# Electronic Database of Non-Native Speakers' Speech Errors as a Scientific and Educational Resource (on the Example of Russian Schools)

**Elena V. Grudeva, Irina A. Buchilova, Alina A. Diveeva, Elena M. Ivanova, Oleg L. Selyanichev, Maxim S. Trapeznikov**

***Abstract*: The article is devoted to the development of an electronic database of speech errors of non-native speakers studying in Russian schools. The database of non-native speakers' speech errors developed by the authors includes the following parameters: 1) name of the non-native speaker, 2) age, 3) gender, 4) native language, 5) target word / phrase, 6) context1 (context that includes an error), 7) normative spelling, 8) context2 (context with a corrected error), 9) type of text in which the error occurred, 10) category of error, 11) type of error, 12) year of the material collection. Database pattern, as well as database program, which are a client-server Web application, are presented in the article. The study involved 87 non-native speakers of 11 nationalities: Azerbaijanis (25), Uzbeks (13), Ukrainians (13), Kyrghyz (11), Armenians (11), Tajiks (5), Dargins (2), Avars (2), Talysh (2), Belarusians (2), and Vietnamese (1). The authors have collected and systematized educational written texts of 4 types (dictations, essays, summaries, and copyings) of non-native speakers studying in general education schools of the city of Cherepovets of the Vologda Region. At the moment, there are 239 texts, among which: dictations comprise 96, summary - 26, essays - 41, copying - 76. In these texts, 2039 non-normative spellings were found and qualified. The authors of the project introduced a multidimensional system for the classification of errors that were identified in the written texts of the non-native speakers. At the first level, the following types of errors were distinguished: 1) nonspelling, 2) graphic, 3) spelling, 4) lexical, and 5) grammatical. Research and educational problems that can be solved by means of this resource are considered.***

**Revised Manuscript Received on December 30, 2019.**
* Correspondence Author

**Elena V. Grudeva***, Doctor of Philology Sciences, Professor, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia. Email: el.grudeva@yandex.ru

**Irina A. Buchilova**, PhD in Psychology Sciences, Associate Professor, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia

**Alina A. Diveeva**, Master of Philology, Post-graduate student, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia

**Elena M. Ivanova**, PhD in Philology Sciences, Associate Professor, Department of Social Communications and Media, Cherepovets State University, Cherepovets, Russia

**Oleg L. Selvanichev**, PhD in Technology Sciences, Associate Professor, Department of Mathematical and Software Computer, Cherepovets State University, Cherepovets, Russia

**Maxim S. Trapeznikov**, Student, Department of Mathematical and Software Computer, Cherepovets State University, Cherepovets, Russia

***Keywords*: Electronic Database, Russian Language, Speech Errors, Non-Native Speakers, Written Texts.**

## I. INTRODUCTION

In recent decades, in corpus linguistics, the researchers have being developing the corpora of language / speech errors, which often belong to the category of educational corpora [13].

On the one hand, the increased interest in development of various kinds of error corpora is caused by the use of corpus data in education. Nowadays, for example, foreign language studying is considered to be insufficient if only traditional teaching aids are applied, namely: vocabulary and grammar. The corpus of the target (studied) language is one of the obligatory teaching means. Implementation of the corpus allows students to verify the lexical and grammatical compatibility of words in their texts, the relevance of a language unit in a particular context. Therefore, students can check with corpus the questions that they used to address to native speakers of the studied language. The founder of corpus linguistics, John Sinclair, once predicted that "with the development of corpus, a native speaker will be displaced from the role of the main model and supreme arbiter in the issues of language practice" [3]. On the other hand, a corpus of errors is necessary not only for those who study a foreign language, but also for those who teach it. Teachers of foreign languages should be aware of the comparative characteristics of the language studied and the mother tongue for the student, the possible interference of the two language systems (contrastive linguistics is devoted to the development of methodological foundations in this area), as well as the typical mistakes made by native speakers of a particular language. Working with this kind of corpus, a teacher could constantly improve the pedagogical competence and teach a foreign language more effectively [9]. Error corpora development has been considered by the applied linguistics since the 90s of the 20th century. Granger, one of the theorists in this branch of corpus linguistics, called the introduction of educational corpora a revolution in applied linguistics (see the characteristic title of his article: "The learner corpus: a revolution in applied linguistics") [1].

