УДК 378.016:811.161.1. ББК Ш141.12-9-99. ГРНТИ 14.35.09. Код ВАК 5.8.2

# RuTOC: A CORPUS OF ONLINE LESSONS IN RUSSIAN AS A FOREIGN LANGUAGE

**Maria Yu. Lebedeva**
Pushkin State Russian Language Institute (Moscow, Russia)
ORCID ID: https://orcid.org/0000-0002-9893-9846

**Antonina N. Laposhina**
Pushkin State Russian Language Institute (Moscow, Russia)
ORCID ID: https://orcid.org/0000-0003-0693-7657

**Natalia A. Alksnit**
Pushkin State Russian Language Institute (Moscow, Russia)
ORCID ID: https://orcid.org/0000-0002-4009-4262

**Tatyana V. Lyashenko**
Pushkin State Russian Language Institute (Moscow, Russia)
ORCID ID: https://orcid.org/0000-0002-5514-5333

*Abstract.* The paper describes the project of RuTOC – a corpus of online lessons in Russian as a foreign language – and presents the first results of corpus analysis. The corpus RuTOC presented in the article is a special type of corpus, specifically, a corpus of classroom academic or educational discourse. Such collections of text data serve as a basis of discursive and sociolinguistic studies of classroom communication and investigation of the second language acquisition and make a certain contribution to the development of the pedagogical theory and practice. The relevance of the study stems from the fact that for the first time it collected, pre-processed and marked samples of classroom communication in Russian language classes; the corpus has created opportunities for evidence-based research in the theory and practice of teaching Russian as a foreign language. In addition, the relevance of the study is related to the increased need to study the peculiarities of online language learning during the pandemic.

The paper describes the process of creating the corpus, which includes the following steps: 1) collecting video recordings of RFL classes; 2) developing a standard for transcribing video recordings and creating a collection of transcripts; 3) developing a corpus marking system; 4) corpus data marking; 5) post-processing and analysis of the corpus. Currently, the corpus consists of 40 transcripts of lessons with a total duration of more than 56 hours and a total volume of 236,400 words; the first version of the corpus includes lessons in the Russian language at different educational levels, from the pre-university to the master»s program, at three Russian universities.

The article presents some difficulties and peculiarities of the transcription and marking of materials and the first results of a corpus analysis aimed at identifying the differences between student talk and teacher talk in RFL classes. It has been found that online RFL classes are generally characterized by high interactivity, understood as the ratio of conversational turns to total speech amount; at the same time, there is a significant imbalance between the amount of teacher talk and student talk. The paper concludes with a suggestion of promising directions for research on the basis of RuTOC.

*Keywords:* educational discourse corpus; educational discourse; educational texts; foreign language corpus; pedagogical corpus; corpus development; Russian as a foreign language; methods of teaching Russian; information and communication technologies; informatization of education; information educational environment; online lessons

*Acknowledgments:* the reported study was funded by Russian Science Foundation (RSF), project number 21-78-00126

*For citation:* Lebedeva, M. Yu., Laposhina, A. N., Alksnit, N. A., Lyashenko, T. V. (2022). RuTOC: A Corpus of Online Lessons in Russian as a Foreign Language. In *Philological Class*. Vol. 27. No. 2, pp. 19–29.

# RuTOC: КОРПУС ОНЛАЙН-УРОКОВ ПО РУССКОМУ ЯЗЫКУ КАК ИНОСТРАННОМУ

**Лебедева М. Ю.**
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
ORCID ID: https://orcid.org/0000-0002-9893-9846

**Лапошина А. Н.**
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
ORCID ID: https://orcid.org/0000-0003-0693-7657

**Алкснит Н. А.**
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
ORCID ID: https://orcid.org/0000-0002-4009-4262

**Ляшенко Т. В.**
Государственный институт русского языка им. А. С. Пушкина (Москва, Россия)
ORCID ID: https://orcid.org/0000-0002-5514-5333

*Аннотация.* В статье описывается опыт создания корпуса онлайн-занятий по русскому языку как иностранному RuTOC; приводятся первые результаты корпусного анализа учебной коммуникации на онлайн-уроках по РКИ. Представленный в статье корпус онлайн-уроков по РКИ относится к таким корпусам специального типа, как корпус учебного, или академического, дискурса. Подобные коллекции текстовых данных служат основой для дискурсивных, социолингвистических исследований учебной коммуникации, исследований по освоению иностранного языка, а также вносят вклад в развитие теории и практики обучения. Актуальность исследования обусловлена тем, что в нем впервые были собраны, предобработаны и размечены образцы учебной коммуникации на занятиях по русскому языку; коллекция собранных данных создает возможности для проведения доказательных исследований в области теории и методики преподавания РКИ. Кроме этого, актуальность исследования связана с возросшей в период пандемии потребностью в изучении особенностей онлайн-обучения языку.

