Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2022"

Moscow, June 15–18, 2022

Lexical and Syntactic Features for Reader Rating Prediction

Alexander Potekhin
HSE University
St. Petersburg, Russia
aapotekhin@edu.hse.ru

Abstract

Finding a correlation between the structural features of the text and its reception has recently become a challenging task of computational linguistics. However, the correlation between the reader's reception of a literary work and its linguistic features suffers from the ambiguity of many textual parameters, which can be considered in calculations. Concerning Russian-language literature, such a process is complicated by the lack of representative databases of reader reviews and by rather noticeable discrepancies in the methods of text data analysis. In this paper, I propose to investigate the possibility of predicting the rating of a text only by its lexical and syntactic features. To design an experiment, four steps were taken: First, a corpus of Russian novels of different genres was built. Next, the literary rating was scraped from bookstore LitRes via a devised parser. Due to the small size of the corpus, the obtained results were manually cleaned to avoid ambiguity of text ratings. Most of data preprocessing was the selection of linguistic features to be considered. 23 different parameters were extracted after designing a proper software to mine those features. The final part of the work was focused on checking whether the lexical and syntactic parameters correlate with the texts rating and setting a proper predictive model. Random Forest, Cat Boost, Logistic, Linear Regression, and K-Nearest Neighbors algorithms were compared. Since the coefficient of determination for the regression approach had a poor value, it was decided to move on to the classification problem, which brought more significant results. The obtained results confirmed the existence of a correlation between the structure of texts and their ratings and shed a light on new prospects in the research of the features of the text and its perception.

Keywords: NLP; Computational Linguistics; Text Evaluation; Stylometry; Lexical Complexity; Lexical Diversity; Machine Learning Methods.

DOI: 10.28995/2075-7182-2022-21-1140-1148

1 Introduction

The present study is focused on investigating the possible correlation of linguistic and syntactic features of a literary work with its reader rating. The paper has a rather interdisciplinary context, since it affects not only the methods of linguistic research coupled with NLP approaches, but also the way of reflecting the reader's perception of the text—the text rating. If the relationship between the rating and the structural parameters of the work exists, then there is a broad prospect of identifying specific genre features of the text, which can be used in literature and language research.

The theoretical part is based on stylometry methods and quantitative linguistic approaches, while the practical stage is derived from the results of exploratory analysis of linguistic data and text evaluations and subsequent experiment using machine learning methods.

Text analysis is always associated with a number of difficulties, among which the main one is the lack of consistency in research methods, whether that be approaches from fundamental and applied linguistics or literary theory. As for the analysis of a whole corpus, such a process is subject to the standard complications of big data exploration. For these reasons, the first part of the paper focuses on the building of a diverse corpus of Russian literature.

The biggest stage of the research was the collection of linguistic and extralinguistic parameters for each of the literary works available in the corpus. To scrape ratings, a parser that collects book evaluation and number of votes from a large online bookstore LitRes was developed. After obtaining the necessary results, the most important part of the work is covered—the number and kind of textual features that are extracted from the corpus for further analysis. All parameters are adjusted and adapted for the Russian language and combine the most well-known stylometry methods.

When all the linguistic features have been extracted via developed software and the data are prepared for analysis, an exploratory data analysis is conducted, a possible correlation between the parameters is checked, and an experiment to predict the rating of texts from the corpus based on the collected parameters is performed.

2 Corpus, parsing and preprocessing

2.1 Corpus

For our purposes, it was decided to assemble a small corpus and, if the results satisfy the initial assumptions, increase its size to make sure that the sample was representative.

Literary texts were collected from online libraries that offer texts in the public domain, namely the "Maxim Moshkov Library" and "Flibusta". 3701 books were downloaded from different sections: "classical literature", "domestic fiction" and "LitRPG".

2.2 Parser and data preprocessing

After corpus is built, the next stage of the research is the development of a parser to scrape ratings for the built corpus. Since in the Russian Federation there is no leading book aggregator such as Amazon.com, and large bookstores most often do not specialize in selling electronic versions of books, there was a risk of not getting a rating for a substantial portion of the corpus. To avoid that, the online store LitRes[1] was chosen, with its huge catalog of literary works. In addition to LitRes rating, there are estimates according to the version of another major book-related social media—LiveLib[2]. Since LitRes provides an opportunity to receive an electronic version of the book, the reader evaluates only its content, not print quality, font size, cover design or paper density.

