Towards Automatic Modelling of Thematic Domains of a National Literature: Technical Issues in the Case of Russian

Tatiana Sherstinova, Anna Moskvina, Margarita Kirina
National Research University Higher School of Economics, Saint Petersburg, Russia
{tsherstinova, admoskvina}@hse.ru, mkirina2412@gmail.com

***Abstract*—A significant part of modern technologies associated with the development of artificial intelligence systems and digital analytics of diverse data relies on methods of computer text processing (NLP, speech technologies). However, NLP methods are applied primarily to specialized texts, such as scientific literature, technical documentation, news, etc., or social media discourse. Fiction texts are usually left out of the focus of NLP practitioners as the fictional world seems to be of less significance or less "information value" from a practical point of view. Moreover, due to the poetic and metaphorical nature of literary texts, the use of some NLP methods (e.g., topic modelling) for fiction analysis turned out to be more complicated. At the same time, the influence of literature both on the consciousness of individuals and on the formation of social values can hardly be overestimated. Besides, making computers "understand" fiction in a similar way as humans do would be a real challenge for artificial intelligence. The article puts forward the idea of modelling thematic areas of literature on a national scale, which should reveal the main thematic domains of national literature as a whole. It will allow a better understanding of the cultural traits of the national consciousness in a given historical period and contribute to either literary studies or practical tasks. Methodological approaches to determining and modelling themes of literary works are considered, technical difficulties arising in the process are described, and the ways to solve them are suggested. The proposed methodology has been implemented in the design of the corpus of Russian short stories of 1900-1930s and can be applied in the development of artificial intelligence systems that process large volumes of literary texts in any language.**

I. Introduction

A significant part of modern technologies associated with the development of artificial intelligence systems and digital analytics of diverse data relies on methods of computer text processing (NPL, speech technologies) [1], [2]. However, NLP methods are often applied just to specialized texts (scientific literature, technical documentation, news, and journalistic texts, etc.) or social media discourse; fiction texts being usually left out of focus of NLP practitioners as their fictional world may seem to be of less significance or less "information value" from practical point of view.

Moreover, due to the imagery, poetic, and metaphorical nature of literary texts, the use of some NPL methods (e.g., for topic modelling) for fiction analysis turned out to be not as effective as for processing of non-fiction texts [3].

However, the influence of literature both on the consciousness of individuals and on the formation of the values of society can hardly be overestimated. "A national literature is an essential element in the formation of national character. It is not merely the record of a country's mental progress: it is the expression of its intellectual life, the bond of national unity, and the guide of national energy" [4]. Besides, the idea to make computers "understand" fiction in the similar way as humans do would be a real challenge for artificial intelligence.

John Sinclair, an outstanding linguist and one of the founders of corpus linguistics, in his book "Trust the Text: Language, Corpus and Discourse" spoke well about the importance and necessity of studying the language of fiction: "Literature is a prime example of language in use; no systematic apparatus can claim to describe language if it does not embrace the literature also; and not as a freakish development, but as a natural specialization of categories which are required in other parts of the descriptive system" [5].

Fiction, especially the genre of short story, quickly reacts to events in social, political, and cultural life, especially in the era of social transformations and crucial historical shifts [6]. On the pages of literary texts, the formation of linguistic norms may be traced, reflecting their change and development. At the same time, for the completeness of the presentation of the general picture, the entire set of works published in the era under study is of interest [7]. An analysis of the thematic component of the national literature in a statistical aspect provides unique information about the aspects of personal and social life which concerns the society more than the others in the historical period under consideration. Therefore, fiction can be considered a reliable source of information, both in quantitative and qualitative aspects.

The importance of literary thematic domain studies is determined not only by the fact that they help to reveal collective self-consciousness of the nation. The innovative potential of this approach lies also in the fact that in case that certain thematic trends will be proved to be associated with determined social changes, they may be considered to be the "weak signals" based on which it may be possible to predict the likelihood of certain social, and even revolutionary,

ISSN 2305-7254

changes in the society. The concept of "weak signals" is now actively used in futurological studies, and in most cases, they are identified through texts of different genres [8], [9].

It should be noted that digital research of literary tests has already been carried out in recent years. However, the goal of many such projects is to describe the language of only one outstanding author (Lermontov, Dostoevsky, Chekhov, Shakespeare, Dickens, Wolf, etc.). Some attempts at building a generalized model are presented in the construction of the CLiC Dickens project with 19th century reference corpus of English Literature [10], they may be also traced in "distant reading" [11] and "cultural analytics" [12] briefly described in the next section. At the same time, most writers — both well-known and, moreover, less well-known or simply forgotten — find themselves outside the field of vision of NLP studies. Thus, we claim that modern NLP methods do not yet fully exploit the potential of fiction.

As for studies of Russian literature, they are usually carried out both on the basis of the well-known National Corpus of the Russian Language [13] and on a number of specialized digital resources. For example, the Corpus and the frequency dictionary of literary works by Anton Chekhov were developed at Moscow State University [14], frequency dictionary of Fyodor Dostoevsky's language was compiled at the Institute of the Russian language [15], and the statistical analysis of the language of Russian prose of 1850–1870s was carried out, too [16]. In the Department of Mathematical Linguistics of St. Petersburg State University, a series of frequency dictionaries for the prominent Russian writers — Anton Chekhov [17], Leonid Andreev [18], Alexander Kuprin [19], and Ivan Bunin [20] has been published. An impressive resource Tolstoy Digital was prepared in the HSE University [21].

There are several successful attempts to apply topic modelling techniques to fiction literature, although not that many. One of the pioneer studies of the field [22] investigates the corpus of English-language literary fiction by extracting the topics with the help of statistical methods, also taking into consideration external factors to evaluate the adequacy of the received themes. This work proves that topic modelling can be used to detect broad themes, but more from the distant reading perspective. The [23] demonstrates the use of topic-based clustering to perform subgenre classification on the French drama corpus as a scientific result. In [24] the authors analyze the correlations between the cleaning, preprocessing, and organization of literary texts for model training with the quality and features of the received topics.

As for Russian literature, in [25] authors apply non-negative matrix factorization (NMF) to trace the thematic dynamics on the corpus of Russian short stories of 1900−1930s. The results show that the received topics are quite different in their internal structure, reflecting various syntagmatic and paradigmatic lexical relations within the corpus. Since the model is susceptible to changes in the number of topics, the optimal number was established individually for each period. The results are interpreted mostly from linguistic rather than literary point of view. There is an important question to keep in mind: what exactly do we obtain by the process, or, in other words, what exactly such models reflect: the narrative structure, the characters, the relations, the actions? Unlike commercial research, literary investigations, even those that employ machine learning techniques, may not have a preset goal. The earlier study [26] discusses what types of information the topics can provide for the corpus of Russian fairytales, including names, typical actions, attributes of specific characters, and how they correlate with the fairytale structure. In this study, the LDA algorithm was used for topic extraction.