В статье описан процесс создания корпуса, включающий в себя следующие этапы: 1) сбор видеозаписей занятий по РКИ; 2) разработка стандарта транскрибации видеозаписей и создание коллекции транскриптов; 3) разработка системы корпусной разметки, учитывающей специфику материала; 4) разметка корпусных данных; 5) постобработка и анализ корпуса. В настоящее время корпус состоит из 40 транскриптов уроков общей продолжительностью более 56 часов и общим объемом 236 400 слов; в первую версию корпуса вошли занятия по РКИ, проведенные в трех вузах России на разных уровнях образования – от подготовительного факультета до магистратуры. В статье приводятся некоторые сложности и особенности транскрибации и аннотации материалов и первые результаты корпусного анализа, направленного на выявление соотношения речи студентов и речи преподавателя на занятиях по РКИ. Было обнаружено, что онлайн-занятия по РКИ характеризуются в целом высокой интерактивностью, понимаемой как соотношение чередований говорящих к общему объему речи; вместе с этим наблюдается существенный дисбаланс между объемом речи преподавателя и речи студентов. В заключении статьи делаются предположения о перспективных направлениях исследований на материале корпуса RuTOC.

*Ключевые слова:* корпус учебного дискурса; учебный дискурс; учебные тексты; корпус иноязычной речи; педагогический корпус; создание корпуса; РКИ; русский язык как иностранный; методика преподавания русского языка; информационно-коммуникационные технологии; информатизация образования; информационная образовательная среда; онлайн-занятия

*Благодарности: Исследование* выполнено при финансовой поддержке Российского научного фонда (РНФ) в рамках научного проекта № 21-78-00126

*Для цитирования:* Лебедева, М. Ю. RuTOC: корпус онлайн-уроков по русскому языку как иностранному / М. Ю. Лебедева, А. Н. Лапошина, Н. А. Алкснит, Т. В. Ляшенко. – Текст : непосредственный // Филологический класс. – 2022. – Том 27, № 2. – С. 19–29.

**Introduction**

During the Covid-19 pandemic, much of the communication has moved to the online environment, and this transition was especially noticeable in education. Classes began to be conducted online, in videoconferencing format, and were often recorded. This expanded the possibilities for collecting corpus of the educational discourse.

The creation and study of corpora of classroom and academic discourse takes a specific place in corpus research. Such datasets are of interest to research of spontaneous or planned academic speech; they enable the various discourse and sociolinguistics studies [Biber 2006; Walsh 2006; Csomay 2012; Evison 2013; Sung, Kim 2020]. On the other hand, studies of such corpora have implications for educational theory and practice. The lectures or lessons transcripts collections are used to explore teachers' approaches, strategies, and tactics [Atwood, Turnbull, Carpendale 2010; Farr, Riordan 2015; Betz et al. 2019], and the content aspects of teaching [Biber, Conrad, Cortes 2004]. A special type of instructional discourse corpus is the collection of recordings of foreign language lessons; such corpora are valuable for language development research and L2 pedagogy.

The paper presents the project of the RuTOC – corpus of online lessons in Russian as a foreign language. The data collection, the current structure of the corpus, procedures of speech transcription and annotation are described. The first results of the corpus analysis are presented, and assumptions are made about the directions of research on the material of the corpus.

**Related work:**
**corpora of classroom discourse**

As D. Biber notes, much of research of academic discourse has been motivated by applied concerns about specific kinds of language a L2 learner will need [Biber 2006: 6].

One of the first corpora of spoken academic speech is Michigan Corpus of Academic Spoken English (MICASE; 1,7 million words; approximately 200 hours), which represents contemporary university speech in 15 different types of speech events such as small/large lectures, study groups, student presentations, etc. that have taken place in Michigan University [Simpson et al. 2002: 4–5]. The MICASE is open source, equipped with web search tools. A number of studies have been conducted on the MICASE materials, e. g. study of the academic vocabulary [Fortanet-Gomez 2004], implication of corpus methods in language teaching [Tehseen, Abbas 2018].

British analog to MICASE – BASE (British Academic Spoken English; 1,6 million tokens; approximately 200 hours) is a multimodal corpus that contains 160 lectures and 39 seminars transcriptions and recordings representing undergraduate and postgraduate level of education [Nesi, Thompson 2006]. BASE enabled a study of educational communication [Vodyanitskaya, Yaremenko 2020; Nergis 2021] and The BASE allowed for research on pedagogical communication and became the basis for practical guidelines for English teachers [Breeze, Sancho Guinda 2021].

Another example of an academic speaking corpus is the Limerick-Belfast Corpus of Academic Spoken English (LIBEL; approximately 500,000 words) [O'Keeffe and Walsh 2012]. E.g. based on LIBEL, the research of how the forms and frequency of some vague language expressions change in everyday and formal/institutional contexts was conducted [O'Keeffe 2008: 9].

Non-open access TOEFL 2000 Spoken & Written Academic Language Corpus (T2K-SWAL Corpus; 2,7 million words) was created purposefully for solving language learning issues. Containing spoken and written samples of American academic discourses, it has been used to design receptive components of the TOEFL 2000 exam [Biber et al. 2004: 7].

There is also a separate group of corpuses collected to study certain features of the L2 educational process in specific conditions. LEarning and TEaching Corpus (LETEC) contains interaction data of global simulation in French as foreign language [Wigham, Chanier 2013]. The Primary English Classroom Corpus (PECC) consists of 30 transcripts of primary school lessons in EFL classrooms in Germany [Limberg 2019]. SEN Classrooms Corpus (52,813 words) was created to investigate teacher discourse in special educational needs classrooms [Smith 2020]. Teacher-Student Chat Corpus (TSCC; approximately 133,000 words) contains written conversations from 102 one-to-one online English lessons [Caines 2020].

