Georgy Vekshin, Egor Maximov, and Marina Lemesheva

Poeticisms and Common Poetic Discourse in the Digital *Russian Live Stylistic Dictionary*

Abstract: The use of a word in a specific sociocultural environment makes it a marker of that context and of the corresponding typical speech role. Is it possible to create an automatic detector of the poet's role in a text? The Russian poeticisms discussed in this chapter constitute a layer of vocabulary and phraseology that is optional for poetry but indispensable for authors who position themselves as poets and try to make their texts sound as poetry-like as possible. In Russian culture, this stratum is mainly used in common poetic discourse, the popular tradition of naive versification. The technology for poeticism detection implemented in the *Russian Live Stylistic Dictionary* and described in this chapter opens up possibilities for the essential stylistic differentiation of poems and the preliminary assessment of their aesthetic quality.

1 Introduction

The rapid advances made by corpus linguistics in recent years have allowed us to set ourselves the task of creating electronic dictionaries that automatically

Acknowledgments: We are grateful to all of those with whom we have had the pleasure to work on this project: Polina Morozova, one of the project's initiators, linguists Valentina Ledenyova and Elena Kukushkina, web designers Mikhail Gertzev and Nikolai Tzapkin, and others.

The work has been carried out within the framework of the St. Petersburg University research project, "Study of Vladimir V. Nabokov's Literary Heritage in an Interdisciplinary Perspective Using Information Technology Methods" (ID 72828386).

This project was supported by RFBR, grant no. 17-04-00421: Linguistic Development and Creation of the Electronic *Russian Live Stylistic Dictionary*.

Georgy Vekshin, Department of Russian Linguistics and Literary History, Faculty of Philology St. Petersburg University Saint Petersburg, Russian Federation/Moscow Polytechnic University, Moscow, Russian Federation, e-mail: philologos@yandex.ru
Egor Maximov, Department of Aerophysics and Space Research, Moscow Institute of Physics and Technology, Moscow, Russian Federation, e-mail: egor.maksimov@phystech.edu
Marina Lemesheva, Department of Russian Linguistics and Literary History, Moscow Polytechnic University, Moscow, Russian Federation, e-mail: lemesheva.m@list.ru

Open Access. © 2022 the author(s), published by De Gruyter. BY-NC-ND This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
https://doi.org/10.1515/9783110781502-009

create a multidimensional stylistic portrait of a language unit, taking into account all the features of its sociocultural use. This solves the problem faced by the compilers of traditional dictionaries when it comes to describing the stylistic potential of words. Existing printed dictionaries now provide clues (stylistic marks), which, firstly, do not cover all types of stylistic coloring of the word; secondly, such clues are not based on an objective picture of the communicative practice of society but on the individual vision of the idea held by the compilers; finally, printed dictionaries do not have time to follow the real changes in the sociocultural and affective meaning of words, and are thus unable to quickly represent the social life of a word in its dynamics. This situation could be changed by a modern digital dictionary, representing stylistic variations of a word on the basis of its fixed applications in characteristic contexts, texts, and collocations.

The means to solving this problem from different angles stems from the tradition of the sociolinguistics of genres (M. Bakhtin, A. Wierzbicka); the study of register variation (M. Halliday, D. Biber); the tradition of semantic speech analysis within the framework of Prague functionalism and the Russian "theory of styles," with its emphasis on the sociolinguistics of institutional spheres (K. Hausenblas, M. Kozhina); and the French tradition of stylistic semantics (C. Bally, P. Guiraud); as well as the corpus study of sociolects and ethnolects, and work on corpus research into tonality and topic modeling. At the same time, there is an extremely wide range of methods on offer to describe language sociolinguistically and semantically. We still have no universally accepted criteria for describing and defining the socially and affectively determined semantics of a word or other language units.

This chapter proposes taking an approach to the description of semantic structure and to the automatic identification of lexical units determined by one of the spheres of sociocultural interaction universal to European culture – the field of verbal art, the specificity of which is most clearly presented in the field of poetry, which in the mind of a naive speaker is equated with verse-composing practices. In accordance with the method described below, it is poetic works with their most obvious features of versification (accentual-syllabic meter, rhyme) that will be included in the corpus of the dictionary we propose.

The techniques developed for the automatic identification of poetically determined semantics and pragmatics of a word in the *Russian Live Stylistic Dictionary* (http://livedict.syllabica.com; hereinafter referred to as the "*Live Dictionary*") project may prove useful for the prospect of using corpus methods to identify words and other linguistic units as deictic pointers to typical sociocultural contexts and as markers of communicative image, social status, and the cultural "self" of the author.

2 Theoretical background

2.1 Denotative core and stylistic periphery of meaning

It is well known that meaning as a linguistic phenomenon is the result of the use of a sign in speech contexts: "the meaning of a word is its use in the language" (Wittgenstein, 2009: 25[e]). The speaker is guided by the memory of the sign, extracting the word from its repository as already marked by its typical use and then using it in a real, unique situation. The description of a word's or idiom's semantic features accepted in this work takes into account the fact that the meaning of a word is formed by the restrictions and preferences for its use within corresponding utterances in typical situations, including not only nearby pragmatic contexts but also typical contexts of institutional action and personal emotive condition. This description is based on a three-level model of the semantic structure of linguistic units, which distinguishes between layers 1) denotative-significative semantics (objective logical core), 2) stylistic coloring (semantic periphery of level 1 – sociocultural and affective contextual-role semantics), and 3) the connotative semantic periphery of level 2, which following Apresyan (1995) is understood as associative semantics formed by nationally specific contexts.

The area of semantics that the *Live Dictionary* corpus and the dictionary itself is designed to reveal is an area of stylistic sociocultural and emotive coloring. The stylistic coloring of linguistic units (cf. Bally, 1921; Leech, 1974; Dolinin, 1987; Vekshin, 2017) is formed on the basis of the indexical ability of the linguistic sign under the influence of typical situations and roles of two kinds: typical sociocultural frames and roles (both those universal to culture and more specific), and typical affective states and the corresponding emotive roles ("I am a scientist"; "I am a professional"; "I am a woman"; "I appreciate," etc.).

Using templates for pragmatic word description (Wierzbicka, 1996; Goddard, 2019), information expressed by stylistic coloring can be described as follows using the example of poetry:

1. I know that the same thing can be said in different ways depending on the tasks of the speaker and the conditions of communication.
2. I say this as poets and people who write poetry usually say it.
3. I want you to believe that it is a poet speaking and that we are in a situation of poetic creativity.

For Russian socioculture, the following universal typical contexts (those that inevitably determine the life and behavior of any bearer of a given national culture) are considered the most influential: 1) the context of family relations (intimate communication and cognition) in contrast to distant, societally institutionalized

communication; 2) legal and official relations associated with the state (the exercise of state power is an objective pole of the social space); and 3) political and ideological relations (maintaining and redistributing power – the subjective pole of the space). These are accompanied by three contexts that provide cognitive activity and are formed by it: 4) science (the rational-logical mastery of nature), 5) religion, and 6) art (the objective and subjective "poles" of irrationally exploring the physical and metaphysical world) (cf. Shapir, 1990). The area of semantics that the *Live Dictionary* corpus and the dictionary itself is designed to reveal is its stylistic sociocultural and emotive coloring.