The article puts forward the idea of modelling thematic domains of literature on a national scale, which should reveal the main themes of a national literature as a whole. It will allow better understanding the cultural specifics of the national consciousness in a given historical period and help to solve a number of important humanitarian and practical tasks. In the next section, we consider theoretical background and methodological approaches to determining and modelling themes of literary works, and thereafter we discuss technical difficulties arising in the process of completing this task, as well as some possible ways to solve them.

II. Theoretical Background and the Main Tasks

A. Theoretical Background

The following notions and research directions can serve as a theoretical basis for modelling thematic domains of a national literature:

1) **Literary and artistic system**

The concept of a "literary and artistic system", was proposed by one of the most prominent representatives of the Russian formal school — Yuri Tynyanov. Tynyanov's innovation consisted in the fact that he was first suggested to consider a national literature not as a collection of individual works, but as a wholistic system that includes all the literary works of a particular historical era [27]. Astonishingly, Tynyanov's idea was put forward more than 90 years ago. For obvious reasons, in those days it could not be technically implemented, since its solution requires powerful data processing tools and the ability to access big data, which have appeared just very recently [28].

Another notion that can be considered to be a prerequisite for the proposed approach is the classification concepts by the famous Russian linguist Victor Vinogradov, who put forward the idea of constructing linguistic analogues of literary schools, trends, styles based on the criterion of linguistic proximity of works by various authors [29].

Within the framework of our study, it should also be noted the method proposed by another prominent Russian linguist Victor Zhirmunsky for describing the worldview of a writer through a set of its "verbal themes" [30].

2) **Stylometry studies**

Stylometry is an applied philological discipline, which deals with the measurement of stylistic characteristics in order to organize and systematize texts and their parts. It is often

associated with text diagnostics, taxonomy, attribution, and typology [31].

The first experience of a systematic analysis of the style of Russian literature was a stylometric study carried out by Gregory Martynenko on the literary texts by 100 Russian writers who worked at the beginning of the 20th century [32]. As a result, preliminary quantitative data were obtained on the stylistic closeness and stylistic differences of Russian prose writers, the characteristic features of their individual styles were revealed, and multidimensional classifications of the author's style were constructed [ibid.].

3) **"Distant reading"**

In comparison with stylometry, a relatively new direction is "distant reading", which has been recently appeared within literary studies [33]. Nowadays, this method is usually opposed to "close reading", which mean traditional detailed analysis of some text or its fragment. In distant reading, "in principle, one could study the history of a literary tradition without reading any of literature" [34]. It applies a number of computational methods of literary texts processing, e. g. computational literary studies, quantitative literary studies, and macroanalysis [35].

4) **"Cultural analytics"**

Finally, the other approach which should be mentioned in concern of literary texts processing is "cultural analytics" [36], which studies large volumes of cultural data by means of computational and visualization methods. As the name of the term suggests, its application is somewhat broader than just literary studies. Nevertheless, it seems appropriate to us to use it for literary studies, because through the description of the main topics of literature we achieve a description of the deeper cultural layers of society. Moreover, individual quantitative results arising in the process of processing the literary corpus, especially in combination with biographic information about the authors of literary works, provide results that are extremely important for the cultural analysis of the nation for a given historical period.

Basing on Tynyanov's ideas, and also taking into account the approaches of stylometry, "distant reading" and "cultural analytics", the concept of a digital literary resource of a new type was developed and found its implementation in the corpus of Russian Short stories of 1900–1930s [7, 37]. The developers set a task to include in the corpus texts written by the maximum number of authors who wrote in the given historical period. Such an approach will contribute to the objectivity of literary and linguistic studies carried out on its material, since it will allow processing literary works both by the leading writers of their time and also by many secondary authors. The artistic heritage of the latter will make it possible to enrich our understanding of different aspects of social and cultural life of that time, about the characteristic features of language, as well as about thematic distribution of literary texts. The creation of a representative corpus of Russian short stories, including texts by a large number of writers, will make it possible to carry out a quantitative analysis of their language, style and literary futures in synchrony and diachrony [38].

B. The main tasks to be solved

Currently, the prerequisites for the study of modern literature by means of computer methods have emerged [28, 35]. This task is simplified by the fact that all published data are already stored in digital form. Therefore, in order to collect all existing digital texts into one large resource, the appearance of an organization or a national project that carries out such a consolidation would be sufficient. The methodology for collecting Czech National Corpus, which automatically accumulates all texts published in the country, can serve as an excellent example to follow [39]. But it is obvious that the implementation of a similar project within such a large country as Russia is hardly possible in the coming years.

Speaking about the literature of the past, the possibility of its analysis is limited by the amount of digitized text data. Moreover, until the bulk of the texts has been digitized, it is difficult to predict in advance how many different writers really worked in a given historical period.

Data preparation for computer studies of a national literature may consists of solving the following tasks:

1) Compilation of the complete list of writers (and their works).
2) Cataloging of literary texts written and published during the given period.
3) Search and collection of texts electronic versions.
4) Digitization of texts for which there are no electronic versions.
5) Conversion of texts from the old graphics (spelling) into the new one (when necessary).
6) Proofreading of electronic texts versions.
7) Meta-tagging of electronic texts collections.

At the next level of the literary resource development, it seems appropriate:

8) Compilation of information about each writer (when available).
9) Text segmentation into structural parts (sections, paragraphs, sentences, etc.).
10) Expert data annotation (at least on a representative, "training" sample).

Finally, the decisive factor is the choice of the method for obtaining thematic information. Thus, in particular, the need to perform some of tasks described above will depend just on it.

Let us demonstrate how these tasks are being solved to model the thematic diversity of Russian literature of 1900–1930s on the basis of the corpus of Russian short stories.

III. Methodological and Technical Issues on Data Collection

A. Compilation of the complete list of writers

The modelling of a national literature assumes that texts by the maximum number of writers who worked in a given historical period are subjected to analysis. Therefore, the first important task is to compile such a list of authors.

The main sources for forming a representative list of Russian writers, as well as and their works, are the following:

1) Bibliographies and literary encyclopedias, dictionaries of writers, etc.
2) Library catalogs, including online catalogs (Russian State Library [40], National Library of Russia [41], Scientific Library of St. Petersburg State University, etc.).
3) Periodicals of the given historical period (e.g., “Apollon”, “Shipovnik”, “Niva”, “Ogonek”, “Novyy mir”, etc.).
4) Electronic libraries — Lib.ru: "Classics" [42], Litres [43], Aldebaran [44], LitMir [45], the IMLI RAS electronic library [46], etc.
5) Other Internet resources (Sovlit project [47], “3500 Russian prose works (1800–1940s)” [48], “Almost forgotten” [49], etc.).