The multimodal, multilingual and multidisciplinary corpus of classroom discourse is SCoRE (Multimodal Corpus Database of Education Discourse in Singapore Schools; 500h in total). It in-

cludes the annotated data of Singapore primary and secondary schools classroom lessons [Hong 2005].

As one can easily see, the corpora, with few exceptions, represent either L1 academic speech in English or samples of classroom discourse in L2 English classes. We are not aware of any corpus that would represent samples of L2 Russian learners' classroom discourse.

**Data collection and transcription**

The corpus is currently represented by recordings of classes of Russian as a foreign language held at several Moscow universities from the spring semester of 2020 to the spring semester of 2021. The corpus includes classes for different levels of education: pre-university, bachelor programs, master programs, and internships for international students. Students' Russian language proficiency levels range from A1 to B2 according to the CEFR. The core of the corpus is the practical Russian language course, and courses in the language for academic purposes of specialization and specific courses such as listening class, writing class, etc. are also presented.

The classes that formed the basis of the corpus were conducted in monolingual groups, among which L1 Chinese speakers predominated, or in multilingual groups of varying composition (e. g. L1 Chinese, Arabic, Bulgarian, Farsi, Indonesian students study together in the same group).

All classes were conducted on the Zoom video conferencing platform. Students and instructors were notified that classes were being recorded. According to APA Ethics Code informed consents were not signed, as research involves the study of normal educational practices[1].

The collected recordings were transcribed in semi-automatic mode. L2 speech fragments were transcribed entirely manually, as automatic speech recognition software was not able to recognize the accent. Hesitation and filler words were also transcribed.

Following the example of MICASE, we have applied standard orthography in transcriptions, so phonetic variations of L2 Russian speech are not reflected in RuTOC. At the same time, grammatical and lexical errors in the students' speech were not corrected and were recorded with the accuracy that was available to transcribers, e. g.:

S1: *Как дела?*

S2: *Эээ… Хорошо, спасибо, а как твоя… а как тебе дела?*

In order to respect the anonymity requirements, we replaced personal names with TEACHER or STUDENT N placeholders to protect the privacy of teacher and student participants. No other sensitive data were found in the dataset.

Mark-up conventions of RuTOC are shown in Table 1.

Table 1. Mark-Up Conventions

| CATEGORY | CODE | DEFINITION |
|---|---|---|
| PAUSES | . | Period indicates a brief pause accompanied by an utterance final (falling) intonation contour; not used in a syntactic sense to indicate complete sentences |
| | … | Ellipses indicate a brief (1–2 second) mid-utterance pause with non phrase-final intonation contour |
| | <PAUSE DESC="3"/> | Pauses of 3 seconds or longer are timed to the nearest second |
| LAUGHTER | <EVENT DESC="LAUGH"/> | |
| SMILE | <EVENT DESC="SMILE"/> | |
| ORGANIZING EVENTS | <EVENT DESC="SHOW_SLIDE"/> | |
| UNCERTAIN or UNINTELLIGIBLE SPEECH | <EVENT DESC="UNCERTAIN"/> | |
| NAMES | When participants' names occur in a recording, they are changed to corresponding speakers ID in the transcript | |
| MISTAKES AND MISSPELLING | Standard orthography is applied; grammatical and lexical mistakes are reproduced | |

[1] APA Ethics Code. URL: https://www.apa.org/ethics/code.

**Corpus structure**

The current state of RuTOC consists of transcripts of 40 lessons, with a total duration of 56h 19min. The total volume of the corpus is 236 395 words.

The corpus contains speech samples from 9 instructors. The total number of speakers labeled as students in all transcripts is 392; however, due to anonymization, we cannot currently establish overlap between students in different transcripts.

The current composition of the corpora is shown in Table 2.

Table 2. Current composition of the RuTOC

| Education level | Number of transcripts | Number of words | Duration (min) |
|---|---|---|---|
| Language internship | 13 | 81 962 | 1 365 |
| Pre-university faculty | 10 | 70 818 | 907 |
| Master program | 9 | 45 496 | 601 |
| Bachelor program | 6 | 29 848 | 394 |
| Extra courses | 2 | 8 271 | 112 |
| Total | 40 | 236 395 | 3 379 |

**Corpus Annotation**

RuTOC mark-up provides information about lesson, speaker and speech fragments attributes, relevant for a research of the online language teaching. Annotation scheme is based on the MICASE corpus description where it was possible.

Each lesson of the corpus is categorized according to various attributes, including organizing attributes such as date and name of institution, lesson duration time, number of participants and pedagogical attributes such as educational level of participants (e.g., pre-university program or master program), main topic of a lesson, approximate Russian proficiency level of a students group. These attributes can be found in the header of each transcript. Each speaker is annotated in terms of its role in educational process (e.g., teacher or student), individual speaker ID (which allows us to explore the different pedagogical strategies of one teacher among different lessons and students groups, or compare the speech of some students within the same study group), speaker demographic variables (e.g., gender and age), speaker native language. Finally, each speech fragment includes the information about speaker ID, starting time of a speech fragment and camera setup of a speaker. A description of all the attributes and their corresponding codes is shown in Table 3.