These spheres of the sociocultural space not only relate to the life of every bearer of modern, primarily European culture but together simultaneously form the communicative competence of any person. A person may give preference to some of these areas or specialize in some of them, but they cannot completely avoid activities in at least one of them. A person may not be a scientist, but they cannot but possess basic scientific concepts, for example, they cannot not understand what "temperature above zero" means; they may not be a believer, but they cannot completely isolate their mind from the category of "God." This allows us to speak about the universal nature of these basic contexts. A person may not speak the language of a certain profession, not know the territorial dialect; they cannot be an aristocrat and a peasant, an adult and a child at the same time. However, in order to be a fully-fledged bearer of culture, any bearer of it must, to a greater or lesser extent, live everyday family life and obey the laws of the state, etc. In total there are six such universal spheres. They are grouped in a certain way and make up a system (see Fig. 1). While state and political activity constitute the objective and subjective poles within the single social space, in the same way, religious and aesthetic activity form two the poles of the mythological, extralogical knowledge of absolute and metaphysical reality. This is not the place to talk about the peculiarities of interaction or the axiological properties of these spheres and semiotic systems in different cultures, where their properties differ. It is enough for us to point out that six main institutional contexts determine the universal segmentation of the communicative space of European cultures, which is relatively independent of their more special communicative spheres and is superimposed on them.

In addition to these six basic contexts, the *Live Dictionary* is designed to identify typical, non-universal contexts (those defining social life and behavior but not necessarily encompassing the lives of each member). These are 1) social-estate contexts (bourgeois, peasant, aristocratic, lumpenproletarian, intellectual, etc.), 2) professional (programmers, school-teachers, etc.), 3) geographical (Russian South, St. Petersburg, etc.), 4) gender (contexts of male and female communication),

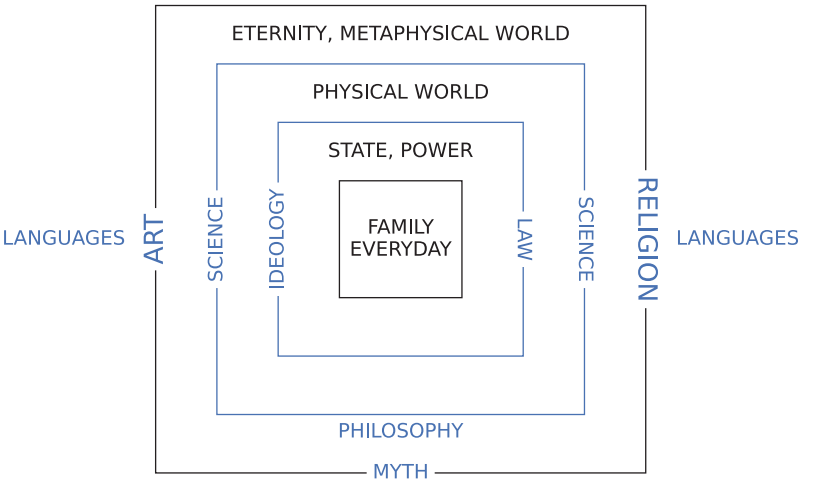

Fig. 1: The basic sociocultural frames and languages of culture.

5) age (contexts of the elderly, children's communication, etc.), 6) chronological (archaic, of the modern era, the latest contexts), 7) xenological (contexts putting in focus the opposition between cultural nativity and alterity – what is one's own, national, and what is alien, foreign), and 8) genre (typical situations in relation to typical goals and typical means and rituals of communication). For modern, at least Russian culture, these eight types of typical non-universal contexts can be considered the most significant, although this number is not finally determined. All of them are revealed as a relatively closed (chronology, gender) or open (profession, genre etc.) series of features.

Moreover, there are four main modal-evaluative contexts that form the affective component of stylistic coloring, which never denotes emotion but does express an emotional assessment as the speaker's point of view, distant from the essence of the subject, the conceptual content of the sign. We conceive of the emotion of pleasure/displeasure as the most universal, linguistically reflected affect forming the core meaning of desirability/undesirability (approval/disapproval) (1) within stylistic coloring (cf. Osgood, 1990; Russell, 1991, 2003; Wierzbicka, 1999). In the usual interaction with this opposition but also independently of it, three more types of affective meaning dimension in Russian are characteristic of stylistic coloring and are realized in the corpus tagging: (2) evaluation (importance/unimportance), (3) distance (intimacy, proximity/detachment), and (4) accommodation (friendliness/aggressiveness).

Thus, the layout of the *Live Dictionary*'s corpora is shaped by 18 basic contextual role parameters – stylistic primitives of Russian speech.

2.2 *To be* and *to appear* in culture and communication: The basic approach to the formation of a stylistically relevant corpus

Humans are characterized by their desire not only to be but also to appear. These two basic sets of communicative behavior are not independent of each other, but they are polar in their extremes. Society is structured in such a way that appearances and even the imaginary, all kinds of relatively or purely formal indicators of the speaker's role, status, and function (not always real or sincerely fulfilled by them), are an integral attribute of communicative behavior in general. When "the actor identifies with the socially objectivated typifications of conduct in actu, but re-establishes distance from them as he reflects about his conduct afterwards," "the roles, objectified linguistically, are an essential ingredient of the objectively available world of any society" (Berger, Luckmann, 1966: 91). Born as socially objectified scenarios of behavior in typified situations of communication, they are further performed as tools of social self-presentation, relatively free from the pragmatics of the activities and contexts that gave rise to them. This property of the role of behavior is emphasized in the definition of the Jungian concept of the persona as "a complicated system of relations between individual consciousness and society, fittingly enough a kind of mask, designed on the one hand to make a definite impression upon others, and, on the other, to conceal the true nature of the individual" (Jung, 1966: 264).

The peripheral social meaning in the semantic system of a language like Russian is primarily formed by roles as images, more than by roles as functions and identities. Fundamentally, a text that upholds the conventions of academic writing is not the same thing as a scientific one, just as a text that rhymes and is saturated by poeticisms does not yet provide an aesthetic effect:

> If the specific properties of scientific speech were entirely derived from scientific needs proper, then, obviously, the articles and books of the most outstanding scientists would be the most typical examples of the academic writing style. Meanwhile, there is more likely an inverse relationship: real scientists are often inclined to violate the unwritten norms of the academic writing style, while works that are weak in terms of content are most often written quite academically. (Dolinin, 1987: 75)

The success of the act of scientific communication is achieved by means of the language of science – as an operational semiotic system of devices, tactics, and strategies – the rigor of logical constructions, and the explanatory power of conclusions. The pragmatics of science may demand a concentration of terminology in the text if it is needed to build a conceptual and categorical system of knowledge. However, young scholars often do not notice that they are only

inventing a complex term to emphasize their innovativeness, or that they are complicating their syntax too much to simply manifest the depth and complexity of their thought. This paradox can be very clearly observed in Russian scholarly practice. (The authors of this chapter, writing in English, are not always able to successfully fight the stylistic inertia of Russian scientific communication – for example, the excessive use of passive constructions to "objectify" knowledge, sadly noticing that they speak more complexly than pragmatics and simple common sense require.) Of course, this does not negate the possibility or even the appropriateness of using stylistically colored units to achieve the aims of scientific knowledge, but stylistically marked academic elements as such (as well as the exclusion of neutral expressions) are required especially where the writer diligently signals scientific discourse or even feigns it. Thus, there is a tendency toward asymmetry between the scientific quality of thought and the academicity of writing. A similar trend can be observed in literary discourse.

So, the stylistic meaning of linguistic units cannot be deduced directly from the essential pragmatics of institutional communicative acts. Moreover, it is impossible to build and tag a stylistically adequate corpus by relying on the formal classification of texts alone according to their qualification by the author or the bibliographer (this is how most corpora with genre-stylistic tagging are arranged). The basis of the stylistically relevant corpus should be formed by the texts and text fragments that most consistently and clearly manifest a typical social role and actively signal the relevance of verbal action in institutional contexts. To identify the sociocultural significance of linguistic units, not all texts that are nominally related, for example, to science or poetry, must be included in the corpora.