It seems appropriate to include in this list all writers whose creative heritage is represented by at least one text of the genre under study. In our case, there is a genre of the short story. The goal is to involve into the research not only “metropolitan” writers, but also regional writers who wrote in Russian and lived on the territory of the Russian Empire (until 1917), and later on the territory of the RSFSR and the USSR.

The primary list of authors for inclusion in the corpus was formed on the basis of encyclopedic information about personalities (for example, [50]), existing bibliographic indexes (for example, [51], [52], [53], [54], etc.), anthologies of Russian story and collections of stories (for example, [6]), as well as publications in authoritative periodicals. For example, according to [50], a list of 273 persons was formed.

However, the list of authors obtained in this way turned out to be both insufficient and redundant at the same time. On the one hand, many peripheral Russian writers were not mentioned in [50], and on the other hand, it is not always clear from the relatively small articles of the encyclopedia whether this author wrote stories and whether they fall within the studied time period.

The other sources listed above were reviewed in a similar manner. Thus, periodicals of 1900–1930s, electronic libraries available on the network [42–45], and some other Internet resources [47–49] were selectively cataloged. It turned out that working with electronic libraries that already present data in a structured form greatly simplifies researchers’ tasks. On the other hand, working with periodicals, even those digitized in the form of scans, is a much more difficult task, therefore, in our study, it was only partially completed.

As a result, the list of Russian prose writers of 1900–1930s, which we have been able to collect from various sources, numbers 2800 persons. This is a fairly significant value. After (or in parallel) with the formation of writhers’ list, the task of cataloging their texts is to be solved.

B. Cataloging of literary texts written and published during the period under study

Obtaining relatively complete texts catalogs can be carried out in different, mutually complementary ways.

First of all, it is worth paying attention to the catalogs of national libraries (for Russian literature, these are, first of all, the Russian State Library [40] and the Russian National Library in St. Petersburg [41]). Library catalogs give a good overview of literary texts published as separate publications (brochures). For example, according to the results of short stories cataloging, the following distribution of texts for a 30-year period was obtained. Fig. 1 shows that there is a completely explainable decline in “publication activity” in the around-revolutionary period of 1916-1924.

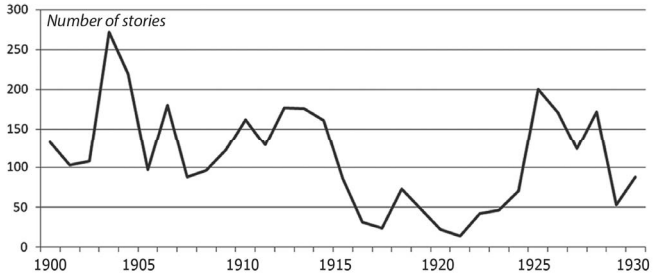

Fig. 1. Dynamics of publications of Russian short stories by years for 1900–1930s (according to the National Library of Russia, for stories published as separate brochures)

However, the catalogs of libraries, even the national ones, do not provide an opportunity to obtain data on how many and which specific works of one or another author are presented in the author’s collection, is books of collected works or in journal publications. Thus, two technical problems arise — identifying short stories in author’s collections and viewing all the periodicals of the corresponding era.

C. Search and collection of electronic versions of texts

To expand the corpus size, we carried out several experiments with downloading texts from Lib.ru [42] online library automatically, which allowed establishing the procedure of such a task. For collecting texts and extracting useful features for data annotation the Beautiful Soup Python library was used.

Since we were primarily interested in short stories we chose the corresponding subset from the Lib.ru classics collection. Next, we limited our search to start with the 1900 year that resulted in more than 6000 URL links, each containing a text marked as a short story of the period in question. By parsing the HTML, we extracted the following features for each text: title, author, year of publication, genre, and ‘miscellaneous’ information based on the internal hierarchy of the library, which sometimes included the name of the collection or a series of books containing the given short story (e.g., *Izobreteniya professora Vagnera*) that could also contain the specific period of time (e.g., Short stories of 1918–1937s).

Each text was saved in a separate plain text file and given a unique id, with metadata saved in a separate table.

All information being obtained as provided by its library annotation, this approach has a risk of inheriting and multiplying the mistakes and discrepancies of the source repository and therefore ideally requires expert validation. By preliminary evaluation we can claim that the metadata was extracted with high precision.

D. Digitization of texts for which there are no electronic versions

In case there is no digitized version of a text, the question of digitization arises. The task includes library work, scanning of the texts, and their subsequent recognition (for example, to plain text *.txt* format). For some texts digitalization is simplified if a pdf scan copy already exists and can be retrieved. Since digitalization of texts is a rather challenging and time-consuming task, when building the corpus of Russian short stories we decided to limit ourselves to digitizing one text for each "forgotten writer". The original plan was to expand this collection only if the corpus contains at least one text for all the writers of the generated list.

E. Conversion of texts from the old graphics (spelling) into the new one

Large-scale spelling reform was carried out in 1917–1918 along with other revolutionary events [55]. The idea behind was to simplify and unify Russian spelling. The four letters were eradicated and substituted with existing analogs. The hard sign (letter "Ъ") lost part of its functions and most of its uses, only remaining as the orthographic means to separate consonant-ending prefixes followed by vowels, while it used to follow each non-palatal consonant at the end of the word. Obviously, these changes constitute a well-known list. To some point, they may be described with a list of replacement rules for automatic translation, including replacing "Ъ", "Ѳ", and "І" ("V") with "Е", "Ф", and "И" correspondingly, removing and inserting hyphens at certain places, some letter combinations substitutions, etc.

Some rules are more complicated and require additional morphological information: for example, the ending *-аго* should be translated to *-ого*/*-его* in case it is indeed adjectival flexion. Removing hyphens after prefixes may require additional checking as well. Even in the case of a simple replacement of the vanished letters, there are exceptions, for example *і* → *ь* (*піеса* → *пьеса*) или *і* → *й* (*іод* → *йод*) either positional and lexicographical.

While such an instruction can be imagined and constructed, however complex and nested the rules will be, we will still have to address the dictionary at some point of this analysis. For now, this procedure still requires expert revision and post-processing correction.

Another issue here is punctuation, which is formalized to a far lesser degree.

F. Proofreading of electronic texts versions

This kind of work is one of the most time-consuming tasks. Unfortunately, there are no effective ways to automatically correct spelling yet, especially considering that authors may deliberately use incorrect spelling in literary texts to add imagery or to emphasize a particular phenomenon (see the section below). It can be assumed that deep learning methods should show the greatest efficiency, however, the current needs of the corpus demand manual proofreading and expert verification.