Class session transcripts and attributes are XML format files and compatible with Text Encoding Initiative (TEI). An example of corpus annotation of RuTOC is presented in the Picture 1.

Table 3. Annotation attributes of RuTOC

| Level of attributes | Category | Code/definition |
|---|---|---|
| Lesson attributes | TITLE | Main topic of the lesson |
| | INSTITUTION | Institution name |
| | DATE | Lesson date |
| | WORDCOUNT | Number of words in transcript |
| | DURATION | Lesson duration in minutes |
| | STUDENTS NUMBER | Number of present students |
| | CEFR LEVEL | |
| | A1–A2 | A |
| | B1–B2 | B |
| | C1–C2 | C |
| | EDUCATION LEVEL | |
| | Pre-university faculty | pu |

| LEVEL OF ATTRIBUTES | CATEGORY | CODE/DEFINITION |
|---|---|---|
| | Bachelor program | bc |
| | Master program | md |
| | Extra courses | c |
| | Language internship | in |
| | INTERACTIVITY RATE | |
| | Low | L |
| | Medium | M |
| | Hight | H |
| Speaker attributes | GENDER | |
| | Female | F |
| | Male | M |
| | AGE GROUP | |
| | 17–23 | 1 |
| | 24–30 | 2 |
| | 31–50 | 3 |
| | 51 and older | 4 |
| | ACADEMIC ROLE | |
| | Teacher | Teacher_ID |
| | Student | Student_ID |
| | NATIVE SPEAKER STATUS | |
| | Native speakers of Russian | NS |
| | Non-native speakers of Russian | NNS |
| | FIRST LANGUAGE | |
| | ISO 639-2 code of speaker`s native language | CHI |
| Speech fragment attributes | TIME | Starting time of speech fragment |
| | SPEAKER | Speaker_ID |
| | SPEAKER'S CAMERA | |
| | Camera is on | ON |
| | Camera is off | OFF |
| | Photo or avatar is seen | AVATAR |

```
<p SPEAKER="Teacher_1" ROLE="TEACHER" LANG="NS" FIRSTLANG="RU" SEX="F" AGE="3" CAMERA="ON" TIME="03:18">
Итак, скажите, пожалуйста, от какого глагола мы образовали слово "знать", то есть слово "знание". Слово "знание". <EVENT DESC="SHOW_SLIDE"/> От какого глагола мы получили?
</p>
<p SPEAKER="Student_1" ROLE="STUDENT" LANG="NNS" FIRSTLANG="CHI" SEX="F" AGE="1" CAMERA="OFF" TIME="03:23">
Э... можно? <EVENT DESC="UNCERTAIN"/>
</p>
<p SPEAKER="Teacher_1" ROLE="TEACHER" LANG="NS" FIRSTLANG="RU" SEX="F" AGE="3" CAMERA="ON" TIME="03:25">
Да, STUDENT_1.
</p>
<p SPEAKER="Student_1" ROLE="STUDENT" LANG="NNS" FIRSTLANG="CHI" SEX="F" AGE="1" CAMERA="OFF" TIME="03:27">
Э... От глагола "знать".
</p>
```

Picture 1. An example of corpus annotation format of RuTOC

**First results**

For the initial analysis of the corpus, we calculated two parameters of foreign language classroom discourse: interactiveness of lessons and ratio of student talking time (STT) to teacher talking time (TTT).

The levels of interactiveness were distinguished in line with T2K-SWAL calculation methodology: a score fewer than 10 turns per 1,000 words corresponds to low interactiveness; a score between 10 and 25 turns per 1,000 words corresponds to medium interactiveness; and a score more than 25 turns per 1,000 words corresponds to high interactiveness [Biber et al. 2004: 9]. We expect to see a high level of interactivity in the foreign Russian language classes we have collected in RuTOC.

Teacher talking time (TTT) and student talking time (STT) are basic categories in a foreign language teaching methodology. In the modern communicative language teaching approach, it is considered that effective L2 lessons keep a balance between TTT and STT. Moreover, some recommendations instruct teachers to increase STT by up to 70–80%[1]. In the case of RuTOC, teacher talking time was calculated as the ratio of the number of words said by a teacher to the total number of words said during the lesson. We hypothesized that the ratio of STT to TTT would vary depending on the group's level of Russian language proficiency.

The interactiveness score and the distribution of lesson word volume between TTT and STT for each level of education are shown in Table 4.

Table 4. Interactiveness and Teacher Talking Time / Student Talking Time over the educational level

| Education level | Russian language proficiency level | Interactiveness ± SD | Teacher talking time (%) | Student talking time (%) |
|---|---|---|---|---|
| Pre-university faculty | A1–A2 | 92±31 | 79 | 21 |
| Bachelor program | B1–B2 | 85±34 | 71 | 29 |
| Master program | B1–B2 | 98±21 | 67 | 33 |
| Language internship | B1–B2 | 52±13 | 61 | 39 |
| Extra courses | A1–A2 | 48±34 | 85 | 15 |
| Average | – | 77±31 | 74 | 26 |

The interactiveness score in our data differs from the T2K-SWAL collection, where sessions with more than 25 turns per thousand words are considered to be highly interactive. In the RuTOC data, this score varies from 24 to 143. This is due to the specifics of the language class, where turn-taking occurs rather frequently, unlike an academic lecture. Therefore, there is an obvious need to modify the scale of interactiveness to suit the peculiarities of the language class.