The fact that the sociocultural pragmatics of the text and the nature of the stylistic coloring of its elements do not depend directly on each other can be observed, for example, in advertising texts. Thus, the task of promoting cosmetics, which has nothing to do with the pragmatics of scientific knowledge, is often performed using academic phraseology and syntax (Diez-Arroyo, 2013). Such a text works as a persuasive advertisement because the image of the speaker is built as the image of a scientist, a professional. Instead of expressions such as "for fine and supple skin," terms and verbal nouns will appear in a text: "Sharp temperature changes, environmental pollution and stress make our skin lose its optimum level of moisturization" (Glacier Essence, Sensilis, leaflet; Diez-Arroyo, 2013: 202); "[t]his eye serum contains marine kelp that's meant to lift skin while retinol stimulates collagen to plump the area" (Murad advertisement). A person buying this product does not need to know what retinol or collagen is, but they should have a feeling that it was a professional who advised them to use it. Such advertising

tactics may even be explicit: “Discover a dermatologist’s way to reveal fresh, new, healthy skin” (L’Oreal advertisement).

If the text is saturated with units of a scientific coloring, sustained in a single academic writing style, and corresponding to an overall image of the author as a scholar, even though it might never be formally attributed to science, it can enter the scientific corpus of the Stylistic Dictionary. Conversely, popular scientific texts, which are as accessible as possible, explaining the nature of things in a trusting, friendly tone, will not be included in the scientific corpus of the *Live Dictionary*, since the role-playing, image side of these texts correlates with the image of a close friend, not a scientist. If any text or fragment of text, regardless of its institutional pragmatics, is kept in a single informal register and embodies the image of a loved one with the help of colloquial markers, it will be included in the conversational corpus. Such a dictionary corpus, for example, in relation to its other corpuses, will automatically detect colloquial markers that tend to be used in everyday contexts.

2.3 Poetry, poeticity, and poetic corpus: The semantic structure of Russian poeticism

The language of literary art as a semiotic operational system, as a technique of “estranging” the cognition of verbal and extra-linguistic reality in its metaphysical perspective, with its unique techniques and tactics (Shklovsky, 1990; Hansen-Löve, 1978, etc.), is not accessible to every native speaker. However, everyone has some idea of what poetry is, of how it differs from other types of speech. The national literary language as a public domain includes elements that native speakers, regardless of their ability to understand art, associate with artistry, poetry as a cultural institution, and with the way they think a typical poet should speak. For example, a composer of amateur congratulatory poems will be guided by this norm, including accentual syllabic meter, rhyme – albeit flawed – and a certain kind of vocabulary and phraseology; these are popular markers of poetic diction. We label these markers poeticisms. Their presence in the text does not necessarily mean that we are dealing with verbal art, although it does not automatically mean that we are dealing with a sample of amateur writing. Modern Russian poetry may use poeticisms as well as other linguistic means within the frame of its artistic tactics, which may include the task of portraying the typical speech role of the author as a poet (along with any other possible roles). However, poetry and the poetic on the one hand and poeticity and the poetical as signals of the poet’s speech role on the other are radically different concepts.

Poeticisms, being part of mass cultural consciousness and containing poetical coloring as a component of their stylistic semantics, are partially taken into account and described as such by traditional dictionaries. Meanwhile, the accuracy and adequacy of their presentation in conventional dictionaries entirely depend on the mindset of their compilers, which is not only subjective but also quickly becomes obsolete. Words marked "poetic" usually include those implying the meaning of "high," "solemn." However, genuine Russian poeticisms as exponents of a poet's speech role go far beyond these stylistic classes (Vekshin, Lemesheva, 2019).

The *Live Stylistic Dictionary*, aimed in particular at the objective automatic recognition of poeticisms, uses a poetic subcorpus that, owing to the style set technique (see 3.2), primarily includes texts that manifest the speech role "I am a poet, a composer of verse" and that are recognized as poetry due to their poeticity. In this regard, the poetic corpus of the *Live Dictionary* covers the widest range of poetic texts – from high poetry to graphomania – but, first and foremost, those that the majority of Russian speakers will qualify as typical verses and lyrics expressed by way of a poet's typical speech, for poetry outside the verse and lyric genre is practically inexistent for naive native speakers.

The poetic subcorpus of the *Live Dictionary* is part of the corpus of fiction, characterized by a combination of tags such as fiction, verse, and lyrics. One necessary and sufficient sign of a poeticism is its poetic social coloring. Such, in particular, are the lexical and phraseological markers of poetry: *безбрежный* (shoreless), *безмолвие* (quietude), *брожу* (I roam), *былое* (yore), *взор* (gaze), *вериги* (chains), etc. The main semantic component of poeticism is actualized, for example, in the following constructions: *У нее не взгляд, а взор; Мы говорили не о прошлом, а о былом; День был не удивительный, а дивный*. Something similar to these expressions can be represented in English sentences: *It was not a holiday, but a feast. He was not crying, but weeping.*

3 Related work

3.1 General approaches

In works on automatic genre identification (cf. in particular Stamatos et al., 2000; Santini, 2007; Sharov, 2018, etc.), a fully automated approach based on the n-grams method has been proposed, which was designed to capture nuances of style, including lexical variation. However, grammatical and formal indicators (verb, substantivity, share of functional words, average word length, sentences,

etc.) are considered the main ones, while lexical indicators (high-frequency words) are treated as non-universal and are used only as an addition to the main set of parameters since they are considered subject-dependent (Ljashevskaja, Sharov, 2009). Contextual role-based sociocultural parameters of speech are obviously in some coordination with the topic, but they act as an independent and powerful factor in style formation. When combined with stylometric data, they can be very useful in the attribution of texts and the determination of authorship and individual style.

In contrast to the approaches adopted in automatic genre identification, the *Live Dictionary* fundamentally distinguishes between speech genres as culturally patterned and rigidly pragmatically determined, textual practices (complexes of typical textual means in typical situations to achieve typical goals) (cf. Bakhtin, 2011; Wierzbicka, 1985; Günthner, Knoblauch, 1995; Vekshin, 2017; and others) on the one hand and the complexes of universal sociocultural role markers (in Russian and Czech traditions often called "functional styles") on the other. Thus, identifying the text as belonging to the genre of the church sermon (an important reference point here is the formal name of the genre of the text and its typical pragmatics), which makes it possible to assign the genre tag "sermon," does not interfere with the text being simultaneously assigned to the religious corpus, if the speech image of the preacher is primarily constructed as "I am a believer," or to the spoken corpus, if the dominant speech role in the text is "I am a person close to you." Genres and style are phenomena of a different order, which is why the genre and style markup of texts for the *Live Dictionary* corpus are carried out independently of one another.

The sociocultural style, with its exceptional contextual role determinant, and the speech genre are phenomena not only of a different hierarchical order but also of a different nature. This is not usually taken into account in works on register analysis (Halliday, Hasan, 1985; Martin, 1993; Biber, 1993) and is also reflected in corpora classification and tagging systems. In the *Russian National Corpus* (RNC; http://www.ruscorpora.ru), prose is included in the main body, and poetry is presented as a separate one, along with dialect (the subcorpus of territorial varieties of the language) and newspapers (the collection of texts of any genre limited to a specific print source). The RNC poetic subcorpus is made up exclusively of high-quality, professional poetry, striving to overcome the canon, often intentionally creating stylistic contrasts, combining elements of different sociocultural styles to implement artistic tasks. This corpus may be a source of data on the frequency of words used in Russian poetry at different times, on the keywords of certain authors, but we can only partially judge the stylistic semantics of a word to the extent that these texts embody a poet's typical, stable speech role (despite the fact that professional poetry normally does not use such make-believe tactics).

We hope that, to understand the sociocultural use of the word as a whole, the *Live Dictionary* corpus, compiled based on the role context factor and using the styleset method, which will be described below, is much more indicative.

