National Research University Higher School of Economics

As a manuscript

Nadezhda Bujlova

Verb constructions as a marker of literary formulas

Dissertation Summary

for the purpose of obtaining academic degree
Doctor of Philosophy in Philology and Linguistics

Academic Supervisor:
Candidate of Sciences
O.N. Lyashevskaya

Moscow, 2023

General characteristics of the study

This PhD thesis addresses verb constructions in four types (or microgenres) of fomulaic (or formula) literature – romance novel, detective fiction, science fiction, and fantasy. Typical constructions were identified for each microgenre using machine learning methods for text classification. These constructions were analyzed quantitatively and qualitatively. High-level features were proved to be effective for machine learning classification.

The dissertation is written within the theoretical paradigm of construction grammar. Verb constructions were examined in the works of L. Tesnière, C. Fillmore, A. Goldberg, L. Talmi, H. Boas and other researchers. For the Russian language, studies of the semantic and syntactic properties of the verb by means of the functional and the cognitive approaches were carried out by Yu.D. Apresyan, I.A. Melchuk, L.L. Iomdin, I.M. Boguslavsky, V.S. Khrakovskiy, E.V. Paducheva, G.I. Kustova, E.V. Rakhilina, Yu.L. Kuznetsova, and others. The definition of formulaic fiction is based on the works of literary critics and culturologists (J. G. Cawelti, A.-M. Boye, S. Bordoni, T.G. Skrebtsova, N.M. Marusenko, etc.). The quantitative methods used in this thesis are based on computational linguistic studies (cf. B. Kesler, J. Karlgren, D. Cutting, H. Schutze, and F. Sebastiani) and quantitative corpus research (J. Bieber, A. Stefanovich, S. Gries and others) on identification of text genre.

The study is based on the syntactically annotated corpus of contemporary Russian mass literature. **The purpose of the study** is to identify the specific traits of construction usage in various microgenres. To achieve this goal, **the following tasks should be accomplished**:

— to analyze the selected microgenres and identify their specific characteristics;

— to formulate the principles of text selection for the experiments;

— to adapt the key concepts of Construction Grammar to microgenre identification with syntactic information;

— to evaluate the scope of variability of verb constructions in closely related microgenres;

— to identify the groups of verbs that play the key role in distinguishing between microgenres;

— to analyze these verbs and their constructional properties.

The present study focuses on the typical syntactic features of contemporary mass literature – which has not been addressed in previous research on syntax and machine

learning. This gap, along with the potential significance of syntactic features for redefining microgenres, provides the basis of the study's **relevance**.

Interpretation of language data is carried out using **a variety of linguistic and computational methods**. In particular, the author of the thesis created an algorithm for identification of verb constructions in fiction, collected lists of genre-specific verbs, and conducted classification experiments using machine learning methods.

The theoretical significance of the study is based on recognition of syntactic construction as an important stylistic marker. The division of verbs into categories according to the number of filled valencies, the rank of constructions that meet a specific threshold, as well as identification of genre-specific clusters are expected to suggest new vistas for the development of stylometric studies of formulaic fiction.

The following theses are submitted for the defense:

1. Genre-specific verb constructions (or marker constructions) are constructions of a verb with its syntactic dependents which predominantly occur in one of the microgenres.

2. The problem of automatic identification of microgenre can be solved at the level of vocabulary and lexical constructions as well as using low-level linguistic features (lengths of words and sentences, etc.).

3. The distribution of the frequences of constructions for individual verbs in the corpus is skewed: on average, two or three most frequent constructions cover 50% of all the occurrences of the verb.

4. Significance and importance metrics used in automatic classification algorithms can be applied in identification of genre-specific verb constructions.

5. There are several groups of verb constructions the variability of which serves as a marker of microgenre (e.g. full vs. incomplete, argumentative vs. constructions with semantically optional modifiers, and constructions of specific thematic groups of verbs).

6. Analysis of the use of verbal constructions in different microgenres suggests that they may be linked with genre features of a higher level (such as theme, trend, and author's restrictions on vocabulary).

The **relevance** and **significance** of the study stem from the practice of the state-of-the-art approach to microgenre identification. In machine learning, the currently most widely used models are based on statistical features, paying little

attention to high-level text parameters, such as the structure of sentence or phrase; thus, putting verb construction into consideration seems to be a promising venue.

Science fiction, detective stories, romance and fantasy are examples of formula fiction that can be used to develop a method for automatic microgenre identification. As part of the study, a corpus of mass literature was created, which can be used as a dataset for assessing the quality of microgenre-determining algorithms.

The **practical significance** of the work consists in development of the algorithm that can improve meta-annotation of Russian-language literature corpora. Most of the currently available corpora contain multiple layers of annotation (morphological, syntactic, and semantic); however, text metatags (e.g. location or time of creation) are assigned manually by literary critics and historians, rather than programmatically. Meanwhile, some metatags can be defined automatically – for example, the genre of the text, certain characteristics of poetic texts, etc. The algorithm for determining text microgenre presented in this dissertation can be applied in corpora annotation; notably, the algorithm is capable of distinguishing between closely related genres. Moreover, the suggested approach shows potential for identifying not only the microgenre, but also other characteristics of a text, for example, authorship and the time of writing.