The parser development was carried out in Python using the BeautifulSoup library[3], which allows to get information from HTML and XML files and the Requests[4] package for sending HTTP requests. The created program receives the name of the work and the author as input and returns an answer from LitRes. If a substring from the query is found, then it is checked whether the found book is electronic in order to avoid noise in data, for example, audiobooks. After that, a request is sent that receives the text in the window with the rating of the work. It was decided to add a test for the found author's name to the request. Since the page markup elements of certain books differed from the rest, a note was added to each output about the necessity to verify its authenticity of authorship if at least one test was not passed. Requests were sent at a set time interval to prevent exceeding the number of allowed requests to the site.

The parser provides information about the ratings for 1411 texts, while the information on the rest was absent in LitRes. After automatic filtering, all estimates marked with a possible error were checked manually. At the moment of final check, there were 1256 texts in the corpus with a confirmed rating and the number of votes.

Number of documents	1256
Approximate number of tokens	896 874 542
Mean rating (1-5)	4.3
Mean votes	541

Table 1: Corpus description.

[1]LitRes bookstore. URL: https://www.litres.ru/
[2]LiveLib book site. URL: https://www.livelib.ru/
[3]Beautiful Soup Documentation. URL: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[4]Requests python package. URL: https://docs.python-requests.org/en/

3 Feature extraction and methods

3.1 Methods

To extract linguistic parameters from the corpus, a custom feature extractor was developed. The program was written in Python and the absolute majority of metrics were implemented without the involvement of external dependencies, since a greater number of text processing software is not adjusted for application to the Russian language. Tokenization was performed using a regular expression that selects all the letters of the Russian and Latin alphabet, lemmatization was carried out through the Pymorphy2[5] package, and in some cases NLTK[6] tools were used.

3.2 Syntax complexity

To analyze the syntactic complexity of the text, a number of traditional stylometry parameters were extracted with minor modifications. To begin with, a standard set of measures of textual complexity was taken, presented by Nikolai A. Rybakin (Solovyev et al., 2018), who proposed three main features among the 100 considered for 10,000 texts: familiarity of words, sentence length, and the length of words, with the last one determining the overall complexity of the syntax structure in a literary work. In addition, the average length of words per paragraph and the dash, colon and comma frequency per 1000 characters were calculated—heuristic parameters empirically selected as possible discursive markers.

3.3 Readability score

The readability index is traditionally calculated using the Flesch Reading Ease and the Flesch-Kincaid Grade level formulas (J.P. et al., 1975), but for the Russian text I. Oborneva's (I.V., 2006) and V. Solovyov's (Solovyev et al., 2018) adaptations will be considered:

$$
\begin{aligned}
FRE_O &= 206.835 - 1.3 \times ASL - 60.1 \times ASW \\
FRE_S &= 208.7 - 2.6 \times ASL - 39.2 \times ASW \\
FKG_O &= 0.5 \times ASL + 8.4 \times ASW - 15.59 \\
FKG_S &= 0.36 \times ASL + 5.76 \times ASW - 11.97,
\end{aligned}
$$

where ASL is average sentence length and ASW is average amount of syllables per word.

3.4 Lexical diversity

In the mentioned readability score studies, that metric was considered in the context of the analysis of schoolchildren's texts. Since that very sample often becomes hugely representative in the development of linguistic features, a rather non-standard decision was made—to take another measure of educational works examination—the complexity of vocabulary.

In order to assess the lexical complexity, reference dictionaries are needed. Frequency dictionaries are often used in such cases, however, it was decided to resort to dictionaries with marked levels of complexity of word forms, since they are based on the usage frequency, but at the same time clearly label groups of words with different frames of use—daily, publicistic and scientific speech in accordance with the order of language acquisition.

Lexical complexity was calculated by the ratio of words of various levels of complexity with normalization per 1000 words. For analysis, 1000 random words were selected from the text and, after lemmatization, each was compared with the existing vocabulary levels dictionary.