The list of the 50 most significant Russian poeticisms obtained on the basis of the *Live Dictionary* is a series in which we do not find a single random element and which includes, in addition to frequency, poetic concepts and formal operators, pure carriers of poetic sociocultural coloring: *ты* (you), *словно* (as if, like [poet.]), *мне* (me), *сердце* (heart), *над* (over, above), *солнце* (the sun), *небо* (sky, heaven), *снег* (snow), *всё* (all, everything), *свет* (light, world), *как* (as, like), *лишь* (only [poet.]), *осень* (autumn), *где* (where), *ночь* (night), *любви* (love [dat., gen., abl. loc.]), *жизнь* (life), *иль* (or [poet.]), *чтоб* (so, for [poet.]), *душа* (soul), *вдруг* (suddenly), *вновь* (again [poet.]), *ветер* (wind), *сквозь* (through), *тобой* (you [abl. instr.]), *будто* (as if, like [poet.]), *дождь* (rain), *ни* (nor), *боль* (pain), *любовь* (love), *душе* (soul [dat., abl. loc.]), *глаза* (eyes, eye [gen]), *снова* (again), *миг* (moment, blink, about time), *тебя* (you [gen.]), *как будто* (as though), *твой* (your), *не* (not), *птицы* (birds), *счастье* (happiness), *мной* (me [abl.]), *моей* (my [abl., fem.]), *ночи* (night [gen., abl. loc.]), *души* (soul [gen.]), *он* (he), *дом* (home), *ль* (if, whether [poet.]), *мой* (my [masc.]), *моя* (my [fem]), *лес* (forest) (see comparative data on frequent lexemes in Russian naive poetry and lexemes dominant in the RNC poetic subcorpus in Bonch-Osmolovskaya, Orekhov, 2013).

3.2 Method of corpora formation

The theoretical apparatus described here is the basis of the methodology for the formation and labeling of the *Live Dictionary* corpus. Eighteen types of Russian elementary stylistic meanings, which reflect the corresponding types of sociocultural contexts and emotional states, require the building of 18 dictionary corpora. For texts reflecting non-universal contexts, subcorpora (for example, professional or genre) are collected. Each of them should include at least 1,000 texts. Crucial for fulfilling the main tasks of the *Live Dictionary* are six universal, basic sociocultural contexts (conversational, administrative, ideological, academic, literary/poetic, and religious). We have assembled these cases most deliberately. Since the markup of any text includes 18 tag types, other corpora will also be formed in the process of compiling the six main corpora; however, the deliberate choice of texts for a particular corpus remains most effective since the principle of maximum stylistic uniformity of the text is being observed here.

To build the corpora, experts are using the "styleset" principle (Avamilova, Vekshin et al., 2019). The main feature of this method is that it excludes certain selections of texts according to their formal classification and explicit attribution

and requires only those texts that most typically exhibit the typical speech role of the speaker. That is why, for example, not all articles published in scientific journals can be selected for the scientific corpus. Only those papers and their fragments that actively use the style of a word to create the typical speech role of a scientist will get into the corpus. And in this case, articles by novice scientists who are very concerned about their speech role and seek to demonstrate their scholarship will make their way into the *Live Dictionary* over texts by major researchers. Thus, the compilers of the *Live Dictionary* are guided in principle by texts where the author seeks "to appear" much more than "to be." These texts turn out to be the most saturated with stylistically specific vocabulary and phraseology. And the frequent appearance of any word in such contexts will ultimately be a guide for the stylistic identification of a unit as a result of machine learning.

The second feature of the styleset method is the expert's work algorithm, which involves the initial formation of search queries consisting of five to seven words or expressions exclusively specific to this context and speech role. The expert's next main partner is then the web search engine, returning texts from which the expert selects those that are stylistically most homogeneous, with the most clearly expressed desire on the part of the author to play a corresponding speech role. The expert, firstly, selects these texts for the corpus (sometimes not the whole texts, because we require the most stylistically typical fragments). Secondly, in these texts, he or she looks for the most striking markers of contexts and roles, then uses them for new queries.

To give an idea of this process, we will try to use the English poetic styleset we have chosen intuitively: *misty purple wane glory light restless*. The algorithm of action will be as follows: after sending the request, poetic texts are returned. These are, in particular:

- The Complete Poems of Emily Brontë (https://en.wikisource.org/wiki/The_Complete_Poems_of_Emily_Bront%2526%2523x00EB)
- Songs of the Sea Children / Bliss Carman [electronic text] (https://quod.lib.umich.edu/a/amverse/BAC8020.0001.001?view=toc)
- Forest Buds: From the Woods of Maine, Elizabeth Akers Allen (https://quod.lib.umich.edu/m/moa/ABK0842.0001.001?rgn=main;view=fulltext);
- Victorian Women Writers Project: The Dream, and Other Poems, Caroline Sheridan Norton, 1808–1877. (http://purl.dlib.indiana.edu/iudl/vwwp/VAB7052);
- Songs – Song – Wedgeblade.net (collection of lyrics); and others.

Further, in Bliss Carman's cycle "Songs of the Sea Children," chosen because of its general stylistic poeticity (regardless of its pragmatics – ironic or serious), we find the most poetical words and phrases: *joyous soul, golden April, fare-*

thee-well, twilight on hills, without thee, rose of dawn, hollow jar, and others. They will fall into the styleset base for the formation of new stylesets and will also be used to expand the further search for texts. Moreover, the most stylistically specific poetic texts will be selected for the poetic corpus. Please note that, as stated above, Russian poeticisms are undoubtedly more active in modern speech as indicators of the role of the poet than in English, and the status of a poeticism in modern English speech differs greatly from its status in Russian – it is more of an exotic element than a fact of modern literary language and mass versification practices. Therefore, in the case of a similar search on the Russian internet, we will receive a large number of today's amateur poems in which the author seeks to sincerely implement his speech role as a poet.

Stylesets include predominantly poetically colored units as well as thematic conceptual words, and, finally, words and phrases that are simply frequent in poetry. To make the styleset base more complete and objective, an expert could resort to the data of lexicography, which widely uses the mark "poetic," as well as "high" (Kourova, 2016). However, it should be noted that the stylistic marks of dictionaries often suffer from inaccuracy and much more subjectivity than the intuition of a modern native speaker, and are also archaic and usually do not take into account many new trends in the use of words (Vekshin, Shilihina, 2017). Therefore, we draw from these sources with great care.

In addition, to add typical poetic concepts and characters to the database, dictionaries of poetic language may serve as a support (*Dictionary of the Language of Russian Poetry*, 2001–; Ivanova, 2004; Pavlovich, 2007) as well as the most significant linguistic studies of poetry and authors' individual style.

The most stylistically homogeneous texts or text fragments selected from the search results are further subjected to double processing. Firstly, units are extracted from them to form new stylesets and further replenish the corpus. In the immediate context (within the limits of one poem), for example, other poetical words and phrases are supposed to appear, obeying the rule of stylistic attraction. An expert can verify the correctness of the "linguistic flair" by using an additional web search, which allows us to understand whether a given word or combination of words is mainly unique to poetic texts or is also regularly found in utterances of other pragmatics, texts of non-poetic genres (for example, religious). When solving this problem, the poetic subcorpus of the *National Corps of the Russian Language* is also of great help.

Secondly, the selected texts are tagged by experts in accordance with the established parameters, which are divided into two blocks: 1) factual information about the text and 2) its stylistic features. Factual information includes attribution: name, source, author (name, gender), and date. These data can serve, in particular, as guidelines for automatically reconstructing the picture of the dynamics of

the use of a language unit. In addition to the main sociocultural features, the stylistic tagging of texts requires us to determine their narrower social and genre specificity: gender, age, profession, estate, areal, xenological, chronological (in relation not to the actual historical period but to the one recreated in the text), and, finally, genre proper. Xenological stylistic coloring is a specific semantic parameter of the Russian language unit, which is used as a deictic indication of its belonging to a foreign cultural environment (first of all, European, due to which the word also forms a modality of importance), when its foreign cultural origin is tangible to native speakers. These include, for example, Gallicisms and the latest Anglicisms in Russian, and Church Slavisms in archaic vocabulary. Semantics of the latter type, combined with the coloring of poeticism, strengthens it in this status and generates the meaning of "high."