A detailed overview of the most famous educational corpora in the world corpus linguistics is presented in the article "Learner corpora: design, development and applications" by Yukio Tono [4]. A remarkable fact is that most of the presented educational corpora are based on the texts of the foreigners studying English.

The development of Russian educational corpora is relatively new [5]. The leader in this area at the moment is the Linguistic Laboratory for Corpus Technologies of the Research Institute "Higher School of Economics". Previously, we conducted a review of the educational corpora, such as: The Corpus of Russian Educational Texts (CRET, https://ling.hse.ru/krut); Russian Learning Corpus RLC (http://web-corpora.net/RLC/) [9], based on, among other things, texts of so-called non-standard speakers of the Russian language (foreigners studying the Russian language; heritage speakers of the Russian language).

Specialists in the field of corpus linguistics and in the development of corpora of the use of language in educational activities have identified different groups of native speakers, depending on their status in the field of language proficiency. Thus, Mustajoki, Protasova and Vakhtin defined the following groups of "non-standard" native speakers of the Russian language: a) students studying Russian as a foreign language (RAF), b) native speakers of regional Russian, c) native speakers of so-called "heritage Russian", d) children and teenagers, e) adult speakers of standard Russian in certain circumstances [2]. Non-native speakers, studying in the Russian schools are considered as a special category of non-standard speakers of the Russian language.

Ideally, error corpora of different categories of non-native Russian speakers should be developed.

In the world and Russian practice there is a steady interest in the development of educational and error corpora, which are based on the texts of non-standard native speakers of the Russian language. There are observed some deviations from the norm. Such products can be applied in solving both research and pedagogical problems.

Due to the increase of multi-ethnic classes in general education schools of the Russian Federation, there is a necessity to individualize the learning means for children whose native language is not Russian. Therefore, it is necessary to develop at least an educational corpus with the texts of non-native speakers, and to study the characteristics of the mistakes.

The purpose of this article is to present the architecture of the electronic database of speech errors of non-native speakers studying in general education schools of the city of Cherepovets of the Vologda Region (Russia), and to show the possibilities of its application for scientific and educational purposes. The city of Cherepovets and the Vologda Region as a whole are among the mono-ethnic regions of the Russian Federation, although the number of non-native speakers with various ethnic and linguistic characteristics in Cherepovets schools is increasing [8].

The electronic database of speech errors of non-native speakers makes it possible to conduct linguistic research, systematize the speech errors of non-native speakers studying in Russian schools, and also develop diagnostic models for the language adaptation of non-native speakers of different age and ethnic categories.

## II. METHODS

The research task was to collect and analyze negative language material of the non-native speakers (written) within the framework of an electronic database of errors.

The research was carried out in four stages.

The first stage presupposed collecting speech material, which consisted of four types of educational written texts (dictations, essays, summaries and copyings) of non-native speakers studying in general education schools of the city of Cherepovets of the Vologda Region.

The material is represented by 239 texts, among which there are 96 dictations, 26 summaries, 41 essays, and 76 copyings.

The study involved 87 non-native speakers of 11 nationalities: Azerbaijanis (25), Uzbeks (13), Ukrainians (13), Kyrghyz (11), Armenians (11), Tajiks (5), Dargins (2), Avars (2), Talysh (2), Belarusians (2), and Vietnamese (1).

The second research stage was aimed at the analysis of the existing approaches to the systematization and classification of speech errors of both standard and non-standard Russian [11]. Multidimensional system for the classification of errors that were identified in the written texts of the non-native speakers was introduced [15].

It led to the third stage of the research, at which 2039 non-normative spellings in the analyzed speech material was identified.

The fourth stage dealt with the development of database parameters for non-native speakers' speech errors: 1) name of the non-native speaker, 2) age, 3) gender, 4) native language, 5) target word / phrase, 6) context$^1$ (context that includes an error), 7) normative spelling, 8) context$^2$ (context with a corrected error), 9) type of text in which the error occurred, 10) category of error, 11) type of error, 12) year of the material collection.

To gain the goal of the research, a database and an application for its operation were designed [12].

## III. RESULTS

The analysis of the non-native speakers' written speech production proved that it has a number of differences in the framework of phonemic and graphemic correspondences. These differences in each case form a different configuration, but linguistic interference can be regarded as the universal reason. In this regard, we have developed an approach that most fully describes the speech errors of non-native speakers, and introduced the multidimensional system for the classification of errors that were identified in the written texts of the non-native speakers.