The **scientific novelty** of the work lies in the fact that, prior to this study, the constructive potential of vocabulary was not viewed as holding promise for microgenre identification. Most of the previous works of this kind utilized lower-level markers, such as part-of-speech characteristics of the text, distribution of function words, or letter bigrams.

Pre-defence public presentation of thesis results. The main results of the work were presented at the following conferences: "The Kolmogorov Readings" (Moscow 2017), ARANEA (Bratislava 2018), and ITiS (Istra 2022), and reported in the following publications: "Scientific and technical information" ser.2; Proceedings of ARANEA 2018; "Vestnik of PSTGU", series III Philology, 2021; and Vestnik of NGU, series: Linguistics and intercultural communication, 2022.

Thesis structure. The work consists of five chapters. Chapter 1 presents the theoretical justifications for using verb constructions as a feature for a machine learning classifier. In particular, Section 1.1 deals with the theoretical foundations of Construction Grammar; Section 1.2. describes approaches to defining the literary formula. The Chapter 2 describes the data and the resources utilized. The corpora used in the study are discussed in Section 2.1; the text preprocessing pipeline – in Section 2.2; and the general characteristics of the corpora – in Section 2.3. Chapter 3 discusses the method of feature engineering, with its scope and limitations. Section 3.1 describes the overall design of the study. Description of feature engineering for

machine learning is available in Section 3.2. Section 3.3 is devoted to the constructions chosen as machine learning features; it also describes identification of constructions at specific thresholds. Chapter 4 describes the machine learning experiments– those involving low-level and high-level features. Chapter 5 presents an analysis of the obtained linguistic evidence. Conclusion summarizes the results of the study.

The content of the work

Study of surface syntax through statistical data became possible due to the advent of large corpora and Universal Dependencies. Still, it is the normative language, the language of “serious fiction” that traditionally enjoys the attention of researchers, while children's literature, mass literature, and nonfiction have remained in the shadows. In our study, we turned to the so-called "formula (or formulaic) literature" and set ourselves the task of finding out, using quantitative methods, how the surface syntax of such texts can determine their "formularity". Therefore, the **Literature Review** consists of three sections that deal with the areas of key interest to this dissertation: Section 1.1 describes research papers and monographs on Construction Grammar; Section 1.2 provides a brief overview of current views on literary formulas and comparable phenomena; Section 1.3 discusses the central works on application of machine learning to classification of texts of various genres.

In our work, we study syntax using the verb-centric Dependency Grammar developed by L. Tesnière and published in his Fundamentals of Structural Syntax [Tesnière, 1988]. Tesnière was the first to introduce the notion of verbal valency and to classify verbs according to the number of actants they can attach; he also described the means of changing valency structure. Tesnière's actant theory established the basis for Construction Grammar (CxG) shaped by C. Fillmore, A. Goldberg, and others. Among the key principles that are shared today by all branches of CxG, one can name the denial of a clear boundary between grammar and vocabulary and the rejection of a static view of the language. Mutual restrictions "equalize" all the constituent parts of the system in their importance, i.e. noun phrases denoting situation participants in verb constructions are regarded on an equal footing with verbs. Semantics, syntax and even morphology are fluid and mutually dependent in such constructions.

A prominent contribution to Construction Grammar was made by the works of Yu. D. Apresyan, e.g. by his "Experimental Study of the Russian Verb" [Apresyan, 1967], which postulates a direct connection between syntax and semantics. The work analyzes about 1,500 most frequent Russian verbs and their constructions. The constructions under examination are selected to meet certain requirements, such as taking no circonstants. Verb constructions are analyzed according to formal principles, using transformation classes, trees, hierarchical classifications, etc. In continuation of research on circonstants, it was established that the circonstants is semantically related to the verb: it cannot contradict any part of the construction in meaning. The semantic

compatibility of actants and circonstants was examined in “Adjuncts in Interpretation?” [Plungyan, Rakhilina, 1990], which showed that certain adjuncts combine with verbs describing concrete situations (e.g. *bežat', rezat'* (*to run, to cut*)), and are incompatible with abstract situations (e.g. *portit', mstit'* (*to spoil, to avenge*)).

In selecting our data, we follow J. G. Cawelti's definition from “The Study of Literary Formulas” [Cawelti, 1976 p.6]: “… formulas are ways in which specific cultural themes and stereotypes become embodied in more universal story archetypes.” In our study, we will examine several of the formulas proposed by Cawelti: *adventure*, *romance*, *mystery*, and *alien beings and states*. It is easy to notice the correlation between the varieties of popular literature selected for analysis and Cawelti's four formulas. Specifically, the *romance* formula corresponds to love story, *mystery* to detective story, *adventure* to fantasy, and *aliens* to science fiction.