G. Meta-tagging of electronic texts collections

Meta-tagging of texts is a rather important, although not an obligatory element for the fiction corpus per se. For digital resources, this is an important feature that provides filtering and processing of data by homogeneous groups. For example, an important parameter in the description of texts is the genre of fictional prose: for example, "short story", "story", "novel", "play", etc. Some genre features can be determined — among other things — by formal criteria (size, text structure, etc.).

For several tasks it seems essential to separate literature intended for an adult reader from literature for children. For example, this is the case of the corpus of the Russian short stories. It is clear, that from the point of view of both text complexity and the depth and elaboration of a story plot, it is expected that texts for adults should be more complex than children's "adapted" literature. Although concerning the imagery nature of the language, in children's literature one may find samples of very high quality as well. However, it is often the case that there is no explicit marker of "childishness" within such texts. Even if some text has a subheading "A fairytale", this does not mean that the text is intended for children.

If texts are collected from a digital library, it is quite possible that they already contain the parameters necessary for meta-annotation.

When working with Lib.ru texts, apart from titles, names, and years, the following tags were gained and included in the annotation: "prose", "children's", "translations", "science fiction", "biography" and various fine-grained genre information, such as "humor and satire" or "historical". Navigating by those tags allows us, for example, automatically exclude from consideration both the juvenile and foreign translated literature as dictated by corpus requirements, and potentially implement the additional extension of the corpus in the process of building a model or interpreting its results.

Some information provided by the library in a less consistent way turned out to be more complicated for automatic extraction. For example, the name and details on the first publication are often mentioned at the end of the text block, but there is no way to efficiently formalize the procedure of their extraction and identification even if this data is present at the time of text collecting.

Further (optional) data annotation may include the following steps:

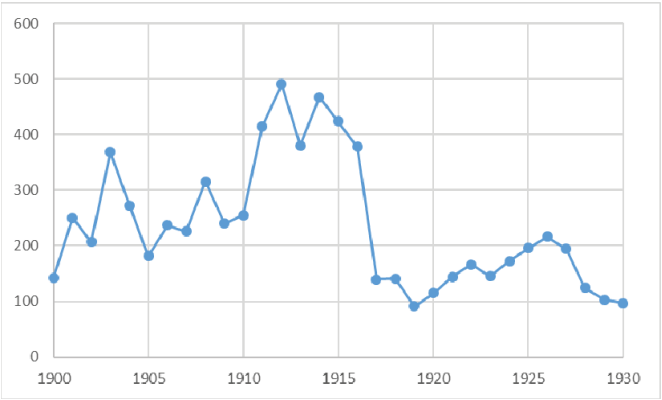

Fig. 2. Distribution of Russian short stories in 1900–1930s by the year of creation.

H. Compiling formalized information about each writer

Formalized information about writers allows grouping texts into qualitatively homogeneous groups according to some criterion, and to obtain the distribution of thematic fields with regard to the different social characteristics of their authors. To achieve that, a database of biographies [56] can be employed by associating some historical or personal facts with the year of writing or publishing of a certain literary work.

I. Texts segmentation into structural parts

This kind of annotation is preferable when text is not homogeneous with regards to its contents or stylistics (for instance, if it combines two narrative plans, as it can be seen in Bulgakov's *The Master and Margarita*). In this case these peculiarities need to be taken into consideration for model specification. Another example is when it is important to consider the development of the narrative over time [57].

J. Expert data annotation

At last, an expert annotation of textual data can be done. Evidently, it is very difficult to process large amounts of textual information manually (though such technologies as crowdsourcing open new possibilities [58]). Nevertheless, in some cases it seems quite reasonable to make expert annotation, at least for the representative "training" sample.

One example of such annotation demonstrates the expert thematic annotation, which is to be described in more detail in the next section.

IV. Methodological Approaches to Determining and Modelling Themes of Literary Works

A. Automatic topic modelling

Automatic topic modelling is one of the most popular tasks in state-of-art NLP technologies. However, like most of these methods, they are mainly applied to specialized texts, whose main purpose is to transfer information. In recent years, they have also gained popularity in research of online media discourse and social network analysis.

There is a diverse range of topic modelling algorithms (LSA, pLSA, LDA, NMF, etc.) [59; 60; 61] as well as their implementations (MALLET, Stanford Topic Modelling Toolbox, gensim, scikit-learn, BigARTM, etc.). In this work, we do not try to provide an overview of such methods, since it has already been done in various works [59; 62].

As we already mentioned in the introduction, fiction topic modelling is a rather difficult and challenging task [63], as unlike nonfiction texts (scientific, mass media, business, etc.), in which several terms or keywords quite unambiguously determine the topics of the document, the topics of literary texts are often hidden [3].

An example of using topic modelling methods on a fragment of national literature corpus is discussed in [25].

B. Expert thematic annotation and its normalization

In case of fiction, topic modelling is used to extract literary themes and motifs [22; 78]. However, as experience shows, the complications may appear when it comes to the assessment of the literary interpretation of topics obtained automatically [3; 24; 78]. Since the latter can be performed only by a human, the need to involve an expert for manual annotation of texts — at least for a limited training text sample — arises.

Indeed, literary corpora rarely provide literary annotation, let alone the thematic one. The main issue here is not only determined by the fact that there are not any tools that allow to automatically implement this kind of mark-up. The problem is that the number of factors that literary annotation of a corpus can contain is, as a matter of fact, unlimited — it is almost impossible to determine which information is more relevant and more needed since it is in many aspects a really fruitful data for various kinds of research.

One of the kinds of literary annotation that we propose to be done in the first place is thematic. Theme extraction from a text is a difficult task, because of that this procedure is rarely used in corpus linguistics [66]. Nevertheless, a number of researchers emphasize its importance for carrying out a full-fledged literary analysis (e.g. [67; 68]). Besides, thematic annotation can be useful for linguistic agenda as well — interpretation of frequency word dictionaries, stylistic specificities of the established groups of authors, etc. [64].

The whole procedure of thematic annotation suggested for the annotated subcorpus of the corpus of Russian short stories 1900–1930s is described in detail in [65]. At that moment, the annotation was done only by one expert. Here, we will briefly consider the main principles of the thematic annotation and its further normalization used in the corpus since they are relevant for better understanding of the experiment which is to be discussed next. The theme-mapping required initial reading of all texts presented in the sample and then the formation of themes list. The resulting list of themes was compiled without taking into account the existing lists of thematic sets like "100 common thematic topics" [3]. The focus on the social upheavals found throughout the period in question was also smoothed. The expertise mainly relied on reading and reader's perception of text, the theme being understood as "all semantic components that contribute to the plot, determine the protagonist's motives and actions, and directly bear on the conflict and its resolution" [69].

