translations of "How the Leopard Got His Spots" are significant: peculiarities of fairytale topography (High Veld) and color of the leopard's skin; color detail of the landscape (prevailing yellow and brown color); the words of Pavian, etc.

In the result of factor analysis of fairytale translations, 20 object-factors (key episodes) connected to the plot structure of the original text were derived. On their basis, assumptions

about text translation strategy (or retelling) choice, and key microplots of the fairytale were characterized.

The result of the comparison is not only a full analysis of advantages and disadvantages of the 11 known translations of Kipling's "How the Leopard Got His Spots", but also a compilation of the hypothetically most complete translation, which includes different fragments of authorial variants. As the experiment of the literary text shows, content analysis enabled identifying of a concealed sense of the text and its translations, and full semantic analysis of the fairytale.

In such a manner, the program for factor analysis was used for the purpose of comparative literary studies. Such an analysis is rational for cycles of short texts, which are not characteristic of semantic polysemy. Studies of computer-based finding of semantic connections in folklore and authorial fairytales in this case can bring interesting results.

AntConc in researches of texts in Russian

AntConc concordancer is easy to operate and requires minimal training for text analysis. AntConc is a free program, which does not require PC installation and helps to perform the easiest types of lexical analysis of the user's text collections. The ability to process common world languages without necessity to type in Latin is an undeniable advantage of this program. AntConc is not able to combine different morphologic word forms into one lexeme (as it was in LEKTA), but this problem can be avoided if combination rules for different lexemes by means of regular expressions are used.

The AntConc tool set is quite varied, but only some of the tools are used in the process of work with a literary text: the program allows deriving a concordance of key words in context (KWIC concordance); choose key words in the text; to derive frequency, alphabet-frequency, and invert-frequency text dictionaries; and make lists of collocations with the required word.

The fantastic tales of V. Krapivin were analyzed by means of a concordance program. Specificlly, four texts were used for analyzing Krapivin's tales: "A Crany and Lightings", "Summer Will Not End Soon", "Grandmother's Grandchild and His Brothers", and "A Plane Named Serezhka". The object of study was the most frequently encountered character of the author, the so-called "Krapivin's boy" (Razumikhin, 1982). An attempt to identify peculiarities of language behavior, character, and appearance of Krapivin's heroes on the basis of co-occurrence of words was completed.

First of all, an analysis by words "*mal'chik*", "*mal'chishka*" (a boy) and by names of the main characters was done. By means of frequency analysis and collocation lists, it was determined that in texts of this author the following word combinations are frequently met: "*veselo skazal mal'chik*" (gaily said the boy), "*mal'chik ulybnulsia*" (the boy smiled). Krapivin's books are mostly targeted at teenagers, and it's no wonder that the name of the main character often stands together with such words as "*obradovalsia*" (brighten up), "*usmekhnulsia*" (smirk), "*zasmeyalsia*" (laughed), *"veselo*" (gaily), "*ulybnulsia*" (smiled) – no matter what problems and unpleasant situations happen to the main character, the author idealizes the world, depicting the boy as joyous and happy.

Texts also contain a lot of descriptions of the character's appearance: "*kudriavyĭ mal'chik*", "*simpatichnyĭ mal'chik, ryzhen'kiĭ*", "*ryzhevatyĭ mal'chik*", "*dlinnovolosyĭ mal'chishka*", "*belobrysyĭ mal'chishka*", "*vesnushchatyĭ mal'chishka*", "*mal'chishka v sineĭ rubashke*", "*mal'chishka v shirokoĭ kepke*", "*zheltoglazyĭ mal'chishka*". Many words surrounding the name of the main hero prove the hypothesis about the character of a typical Krapivin's hero.

Characters of the above-mentioned tales were also considered separately from each other, particularly Zhurka, the hero of "A Crany and Lightings". Analysis done by means of AntConc showed that the most frequent collocations are word combinations, consisting of a noun (hero's name) and a verb (*skazal, podumal, sprosil, udivilsia*, etc.). Word combinations with adverbs (*rasterianno, mashinal'no, medlenno, okhotno*, etc.) also exist. Many word combinations prove the hypothesis about the character of a typical Krapivin hero.