At the first level, the following types of errors were distinguished: 1) nonspelling, 2) graphic, 3) spelling, 4) lexical, 5) grammatical. At the second level, in the framework of **1) nonspelling errors**, the following types of errors were identified:

a) writing together, separately, with a hyphen;

b) orthoepic mistakes (omissions, additions, permutations);

c) Mistakes caused by the insufficient understanding of the textual situation (all cases of the joint writing of prepositions were also referred to this kind).

In the framework of **2) graphic mistakes** the following types of errors were identified:

a) letters distortions (replacement of adjacent letters, mirror letters, etc.);

b) the use of capital and small letters;

c) sound [ʃ̑] and its writing;

d) double / single consonants in terms of longitude;

e) voiceless and voiced consonants in strong position;

f) vowels in strong position;

g) dropping of the separating hard and soft signs;

h) dropping of the soft sign;

i) dropping of the soft sign, used for the grammatical form identification;

j) replacement or omission of the letter Й (short I);

k) replacement of vowels used for the hardness and softness of consonant phonemes, substitutions Э-Е, А-Я, И-Ы, О-Ё and reverse substitutions;

l) non-normative writing after the hush sound and the letter Ц (C);

m) other features of the hardness and softness in writing (assimilative softness, excessive softness, etc.).

In the framework of **3) spelling mistakes** the following types of errors were identified:

a) various non-normative writing of the 1st type, following the principle “is spelled as it sounds”;

b) various non-normative writing of the 2nd type, following the principle “is not spelled as it sounds”;

In the framework of **4) lexical mistakes** the following types of errors were identified:

a) incorrect usage of words;

b) ambiguity causing two interpretations of the statement;

c) incomplete statement;

d) reiteration within a single sentence or between several sentences.

In the framework of **5) grammatical** mistakes the following types of errors were identified:

a) wrong number form;

b) agreement mistakes;

c) verb patterns mistakes;

d) coordination mistakes;

e) change of the word grammatical gender;

f) incorrect of verb form (including participles);

g) participle usage mistakes;

h) other grammatical mistakes (incorrect use of conjunctions, prepositions, case form of a noun).

Such classification contributed to the identification of 2039 non-normative spellings in the analyzed speech material.

According to the research results, the database and application for its operation were designed.

The database pattern is shown in Figure 1. The description of the database is introduced in Table 1.

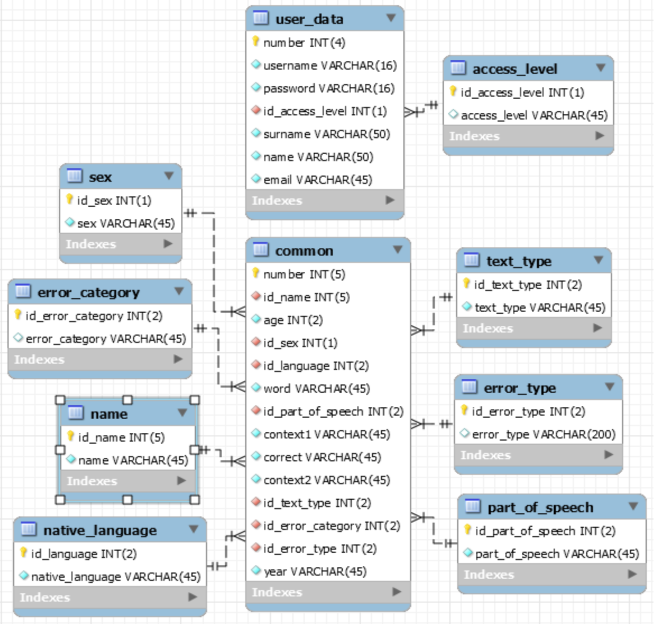


**Fig. 1. Database pattern.**

**Table I: Description of tables and database fields**

| Table name | Field indentifier | Description |
|---|---|---|
| common | number | Contains the record number |
| | id_name | Contains the participant’s id-number |
| | age | Contains the participant’s age |
| | id_sex | Contains id-number of a gender |
| | id_native_lang | Contains id- number of native language |
| | word | Contains a word with a mistake |
| | id_part_of_speech | Contains id-number of a part of speech |
| | context1 | Describes the sentence context with a mistake |
| | correct | Contains normative writing of the word with a mistake |
| | context2 | Contains normative writing of the sentence |
| | id_text_type | Contains id-number of a type of the text with a mistake |
| | id_error_category | Contains id-number of the mistake category |
| | id_error_type | Contains id-number of the mistake type |
| | year | Contains the year of collection |
| name | Id_name | Contains the participant’s id-number |
| | name | Contains the participant’s name |
| text_type | id_text | Contains the id-number of the text type |
| | text_type | Contains the description of the text type |
| native_language | id_native | Contains id-number of the native language |
| | language | Contains name of the native language |
| error_category | id_error_category | Contains id-number of the mistake category |
| | error_category | Contains the name of the mistake category |