Applying literary formulas allows researcher not only to distance oneself from the specific features of the plot, but also to directly access linguistic features and language mechanisms. The recurrent nature of plots and worlds, quite expectedly, leads to thematic predictability of vocabulary and recurrent patterns in choice of lexical means. Standardization of such texts is convenient for a linguist, since it allows us to focus on the linguistic features of the text and to distance ourselves from the plot as much as possible.

There are several approaches to study of popular literature. A ‘pink’ romance novel is often analyzed in the context of the current socio-cultural paradigm, by studying the reasons of its popularity and the "evolution" of the formal features of the genre. The majority of Russian-language romance novels are calques from English-language originals, with a minimal infusion of current realities. It has been noted that, according to the canon of the genre, the text cannot be first-person narration, and, consequently, there will be no verbs in this form.

Another approach, adopted in analysis of detective stories, has to do with studying the origins of the two major subtypes of contemporary detective: the British one (with a primary focus on the investigation, i.e. largely "female” detective story), and the American one (with an emphasis on pursuit and violent encounters, i.e. more "masculine” detective story).

Fantasy and science fiction are two genres with similar formulas – “adventure” + “alien creatures and states” – albeit with different accents. According to most researchers, the feature of fantasy that distinguishes it from science fiction is the ultimate impossibility of the story to happen in reality. This invites occasional remarks about the possibility of constructing a dictionary of this genre (which can include both idiosyncratic neologisms and archaisms). Space wars, a microgenre of science fiction, is characterized by the fundamental explainability of the world. At the same time, this subgenre also contains features of adventure (shootouts, solving problems by brute force) and alien beings and states (the action takes place in the outer space and/or on

other planets (within the solar system or beyond), in an unrealistic (usually exotic) setting).

It is obvious that the existing studies of formulaic literature devote considerably more attention to the literary and cultural features of texts, while the overall linguistic characteristics and the constructive potential of genres has been largely overlooked.

Specific features of microgenres can be used to automatically distinguish between them. Current research on differential models, collection of homogeneous corpora, and text classification are faced with problems arising from the fact that features are determined not only by the genre, but also by additional factors (annotator disagreement, a large range of genres, etc.), which causes difficulties in defining the boundaries of even major genres. Attempts are under way to suggest new genre-distinguishing features, including the surface syntax of the verb.

Early work on genre identification used discriminant analysis and logistic regression, including neural networks. Part-of-speech characteristics, as well as various measures of readability, were used as basic descriptors of texts. Later on, bag-of-words models and Decision Trees on part-of-speech tags arrived in the genre identification scene. A major area of document classification was the HTML-based classification, which allows combining quantitative methods of describing the text with the non-textual elements of hypertext markup. In addition, mention should be made of syntactically tagged corpora, known as treebanks, which enable analysis of discourse relationships.

Chapter 2 describes the data used in the study – the genre-specific subcorpora of detective, romance novels, fantasy, and science fiction, each of which contains more than 280 texts. The texts were selected based on user-assigned genre tags; the total volume of the corpus is 104,919,587 tokens (romance novels – 18,205,059; detective fiction – 24,038,408; science fiction – 30,086,136; and fantasy – 32,589,984).

Since genre tags assigned by users of open sources may contain errors, all texts were manually examined to ensure congruence with the specified genre. Due to considerable diversity within the genres (i.e. romantic detectives, science fantasy, etc.), our work concentrates on the following subgenres:

1. “Pink” romance novel, focused on romantic and erotic descriptions.

2. “Hardboiled” detectives, i.e.detective fiction in which, unlike in classic detectives, less weight is given to the protagonist’s ability to draw logical conclusions; instead, the story emphasizes their endurance and proficiency in sharpshooting or hand-to-hand combat).

3. “Portal” science fiction or space war science fiction, describing confrontations with alien lifeforms. Here, the scientific achievements which propel the development of the world often have nothing to do with reality and are nothing but a product of the author’s creative will.

4. “Portal” fantasy or high fantasy, set in magical worlds. The protagonist is invariably “the chosen one", possessing extraordinary abilities that make him or her stand out from the crowd.

No more than five texts from each author were sampled in each category, which, in our opinion, is expected to significantly reduce the author-specific bias. After the filtering, the total volume of the corpus was reduced to 1201 texts (280 romance novels, 319 detective novels, 304 fantasy novels, and 321 science fiction novels).