The performed analysis confirmed characteristics of Krapivin's heroes, outlined in the critical studies of the author's texts. The main traits of "Krapivin's boy" were confirmed by quantitative indicators obtained in the result of literary text analysis done by the AntConc program.

AntConc in researches of texts in English

A frequency analysis of words in a text can be done not only for the Russian language. AntConc was used for identification of verbal representatives of key concepts in T. Dreiser's "Trilogy of Desire" ("The Financier", "The Titan", "The Stoic"). In particular, the correctness of evaluations of Frank Cowperwood's character in Russian and foreign critical studies was verified. In the result, with help from statistical analysis of the writer's language, mental concepts important for his literary world that did not previously attracted attention of researches

were detected. Besides that, a restructuring of existing concepts identified by literature theorists was implemented.

A controversy of Frank Cowperwood's character and Dreiser's attitude towards him were clearly confirmed by a statistical analysis of key words. A treatment of results always contains an element of subjectivity. That is why key words were selected in accordance with strict criteria, mainly consisting of frequency of words from one semantic field. The result was extraction of four global concepts of the literary world of "The Financier": "Time", "Financier", "Winner", and "Senses".

The concept of "Time" appeared to be the most meaningful for the novel's literary world, since the word "time" itself is the most frequently encountered word in the text (479 citations). Besides this, many other verbal representatives of this concept were detected in the text: *now* (610), *once* (195), *never* (192), *always* (92), *today* (34), *tomorrow* (17), *day* (196), *years* (152), *moment* (89), *times* (71), *already* (70), *finally* (70), *soon* (60), *morning* (59), etc.

Cowperwood strives to receive as much as possible here and now, which is proven by the statistics, where the adverb "*now*" is the most frequent one (610 citations):

"*What am I bid for this exceptional lot of Java coffee, twenty-two bags all told, which is now selling in the market for seven dollars and thirty-two cents a bag wholesale*?" [Dreiser, 9], "*I am going to offer you now a fine lot of Castile soap-seven cases, no less-which, as you know, if you know anything about soap, is now selling at fourteen cents a bar*" [Dreiser, 11], "*I want to pay for that soap," he suggested. Now? Yes. Will you give me a receipt?*" [Dreiser, 12], "*Now, Frank, if you're ready for it, I think I know where there's a good opening for you*" [Dreiser, 17].

It is also worth mentioning that the adverb "*now*" (610) is mentioned several times more frequent than "*never*" (192).

The main character is a spectacular example of a true winner. This person is self-confident and he is steady in his purpose. No wonder that the modal verb "*can*" is statistically encountered in the novel 486 times: "*And, meantime, keep your health and learn all you can*" [Dreiser, 9], "*He's already offered me sixty-two for it. I can get it for thirty-two*" [Dreiser, 12], "*Yes, I can*" [Dreiser, 15]. A widespread concept of the American nation's "can-do attitude" finds expression here. Moreover, Frank's motto that his "wishes are first of all" also confirms his nature of a winner, and it is brightly illustrated by the frequency of the verb "want", which appears 307 times.

Statistics also offer a number of adjectives that characterize a true winner: "*great*" (240), "*best*" (85), "*strong*" (47), "*rich*" (35), "*eager*" (21), "*individual*" (21), "*forceful*" (20), "*determined*" (20), "*successful*" (19), "*powerful*" (15), "*superior*" (13), and "*famous*" (12). A

winner is bound to be a man of spirit, energetic, courageous, and firm. He highly values his own individuality, power, and dominance over other people. Besides this, statistics contain words that describe indispensable parts of the life of a true winner.

The performed statistical analysis of words gives an opportunity to discover the author's view of Frank Cowperwood's character. Literary theorists tend to think that his character is negative, that he is an egoist to the marrow of his bones, that he has no values, besides material benefits. However, statistics of the key words in "The Financier" enables to see the other side of Frank's character. It turns out that Dreiser gifted the character with a positively viewed American conscience – traits of a true winner.