| | | |
|---|---|---|
| error_type | id_error_type | Contains id-number of the mistake type |
| | error_type | Contains the name of the mistake type |
| part_of_speech | id_part_of_speech | Contains id-number of the part of speech |
| | part_of_speech | Contains the name of the part of speech |
| sex | id_sex | Contains id-number of the gender |
| | sex | Contains name of the gender |
| user_data | number | Contains the user's sequential number |
| | username | Contains the user's name |
| | password | Contains a password |
| | id_access_level | Contains id-number of the access level |
| | surname | Contains the user's surname |
| | name | Contains the user's name |
| | email | Contains the user's email |
| access_level | id_access_level | Contains id-number of the access level |
| | access_level | Contains the access level |

In the database, the main table "common" is linked with tables "name" for storing the names of children under the survey, "native_language" containing the name of the child's mother tongue, "text_type" with the types of the studied written text, "error_category" with the identified error categories and "error_type" with types of errors, "part_of_speech" with parts of speech with errors, and "sex" containing the data about the participants' gender. The "user_data" table shows users' data. All the fields except "id_access_level" are variable, as they are filled in by users. Thus, no optimization is possible. The "access_level" table contains a description of a user access level. In the main table "common", the fields "context1", "correct" and "context2" are not displayed in the corresponding tables because their values will take various values.

The program for operating the database is a client-server Web application.

The server part controls the database and users' work, and also transmits, receives and processes client data. The REST API architecture selected for the server allows using the server application with any device capable of sending requests, such as desktop computers and mobile devices.

For the client application, the React JavaScript library was selected, which will contribute to the compatibility with older browsers, and provide the key features for high-quality web application development. Figures 2-4 show the application windows.

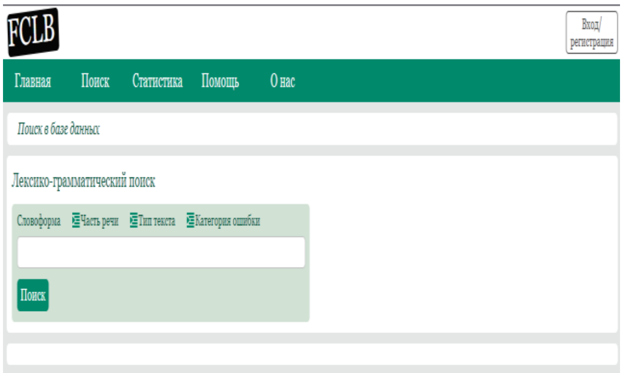
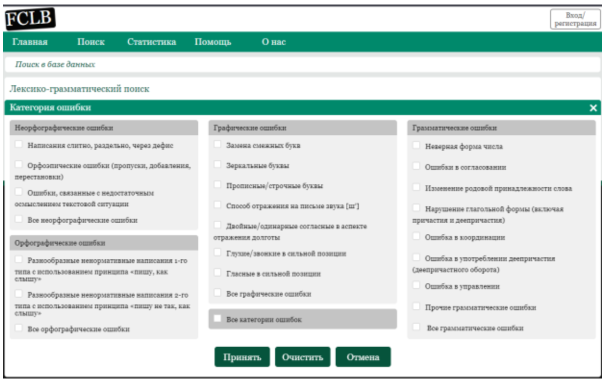


Fig. 2. General view of the search page



Неорфографические ошибки

□ Написания слитно, раздельно, через дефис
□ Орфоэпические ошибки (пропуски, добавления, перестановки)
□ Ошибки, связанные с недостаточным осмыслением текстовой ситуации
□ Все неорфографические ошибки

Орфографические ошибки

□ Разнообразные ненормативные написания 1-го типа с использованием принципа «пишу, как слышу»
□ Разнообразные ненормативные написания 2-го типа с использованием принципа «пишу не так, как слышу»
□ Все орфографические ошибки