The collected texts was processed using the morphological and syntactic parser UDPipe 2.6 with the UD-SynTagRus 2.6 model, in the R environment. As a result, for each token we obtained its lemma, part-of-speech, and grammatical characteristics, as well as the tree of its syntactic relations with nodes and dependencies (Fig. 1). In the example below, the verb *načalas'* (*started*) is the root of the sentence which has the following arguments: *istorija* (*story*) (nsubj) and *avarii* (*accident*) (obl). Next, the the parses were processed to extract collocations of verbs with specific dependency projections such as *root-nsubj* and *root-obl-obl* (where **root** is the pinnacle of the sentence).

```
# newdoc
# newpar
# sent_id = 1
# text = Эта история началась с вполне заурядной автомобильной аварии.
1   Эта  этот DET     _   Case=Nom|Gender=Fem|Number=Sing     2   det  _   _
2   история  история  NOUN   _   Animacy=Inan|Case=Nom|Gender=Fem|Number=Sing     3   nsubj   _   _
3   началась     начаться     VERB   _   Aspect=Perf|Gender=Fem|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Mid     0   root _   _
4   с   с   ADP    _   _   8   case     _   _
5   вполне   вполне   ADV    _   Degree=Pos   6   obl   _   _
6   заурядной    заурядный    ADJ _   Case=Gen|Degree=Pos|Gender=Fem|Number=Sing   8   amod    _   _
7   автомобильной    автомобильный    ADJ _   Case=Gen|Degree=Pos|Gender=Fem|Number=Sing   8   amod    _   _
8   аварии   авария   NOUN   _   Animacy=Inan|Case=Gen|Gender=Fem|Number=Sing 3   obl   _   SpaceAfter=No
9   .   .   PUNCT  _   _   3   punct    _   _

# sent_id = 2
# text = Так будет и мне легче, и вам понятнее.
1   Так  так  ADV    _   Degree=Pos   2   advmod  _   _
2   будет    быть     VERB   _   Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act     0   root _   _
3   и   и   PART   _   _   4   advmod  _   _
4   мне  я   PRON   _   Case=Dat|Number=Sing|Person=1     5   iobj  _   _
5   легче    легкий   ADJ _   Degree=Cmp  2   nsubj    _   SpaceAfter=No
6   ,   ,   PUNCT  _   _   9   punct    _   _
7   и   и   CCONJ  _   _   9   cc   _   _
8   вам  вы  PRON   _   Case=Dat|Number=Plur|Person=2 9   iobj  _   _
9   понятнее     понятный     ADJ _   Degree=Cmp  2   conj _   SpaceAfter=No
10  .   .   PUNCT  _   _   2   punct    _   SpaceAfter=No
```

Figure 1. An example of UDPipe annotation.

We expect that the randomized selection of texts for our corpus reflects the actual state of affairs in literary practice; therefore, the texts were not normalized by length. Figure 2a shows the density of distribution of text lengths across the subcorpora. Since the main goal of the experiment was to identify the verb constructions which are the most informative for a machine learning classifier, the text length feature was not used in subsequent experiments. Figure 2b demonstrates the distribution of sentence lengths (in tokens). Since the distributions heavily overlap,

this feature also cannot be used to effectively distinguish between genres. The selected microgenres were also shown to have little difference in the distribution of POS-tags. Among the syntactic dependencies, punctuation was the most frequent relation, followed by subject and indirect object. In most cases, the statistics did not vary much across the four genres.

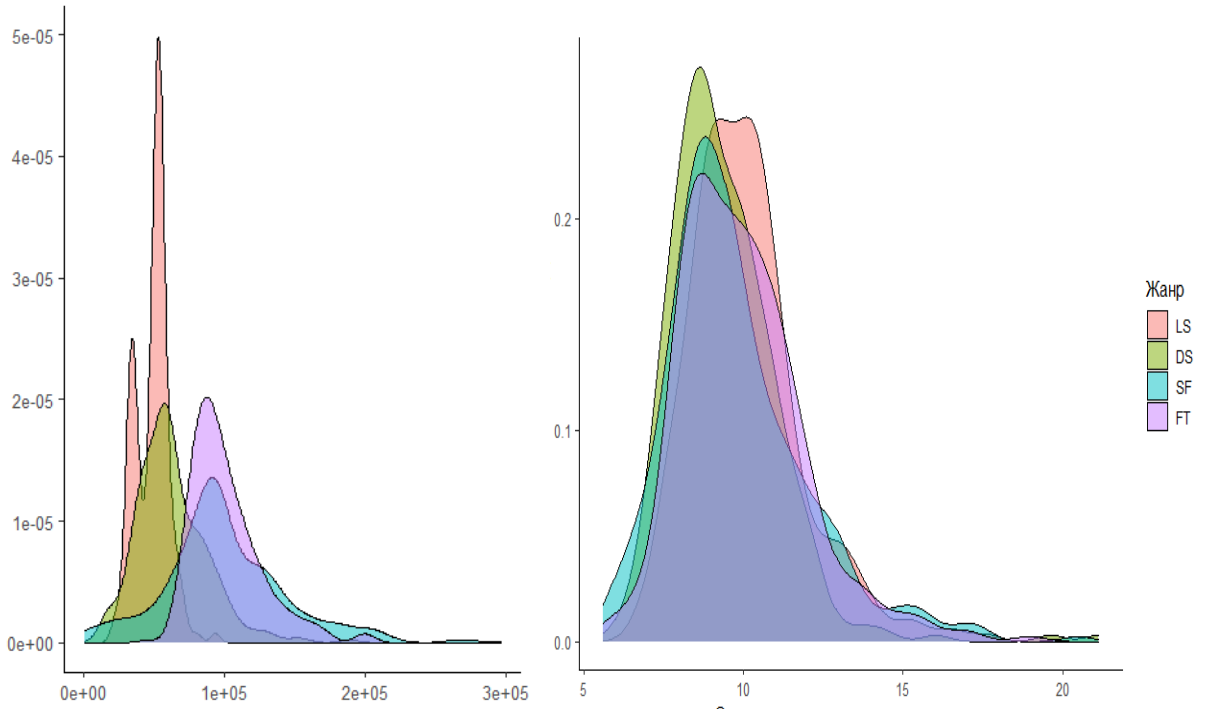

Figure 2. Descriptive statistics of the subcorpora. **a.** Distribution of token frequencies. **b.** Distribution of sentence lengths.