Words that represent the author's estimation of the main character's traits and have a positive key prevail over words with a negative meaning.

The frequency of key words disproves the single-value estimate of Frank Cowperwood's character being viewed as negative, which is widely accepted by literature theorists. The traits of his character are ideal for any American: diligence, firmness in achieving goals, determination, power, and ability to act. At the same time, such negative traits as egoism, indifference, cruelty, dishonor, deception, and a tendency to dominate and manipulate are represented in the text of the novel by singular citations, which are not sufficient to unmask the hero.

Conclusion

One of the main tasks of this review is to urge to change the false idea about the comparison of "literary text and computer analysis", because this delusion is based on two preposterous thesis: that the "computer is able to estimate everything" and that "literary studies do not require computer analysis".

This review deliberately describes free and conditionally free programs. The necessity to raise funding always hinders experimental research work, especially for students.

It can be observed that the discussed programs were not created purposely for literary analysis and are not striving to replace the researcher scientist. Each program requires certain skills when working with it. A common advantage of all similar programs is the ability to process big volumes of texts in considerably less time than if working with hardcopies of texts.

In the end, the variability of text processing (involving computer programs or not) depends only on the researcher.

One should not forget that computer-based text analysis should be practiced only with further literary transcription and interpretation. This comparison of data, retrieved in the process of computer analysis with existing traditional researches, may mark the dawn of a new stage of literary text analysis.

This study was carried out within The National Research University Higher School of Economics' Academic Fund Program in 2013-2014, research grant No. 12-01-0129.

References

1. AntConc http://www.antlab.sci.waseda.ac.jp/software.html.
2. Concordance vocabulary of journalism of F.M. Dostoevsky http://dostoevskii.karelia.ru/author.phtml
3. Dobrushina Nina. Corpora methods in education Russian // Russian National Corpus 2006-2008, 2009. SPb (in Russian).
4. Dreiser Theodor. Trilogy of Desire. Great Britain, 1994.
5. Frank Il′ia, Kuznetsova Oksana. Pol′skiĭ iazyk s Ioannoĭ Khmelevskoĭ "Vse krasnoe". AST, 2008 (in Russian).
6. Khmelevskaia Ioanna. Vse krasnoe (per. M.Krongauz). Ekaterinburg, U-Faktoriia, 2001 (in Russian).
7. Khmelevskaia Ioanna. Vse krasnoe (per. V.Selivanovoĭ). HarryFan SF&F Laboratory. http://www.lib.ru/DETEKTIWY/HMELEVSKA/all_red.txt.
8. Krapivin Vladislav. Sobranie sochineniĭ. Kniga 18. Babushkin vnuk i ego brat′ia. M.: TSentrpoligraf, 2001 (in Russian).
9. Kuvshinskaya Yulia. http://studiorum.ruscorpora.ru/stylistics/
10. LEKTA. http://www.nisoc.ru/lekta.html.
11. Levinson Anna. Using the RNC in Teaching "Rhetoric" in high school // Russian National Corpus and the problems of humanitarian education. Moscow, 2007 (in Russian).
12. LF Aligner. http://sourceforge.net/projects/aligner/
13. Russian National Corpus. http://ruscorpora.ru
14. Razumikhin Andreĭ. Pravilo bez iskliucheniĭ, ili Prozrachnaia zlost′ i intelligentnye mal′chiki Vladislava Krapivina // Ural - 1982. - № 8. - P. 149-152 (in Russian).
15. Russian-French poetry corpus of the first third of XIX century http://nevmenandr.net/fr/index.php?go=head.

16. Sichinava Dmitriĭ. K zadache sozdaniia korpusov russkogo iazyka (in Russian). http://www.mccme.ru/ling/mitrius/article.html
17. The parallel corpus of The Tale of Igor's Campaign. http://nevmenandr.net/slovo/main-en.html.

Vera G. Sibirtseva
National Research University Higher School of Economics. Department of Applied Linguistics and Intercultural Communication, associate professor;
E-mail: vsibirtseva@hse.ru