Графические ошибки

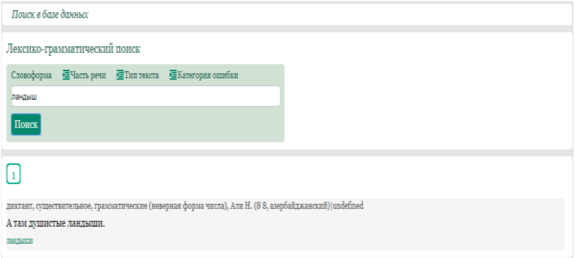
□ Замена смежных букв
□ Зеркальные буквы
□ Прописные/строчные буквы
□ Способ отражения на письме звука [ш']
□ Двойные/одинарные согласные в аспекте отражения долготы
□ Глухие/звонкие в сильной позиции
□ Гласные в сильной позиции
□ Все графические ошибки

□ Все категории ошибок

Грамматические ошибки

□ Неверная форма числа
□ Ошибки в согласовании
□ Изменение родовой принадлежности слова
□ Нарушение глагольной формы (включая причастия и деепричастия)
□ Ошибка в координации
□ Ошибка в употреблении деепричастия (деепричастного оборота)
□ Ошибка в управлении
□ Прочие грамматические ошибки
□ Все грамматические ошибки

Fig. 3. The window for the "Error Category" filter definition



ландыш

диктант, существительное, грамматические (неверная форма числа), Али Н. (8 8, азербайджанский)|undefined

А там душистые ландыши.

ландыши

Fig. 4. Search result window

## IV. DISCUSSION

The developed electronic database of speech errors of non-native speakers has great information potential for solving various research and educational problems.

Thus, a preliminary analysis of the collected empirical material, correlated with the classification of non-normative spelling in terms of the Russian writing system, showed that the question of the specificity of errors in the written activity of the non-native speakers remains highly controversial and requires a further study. Modern scientific literature on Russian speech production by the non-native speakers regards the non-normative spellings related to the positional principle of Russian writing, and, first of all, the hardness and softness of consonant phonemes in writing, as specific errors. They are characteristic for native speakers of the language in which there is no phonological contrast between hard and soft consonants [14].

According to the behavior strategy, children, who are non-native speakers, can be divided into two groups:

1) those who apply a graphic (visual) image in the designation of the softness of consonant phonemes;

2) those who apply the sound (acoustic) images and associations.

In the first case, the use of a soft sign to indicate softness (including assimilative) will be super-generalized. Thus, for example: [stol'ko rados'te] – "so much joy". In the dictations texts of the non-native speakers, there are cases of the absence of a soft sign for softness indication (for example, [vyisunulas] instead of [vyisunulas'] - "popped out", [los] instead of [los'] - "moose").

It is worth highlighting that in the article by Bogomazov "Classification of non-normative spelling in terms of the Russian writing studying by children", similar examples are identified in the writing of Russian-speaking children. The researcher declares that such spelling does not indicate the indistinctions between hard and soft phonemes in the child's linguistic consciousness [6]. All the noted features confirm the relevance of the classification of the speech errors of children, who are non-native speakers, especially the specificity of these errors. In this regard, the study of the speech errors specifics depending on the native language and on the text type was determined as the priority for further research. Two linguistic systems (mother tongue and Russian as non-native) interact in the minds of non-native speakers studying the Russian language. This causes cross-language errors. Usually, the following factors that impact the mastering of the Russian language are distinguished:

- the degree of similarity of the native language with Russian;
- graphic writing type of the native language;
- the accessibility or inaccessibility of a natural bilingual environment.

The Russian language studying in the context of natural bilingualism is influenced by the native language of a non-native speaker on the phonetic, lexical-semantic and grammatical levels.

## V. CONCLUSION

Development of learning corpora, which scientists called a revolution in applied linguistics, has great prospects for solving various research and educational problems. A systematic collection of the negative language material and the development of corpora of non-native speakers' errors can contribute to a deeper study of the influence of the native language on the mastering of the Russian language system in general and Russian writing in particular. The Information system for collecting, storing and processing data on speech errors of non-native speakers studying in general education schools in the city of Cherepovets, the Vologda Region is regarded as one of such resources.

## ACKNOWLEDGMENT

The study is funded by the Russian Foundation for Basic Research (project No. 18-412-350003 "Development of an electronic database of speech errors of non-native speakers studying in general education schools in the city of Cherepovets of the Vologda Region").