Considering lexical diversity, the corpora show increasing numbers of unique lexemes along with the increase in corpus size, so that science fiction has the richest vocabulary.

Thus, our microgenres are quite alike in terms of several parameters: sentence length, part-of-speech distribution, and lexical diversity. The lengths of the texts do vary to some extent, which can be explained by the fact that science fiction traditionally favors longer forms (trilogies, tetralogies, etc.), which were included in our sample.

Chapter 3 presents the general design of the study. As shown above, individual literary genres are weakly differentiated in terms of low-level features, which may impede performance of the classifier. A feature of a different level –*bag of words* – is also unsuitable for classifiers due to the low lexical diversity of mass fiction. In our study, low-level features were supplemented with verb lexemes and verb

constructions, which are expected to be unequally distributed in the subgenres under examination.

As mentioned above, verb constructions were extracted from the subcorpora using the UDPipe parser. It should be noted that the syntactic annotation is not free of errors; however, the analysis of randomly selected sentences showed that the annotation quality was not lower than the results reported in the CoNLL 2018 Shared Task. The following features were used to classify the genres:

1. low-level features (frequency vectors of word-form lengths and frequency vectors of sentence lengths);

2. verb lexemes (frequency vectors of verb lemmas);

3. syntactic constructions of individual verbs (frequency vectors of the syntactic constructions of individual verbs).

The frequency vectors were produced using the frequency normalization method based on the LL (log-likelihood) score. This metric identifies the units that are indicative of each subgenre (*kill* and *fear* in typical detective stories, *marry* and *love* in romance novels, etc.). Verbs and constructions with an LL-score above 3.5 were considered genre-specific.

A syntactic construction is defined as a set of syntactic dependents which cooccurwith a verb in texts, e.g. <*prevraščat'sja*;nsubj;obl> (<*to turn into;nsubj;obl*>) in the sentence *On prevratilsja v čudovišče.* (*He turned into a monster*). We selected six types of relations – four actant relations (which are obligatory for a grammatically and semantically valid sentence) and two circonstant (optional) relations, along with their combinations: *nsubj* (object, subject); *obj* (subject, direct object); *ccomp* (clausal complement); *xcomp* (open clausal complement); *obl* (*oblique object*); *advcl* (*adverbial clause modifier*).

The analysis took into account only the presence of relations, not their linear order, i.e. <*nsubj;obj*> and <*obj;nsubj*> were considered equivalent constructions. Analysis of the frequencies of constructions revealed that certain syntactic relations have a higher frequency in the detective fiction corpus (e.g. <*ccomp;nsubj*>, <*obl;xcomp*>) or the romance novel corpus (e.g. <*nsubj;obl;obl*>), which may be taken to be a genre marker. However, such relations are quite few, therefore it is necessary to look at individual verbs that occur in different constructions depending on the genre. To identify such verbs, we introduced the term “construction at specific threshold” (CST), i.e. the rank of a construction in the ranking of verb constructions, sorted by descending frequency, at which the total number of the verb’s occurrences exceeds N% of the total number. Our results showed that, firstly, the number of verb constructions required to cross the threshold rarely changes with the genre, and secondly, that the three most frequent verb constructions in most cases coincide, but their frequency rank may vary.

Chapter 4 presents the results of applying machine learning algorithms to subgenre classification; three algorithms and three feature sets were explored (Table 1).

The following machine learning algorithms were experimented with: Naive Bayes, Decision Trees, and Random Forest. The computations were preformed using the Python 3.1 programming language and the os module. The hyperparameters of the classifiers were not fine-tuned, since we were primarily interested in contribution of linguistic parameters rather than in optimizing the algorithms. Naive Bayes served as a baseline against which the increase in the quality of classification was measured. The choice of Decision Trees and Random Forest was prompted by their simplicity and the interpretability of the results.

Three sets of features were tested: the first (baseline) one consisted of low-level features (word length in characters, sentence length in words, and POS characteristics); the second set added a weighted list of verb lexemes computed with similarity coefficient; and the third set was a combination of the first two plus the lexico-syntactic features (i.e. the list of verbs with their constructions).

Table 1. Results of machine learning experiments (P – precision; R – recall; F – f-score).