Since a number of themes per story was not limited and several themes could have been assigned to one story simultaneously, in the final list of expert themes 89 elements were included. For the tasks of further automatic data processing, this number of themes seemed to be excessive, therefore this list was reduced via normalization. The suggested procedure was in combining the thematic elements that are semantically related into one tag. For instance, such themes as *natural death*, *murder*, *death at war*, *execution*, resulted in a single tag DEATH. The total number of resulting tags equals to 30. These tags are FUTURE, MODE OF LIFE, RELATIONS, WAR, CITY LIFE, MONEY, CHILDREN, VIRTUE, LEISURE TIME, ART, BEAUTY, LOVE, VICE, DREAM, YOUNG PEOPLE, VIOLENCE, POLITICAL STRUGGLE, NATURE, PROGRESS, MENTAL STATE, REVOLUTION, RELIGION, FREEDOM, FAMILY, SLEEP,

SOCIAL GROUPS, SOCIAL PROCESSES, LABOR, and FANTASY.

Fig. 3 illustrates the frequency of these tags counted for the stories in the annotated corpus. As it can be seen, some of the tags are more "popular" than the others, these are RELATIONS (115 stories), DEATH (106 stories), LOVE (103 stories). Their high frequency is quite understandable since they refer to the "common" literary themes. The next tags that occur relatively often are SOCIAL GROUPS (79 stories) and SOCIAL PROCESSES (75 stories). They are the ones that are specific to the corpus in the whole and, probably, reflect the social changes in the country and people's lives that happened in Russia at the early 20th century.

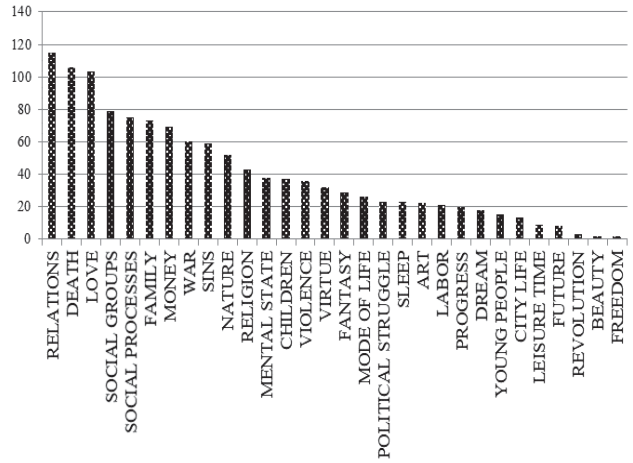

Fig. 3. Frequency distribution of topic tags.

Fig. 4 shows the frequency of distribution of the two top themes DEATH and LOVE in dynamics.

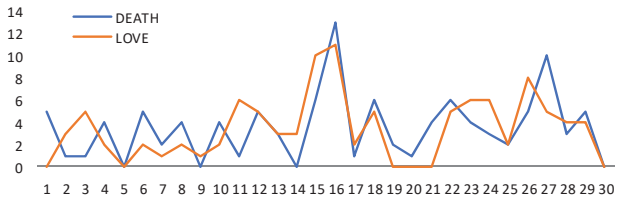

Fig. 4. Frequency distribution of tags DEATH and LOVE

C. Comparison of the results of automatic and expert thematic annotation and hybrid thematic annotation

Finally, the comparison of two types of thematic annotation — automatic and expert ones — may be done. Here, we need to distinguish two terms that we use — "topic" and "theme". First of all, accepting Jocker and Mimno's idea, we do believe that "topic" and "theme" are aimed at presenting "a type of literary content that is semantically unified and recurs with some degree of frequency or regularity throughout and across a corpus" [22]. However, "topic" is more of a technical term referring to a set of keywords obtained automatically which characterizes a separate document or a set of documents [61]. The term "theme", on the other hand, is considered to be a conclusion regarding the semantic core of the fictional story or poem after the actual reading of it.

The first example of such a study was described in [3], where a novel type of text annotation was proposed and implemented. Thus, a hybrid (both expert and automated) topic annotation was introduced. This type of annotation may be recommended specially for fiction texts as it allows to describe relevant aspects of the plot, features characterizing main characters of the text, etc. Topic annotation proves to be a significant extension to tradition multilevel annotation schemes accepted in contemporary corpus linguistics [3].

The experiment was conducted based on two types of annotation, manual and automatic, made for the corpus. For manual annotation, the normalized list of themes was used whilst the automatic one was obtained using topic modelling via non-negative matrix factorization (NMF). As the comparison has shown, automatically extracted themes tend to reveal the background of stories (where they take place). The themes of the better correlation are *war*, *love*, *nature*, and family. However, such themes as, for instance, *death* and *relations*, which are considered as relevant to the plot and the description of interaction between characters, were not detected automatically.

In this experiment we largely rely on the literary assessment of themes because it can shed some light on the applicability of topic modelling and the ways it treats data.

V. Technical Difficulties and the Ways of Their Solution

In this section we will consider the main technical difficulties we faced while preparing the representative literary resource of Russian short prose and share some thoughts on how these obstacles can be overcome in future. As we think, our experience may appear useful for the developers of literary resources for other languages as well.

1) Despite the fact that we compiled a quite representative list of Russian writers of the beginning of the 20th century, it seems that it should not be considered as final. This conclusion is based on the following factors:

1.1) There are authors published in 1900−1930s for whom we were not able to find any data or bibliographical information, that could have confirmed that this writer indeed lived and worked in the epoch under study. It can be said that these authors are "totally forgotten"; they were not popular enough at the time they created to be included in the dictionaries of writers and literary encyclopaedias or to be analysed by literary critics. Besides, regarding these fameless authors we cannot be sure that their names are not, in fact, one of the unknown pseudonyms of another author.

In this aspect, a database of writers' pseudonyms can greatly facilitate the work of researchers. For the corpus of Russian short stories, the verification of the authors' names is being conducted on the base of the dictionary of pseudonyms [71]. It is suggested that, as a result, a total number of unique writers will be slightly reduced. However, the difficulties of interpretation are inevitable, as there are cases when different authors used almost or totally the same pseudonyms.

Sometimes biographical data turns out to be crucial for the determination of some quality text characteristics. For example, while implementing the project regarding the language and style of Russian literature in the epoch of various social upheavals, it was decided to exclude from the corpus of Russian short stories those texts of Russian writers that were written while the authors were in emigration [7]. The logic behind this decision was obviously motivated by the fact that if the research objective is to study linguistic changes caused by social changes inside the country, then texts written in foreign environment should not be taken into account for further analysis. The absence of biographical data for a large number of Russian writers leads to the point when we cannot guarantee that this requirement is always properly satisfied.

2) The problem concerning genre attribution is also on agenda. Though for some literary genres genre definition does not seem difficult according to formal criteria (for example, for a novel or a dramatic work), for particular genres it may be difficult. For example, it is not always easy to distinguish a short story from an essay or a novella.