Thus, a combined stylistic portrait of a linguistic unit, which can be compiled with the help of a dictionary, will reveal not only poeticisms in general but also those that are characteristic, for example, of female poetry or a folkish poetic style.

3.3 Style identification

The word classification problem can be considered a word representation learning procedure. The majority of modern word representation algorithms are based on neural networks (Joulin et al., 2016; Mikolov et al., 2013; Devlin et al., 2019; Peters et al., 2018). These algorithms are able to learn word representation using the context. However, the aforementioned methods are unsupervised. That means we would not be able to obtain the necessary style features of the word from its representation. Due to this disadvantage, we cannot use these methods in our dictionary. At the same time, every single word may be considered as a text that contains only one word. This allows us to use a text classification model to predict word labeling.

The majority of modern approaches to the text classification problem are based on recurrent or convolutional neural networks (Zhang et al., 2015; Liu et al., 2016; Conneau et al., 2016; Howard, Ruder, 2018; Lai et al., 2015). This means that the model takes into account not only a single word but also the word order. For this reason, the model cannot be used in our case.

Another group of algorithms uses topic modeling as a preprocessing step for text classification (Neogi et al., 2020; Li et al., 2018; Pavlinek, Podgorelec, 2017). In these approaches, a text is represented as a vector in a low-dimensional feature space. Some topic modeling algorithms, like LDA (Blei et al., 2003) or PLSA (Hofmann, 2013) are based on the co-occurrence of words in texts. These approaches

are implemented in different libraries for NLP (Vorontsov et al., 2015; Egorov et al., 2019; Loper, Bird, 2002; Rehůřek, Sojka et al., 2011). However, topic models are also unsupervised, so we cannot control the result of the representation.

To solve our problem, we require supervised text classification approaches that take into account the word presence but not the word order. The classical approaches based on the bag-of-words model (Zhang et al., 2010; Harris, 1954) have these properties. The more advanced approach uses TF-IDF (Spärck, Jones, 1972).

4 Proposed corpus

The first important task in creating such a complicated computer system as the *Live Stylistic Dictionary* is to collect data on which the machine learning model can later be trained. To solve this problem a new corpus needs to be created. This corpus must contain texts of different styles and genres, written by different authors in different periods. One can download an up-to-date version of the *Live Dictionary* corpus from our official website (https://livedict.syllabica.com).

4.1 Overview

Each text should be properly labeled according to the following features: title; source; date of writing; type of source (e.g., internet, newspaper, etc.); gender of the author; typical social and pragmatic context affiliation (style); social stratum; age; occupational, regional, gender, xenological, and chronological specificity; speech forms (dialogue/monologue; verse/prose; phrase/text); genre characteristics.

The corpus was created and labeled by our group of linguists – experts in stylistics. It currently contains more than 8,000 texts. A labeling process is carried out on a website. An expert pastes in the text and then fills out a simple form where the correct category for each feature must be selected. For some features, such as occupational specificity, a single text may belong to multiple categories. Such texts, for example, may be written by representatives of several professions. The "style" feature is considered essential to the *Live Dictionary*. This feature has been labeled more accurately by the experts, and all the algorithms will be tested on it first. According to this feature, the text may be classified by six categories: colloquial, business, ideological, scientific, religious, fiction (Fig. 1). The latter includes a poetic subcorpus, which is formed as a combination of texts with the following features: fiction + verse + lyrics. Other features also contain different subcategories, but their labeling is a work in

progress. The number of texts in the corpus so far has been modest. However, the corpus is constantly growing, and more ideological and fictional texts, as well as texts with different genre tagging, will soon be added. Therefore, all the advantages will be shown below on the example of text "style" feature.

4.2 Structure

All the data is stored in the SQLite database file. The data is stored in the "question" table. Each text is described with 23 features from "field0" to "field22." Each categorical feature has its own description in the appropriate table.

To speed up computation, each text is stored in its own separate file. The filename is stored in the "question" table in the "field6" column. In the event

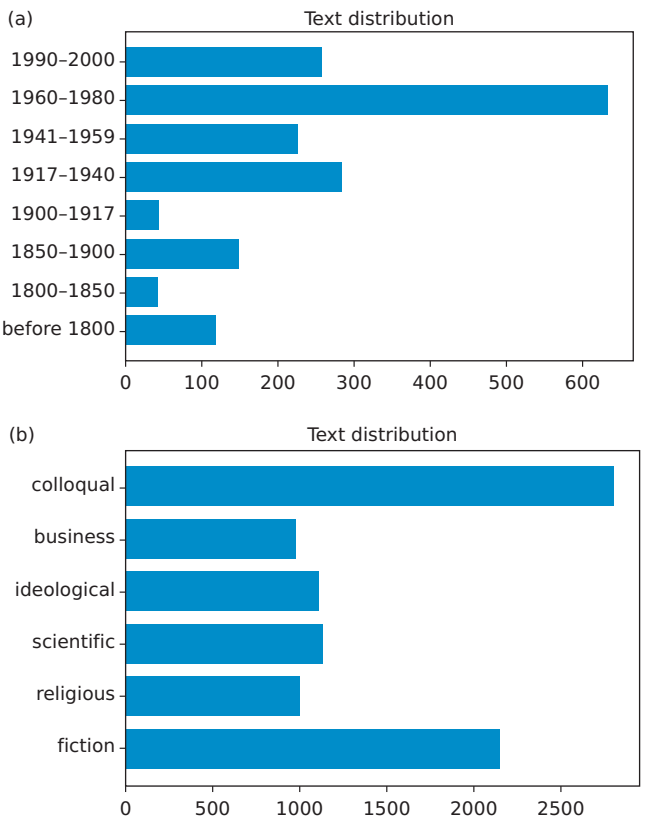

Fig. 2: Distribution of texts in the corpus by a) date and b) style.

that a single text has multiple labels for some features, these labels are separated by a comma.

5 Method

The *Live Stylistic Dictionary* service is based on two different assignments: text style identification (a multitask text classification problem) and single-word style identification. While we do have a labeled corpus to solve the text classification problem, we do not have any labeling for single words, so we have to take a kind of semi-supervised approach to word classification.

There is a huge variety of approaches to perform text classification (Zhang et al., 2015; Kowsari et al., 2019; McCallum et al., 1998; Ikonomakis, Kotsiantis, Tampakas, 2005). These methods perform rather well, and we will not discuss them any further. The main goal for us was to build a word classification pipeline.

5.1 Basic approach

We chose a classic approach based on the bag-of-words model and TF-IDF. TF-IDF is used for feature selection. We trained a logistic regression algorithm to predict the text category. As the algorithm is trained on a word presence vector, the weights of the model indicate the importance of a single word for obtaining a classification result. The probability that the text belongs to a certain category may be described using the following formula:

$$P(T) = \sigma\left(\sum_{i=1}^{N} \omega_i I(\omega_i \in T) + b\right)$$

In this formula, T denotes the text, ω_i is the weight of the words ω_i, $I(\cdot)$ is the indicator function for a word ω_i to appear in the text T, $\sigma(x) = \frac{1}{1+e^{-x}}$.

If ω_i is positive, the i-th word is more likely to appear in the text. We use the *Live Dictionary* to store the information about each term. This ID determines where the term's weights are held among all model weights (Fig. 2). For each term there is a weight corresponding to the specific category of a certain feature.

We also use the 2-3-gram model (Broder et al., 1997). This may help to improve the classification quality and allow us to classify different word combinations such as *в шаговой доступности* (a stone's throw), *превзойти ожидания* (to exceed expectations), etc.