The dictionary was compiled on the basis of lexical minima for each CEFR (CEFR,) degree of the Ministry of Education of the Russian Federation (N.P and T.V., 2015a; N.P and T.V., 2015b; N.P and T.V., 2015c; N.P and others, 2019) and part of the database of the project "Visualizing Russian" created by

[5]Pymorphy2 parser. URL: https://pymorphy2.readthedocs.io/en/stable/

[6]Natural Language Toolkit. URL: https://www.nltk.org/

Steven Clancy (S, 2014 2022). Since the data of the "Visualizing Russian" project were presented in the ACTFL assessment system, it was decided to transfer all levels to the CEFR format for the convenience of the European developer.

3.5 Lexical complexity

Even nowadays the most common metric of lexical diversity is still the TTR (Type-Token Ratio) formula, presented by M. Templin (Templin, 1957):

$$TTR = \frac{L}{T},$$

where L is the number of lemmas, and T is the total number of tokens. However, such a formula has quite obvious drawbacks—the more words in the text, the less indicative its indicator becomes. That is explained by an unlimited number of lexemes, many of which have a tendency to be used with the certain phrase constructions, which does not allow comparing TTR of texts with the different lengths.

Researchers assessing lexical diversity have proposed many approaches to correct this shortcoming. For example, Guiraud-index (RTTR) was proposed by Pierre Guiraud (Guiraud, 1954):

$$RTTR = \frac{L}{\sqrt{T}}$$

The author suggested that such a formula would most accurately follow Zipf's law (G.K., 1949), which determines that the ordinal number of a word form is proportional to its frequency of use in a descending list of lemmas. The corrected TTR(c) formula was subsequently proposed by Carroll (Carroll, 1964):

$$TTR_c = \frac{L}{2T},$$

and the Herdan index (A., 1955):

$$TTR_{log} = \frac{\log L}{\log T}$$

Those two are likewise traditional metrics, but they do not solve the problem of the original formula completely. "Hypergeometric distribution D" (HD-D) and "Measure of lexical textual diversity" (MTLD), designed by McCarthy and Jarvis (M. and S., 2010), were used in the study to obtain an estimate of lexical diversity most independent of the length of the text. The first calculates as the sum of the probabilities of meeting each token in random 42 words of the text and in the second the average length of text lines with a given TTR is computed. In the last one, the TTR is calculated sequentially for each word, but when a certain threshold is reached, the value resets to zero. These metrics were implemented using the methods of the Lexical diversity[7] package.

4 Results and discussion

4.1 Exploratory data analysis

Data analysis was carried out in the IPython environment using Pandas[8] and Numpy[9] libraries for tabular computing, Matplotlib[10] and Seaborn[11] for plotting, and Sklearn[12] for applying machine learning models. To begin with, all the data from the previous stages of the study were obtained and a Pandas dataframe was formed, where texts were located along with their rating, the number of people who evaluated them, and all their extracted lexical and syntactic features. After that, the missing values were processed and the main statistical measures for all parameters were calculated. For example, the table below presents the median value, standard deviation, and the distribution of values by quartiles.

[7] Lexical diversity package. URL: https://github.com/kristopherkyle/lexical_diversity
[8] Pandas package. URL: https://pandas.pydata.org/docs/
[9] NumPy package. URL: https://numpy.org/
[10] Matplotlib data visualization package. URL: https://matplotlib.org/
[11] Seaborn data visualization package. URL: https://seaborn.pydata.org/
[12] Scikit learn package. URL: https://scikit-learn.org/stable/

Measure	LitRes mark	LitRes votes	Average word length	MTLD	FKG (Soloviev)
Mean	4.37	354.66	5.26	347.23	4.87
Standard deviation	0.56	1088.09	0.26	113.69	1.41
25% quantile	4.20	19.00	5.10	267.64	4.00
50% quantile	4.50	77.00	5.25	328.01	4.68
75% quantile	4.70	267.00	5.41	407.73	5.51

Table 2: General descriptive statistics.