The similar problem emerges when one tries to distinguish “children’s” and “adult” literature. As is has been said, both language and themes of these two groups should differ. It seems that text publication by the publishing house which is specialized in children’s literature or in children’s magazine can be the most effective marker. The useful information can be found in dictionaries of writers that mark those authors, who write predominantly for children as “children’s authors”. However, it cannot be guaranteed that an “adult” writer will not write a text for children, or vice versa.

3) The special question refers to translated literature. It seems that one of the most obvious solutions is not to include literary translations from foreign languages into a national corpus. For the automatic selection of such texts an inner search function “Translated from...” or “translated by…” can be used. However, for some editions the information about the interpreter is not easily found. On the other hand, it should be noted that some texts that have been translated by prominent authors certainly can be considered as a part of the national literature. First of all, it regards poetry.

4) As it has been already mentioned, a separate major issue is the old orthography. For any corpus containing texts in different types of spelling meta-information on what particular type of spelling is used in each text seems to be obligatory. Special utilities can help to obtain this data automatically. For example, for Russian the following rule may be used if there is “ъ” followed by a space in the first few lines, the text can be considered as the one written in the old orthography.

However, if the identification of text in old orthography is possible, there is another problem which relates to translation from the old orthography into the modern one. This issue only partially can be solved via application of certain scripts and macros for autocorrection. One of the main reasons for that is non-existence, at least to our knowledge, of sufficient software for such translation adjusted for massive corpus processing with open source code to experiment and modify. What is more, as our experience shows, such a process still ends up with errors and requires additional manual checking. Even if we suggest that these errors are quite rare and texts can be processed as they are after “autotranslation”, doing so may interfere with the process of topic identification — for instance, a model might result having a topic comprising “wrongly”-spelled words. Finally, when applying or developing software for old-new orthography translation, one may face OCR mistakes as well (e.g., Latin "i" and "I" instead of Cyrillic "і" and "І" and vice versa).

Besides, one should also keep in mind that writers may deliberately use non-standard spelling to reflect some stylistic or content feature, for instance, dialect speech of characters. It has also been noticed that abbreviations, neologisms, bureaucratese, and slang words are very characteristic for this historical period [72-75], partly due to the fact of so-called “popular speech” invading literature [76] which may lead to orthographic inconsistencies as well. All these factors should be considered while processing of texts, since they may interfere with lemmatization and parsing. To solve this problem, one can go at least in two ways: 1) to add irregular word spelling into dictionaries used by parsers, 2) to make normalized corrections in the copies of texts, thereby providing a “standard”-spelling double for each processed text. The second option can be considered more preferable if different software products are used for data processing.

5) The important information about literary text is the year of its creation, as it allows to set the literary work in the context of the certain period in the life of society. However, this kind of information about a text is not always available. Besides, in digital libraries the year of text creation is often substituted by that of its first publication, or by the year of publication of the given text version. To find the original year/date of the publication of a literary text, when it was not indicated by the author himself or not given in the current publication, is a quite serious problem which requires expert investigation. For that reason, it is expected that a certain number of texts, even if a digital resource is of the highest quality, may contain inaccurate or false metadata.

It also seems crucial to preserve information about the year of the first publication of each text in the database, when available. With regards to this type of metainformation, there is less complications. For instance, in the library *Lib.ru*, it is usually given right after the text of a literary work. Besides, it seems worthy to add to the database not only the information about location and year where and when the text was created, but also the publisher or magazine where it was published. Thus, unique data for digital analytics can be obtained.

6) In libraries — both traditional and digital — the unit of description is a book. At the same time, books can contain as just one literary work (e.g., a novel or a novella) or some number of texts (e.g., a short story collection). Even if these books are digitalized, the information system usually does not include any additional data about individual texts that contribute to the collection. For this reason, catalogization of such texts and their segmentation is required. Since the books of the past epochs may be quite heterogeneous regarding the structure of inside literary material, automatic segmentation of these random text collections into individual literary works

that constitute it is one of the unresolved tasks. Similar problem concerns digitalization of periodical literature. Here, one more difficulty arises — one text can be published in fragments in different issues of a magazine or newspaper. In that case it becomes necessary to refer to several library units to access a complete text.

7) As for different methodological approaches to determining themes for literary texts, a recent study [3] has proved that expert thematic annotation provides a comprehensive description of text thematic distribution. However, it has one essential drawback, apart from an obvious factor of labour intensity. We mean the subjectivity of thematic annotation made by an expert. The only possible way to reduce this subjectivity is to engage several independent experts into thematic annotation with a follow-up comparison of the results of their work and detection of the strengths and weaknesses of expert annotation. The aim of such an analysis is to distinguish both texts for which there is maximum agreement in the expert assessments and texts on which the experts disagreed. The latter cases seem quite interesting and worthy of further research. The first experience of independent expert annotation of fiction texts by three independent experts was recently conducted with the involvement of students studying philology is HSE University in St. Petersburg.

8) Finally, one of the most challenging tasks of digital text processing is interpreting the received data. We can say that this is one of the key problems in digital humanities actively discussed by its opponents [77]. Indeed, when we shift the focus of our research from individual works to literature in general on a big data scale, in some sense, we lose support of reference expert knowledge about the object of research. There is no such person who could have read all literary texts of the epoch and therefore could accurately assess the quality of the results obtained as well as their acceptability for solving problems in humanities. For the reason above, we leave this question open for now.

VI. Conclusion

In this paper, we discussed methodological approaches to determining and modelling themes of literary works on the national scale, paying special attention to data collection and texts preparing for automatic processing. Besides, we consider technical difficulties arising in the process of completing this task and suggests ways to solve them.

The proposed methodology of data collection and organization has been approved on the base of Russian short stories of 1900-1930s. By the time of preparing this paper, the statistical characteristics of the corpus are as follows: the list of prose writers' names contains about 2800 persons, the authors with at least one digitalized short story numbers 850, texts for more than 600 of writers were specially digitized for the corpus. The total number of stories in digital form is more than 8000.

Both automated and expert topic modelling was approved on the annotated subcorpus including literary texts by 300 writers. Our next step is to perform automatic topic modelling for the whole collected data and to compare the results with that obtained in the process of the pilot experiment described in [3]. Thus, we obtain representative data on the distribution of the main topics of Russian literature in the given historical period and can assess the effectiveness of expert thematic markup in a training sample for machine learning.