	Naive Bayes			Decision Trees			Random Forest		
	P	R	F	P	R	F	P	R	F
Baseline feature set (word length + sentence length)	0.43	0.47	0.43	0.47	0.47	0.46	0.5	0.53	0.51
Baseline feature set + lemmas (normalized LL-score)	0.6	0.58	0.57	0.73	0.8	0.79	0.82	0.85	0.83
+ verbs with constructions (normalized LL-score)	0.66	0.66	0.66	0.86	0.86	0.85	0.88	0.88	0.88

The lowest results were demonstrated by the Naive Bayes classifier on the baseline feature set (POS-tags, word length, and sentence length). The Random Forest classifier proved the most successful at the task. Quite expectedly, adding lexical and syntactic features to the model improved the performance of all the three classifiers. We used the *var_imp* function of the SKlearn package to obtain lists of each verb's constructions that proved to be the most important features in machine learning. We calculated their LL-score in each subcorpus and selected constructions with the LL-score > 3.5; these constructions are considered to be genre markers. Their analysis was carried out according to the following algorithm:

1. For each construction, the semantic class of the root verb was identified.
2. The nuclear (prototypical) construction was determined.
3. If the construction observed in the text differs from the nuclear one, this may indicate that the construction underwent truncation or extension.
4. Sentences corresponding to the verbal scheme were extracted from the bank of examples.
5. The examples were analyzed, taking into consideration the topic of the microgenre.

Chapter 5 presents the results of the analysis of the marker constructions. We formulated a number of hypotheses regarding the shifts in distribution of verb constructions across microgenres:

1. Various topics are present in a genre, which places certain restrictions on the choice of vocabulary and lexical constructions. This is manifested in the high frequency of genre-specific verbs; additionally, the topic of the text is linked with individual meanings (frames) of a lexical unit, as well as with the choice of its syntactic construction.
2. Authorship also affects the choice of vocabulary and constructions. In our study, we tried to minimize the impact of author-specific style by limiting the number of texts per author to five books.
3. Genres follow different schemata of plot development. For example, romance novels typically have linear narrative with key scenes and filler scenes. As a result, two verb groups with different syntactic constructions are present. Remarkably, the lexical constructions in key scenes are recurrent from novel to novel, thus constituting a "formula", while filler scenes do not necessarily conform to this rule.

In a broader perspective, it can be hypothesized that genres vary in the structure of the narrative, dialogue, and other literary forms, which reflects itself in the observed shifts in distribution of lexico-syntactic units.

We described the verb constructions that came to be viewed as markers of literary formulas along the following lines:

• lexical-semantic group(s);

• complete construction (CC), reflecting the conventional expression of all the mandatory participants of a semantic frame;

• comparison of the marker construction with the complete construction in terms of:

◦ completeness,

◦ presence of new (circonstant) participants,

◦ peculiarities of the morphosyntactic expression of the participants,

◦ peculiarities in the lexical profile of filling the construction's slots.

The marker verb constructions (81 in total) were divided into 13 lexico-semantic groups, the largest being mental verbs, verbs of movement, and speech. Notably, the lexical markers of subgenre can be expressed both by genre-specific verbs and by entire groups (e.g. verbs of social interactions can be prominent in psychological prose, verbs of movement can be characteristic for action stories, etc.).

The complete construction (CC) is a conventional expression of all the key participants of the semantic frame (i.e., a construction that realizes all the valencies of the verb in a given meaning); CCs were determined using the Minor Academic Dictionary of the Russian Language, FrameBank, and The Active Dictionary.

After that, the contexts containing the marker constructions were compared against their CCs in terms of the number of arguments (i.e. whether the construction in the example is complete, incomplete, or extended) and the content of the dependents (i.e. the vocabulary that fills the valencies); the systemically occurring divergencies were recorded and analyzed. In total, we conducted in-depth analysis of 151 constructions marking one or more microgenres. Below is an example of analyzing the < *stanovitsya nsubj;obl* > (<*become nsubj;obl* >) construction, which is significant in love stories.

Marker construction: *stanovitsya (become) nsubj;obl*

lexico-semantic class: existential

CC: V + nsubj (who/what?) + obl (with whom/what?)

Stoit emu nemnogo vypit', i <u>on</u> ***stanovitsya*** *ves'ma n<u>elyubeznym</u>.*

When he drinks a little, <u>he</u> ***becomes*** *quite <u>unkind</u>.*

Tol'ko <u>prihod</u> Karayushchej ***stanovilsya*** *chut' bolee zametnym <u>sobytiem</u>.*

Only the <u>arrival</u> of the Punisher ***became*** *a slightly more noticeable <u>event</u>.*

In love stories, a verb in this construction is more often used to describe a character (character X becomes Y under a certain conditions, see the example). The same construction in fantasy tends to describe changes in the outside world (situation

X becomes Y). At the same time, it should be noted that in this case there is a systemic failure in the tagging, when the direct and the indirect objects are reversed.