It is noticeable that despite the high interactiveness score, compared with T2K-SWAL data, in most of the corpus classes, the STT to TTT ratio is far from balanced. In the sessions we analyzed, the teacher's speech tends to be monological, while the student's speech may be limited to 1–2 words (*yes, I understand*, a short answer to a direct question, etc.). This is confirmed by the number of words spoken by students in one speech fragment: more than 50% of all students' speech fragments are between 1 to 3 words (Picture 2). In some cases this is motivated by the goal of the class (e.g., to teach students how to listen to an academic lecture). However, on the whole, we find that even in classes for high levels of Russian language proficiency, the student's role is only that of respondent. These findings also indicate the need for further development of ways to measure interactivity and students' learning activity, depending on this type of language lessons.

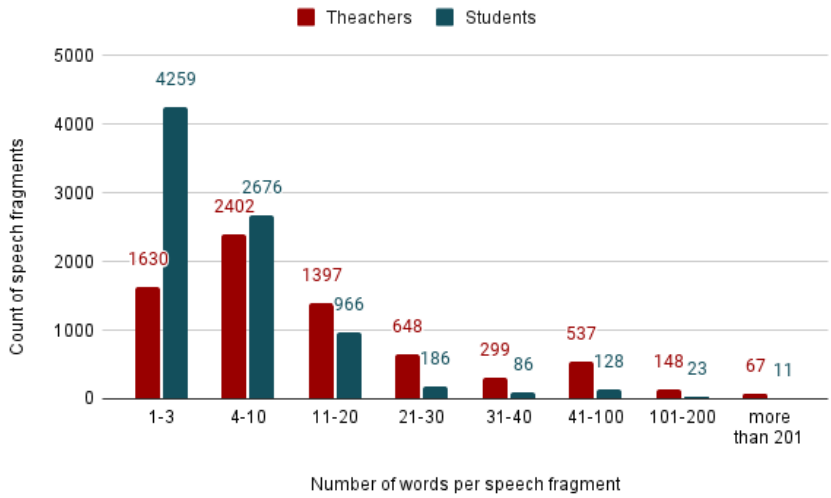
**Conclusion**

We have described the RuTOC – corpus of online lessons in Russian as a foreign language, the first resource of classroom discourse of L2 Russian learners available for research use. RuTOC currently contains 40 class session transcripts, totalling 236 400 words, in the future it will be expanded with new data.

Annotation of RuTOC is intended to enable research of the educational discourse in online learning settings, language teaching process, and Russian as a foreign language development.

A short initial analysis we conducted on RuTOC data revealed that interactiveness of the classroom discourse, calculated as the number of communicative turns per word count, doesn't imply an active learning environment in which students are able to maximize their use of the target language.

[1] E.g. American Council on the Teaching of Foreign Languages recommends following the "80:20 rule in the language classroom". URL: https://www.actfl.org.




Picture 2. Number of words per students' VS teachers' speech fragments

The presented version of the corpus is the result of the first stage of work on it. We plan to develop this resource in several directions. First, a promising development of the corpus may be the addition of error markup, in which case RuTOC may become an error-annotated learner corpus with focus on the oral speech of L2 students. Second, for pedagogical discourse research it might be valuable to present the corpus (or parts of it) as multimodal, with markup by gestures, teacher strategies, and demonstrated materials.

In the next iteration, the collection of material for the corpus is intended to be longitudinal in order to track changes in students' speech at different stages of Russian language learning.

Finally, future research will involve a more detailed examination of L2 classroom communication, particularly the distribution of each individual student's speech time in the total group STT, as well as an analysis of what teacher strategies lead to increased STT in the classroom.

**Литература**

Atwood, S. The construction of knowledge in classroom talk / S. Atwood, W. Turnbull, J. I. M. Carpendale // Journal of the Learning Sciences. – 2010. – Vol. 19 (3). – P. 358–402.

Barker, F. How can corpora be used in language testing? / F. Barker // The Routledge Handbook of Corpus Linguistics. – 1st ed. – London : Routledge, 2010. – Vol. 34 (3). – P. 637.

Betz, N. Cognitive Construal-Consistent Instructor Language in the Undergraduate Biology Classroom / N. Betz, J. S. Leffers, E. E. D. Thor, M. Fux, K. de Nesnera, K. D. Tanner & J. D. Coley. – Text : electronic // CBE – Life Sciences Education. – 2019. – Vol. 18 (4), ar63. – P. 1–16. URL: http://www.sfsusepal.org/wp-content/uploads/2021/10/cbe.19-04-0076.pdf.