## REFERENCES

1. S. Granger, "The learner corpus: a revolution in applied linguistics," in *English Today*, 1994, vol. 39, no. 10/3, pp. 25-29.
2. A. Mustajoki, E. Protassova, N. Vakhtin, *Instrumentarium of linguistics: sociolinguistic approach to non-standard Russian*. Helsinki: Publishing House, 2010.
3. J. Sinclair, *How to use corpora in language teaching*. Amsterdam: Benjamins, 2004.
4. Yukio Tono. (n.d.). Learner corpora: design, development and applications [Online]. Available: http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf.
5. N.A. Zevakhina, S.S. Dzhakupova. (n.d.). Corpus of Russian student texts: design and prospects (Online). Available: http://www.dialog-21.ru/media/1144/zevakhinanadzhakupovass.pdf.
6. G.M. Bogomazov, "Classification of non-normative spelling in terms of the Russian writing studying by children," in *Language and Speech Activity*, 2002, pp. 83–96.
7. I.A. Buchilova, "Specific errors in the written works of the students, who are non-native speakers," in *Bulletin of Cherepovets State University*, 2018, vol. 4, no. 85, pp.39-47.
8. I.A. Buchilova, *Linguistic and ethnic identity of bilingual children studying in Cherepovets. Problems of the speech production and perception: Materials of the 14th distant school-seminar (Cherepovets, December 1-3, 2016)*. Cherepovets: Cherepovets State University, 2017.
9. E.V. Grudeva, I.A. Buchilova, N.A. Volkova, N.A. "Errors Corpus: target audience, possible architecture of the corpus," in *Bulletin of Cherepovets State University*, 2018, no. 5. pp. 63–72.
10. E.V. Grudeva, I.A. Buchilova, *The typology of written texts errors of non-native speakers of secondary school age. Problems of Developmental Psycholinguistics. Materials of the annual international scientific conference. March 20-23, 2018*. St. Petersburg: The Herzen *State Pedagogical University* of *Russia*, 2018.
11. E.V. Grudeva, E.M. Ivanova, I.A. Buchilova, A.A. Diveeva, "Typology of errors in the context of foreign children's speech activity," in *Bulletin of Cherepovets State University*, 2018, vol. 6, no. 87, pp. 78–89.
12. E.V. Grudeva, O.L. Selyanichev, M.S. Trapeznikov, *Students' activity for a research project implementation. Information and pedagogical technologies in a modern educational institution. Materials of the 10th All-Russian Scientific and Practical Conference (Cherepovets, April 5, 2019)*. Cherepovets: ChSU, 2019.
13. E.V. Rakhilina, On new tools for Russian grammar summary: errors corpus. Symposium "Russian Grammar 4.0": Linguistics [Online]. Available: http://www.rakhilina.ru/file/rahilina_foreign_err.pdf.
14. S.N. Tseytlin, *Azerbaijani-Russian bilinguals' specific errors in written speech. Russian Studies and the Present: Materials of the 7th International Scientific and Practical Conference. September 17-18, 2004*. St. Petersburg: Sudarynya, 2005.
15. R.H. Abbas, F.A.E.A. Kareem, "Text language identification using letters (frequency, self-information, and entropy) analysis for English, French, and German Languages," in *Journal of Southwest Jiaotong University*, 2019, vol. 54, no. 4. Available: http://jsju.org/index.php/journal/article/view/334

## AUTHORS PROFILE

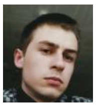
**Elena Valeryevna Grudeva** is a Doctor of Philology Sciences, Professor, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia.

**Irina Anatolyevna Buchilova** is a PhD in Psychology Sciences, Associate Professor, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia.

**Alina Albertovna Diveeva** is a Master of Philology, Post-graduate student, Department of National Philology and Applied Communications, Cherepovets State University, Cherepovets, Russia.

**Elena Mikhailovna Ivanova** is a PhD in Philology Sciences, Associate Professor, Department of Social Communications and Media, Cherepovets State University, Cherepovets, Russia.

**Oleg Leonidovich Selyanichev** is a PhD in Technology Sciences, Associate Professor, Department of Mathematical and Software Computer, Cherepovets State University, Cherepovets, Russia.



**Maxim Sergeevich Trapeznikov** is a Student, Department of Mathematical and Software Computer, Cherepovets State University, Cherepovets, Russia.