Then a heatmap of all available features was constructed, in which the maximum coefficient of correlation of literary rating with parameters was only 0.1. Although this value already allows us to make a number of judgments about the mutual influence of features, for unambiguous conclusions it is necessary to obtain more information about the data and confirm them experimentally. For the target variable of

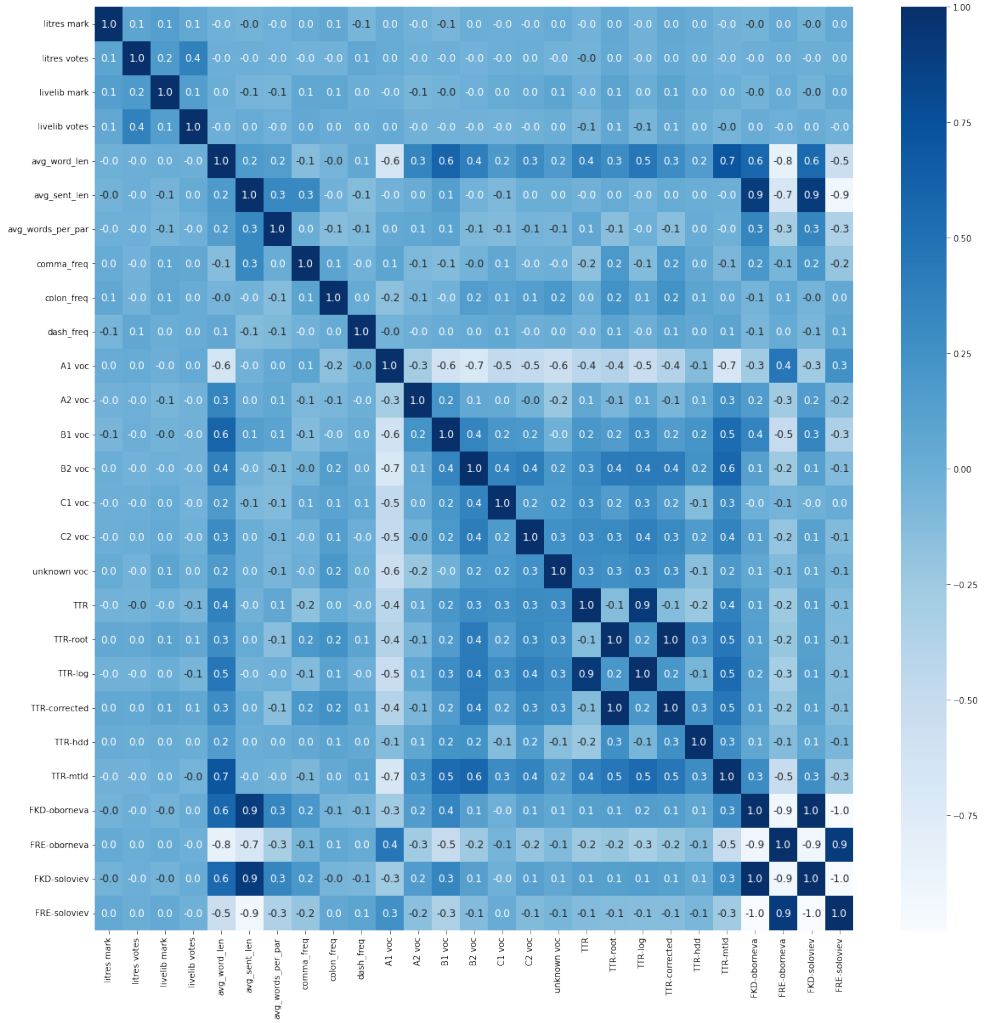

Figure 1: The feature correlation heatmap.

ratings from the LitRes and LiveLib sites, their weighted average was calculated taking into account the number of people's votes. In the graphs below, the distribution of values across all available parameters could be observed. The main conclusion that can be drawn from the boxplot graphs is that the ratings have a distribution close to gaussian only in the range from 3.7 to 5.0 points:

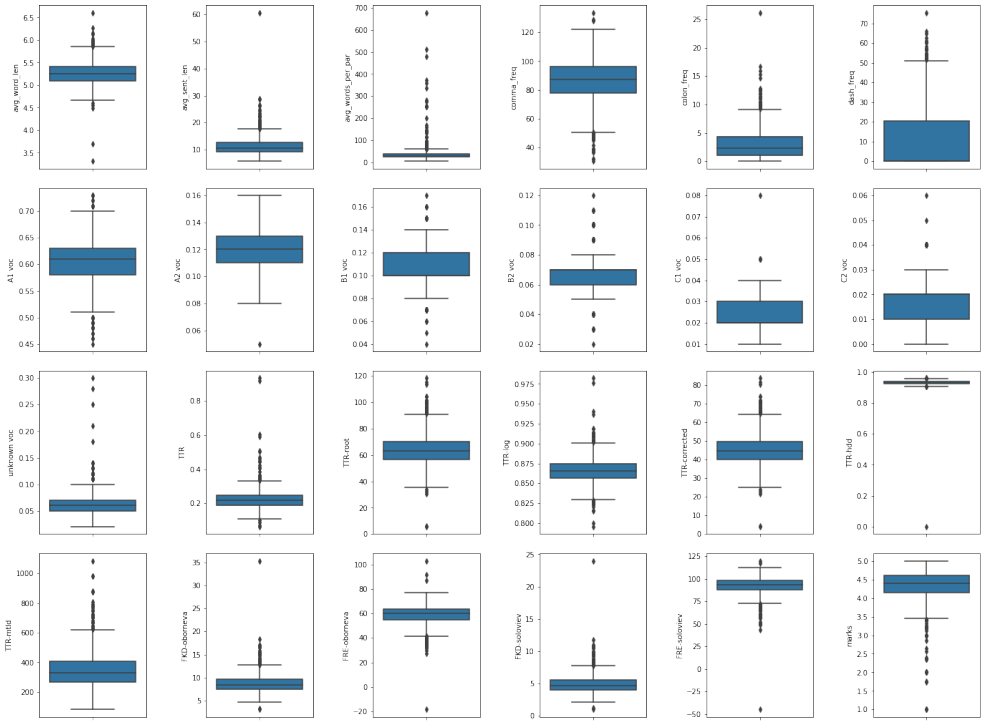

Figure 2: The feature distribution.

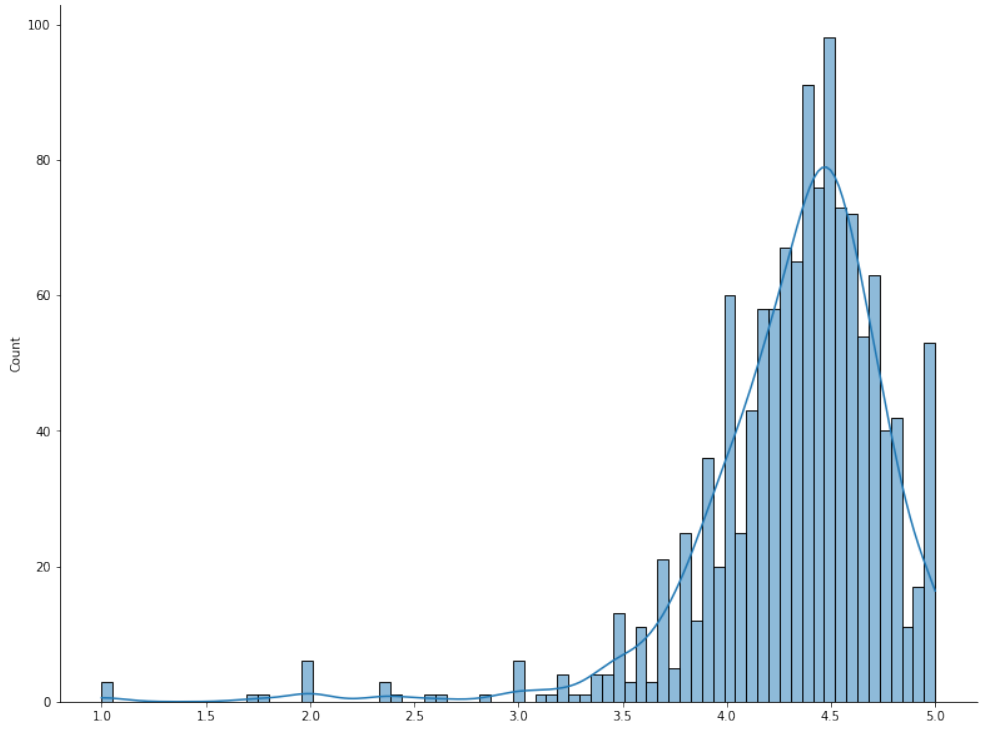

Figure 3: Distribution of the marks.