The constructions were divided into three groups: markers of one genre, two, or three genres. We noted the following features of genre-specific constructions:

– verbs whose constructions serve as markers for love stories belong to different lexical-semantic classes, with verbs of speech and motion being represented most extensively *otvechat'* (to *answer), vskriknut' (to scream), rasskazyvat' (to tell); uezzhat' (to leave), brodit' (to wander), napravlyat'sya (to head)*. Another five verbs constitute the classes of physiology, contact, and interaction. It is worth noting that the impact of microgenre is twofold: it promotes semantic verb classes to become marker constructions, as well as activates specific meanings of polysemous verbs. Thus, such verbs as *izbegat', otvečat'* (*to avoid* and *to answer*) will occur more frequently in the meaning of social interaction.

–the subcorpus of detective stories contains a slightly larger number of marker constructions with speech verbs, as compared to love stories – and, more generally, larger numbers of interaction verbs (expressing engagement in communication, social interaction, etc.). Such verbs may both introduce the characters' text and signal a change of microthemes. Subjectless constructions, especially adverbial participle phrases function as fillers needed to "fill in the background" and to refine the narrative.

– marker constructions of science fiction demonstrate a variety of semantic classes. It is noteworthy that only one marker construction serves to accompany the speech of characters – the verb *vzdyxat'* (*to sigh*). The other constructions tend to denote the observer's perception and interpretation of the worlds described in the story.

– in fantasy, the markers are the constructions of genre-specific verbs. These are verbs of perception, motion, and some other lexico-semantic classes. Many constructions describe events, characters, and their interactions. Besides, such constructions as *<vybirat' obj;> <choose obj;>, <reshat' nsubj; xcomp;> <decide nsubj; xcomp;>, <podhvatyvat' obj;> <pick up obj;>, <uchityvat' obj;> <take into account obj;>, <oborachivat'sya obl;> <turn around obl;>* promote cohesion of the narrative, make it more dynamic and detailed, add new dimensions to the world.

Analysis of verbs' marker constructions identified complete constructions (with all obligatory valencies filled) and the following incomplete constructions (with the omission of one or more participants): subjectless (with the subject omitted), objectless (with the direct object omitted), and others (with the omission of other obligatory actants). In addition, extended constructions were analyzed, i.e. those expressing circonstant participants in addition to the obligatory actants. Complete constructions are more common in fantasy and science fiction; this may be due to the

higher descriptiveness of these genres, which demands greater coherence. Incomplete constructions, both subjectless and objectless, are more common in love stories and detective stories.

The majority of subjectless constructions occur when the verb is within adverbial participle phrases or coordinative noun phrases. Omission of the actor is a serious modification of the verb construction, which may be indicative of the relative complexity of the syntax: in a simplest case, such sentences contain several verbs that complicate the syntax tree, and therefore the actor is often syntactically controlled by another predicate. Consider the subjectless construction in the following examples:

*(Love story) My partner, without **removing** his hand from my waist, leaned over to my hand and kissed it.*

*Moj partner, ne **ubiraya** ruki s moej talii, nagnulsya k moej kisti i poceloval.*

*(Detective story) "This is Rip," I said, **removing** my hand from the handle of the Zig Sauer.*

*– Eto Rip, – skazal ya, **ubiraya** ruku s rukoyatki «zig-zauera».*

*(Fantasy) "And who's the big-eyed one here," the archer muttered, **putting** her arrow **back** in her quiver.*

*– I kto u nas tut samyj glazastyj, – burknula luchnica, **ubiraya** strelu v kolchan.*

Unlike subjects, objects are seldom omitted, and the corresponding incomplete construction is hardly encountered among the marker constructions of formula literature. The only exception is the construction of the verb *glotat'* (*to swallow*) with an incorporated object (this is a marker of detective stories). Non-expression of an object can be explained by the competition among the available expressions of the second semantic actant, namely, constructions with a direct object and constructions with a complementary clause, cf. *bojalsja nasmešek* and *bojalsja pokazat'sja smešnym / čto ego djadja-lesorub posčitaet plemjannika plaksoj* (*was afraid of ridicule* and *was afraid of appearing ridiculous / that his uncle the lumberjack would consider his nephew a crybaby)*. In the Russian language, it is extremely rare for a direct object and a complement to refer to different semantic actants and to be expressed simultaneously; typically, these are two expressions of the same participant. For example, a direct object can be a compressed expression of a proposition which otherwise, in its extended form, would be expressed by a complement, cf. *rešit' vopros* (*to solve an issue*).

The following groups were observed among the constructions with an unexpressed object: constructions with the competing *xcomp* and *obj* subcategorization frames displacing the direct object (typical for fantasy; these are primarily modal and existential verbs); constructions where the object is marginalized (the object is present, yet it has a low communicative ranking, cf. *rasskazyvat' o proisšestvii* vs. *rasskazyvat' skazku* (to *talk about an incident* vs. *to tell a fairy tale*), where the prepositional construction may be perceived as more neutral));

constructions accompanying direct speech (typical for romance novels and detective stories, where the topic of conversation is embedded into direct speech).