Biber, D. If you look at …: Lexical Bundles in University Teaching and Textbooks / D. Biber, S. Conrad, V. Cortes // Applied Linguistics. – 2004. – Vol. 25, Issue 3. – P. 371–405. – https://doi.org/10.1093/applin/25.3.371.

Biber, D. Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. Report Number: RM-04–03, Supplemental Report Number: TOEFL-MS-25 / D. Biber, S. Conrad, R. Reppen, P. Byrd, M. Helt, V. Clark, V. Cortes, E. Csomay and A. Urzua. – Princeton, NJ : Educational Testing Service, 2004. – URL: https://www.ets.org/Media/Research/pdf/RM-04-03.pdf. – Text : electronic.

Biber, D. Stance in spoken and written university registers / D. Biber // Journal of English for Academic Purposes. – 2006. – Vol. 5, Issue 2. – P. 97–116.

Biber, D. University Language: A corpus-based study of spoken and written registers / D. Biber. – J. Benjamins, 2006. – 261 p.

Breeze, R. Teaching English-Medium Instruction Courses in Higher Education: A Guide for Non-Native Speakers (Chapter 3: Lecturing in English) / R. Breeze & Sancho C. Guinda. – London : Bloomsbury, 2021.

Caines, A. The Teacher-Student Chatroom Corpus / A. Caines et al. – Text : electronic // Proceedings of the 9th Workshop on NLP4CALL. – 2020. – P. 10–20. – URL: https://aclanthology.org/2020.nlp4call-1.2.pdf.

Dapeng, W. On the Significance of English Classroom Discourse Corpus Construction / W. Dapeng. – Text : electronic // Proceedings of the 2014 Conference on Informatisation in Education, Management and Business. Vol. 7. – Atlantis Press, 2014. – P. 376–378. – URL: https://www.atlantis-press.com/proceedings/iemb-14/13752.

Evison, J. Turn openings in academic talk: where goals and roles intersect / J. Evison // Classroom Discourse. – 2013. – Vol. 4 (1). – P. 3–26.

Farr, F. Tracing the reflective practices of student teachers in online modes / F. Farr & E. Riordan // ReCALL. – 2015. – Vol. 27 (1). – P. 104–123. – doi: 10.1017/S0958344014000299.

Fortanet-Gomez, I. I think: opinion, uncertainty or politeness in academic spoken English? / I. Fortanet-Gomez // RAEL: revista electrónica de lingüística aplicada. – 2004. – No. 3. – P. 63–84.

Gillian, S. Using corpus methods to investigate classroom interaction and teacher discourse in special educational needs (SEN) classrooms: an investigation of methodological possibilities / S. Gillian. – Lancaster University, 2020. – URL: https://eprints.lancs.ac.uk/id/eprint/145660/1/2020SmithPhD.pdf. – Text : electronic.

Hong, H. Q. SCoRE: A multimodal corpus database of education discourse / H. Q. Hong // Proceedings from the Corpus Linguistics Conference Series. – Birmingham, 2005. – Vol. 1 (1).

Ishikawa, S. The ICNALE Spoken Dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews / S. Ishikawa // English Teaching. – 2019. – Vol. 74 (4). – P. 153–177.

Koester, A. Building small specialised corpora / A. Koester // The Routledge Handbook of Corpus Linguistics. – 1st ed. – London : Routledge, 2010. – Vol. 34 (3). – P. 66–78.

Limberg, H. The Primary English Classroom Corpus (PECC) / H. Limberg. – FLENSBURG UNIVERSITY PRESS, 2019. – Vol. 1. – 450 p. – URL: https://www.uni-flensburg.de/fileadmin/content/projekte/pecc/bilder/pecc/9783939858379.pdf. – Text : electronic.

Nergis, A. Can explicit instruction of formulaic sequences enhance L2 oral fluency? / A. Nergis // Lingua. – 2021. – Vol. 255. – https://doi.org/10.1016/j.lingua.2021.103072.

Nesi, H. British Academic Spoken English corpus / H. Nesi, P. Thompson. – Text : electronic // Oxford Text Archive. – 2006. – URL: http://hdl.handle.net/20.500.12024/2525.

O'Keeffe, A. Applying corpus linguistics and conversation analysis in the investigation of small group teaching in higher education / A. O'Keeffe, S. Walsh // Corpus Linguistics and Linguistic Theory. – 2012. – Vol. 8 (1). – P. 159–181.

O'Keeffe, A. Post-colonialism, multi-culturalism, structuralism, feminism, post-modernism and so on so forth – vague language in academic discourse, a comparative analysis of form, function and context / A. O'Keeffe, M. McCarthy & S. Walsh // Corpora and Discourse (SCL31) / ed. by R. Reppen and A. Ädels. – Amsterdam : John Benjamins, 2008. – P. 9–29.

Simpson, R. C. MICASE manual / R. C. Simpson, D. Y. W. Lee & S. Leicher. – MI: English Language Institute, The University of Michigan, 2002. – URL: https://ca.talkbank.org/access/0docs/MICASE.pdf. – Text : electronic.

Smith, G. Using corpus methods to investigate classroom interaction and teacher discourse in special educational needs (SEN) classrooms: an investigation of methodological possibilities / G. Smith. – Lancaster University, 2020. – 342 p. – URL: https://aclanthology.org/2020.nlp4call-1.2.pdf. – Text : electronic.