4.2 Designing an experiment

Based on the data analysis, I will formulate a null hypothesis—linguistic and syntactic features of the literary work can be used in predicting its reader rating. Since the ratings are distributed abnormally, it will be difficult to accurately predict the real rating in the regression problem, since it will be advantageous for the model to give a prediction of 4 to 4.5 points on all objects, and the corpus is too small for undersampling. It is necessary to check the statement experimentally. Linear Regression algorithms, Random Forest ensembles and Gradient Boosting were used to predict the estimate.

Model name	Coefficient of determination	MSE	MAE
Linear Regression	-0.14	0.35	0.29
Random Forest Regression	0.02	0.34	0.25
Cat Boost Regression	0.03	0.33	0.25

Table 3: Regression task results.

As expected, the coefficient of determination is extremely small, although the mean square (MSE) and the mean absolute error (MSA) are quite reasonable, which confirms nearly constant predictions of models. The problem could be reformulated. Let's propose there is a selection of works and their evaluations and it is necessary to determine which of them have a good rating. Heuristically, it will be logical to consider the median value above which all texts are conditionally successful. Dividing the corpus by the median, 636 and 619 works were obtained, which will be indicated by labels 0 and 1 for the binary classification problem. The ROC AUC score (Receiver Operating Characteristic Curve) will be considered as a target metric to establish the quality of the assignment of consecutive labels. After fitting the Logistic Regression, Random Forest Classifier, Cat Boost Classifier, and K-Nearest Neighbors models on a new dataset with the configured parameters, the following results were obtained on a test sample:

Model name	ROC AUC score on test
Logistic Regression	0.62
Random Forest Classifier	0.56
Cat Boost Classifier	0.55
K-Nearest Neighbors Classifier	0.56

Table 4: Classification task results.

Already by this data, it can be established that the models find a correlation between the target and the features. The results could be improved by selecting parameters for models.

Model name	ROC AUC score on test
Logistic Regression	0.63
Random Forest Classifier	0.61
Cat Boost Classifier	0.61
K-Nearest Neighbors Classifier	0.60

Table 5: Classification task results after hyperparameter tuning.

Thus, it was possible to improve the predictions of models, achieving a better consistent prediction of labels, although the result is compared on a test sample, whereas on cross-validation on all models it does not exceed 0.57.

Considering other metrics, can assume that we are solving a task of selecting the largest number of potentially successful texts from a certain number. In this case, it is necessary to use the Recall formula:

$$Recall = \frac{Successful\ documents\ predicted}{All\ successful\ documents}$$

Let's compare the results on models with the best quality:

Model name	Recall
Logistic Regression	0.69
Random Forest Classifier	0.69
Cat Boost Classifier	0.73

Table 6: Classification task results on Recall metric.

So, it can be noted that the use of linguistic parameters in predicting the rating of a text possesses a practical application.

Thus, it follows from the results obtained that it is possible to establish a minimal correlation between the syntactic and lexical features of the text with their user ratings—the null hypothesis is confirmed.

4.3 Discussion and further improvement

Since the null hypothesis has been confirmed, but the quality of the models does not allow using certain linguistic parameters to unambiguously establish the literature rating, there are several vectors of research development.

The first weak point of the experiment can be considered the composition of the corpus. It is desirable to increase its size several times and strengthen the genre diversity. In addition, a more thorough check of the texts inside the corpus is possible, since according to the results of the exploratory analysis, noises in the distributions were noticed, which signal possible failures in file formatting. It is also worth considering updating the developed parser, adding more tests for the correctness of the estimates found in order to reduce the number of manual checks and eliminate outliers in the data.