This approach poses some challenges. The first problem is that, if we face an unknown word, we are not able to say anything. The second problem is the multiple forms of single words. For this approach, no stemming or lemmatization is used, because each form of the word may contain extra information that could help to classify the text. Nevertheless, it is also a problem, for the *Live Dictionary* becomes extremely large. A dataset of 4,000 texts contains more than 8 million unique terms. Moreover, in that case, we cannot say anything at all about some rare forms of a common word.

5.2 Morpheme-based approach

To overcome the difficulties discussed above, we use another approach to text preprocessing inspired by a number of authors (Joulin et al., 2016; Schütze et al., 1993; Sennrich, Haddow, Birch, 2016).

Every single word consists of letters and combinations of letters (character n-grams). These character n-grams form larger segments of the word that are called morphemes. There are several kinds of morphemes in the Russian language (prefix, root, suffix, postfix, and flection), which are located differently within the word. Moreover, a word might not contain any morphemes except the root or may have more than one morpheme of the same type (excluding flection and postfix).

Different morphemes can carry some stylistic information. Let us look at different forms of a single word. The word *кот* (cat) has many different derivatives such as *котёнок* (kitten), *котик* (pretty cat), *котейка* (nice cat, mostly used on the internet and in feminine discourse), *котэ* (cat, used on the internet by young people), *котяра* (something akin to a large, old cat, mostly used in masculine discourse), *котенька* (lovely little cat, mostly used in feminine discourse or in folklore), *котофей* (cat, in folklore), *котище* (large cat, used in common speech), etc. Some of the morphemes used in these words are absent in all morpheme dictionaries as they have emerged on the internet, where the language used is quite different to ordinary language.

Using morpheme features in classification may give us more accurate classification results. To find all the possible morphemes, we count all character n-grams of length three to six presented in the word. We also include prefixes and suffixes of lengths of up to four in the model as we have assumed that prefixes and suffixes contain important information. The morpheme length parameters are chosen on cross-validation by grid search.

This approach allows us to reduce the dimensionality of a feature space from more than 8 million terms to about 1 million terms. This makes the learning

process much quicker and also improves the quality of the text classification algorithm to 87% accuracy on the style feature.

6 Evaluation and discussion

6.1 Word style identification

There are two different variants of a single-word classification task for a model trained on morphemes. The first approach is to make a prediction for a text of a single word and then subtract the result from the classification result of a text with no words. Here, a classification result for an empty text represents the prior distribution of classes in the corpus learned by the model, and the difference represents the influence of a text on a classification result. This approach also helps us to classify word combinations.

The second approach is to take the weighted sum of the model's weights as it is done in the bag-of-words approach. Here, another hyperparameter appears – the weights of the character n-grams. If the weights are equal, we take an average model. It is rather simple, but it considers all morphemes and parts of words to be equally valuable for classification, which can hardly provide good results. Therefore, another rule may be used: the longer the character n-gram is the more valuable it is. That is better because having some information about the whole word is far more important than having some information about the suffix. On the other hand, if the word is unknown to the system, and we do not know anything about its major part, it may be classified according to its morphemes.

A comparison of these approaches to word classification is presented in Fig. 3. Here, we show the comparison of three different approaches to word classification by columns: character n-gram (Char n-gram), TF-IDF vectorizer without stemming (TF-IDF), and TF-IDF vectorizer with stemming (TF-IDF+stem). The first word was classified quite correctly using all three approaches. The only model to classify the second word correctly was the TF-IDF vectorizer as this model had already seen the word in this form. The third word was classified correctly by the first two approaches, and the third gave us a rather uncertain result. In the fourth word, there was an error, and neither the second nor the third model was able to overcome it. However, the first classifier made the correct decision. The first classifier interpreted the fifth word incorrectly, but the second one was not able to give any result. This is because the required form of the word was not presented in the dataset. Therefore, stemming generally gives us the worst result among all the classifiers. We can only apply it to classify the words occurring in the different forms,

but it may lead to errors. The second approach performs rather well, but only with words it has encountered before. The first approach gives us some misinterpretations, but it can work with words it has never encountered before.

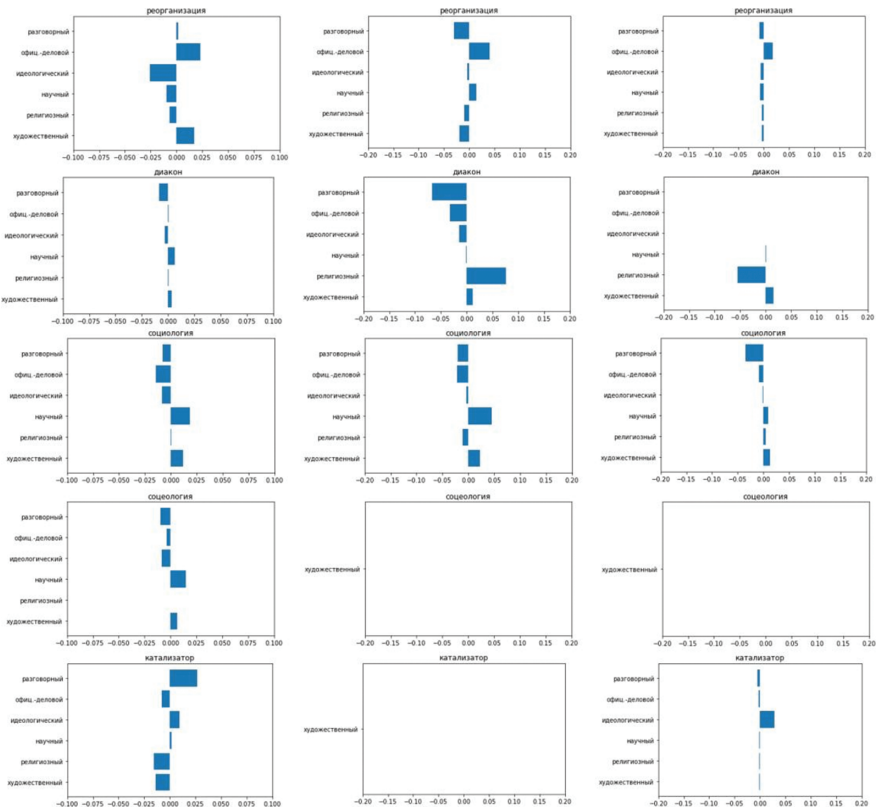

Fig. 3: Predictions of different models by columns: 1. Char n-gram; 2. TF-IDF; 3. TF-IDF+stem Glosses: реорганизация (reorganization), диакон (deacon), социология (sociology), катализатор (catalyst).

6.2 Text classification

As previously mentioned, linear models trained on a bag-of-words model or character n-grams can be applied in text classification. Models trained on character n-grams show higher accuracy on cross-validation as they count not only the cases of word presence but also the forms of words. Tab. 1 shows a comparison of the accuracy of these methods. However, these methods cannot provide

Tab. 1: Comparison of different word style identification models in terms of text classification.

Model	Accuracy score on 5-fold cross-validation
Count vectorizer 1-gram	0.8662
Count vectorizer 1-2-gram	0.8574
Count vectorizer 1-3-gram	0.8514
TF-IDF vectorizer 1-gram	0.8033
TF-IDF vectorizer 1-2-gram	0.7635
TF-IDF vectorizer 1-3-gram	0.7396
TF-IDF vectorizer 1-gram+stemming	0.8423
TF-IDF vectorizer 1-2-gram+stemming	0.7920
TF-IDF vectorizer 1-3-gram+stemming	0.7767
Character 3-7-gram+1-4 prefix+1-3 suffix	0.8751

superior quality as they are based exclusively on word presence but do not take into account the sequence of words.