Sung, M. C. Spontaneous motion in L1- And L2-english speech: A corpus-based study / M. C. Sung, K. Kim. – Text : electronic // English Teaching. – 2020. – Vol. 75, No. 1. – P. 49–66. – URL: https://eric.ed.gov/?id=EJ1274540.

Tehseen, Z. Pedagogical Implications of Corpus-based Approaches to ELT in Pakistan / Z. Tehseen & A. Akhta // Journal of Education and Educational Development. – 2018. – No. 5. – P. 259. – 10.22555/joeed.v5i2.1565.

Vodyanitskaya, A. What is valuable in the academe: Corpus-based analysis. Society. Integration. Education / A. Vodyanitskaya, V. Yaremenko // Proceedings of the International Scientific Conference. – 2020. – Vol. II. – P. 437–455.

Wigham, C. R. LEarning and TEaching Corpora (LETEC): data-sharing and repository for research on multimodal interactions / C. R. Wigham, T. Chanier. – Text : electronic // WorldCALL. 10–13 juillet 2013. – Glasgow : Royaume-Uni., 2013. – URL: http://edutice.archives-ouvertes.fr/edutice-00778274.

**References**

Atwood, S., Turnbull, W., Carpendale, J. I. M. (2010). The Construction of Knowledge in Classroom Talk. In *Journal of the Learning Sciences*. Vol. 19 (3), pp. 358–402.

Barker, F. (2010). How Can Corpora Be Used in Language Testing? In *The Routledge Handbook of Corpus Linguistics*. 1st ed. London, Routledge. Vol. 34 (3), p. 637.

Betz, N., Leffers, J. S., Thor, E. E. D., Fux, M., de Nesnera, K., Tanner, K. D. & Coley, J. D. (2019). Cognitive Construal-Consistent Instructor Language in the Undergraduate Biology Classroom. In *CBE – Life Sciences Education*. Vol. 18 (4), ar63, pp. 1–16. URL: http://www.sfsusepal.org/wp-content/uploads/2021/10/cbe.19-04-0076.pdf.

Biber, D. (2006). Stance in Spoken and Written University Registers. In *Journal of English for Academic Purposes*. Vol. 5. Issue 2, pp. 97–116.

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. J. Benjamins. 261 p.

Biber, D., Conrad, S., Cortes, V. (2004). If You Look at …: Lexical Bundles in University Teaching and Textbooks. In *Applied Linguistics*. Vol. 25. Issue 3, pp. 371–405. https://doi.org/10.1093/applin/25.3.371.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E. and Urzua, A. (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus. Report Number: RM-04-03, Supplemental Report Number: TOEFL-MS-25*. Princeton, NJ, Educational Testing Service. URL: https://www.ets.org/Media/Research/pdf/RM-04-03.pdf.

Breeze, R. & Sancho Guinda, C. (2021). *Teaching English-Medium Instruction Courses in Higher Education: A Guide for Non-Native Speakers (Chapter 3: Lecturing in English)*. London, Bloomsbury.

Caines, A. et al. (2020). The Teacher-Student Chatroom Corpus. In *Proceedings of the 9th Workshop on NLP4CALL*, pp. 10–20. URL: https://aclanthology.org/2020.nlp4call-1.2.pdf.

Dapeng, W. (2014). On the Significance of English Classroom Discourse Corpus Construction. In *Proceedings of the 2014 Conference on Informatisation in Education, Management and Business*. Vol. 7. Atlantis Press, pp. 376–378. URL: https://www.atlantis-press.com/proceedings/iemb-14/13752.

Evison, J. (2013). Turn Openings in Academic Talk: Where Goals and Roles Intersect. In *Classroom Discourse*. Vol. 4 (1), pp. 3–26.

Farr, F. & Riordan, E. (2015). Tracing the Reflective Practices of Student Teachers in Online Modes. In *ReCALL*. Vol. 27 (1), pp. 104–123. doi: 10.1017/S0958344014000299.

Fortanet-Gomez, I. (2004). I Think: Opinion, Uncertainty or Politeness in Academic Spoken English? In *RAEL: revista electrónica de lingüística aplicada*. No. 3, pp. 63–84.

Gillian, S. (2020). *Using Corpus Methods to Investigate Classroom Interaction and Teacher Discourse in Special Educational Needs (SEN) Classrooms: An Investigation of Methodological Possibilities*. Lancaster University. URL: https://eprints.lancs.ac.uk/id/eprint/145660/1/2020SmithPhD.pdf.

Hong, H. Q. (2005). SCoRE: A Multimodal Corpus Database of Education Discourse. In *Proceedings from the Corpus Linguistics Conference Series*. Birmingham. Vol. 1 (1).

Ishikawa, S. (2019). The ICNALE Spoken Dialogue: A New Dataset for the Study of Asian Learners' Performance in L2 English Interviews. In *English Teaching*. Vol. 74 (4), pp. 153–177.

Koester, A. (2010). Building Small Specialised Corpora. In *The Routledge Handbook of Corpus Linguistics*. 1st ed. London, Routledge. Vol. 34 (3), pp. 66–78.