However, the success of the assigned tasks largely depends on the methodology of identifying themes in literary texts. From our point of view, for fiction texts a hybrid topic annotation, combining both expert theme attribution and automated topic modelling is the most promising. We may claim that this type of annotation should be used primarily for literary texts as it allows as well to describe relevant aspects of the plot, features characterizing main characters of literary texts, etc. Hybrid topic annotation proves to be a significant extension to tradition multilevel annotation schemes accepted in contemporary corpus linguistics. We suggest that the proposed methodology can be successfully used in the development of artificial intelligence systems that process large volumes of literary texts in any language.

Acknowledgment

The publication was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE) in 2021 (grant # 21-04-053 'Artificial Intelligence Methods in Literature and Language Studies').

References

[1] J. Hirschberg and C. Manning, "Advances in natural language processing", *Science*, vol.349(6245), 2015, pp. 261-266.
[2] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit", *in Proc. of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55-60.
[3] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, and M. Kirina, "Topic Modelling with NMF vs. Expert Topic Annotation: The Case Study of Russian Fiction", *in Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. Part II*, vol.12469, 2020, pp. 134-152.
[4] E.H. Dewart, "Introductory Essay on Canadian Poetry", *in Selections from Canadian*, 1864.
[5] J. Sinclair, Trust the Text: Language, Corpus and Discourse. London and Ney York: Routledge, 2004.
[6] Yu.M. Nagibin (comp.), *Anthology of the Russian Soviet Story*. Moscow: Book Review, 1987.
[7] G.Ya. Martynenko, T.Yu. Sherstinova, A.G. Melnik, and T.I. Popova, "Methodological problems of creating a Computer Anthology of the Russian story as a language resource for the study of the language and style of Russian artistic prose in the era revolutionary changes (first third of the 20th century)", *in Computational linguistics and computational ontologies IMS-2018,* issue 2, May-June 2018, pp. 99-104.
[8] S.D. Harris and Steven Zeisler, "Weak signals: Detecting the next big thing", *The Futurist*, vol. 36 (6), 2002, pp. 21-29.
[9] E. Hiltunen, "Good Sources of Weak Signals: A Global Study of Where Futurists Look for Weak Signals", *Journal of Futures Studies*, vol.12(4), 2008, pp. 21-44.
[10] M. Mahlberg, P. Stockwell, J. de Joode, C. Smith, and M.B. O'Donnell, "CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics", *Corpora*, vol. 11(3), 2016, pp. 433-463.
[11] F. Moretti, *Distant Reading*. London: Verso, 2013.
[12] L. Manivoch, *Cultural Analytics*. Cambridge, Massachusetts: The MIT Press, 2020.
[13] Naczional'nyj korpus russkogo yazy'ka [Russian National Corpus], Web: https://ruscorpora.ru.

[14] A.A. Polikarpov and E.V. Surovtseva, “About Chekhov’s concepts”, *in Izvestia of the Southern Federal University, Philological sciences*, 1 (2009), 218-219.

[15] Shaykevich A.Ya., Andryushchenko V.M., Rebetskaya N.A. Statistical Dictionary of the Dostoevsky Language. M.: Yazyki slavianskoy cultury, 2003

[16] A.Ya. Shaikevich, V.M. Andryushchenko, and N.A. Rebetskaya, “Distributive-statistical analysis of the language of Russian prose 1850-1870s”, *Languages of Slavic culture,* 2013.

[17] A.O. Grebennikov (comp.), A.O. Martynenko (ed.), “Frequency dictionary of short stories by Anton Chekhov”, *Publishing house of St. Petersburg. University*, 1998.

[18] A.O. Grebennikov (comp.), A.O. Martynenko (ed.), “Frequency dictionary of short stories by Leonid Andreev”, *Publishing house of St. Petersburg. University*, 2003.

[19] A.O. Grebennikov (comp.), A.O. Martynenko (ed.), “Frequency dictionary of short stories by Alexander Kuprin”, *Publishing house of St. Petersburg University*, 2006.

[20] A.O. Grebennikov (comp.), A.O. Martynenko (ed.), “Frequency dictionary of stories by Ivan Bunin”, *Publishing house of St. Petersburg University*, 2012.

[21] A.A. Bonch-Osmolovskaya, D. Skorinkin, B. Orekhov, I.S. Pavlova, and M.G. Kolbasov, “Tolstoy semanticized: Constructing a digital edition for knowledge discovery”, *Web Semantics*, 59 (2019).

[22] M. Jockers and D. Mimno, “Significant Themes in 19th-Century Literature”, *Poetics*, 41 (2013), 750–769.

[23] C. Schöch, “Topic "Modeling Genre: An Exploration of French Classical and Enlightenment Drama”, *Digital Humanities Quarterly*, 2 (2017).

[24] I. Uglanova and E. Gius, “The Order of Things. A Study on Topic Modelling of Literary Texts”, *CHR 2020: Workshop on computational Humanities Research*, Nov. 2020.

[25] E. Zamiraylova, O. Mitrofanova, “Dynamic topic modelling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization”, *in R. Piotrowski’s Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conf. on Language Engineering and Applied Linguistics (PRLEAL-2019), CEUR Workshop Proceedings*, vol. 2552, 2020, pp. 321-339.

[26] O.A. Mitrofanova, “Topic modeling of A.N. Afanasjev’s Russian Fairytales”, *Corpus Linguistics*, 2015.

[27] Yu.N. Tynyanov, *Arkhaisty i novatory [Archaists and Innovators]*. Leningrad: Priboi Publ., 1929.

[28] G.Ya. Martynenko, *Metody matematicheskoj lingvistiki v stilisticheskikh issledovanijakh [Methods of mathematical linguistics in stylistic studies]*. St. Petersburg: Nestor-Istoriya, 2019.

[29] V.V. Vinogradov, *On the language of fiction*. M.: Goslitizdat, 1959.

[30] V.M. Zhirmunsky, *Tasks of poetics* / Zhirmunsky V. M. Theory of literature. Poetics. Stylistics. L .: Nauka, 1977.

[31] G.Ya. Martynenko, “Stylometry: Emergency and Evolution in Context of Interdisciplinary Interaction. Part II. The First Half of the 20th Century: Expansion of Interdisciplinary Contacts”. *Strukturnaya i prikladnaya lingvistika*, 12 (2019).

[32] G.Ya. Martynenko, “Osnovy stilemetrii [The Foundation of Stylometics]”, *St. Petersburg State University, St. Petersburg*, 1988.

[33] F. Moretti*, Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso, 2007.

[34] Martindale C., *The clockwork muse: The predictability of artistic change*. New York, USA: Basicbooks, 1990.

[35] M.L. Jockers, “Macroanalysis: Digital Methods and Literary History (Topics in the Digital Humanities)”, *University of Illinois Press*, 2013.

[36] A.M. Ronchi, “eCulture. Cultural Content in the Digital Age”, *Springer, Berlin*, *Heidelberg*, 2009.