7 Conclusion

In the *Russian Live Stylistic Dictionary*, we offer a number of approaches that can significantly improve the identification of words and phrases as holders of social stylistic meaning, particularly poeticisms – words and expressions with a poetic social coloring, typical of the poetry subcorpus in the *Live Dictionary*'s fiction corpus. From a linguistic perspective, the basis for effectively recognizing Russian poeticisms is the criterion of the unity of the contextual role of the texts included in the poetic corpus. The styleset method helps to extract such texts from online resources and could be the basis for the future automatic detection and parsing of stylistically homogeneous texts and for replenishing corpora. It thus becomes possible to monitor changes in the use of the word and to trace the dynamics of its stylistic meaning, which has not been possible for traditional dictionaries. At the same time, the *Live Dictionary* aims to define the stylistic dominant of a text (the Style Prompter option – https://livedict.syllabica.com/text). A high concentration of poeticisms allows us to speculate that a text is of low artistic value. Combining the features of such texts could present

a universal, very stable repertoire of people who exhibit their sociocultural status and roles, such as poets and scientists. It is thus becoming possible to make a composite sketch of a typical poet, a portrait of the author of mass poetry or naive literature. On this basis, we can carry out primary diagnostics of a poem's artistic merit: compliance with the "zero idiostyle" norm (a lack of personality in style) can point out the mediocrity of the poem, and deviations from it can suggest originality and even a text's uniqueness.

References

Apresyan JD. Konnotatzii kak chast pragmatiky slova [Connotations as Part of Word Pragmatics]. In: Apresyan JD. Izbrannye trudy [Selected Writings], vol. 2. Moscow: Jazyki russkoj kultury, 1995: 156–177.

Avamilova EA, Vekshin GV, Kretov A, Maksimov ES. Algoritmy sostavlenija korpusov "Zhivogo stilisticheskogo slovarja russkogo jazyka" i ego programmnaja razrabotka [Algorithms for Building the Corpora of the Russian Live Stylistic Dictionary and its Software Development]. In: Kniga v sovremennom mire: mesto v kilturnoj paradigme obschestva v uslovijakh tzifrovoj revolucii. Voronezh: VGU, 2019: 4–19.

Bakhtin M. The Problem of Speech Genres. Literary Criticism 2011; 4(15): 114–136.

Bally C. Traité de stylistique française, vol. 1. Heidelberg: C. Winter, 1921.

Berdyaev NA. Sudba Rossii [The Fate of Russia]. Mosscow: Filosofskoye obschestvo SSSR, 1990 [1918]. https://imwerden.de/pdf/berdyaev_sudba_rossii_1918_1990__ocr.pdf (accessed April 1, 2022).

Berger PL, Luckmann T. The Social Construction of Reality: A Treatise in the Sociology of Knowledge. Garden City, NY: Doubleday, 1966.

Biber D. Representativeness in Corpus Design. Literary and Linguistic Computing 1993; 8(4): 243–257.

Biber D, Conrad S. Register, Genre, and Style. Cambridge: Cambridge University Press, 2009.

Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research 2003; 3: 993–1022.

Bonch-Osmolovskaya A, Orekhov B. Nekotoryye primeneniya korpusnykh metodov r naivnoy poezii [Some Applications of Corpus Methods to Naive Poetry]. 2013. http://www.ruthenia.ru/leibov_50/article_b-osm_orexov.html (accessed April 1, 2022).

Broder AZ, Glassman SC, Manasse MS, Zweig G. Syntactic Clustering of the Web. Computer Networks and ISDN Systems 1997; 29(8–13): 1157–1166.

Conneau A, Schwenk H, Barrault L, Lecun Y. Very Deep Convolutional Networks for Text Classification. Preprint, submitted in 2016. https://arxiv.org/abs/1606.01781 (accessed April 1, 2022).

Chloupek J, Nekvapil J, editors. Studies in Functional Stylistics, vol. 36: Linguistic and Literary Studies in Eastern Europe. Amsterdam: John Benjamins Publishing, 1993.

Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language

Technologies, vol. 1. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171–4186.
Diez-Arroyo M. Scientific Language in Skin-care Advertising: Persuading Through Opacity. Revista Espanola de Linguistica Aplicada 2013; 26: 197–214.
Dolinin KA. Stilystika frantzuzskogo yazyka [The Stylistics of the French Language]. Moscow: Prosveschenije, 1987.
Egorov E, Nikitin F, Alekseev V, Goncharov A, Vorontsov K. Topic Modelling for Extracting ´Behavioral Patterns from Transactions Data. In: International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI) Artificial Intelligence: Applications and Innovations (IC-AIAI) 2019: 44–49. http://www.machinelearning.ru/wiki/images/6/69/Egorov19behavioral.pdf (accessed April 1, 2022).
Fedotov GP. Stikhi dukhovnye [Spiritual Poems]. Moscow: Progress, Gnozis Publ., 1991.
Galitsky B, Ilvovsky D, Kuznetsov SO. Style and Genre Classification by Means of Deep Textual Parsing. In: Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue 2016: 171–181. https://www.dialog-21.ru/media/3390/galitskyba.pdf (accessed April 1, 2022).
Goddard C, Taboada M, Trnavac R. The Semantics of Evaluational Adjectives: Perspectives from Natural Semantic Metalanguage and Appraisal. In: Functions of Language 2019; 26(3): 308–342.
Grice P. Logic and Conversation. In: Cole P, Morgan J, editors. Syntax and Semantics 3: Speech Acts. New York: Academic Press, 1975: 41–58.
Grigor'ev VP, Shestakova LL. Slovar' yazyka russkoy poezii XX veka [Dictionary of the Language of Russian Poetry of the Twentieth Century]. Moscow: Znak, 2001.
Guiraud P. Essais de stylistique. Paris: Éditions Klincksieck, 1969.
Günthner S, Knoblauch H. Culturally Patterned Speaking Practices – The Analysis of Communicative Genres. Pragmatics 1995; 5: 1–32.
Halliday MAK, Hasan R. Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective. Geelong: Deakin University Press, 1985.
Hansen-Löve AA. Der Russische Formalismus: Methodologische Rekonstruktion seiner Entwicklung aus dem Prinzip der Verfremdung. Vienna: Austrian Academy of Sciences Press, 1978.
Harris ZS. Distributional Structure. Word 1954; 10(2–3): 146–162.
Hausenblas K. On the Characterization and Classification of Discourses. In: Travaux Linguistiques de Prague 1966; 1: 67–83.
Hofmann T. Probabilistic Latent Semantic Analysis. Preprint, submitted in 2013. https://arxiv.org/abs/1301.6705 (accessed April 1, 2022).
Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. Preprint, submitted in 2018. https://arxiv.org/abs/1801.06146 (accessed April 1, 2022).
Ikonomakis M, Kotsiantis S, Tampakas V. Text Classification Using Machine Learning Techniques. WSEAS Transactions on Computers 2005; 4(8): 966–974.
Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of Tricks for Efficient Text Classification. Preprint, submitted in 2016. https://arxiv.org/abs/1607.01759 (accessed April 1, 2022).
Jung CG. Collected Works, vol. 7: Two Essays on Analytical Psychology, transl. by Hull RFC. Princeton: Princeton University Press, 1966.
Kourova OI. Slovar' tradicionno-pojeticheskoj leksiki i frazeologii pushkinskoj epohi [The Dictionary of Traditional-Poetic Words and Phrases in the Age of Pushkin]. Shadrinsk: Shadrinskij gos. ped. in-t, 2001.

Kourova OI. Tradicionno-poeticheskaja leksika i frazeologija kak termin stilistiki [Traditional-Poetic Vocabulary and Phraseology as Stylistic Concept]. Vestnik Cheljabinskogo gosudarstvennogo pedagogicheskogo universiteta 2016; 8: 167–170.

Kowsari K, Meimandi KJ, Heidarysafa M, Mendu S, Barnes L, Brown D. Text Classification Algorithms: A Survey. Information 2019; 10(4): 150. https://www.doi.org/10.3390/info10040150.