We analyzed the omissions of other semantic actants, dividing them according to the type of semantic role. The first group is formed by verbs with an unexpressed valency of the initial/final point or location, i.e. verbs of motion and transformation, and existential verbs. Here, the marker constructions expectedly express one of the peripheral valences which is emphasized in the semantic frame or in communication. The next group is comprised of verbs with an unexpressed valency of the addressee, the topic, or the content of the utterance, namely, verbs of speech where the semantic actants "addressee" and "topic" have been expressed in the character's speech and are not repeated in the reporting phrase introducing the direct speech with the target verb. Among the other marker constructions in which the indirect object is not expressed, there are constructions with the participants"method", e.g. *zvonit' (po telefonu* or *v dver' (to call (on the phone* or *at the door))*; "point of contact", e.g. *podxvatyvat' (na ruki, pod lokotok* (*to pick up (in one's arms, by the elbow)*); "stimulus", e.g. *smejat'sja* (*nad kem* or *ot čego* (*to laugh (at smb.* or *from smth.*) and *krasnet', rozovet' (ot udovol'stvija) to blush, turn pink (from pleasure)*); "the content of the proposition, e.g. *podozrevat' v čem* (*to suspect of smth.*). In our opinion, there are several reasons for the omission, and they do not form a systematic pattern.

Another type of constructions that is distinct from the prototypical ones is "extended" constructions, in which an circonstant participant is expressed. The *obl* relation can express both obligatory and circonstant members of the frame; however, according to our estimates, the proportion of circonstant expressed in marker constructions containing *obl* is rather small. The *advcl* relation is almost always connected with an circonstant. The verbs *prosit'* and *bojat'sja* (*to ask* and to *be afraid*) should be mentioned in this regard, in which *advcl* encodes both an obligatory participant (the content expressed by a subordinate clause with conjunctions *čtoby and esli* (*in order to* and *if*) and an adverbial participant (time, purpose, etc.)). There are three constructions in which the *advcl* participant is always an circonstant: *<brodit' nsubj;obl;advcl>* (*<to wander nsubj;obl;advcl>*), *<podnimat'sja nsubj;obl;advcl>*,(*<to rise nsubj;obl;advcl>*), and *<stat' nsubj;xcomp;advcl>* (*<to become nsubj;xcomp;advcl>*). In formula literature, these verbs in their concrete meanings can serve as a kind of "access points" to descriptions of emotional experiences and interpretations of the characters' actions (which allows the author to more fully express the descriptive, emotional, or intellectual components without using abstract vocabulary). In order to enhance emotional engagement, the physical verb functions as the syntactic center of the sentence, and the abstract verb is placed in the dependent clause.

*Having loaded provisions into the car, I **began** planning a menu for the evening.*
*Zagruziv proviziju v mashinu, ja **stala** pridumyvat' menju na vecher.*

*Upstairs, Ren **wandered** around the room, wringing her hands and wiping her running tears.*

*Naverhu Rjen **brodila** po komnate, zalamyvaja ruki i vytiraja begushhie slezy.*

*Her heart seemed to have been torn out of her chest, a wave of tears **rising** to her eyes, threatening to flood her with sadness.*
Serdce slovno vyrvali u nee iz grudi, volna slez podnimalas' k glazam, grozja zatopit' pechal'ju.

Conclusion

The present work examined verb constructions as a genre-distinctive feature. We collected and syntactically annotated a corpus of mass literature (including romance novels, detective stories, fantasy, and science fiction). This annotation was used in experiments for genre identification which showed that verb constructions can be used as features in machine learning. The constructions with the largest genre-distinctive contribution were analyzed from the point of view of the genre in which they function as markers. The work also discussed the limitations of the method and the directions for further research.

Publications

Bujlova N. N. Klassifikacija tekstov po zhanram pri pomoshhi algoritmov mashinnogo obuchenija // Nauchno-tehnicheskaja informacija. Serija 2: Informacionnye processy i sistemy. 2018. № 8.— S. 34-38.

Bujlova N. N., Ljashevskaja O. N. Leksiko-sintaksicheskie markery malyh literaturnyh zhanrov // Vestnik Pravoslavnogo Svjato-Tihonovskogo gumanitarnogo universiteta. Serija 3: Filologija. 2021. T. III. № 66. — S. 11-23.

Builova N. Verb constructions as a feature of genre classification, in: CLLS 2018. Computational Linguistics and Language Science. Proceedings of the Workshop on Computational Linguistics and Language Science. Moscow, Russia, April 25, 2018 / Ed. by E. L. Chernyak. CEUR Workshop Proceedings, 2018. pp. 108-113.

Bujlova N. Amateur Prose on the Web: Verb Construction as a Feature of Genre Classification, in: Anna Butašová, Vladimír Benko, Zuzana Puchovská [eds.), Proceedings of ARANEA 2018. Web Corpora as a Language Training Tool, 2018. Bratislava: Univerzita Komenskeho v Bratislave. ISBN: 978-80-223-4597-2. p 25-30.

Bujlova N.N., Ljashevskaja O.N. K voprosu o chastotnostjah glagol'nyh konstrukcij v nekotoryh zhanrah massovoj literatury // Vestnik NGU. Serija: Lingvistika i mezhkul'turnaja kommunikacija. T 20, №3., 2022. – S. 64-75.