Next, it is necessary to significantly expand the number of extracted linguistic parameters. A deeper syntactic analysis could be used, and an algorithm for processing morphological features of the text could be developed. Algorithms for lemmatization and calculation of lexical diversity should be optimized, since they require significant computational power and processing a larger corpus can take several weeks.

The involvement of machine learning methods can also be expanded and attracts the latest approaches in the field of deep learning. Modern neural networks are significantly superior in quality to classical algorithms, which can improve the obtained prediction results.

In the context of the prospects for further exploration, the results of the study allow to link the structure and reception of the text by the reader. In this regard, many subsequent questions and assumptions arise. It is now possible, that linguistic and syntactic parameters of the text form a specific stylistic coloring, which is perceived differently by the subject of reading. On the other hand, a certain combination of features of the text may have a more or less predictable effect on the subsequent rating of the work. Then, with the particular set of parameters, it can be used in the preprocessing and evaluation of books, papers and, most interestingly within the framework of computational linguistics, in natural language generation studies.

5 Conclusion

Thus, in this study, the possible correlation of lexical and syntactic features of fiction with their reader rating was considered. To develop the experiment, a corpus of texts in the public domain was built and their estimates were parsed from the LitRes website using the developed parser. Then, an extractor of linguistic parameters was created in which many methods of stylometry and quantitative linguistics were implemented with the certain additions and adjustments.

After analyzing the data obtained, a hypothesis was put forward stating the possibility of using linguistic features of fiction to predict its reader rating. To confirm the hypothesis, an experiment based on machine learning methods was developed and implemented. Since during the experiment it was found out that on the basis of existing data it is not possible to predict an accurate number of ratings due to the nearly zero values of the determination coefficient, it was decided to refactor the experiment to the

classification problem. Dividing the sample of texts into two classes by median of their ratings and assuming that the upper part has a more successful rating than the other, it becomes possible to predict with acceptable quality whether the certain literary work will be evaluated with a high reader score.

The obtained results open up prospects for further research of the interrelationships of the text and the reader, which can be developed both in the field of linguistics and literature or natural language processing studies.

References

Herdan G. A. 1955. New derivation and interpretation of yule's 'characteristic' k. *Zeitschrift für angewandte Mathematik und Physik*, 6:332–334.

John Bissell Carroll. 1964. *Language And Thought*. Prentice-Hall.

CEFR. Common european framework of reference for languages.

Zipf G.K. 1949. Human behavior and the principle of least effort. *Addison-Wesley Press*, P 484–490.

Pierre Guiraud. 1954. *Les caractères statistiques du vocabulaire essai de méthodologie*. Presses universitaires de France, Paris.

Oborneva I.V. 2006. *Automated complexity evaluation of academic texts based on statistical metrics*. Inst. of Simulation a. Training, Univ. of Central Florida, Moscow, Russia.

Kincaid J.P., Fishburne R.P.Jr., Rogers R.L., and Chissom B.S. 1975. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*. Inst. of Simulation a. Training, Univ. of Central Florida, Millington, TN.

McCarthy P. M. and Jarvis S. 2010. Mtld, vocd-d, and hd-d: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Andryshina N.P et al. 2019. *Lexical minimum of Russian as a foreign language. Level C1*. Zlatoust, St. Petersburg.

Andryshina N.P and Kozlova T.V. 2015a. *Lexical minimum of Russian as a foreign language. Level A1*. Zlatoust, St. Petersburg.

Andryshina N.P and Kozlova T.V. 2015b. *Lexical minimum of Russian as a foreign language. Level B1*. Zlatoust, St. Petersburg.

Andryshina N.P and Kozlova T.V. 2015c. *Lexical minimum of Russian as a foreign language. Level B2*. Zlatoust, St. Petersburg.

Clancy S. 2014-2022. Visualizing russian: Teaching vocabulary at the intersection of frequency, grammar, and communication. *Harvard University*.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of Intelligent and Fuzzy Systems*, 34:1–10, 04.

M. C. Templin. 1957. *Certain Language Skills in Children: Their Development and Interrelationships*, volume 26. University of Minnesota Press, ned - new edition edition.