Lai S, Xu L, Liu K, Jun Zhao J. Recurrent Convolutional Neural Networks for Text Classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence, 2015. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/download/9745/9552 (accessed April 1, 2022).

Leech G, Garside R, Bryant M. CLAWS4: The Tagging of the British National Corpus. In: COLING, vol. 1: The 15th International Conference on Computational Linguistics. 1994. https://www.doi.org/10.3115/991886.991996.

Leech G. Semantics. Suffolk: Richard Clay, 1974.

Li X, Li C, Chi J, Ouyang J, Li C. Dataless Text Classification: A Topic Modeling Approach with Document Manifold. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management 2018: 973–982. https://www.doi.org/10.1145/3269206.3271671.

Liu P, Xipeng Qiu X, Huang X. Recurrent Neural Network for Text Classification with Multi-task Learning. Preprint, submitted in 2016. https://arxiv.org/abs/1605.05101 (accessed April 1, 2022).

Ljashevskaja ON, Sharov SA. Chastotnyj slovar' sovremennogo russkogo jazyka (na materialah Nacional'nogo korpusa russkogo jazyka) [Frequency Dictionary of Modern Russian (based on the Russian National Corpus)]. Moscow: Azbukovnik, 2009.

Loper E, Bird S. Nltk: The Natural Language Toolkit. Preprint, submitted in 2002. https://arxiv.org/abs/cs/0205028 (accessed April 1, 2022).

Martin JR. A Contextual Theory of Language. In: Cope B, Kalantzis M, editors. The Powers of Literacy: A Genre Approach to Teaching Writing. London: Falmer Press, 1993: 116–136.

McCallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI-98 Workshop on Learning for Text Categorization, vol. 752. Citeseer, 1998: 41–48.

Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed Representations of Words and Phrases and Their Compositionality. In: Advances in Neural Information Processing Systems 2013: 3111–3119.

Neogi PPG, Das AK, Goswami S, Mustafi J. Topic Modeling for Text Classification. In: Mandal JK, Bhattacharya D, editors. Emerging Technology in Modelling and Graphics. Singapore: Springer, 2020: 395–407.

Osgood CE, Tzeng O. Language, Meaning, and Culture: The Selected Papers of CE Osgood. New York: Praeger, 1990.

Pavlinek M, Podgorelec V. Text Classification Method Based on Self-training and lda Topic Models. Expert Systems with Applications 2017; 80: 83–93.

Pavlovich NT. Slovar' pojeticheskih obrazov: Na materiale russkoj hudozhestvennoj literatury XVIII–XX vv. T. 1–2 [Dictionary of Poetic Images: Based on Russian Fiction, vol. 1–2]. 2nd ed. Moscow: Editorial URSS, 2007.

Peters ME, Neumann M, Lyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep Contextualized Word Representations, 2018. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1. New Orleans, LA: Association for Computational Linguistics, 2018: 2227–2237.

Rehůřek R, Sojka P. Gensim Statistical Semantics in Python. In: EuroScipy 2011, Paris, Aug. 25–28. 8. 2011. https://www.fi.muni.cz/usr/sojka/posters/rehurek-sojka-scipy2011.pdf (accessed April 1, 2022).
Russel JA. Culture and the Categorization of Emotion. Psychological Bulletin 1991; 110: 26–450.
Russel JA. Core Affect and the Psychological Construction of Emotion. Psychological Review 2003; 110(1): 145–172.
Santini M. Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In: 40th Hawaii International Conference on Systems Science (HICSS-40 2007). Abstracts Proceedings, 3–6 January 2007, Waikoloa, HI: IEEE, 2007: 71. https://www.doi.org/10.1109/HICSS.2007.124.
Schütze H. Word Space. In: Advances in Neural Information Processing Systems. 1993: 895–902. https://proceedings.neurips.cc/paper/1992/file/d86ea612dec96096c5e0fcc8dd42ab6d-Paper.pdf (accessed April 1, 2022).
Sennrich R, Haddow B, Birch A. Neural Machine Translation of Rare Words with Subword Units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers. Berlin: Association for Computational Linguistics, 2016: 1715–1725. https://www.doi.org/10.48550/arXiv.1508.07909.
Shapir MI. Yazyk byta / yazyki dukhovnoy kultury [The Language of Everyday Life / Languages of the Spiritual Culture]. Russian Linguistics 1990; 14(2): 129–146.
Sharoff S. Russian Frequency Lists. June 2008. http://corpus.leeds.ac.uk/serge/frqlist/ (accessed April 1, 2022).
Sharov SA. Using Machine Translation for Automatic Genre Classification in Arabic. In: Computational Linguistics and Intellectual Technologies 2018 Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2018” Moscow, May 30–June 2, 2018: 153–163. https://www.dialog-21.ru/media/4292/bulyginmv_sharoffsa.pdf (accessed April 1, 2022).
Shklovsky V. Art as Device. In: V. Shklovsky. Theory of Prose. Elmwood Park: Dalkey Archive Press, 1990 [1919]: 1–14.
Spärck Jones K. A Statistical Interpretation of Term Specificity and its Application in Retrieval. Journal of Documentation 1972; 28(1): 11–21.
Stamatos E, Fakotakis N, Kokkinakis G. Automatic Text Categorization in Terms of Genre and Author. Computational Linguistics 2000; 26(4): 471–495.
Vekshin GV. Osnovy stilisticheskoi semantiki [Basics of Stylistic Semantics]. Moscow: MUT Publ., 2017.
Vekshin GV, Shilihina KM. Ob istochnikah slovnika “Zhivogo stilisticheskogo slovarja russkogo jazyka” [About the Sources of Glossary of the Russian Live Stylistic Dictionary]. In: Vestnik VGU (Lingvistika i mezhkul’turnaja kommunikacija). Voronezh: VGU, 2017; 3: 16–20. http://www.vestnik.vsu.ru/pdf/lingvo/2017/03/2017-03-02.pdf (accessed April 1, 2022).
Vekshin GV, Lemesheva MM. Poet kak rechevaya rol: k semantike I pragmatike russkogo poetizma [Poet as a Role: On the Semantics and Pragmatics of Russian Poeticism]. In: Vestnik RUDN. Ser.: Teoriya yazyka. Semiotyka. Semantika 2019; 10(4): 1067–1087. https://www.doi.org/10.22363/2313-2299-2019-10-4-1067-1087.
Vorontsov K, Frei O, Apishev M, Romov P, Dudarenko M. Bigartm: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. In: International Conference on Analysis of Images, Social Networks and Texts. Berlin: Springer, 2015: 370–381.

Wierzbicka A. A Semantic Metalanguage for a Crosscultural Comparison of Speech Acts and Speech Genres. Language in Society 1985; 14(4): 491–514.
Wierzbicka A. Emotions Across Languages and Cultures. New York: Cambridge University Press, 1999.
Wierzbicka A. Semantics: Primes and Universals. New York: Oxford University Press, 1996.
Wittgenstein L. Philosophische Untersuchungen = Philosophical Investigations, transl. by Anscombe GEM, Hacker PMS, Schulte J, rev. 4th ed. Oxford: Wiley–Blackwell, 2009.
Zhang X, Zhao J, LeCun Y. Character-level Convolutional Networks for Text Classification. In: Advances in neural information processing systems 2015. Cambridge: MIT Press, 2015: 649–657. https://www.doi.org/10.48550/arXiv.1509.01626.
Zhang Y, Jin R, Zhou Z-H. Understanding Bag-of-Words Model: A Statistical Framework. International Journal of Machine Learning and Cybernetics 2010 (1): 43–52.
Zhivoj stilisticheskij slovar' russkogo jazyka [The Russian Live Stylistic Dictionary]. Vekshin GV, Gertsev MN, Maksimov ES. 2020. http://livedict.syllabica.com (accessed April 1, 2022).