Limberg, H. (2019). *The Primary English Classroom Corpus (PECC)*. FLENSBURG UNIVERSITY PRESS. Vol. 1. 450 p. URL: https://www.uni-flensburg.de/fileadmin/content/projekte/pecc/bilder/pecc/9783939858379.pdf.

Nergis, A. (2021). Can Explicit Instruction of Formulaic Sequences Enhance L2 Oral Fluency? In *Lingua*. Vol. 255. https://doi.org/10.1016/j.lingua.2021.103072.

Nesi, H., Thompson, P. (2006). British Academic Spoken English corpus. In *Oxford Text Archive*. URL: http://hdl.handle.net/20.500.12024/2525.

O'Keeffe, A., McCarthy, M. & Walsh, S. (2008). Post-colonialism, Multi-culturalism, Structuralism, Feminism, Post-modernism and So on So Forth – Vague Language in Academic Discourse, a Comparative Analysis of Form, Function and Context. In Reppen, R. and Ädels, A. (Eds). *Corpora and Discourse (SCL31)*. Amsterdam, John Benjamins, pp. 9–29.

O'Keeffe, A., Walsh, S. (2012). Applying Corpus Linguistics and Conversation Analysis in the Investigation of Small Group Teaching in Higher Education. In *Corpus Linguistics and Linguistic Theory*. Vol. 8 (1), pp. 159–181.

Simpson, R. C., Lee, D. Y. W. & Leicher, S. (2002). *MICASE Manual*. MI, English Language Institute, The University of Michigan. URL: https://ca.talkbank.org/access/0docs/MICASE.pdf.

Smith, G. (2020). *Using Corpus Methods to Investigate Classroom Interaction and Teacher Discourse in Special Educational Needs (SEN) Classrooms: An Investigation of Methodological Possibilities*. Lancaster University. 342 p. URL: https://aclanthology.org/2020.nlp4call-1.2.pdf.

Sung, M. C., Kim, K. (2020). Spontaneous Motion in L1- And L2-English Speech: A Corpus-Based Study. In *English Teaching*. Vol. 75. No. 1, pp. 49–66. URL: https://eric.ed.gov/?id=EJ1274540.

Tehseen, Z. & Akhta, A. (2018). Pedagogical Implications of Corpus-based Approaches to ELT in Pakistan. In *Journal of Education and Educational Development*. No. 5, p. 259. 10.22555/joeed.v5i2.1565.

Vodyanitskaya, A., Yaremenko, V. (2020). What Is Valuable in the Academe: Corpus-Based Analysis. Society. Integration. Education. In *Proceedings of the International Scientific Conference*. Vol. II, pp. 437–455.

Wigham, C. R., Chanier, T. (2013). LEarning and TEaching Corpora (LETEC): Data-Sharing and Repository for Research on Multimodal Interactions. In WorldCALL. 10–13 juillet 2013. Glasgow, Royaume-Uni. URL: http://edutice.archives-ouvertes.fr/edutice-00778274.

**Данные об авторах**

Лебедева Мария Юрьевна – кандидат филологических наук, доцент кафедры методики преподавания русского языка как иностранного, ведущий научный сотрудник лаборатории когнитивных и лингвистических исследований, Государственный институт русского языка им. А. С. Пушкина (Москва, Россия).

Адрес: 117485, Россия, Москва, ул. Академика Волгина, 6.

E-mail: m.u.lebedeva@gmail.com.

Лапошина Антонина Николаевна – ведущий эксперт лаборатории когнитивных и лингвистических исследований, Государственный институт русского языка им. А. С. Пушкина (Москва, Россия).

Адрес: 117485, Россия, Москва, ул. Академика Волгина, 6.

E-mail: antonina.laposhina@gmail.com.

**Authors' information**

Lebedeva Maria Yurievna – Candidate of Philology, Associate Professor of Department of RFL Teaching Methodology, Senior Researcher of Language and Cognition Laboratory, Pushkin State Russian Language Institute (Moscow, Russia).

Laposhina Antonina Nikolaevna – Leading Expert of Language and Cognition Laboratory, Pushkin State Russian Language Institute (Moscow, Russia).

Алкснит Наталья Антоновна – аспирант кафедры методики преподавания русского языка как иностранного, Государственный институт русского языка им. А. С. Пушкина (Москва, Россия).

Адрес: 117485, Россия, Москва, ул. Академика Волгина, 6.

E-mail: n.a.alksnit@gmail.com.

Ляшенко Татьяна Васильевна – аспирант кафедры методики преподавания русского языка как иностранного, Государственный институт русского языка им. А. С. Пушкина (Москва, Россия).

Адрес: 117485, Россия, Москва, ул. Академика Волгина, 6.

E-mail: mrtanya97@gmail.com.

Alksnit Natalia Antonovna – Postgraduate Student of Department of RFL Teaching Methodology, Pushkin State Russian Language Institute (Moscow, Russia).

Lyashenko Tatyana Vasilievna – Postgraduate Student of Department of RFL Teaching Methodology, Pushkin State Russian Language Institute (Moscow, Russia).

Дата поступления: 11.05.2022; дата публикации: 29.06.2022

Date of receipt: 11.05.2022; date of publication: 29.06.2022