[37] G.Ya. Martynenko, T.Yu. Sherstinova, T.I. Popova, A.G. Melnik, and E.V. Zamirajlova, “On the principles of creation of corpus of Russian short stories of the first third of the 20th century”, *in Proc. of the XV Int. Conf. on Computer and Cognitive Linguistics ‘TEL 2018’*, 2018, pp. 180–197.

[38] G.Ya. Martynenko, T. Sherstinova, “Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century”, *in CEUR Workshop Proceedings*, vol. 2552, 2020, pp. 105-120.

[39] Czech National Corpus, Web: https://ucnk.ff.cuni.cz/cs/.

[40] Russian State Library, RGB, Web: https://www.rsl.ru.

[41] National Library of Russia, Web: http://nlr.ru.

[42] Lib.ru: "Classics" (Maxim Moshkov’s Library), Web: http://az.lib.ru.

[43] Digital library Litres, Web: https://www.litres.ru.

[44] Digital library of books Aldebaran, Web: https://aldebaran.ru.

[45] Digital library Litmir, Web: https://www.litmir.me.

[46] Digital Library of IMLI RAS, Web: http://biblio.imli.ru.

[47] Sovlit project, Web: http://www.ruthenia.ru/sovlit.

[48] 3500 Russian prose works (1800-1940), Web: https://doxie-do.livejournal.com/260117.html.

[49] Pochti zabytye (Almost Forgotten), Web: https://slovesnik.org/chto-chitat/pochti-zabytye.html.

[50] Brief literary encyclopedia in 9 volumes. M.: Soviet encyclopedia, 1962-1978, Web: http://feb-web.ru/feb/kle/kle-abc/default.asp.

[51] I.V. Vladislavlev, Russian writers of the XIX-XX centuries. M., 1924.

[52] O.D. Golubeva, From the history of the publication of Russian almanacs of the early XX century. M., 1960.

[53] K.D. Muratova, History of Russian literature of the late XIX – early XX century. M.: AnSSSR, 1963.

[54] P.A. Nikolaev, *Russian Writers 1800–1917, Biographical Dictionary*. M.: Scientific publishing house “Great Russian Encyclopedia”, 2000.

[55] F. Abramenko, Novoe russkoe pravopisanie [New Russian orthography], Polnyj sbornik pravil pravopisanija s uprazhnenijami i kratkimi svedenijami o znakah prepinanija [The compilation of orthographic rules with exercises], Vol. 1, Moscow, 1918.

[56] T. Sherstinova, “Bibliographic database of Russian writers: towards creation of the Russian short stories corpus of the 20th century”, *in Proc. Corpus linguistics-2019,* 2019, pp. 439-447.

[57] G. Martynenko and T. Sherstinova, “Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture”, *Digital Transformation and Global Society. DTGS 2018. Communications in Computer and Information Science*, vol. 859, 2018, pp. 299-309.

[58] Tolstoy Digital, Web: http://tolstoy.ru/projects/tolstoy-digital/

[59] D.M. Blei, A.Y. Ng, and M.I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research,* 3 (2003), 993-1022.

[60] A. Daud, J. Li, L. Zhou, and F. Muhammad, “Knowledge discovery through directed probabilistic topic models: a survey”, *Frontiers of Computer Science in China*, 4 (2010), 280-301.

[61] O.A. Mitrofanova, “Topic modelling of special texts based on LDA algorithm”, *in*: *Proc. XLII International Philological Conf. Selected works*., 2014, pp. 220–233.

[62] H.M. Wallach, “Topic modeling: beyond bag-of-words”, in Proc. 23rd Int. Conf. Mach. Learn., ACM, 2006, pp. 977-984.

[63] L.M. Rhody, “Topic Modelling and Figurative Language”, *Journal of Digital Humanities,* 2 (2012).

[64] T. Sherstinova, A. Grebennikov, T. Skrebtsova, A. Guseva, M. Gukasian, I. Egoshina, and M. Turygina, “Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900-1930)”, *in: Proc. of the 27th Conf. of Open Innovations Association FRUCT*, 2020, pp. 366-373.

[65] T.G. Skrebtsova, “Thematic Tagging of Literary Fiction: The Case of Early 20th Century Russian Short Stories”, *in: Proc. CompLing Conf.*, 2021, pp. 265-276.

[66] A. Esin, *Prinzipi i Priyemi Analyza Litaraturnogo Proizvedeniya* [The Principles and Techniques of Analysis of Literary Text]. Moscow: Flinta, Nauka [Science], 2000.

[67] B. Tomashevsky: *Teoriya literatury* [The Theory of Literature]. Aspekt Press [Aspect Press], 1996.

[68] A. Zholkovsky and Yu. Shcheglov, “K Ponyatiyam ‘Tema’ i ‘Poeticheskiy Mir’”, *Trudy po znakovym systemam*, 7 (1975), pp. 143-167.

[69] T. Sherstinova and T. Skrebtsova, “Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930”, in: Proc. CompLing Conf., 2020, pp. 117-128.

[70] T.G. Skrebtsova, “Struktura narrativa v russkom rasskaze nachala XX veka” [Narrative structure of the Russian short story in the early XX century], *in: Proc. Corpus Linguistics-2019 Int. Conf*., 2019, pp. 426-431.

[71] Slovar’ psevdonimov [Dictionary of pseudonyms], Web: http://feb-web.ru/feb/masanov.

[72] R. Jacobson, *Vliyanie revolyucii na russkij yazyk* [The influence of the revolution on the Russian language]. Prague, 1921.

[73] S.O. Kartsevsky, *Yazyk, vojna i revolyuciya* [Language, war and revolution]. Berlin, 1923.

[74] A.M. Selishchev, *Yazyk revolyutsionnoy epokhi: Iz nablyudeniy nad russkim yazykom poslednikh let (1917–1926)* [The language of the revolutionary era: From observations of the Russian language of recent years. (1917–1926)]. Moscow: Rabotnik prosveshcheniya, 1928.

[75] E.D. Polivanov, "Revolyuciya i literaturnye yazyki soyuza SSR" [The revolution and the literary languages of the USSR], *Za marksistskoye yazykoznaniye*, Moscow: Federatsiya, 1931, pp. 73–94.

[76] G.Ya. Martynenko, "Stilizovannyye sintaksicheskiye triady v russkom rasskaze pervoy treti XX veka" [Stylized syntactic triads in a Russian short story of the first third of the 20th century], *in Proc. Corpus Linguistics – 2019. Int. Conf.*, 2019, pp. 395-404.

[77] S. Fish, "Mind Your P's and B's: The Digital Humanities and Interpretation", *in The New York Times* (23 January 2012).

[78] B. Navarro-Colorado, "On poetic topic modeling: extracting themes and motifs from a corpus of Spanish poetry", *Frontiers in Digital Humanities*, 5 (2018).