*Вестник Челябинского государственного университета. 2023. № 2 (472). С. 12–19.*
*ISSN 1994-2796 (print). ISSN 2782-4829 (online)*
*Bulletin of Chelyabinsk State University. 2023;(2(472):12-19. ISSN 1994-2796 (print). ISSN 2782-4829 (online)*

Original article

# CORPUS-DRIVEN LEXICOGRAPHY: HISTORICAL REVIEW AND RECENT APPLICATIONS TO HIGHER EDUCATION TERMINOLOGY

**Vera G. Budykina**[1], **Oksana Polyakova**[2]

[1] Chelyabinsk State University, Chelyabinsk, Russia, vbudykina@gmail.com, ORCID 0000-0003-1488-0769

[2] Universitat Politècnica de València, Valencia, Spain, okpolnes@upv.es, ORCID 0000-0003-0575-2386

***Abstract.*** The article highlights the advancements of corpus linguistics that have contributed to the development of data-driven learning, computer-assisted L2 learning, translation studies, ways of knowledge assessment, among others. It mostly focuses on the developments in the area of corpus-driven lexicography. The main aim is to reflect on the current use of corpora and lexicographic tools built from corpora as well as on future perspectives in the improvements of corpora and corpus tools. Recent applications of those tools to the refinement of bilingual educational terminology in Spanish and Russian are also discussed in the paper. The authors describe the methodology of compilation of the corpus-based glossary of Spanish-Russian higher education. The architecture for analyzing professional terminology consisting of four components is proposed in the paper. The quantitative method is used to prove the effectiveness of the corpus-driven approach.

***Keywords:*** corpus linguistics, corpus-driven lexicography, corpora, corpus tools, English-Russian dictionary of higher education, corpus-based Spanish-Russian glossary on higher education, higher education terminology

***For citation:*** Budykina VG, Polyakova O. Corpus-driven lexicography: historical review and recent applications to higher education terminology. *Bulletin of Chelyabinsk State University*. 2023;(2(472):12-19.

Научная статья

УДК 811.11

# КОРПУСНАЯ ЛЕКСИКОГРАФИЯ: ИСТОРИЧЕСКИЙ ОБЗОР И СОВРЕМЕННЫЕ ПОДХОДЫ К ТЕРМИНОЛОГИИ ВЫСШЕГО ОБРАЗОВАНИЯ

**Вера Геннадьевна Будыкина**[1], **Oksana Polyakova**[2]

[1] Челябинский государственный университет, Челябинск, Россия, vbudykina@gmail.com, ORCID 0000-0003-1488-0769

[2] Политехнический университет Валенсии, Валенсия, Испания, okpolnes@upv.es, ORCID 0000-0003-0575-2386

***Аннотация.*** В статье освещаются достижения корпусной лингвистики, которые способствовали развитию обучения на основе данных, компьютерного обучения иностранным языкам, переводоведения, способов оценки знаний и т. д. Основное внимание в статье уделяется исследованиям в области лексикографии, в ходе которых применяются корпусные технологии (corpus-driven lexicography). Основная цель состоит в том, чтобы проанализировать современные способы применения корпусов и лексикографических инструментов, построенных на их основе, а также рассмотреть перспективы их развития. В статье также обсуждаются современные способы применения этих инструментов для лексикографического описания двуязычной терминологии на испанском и русском языках. Авторы описывают методологию составления корпусного испанско-русского глоссария высшего образования. В статье описывается структура анализа профессиональной терминологии, состоящая из четырех компонентов. Количественный метод используется для доказательства эффективности применения корпусного подхода.

***Ключевые слова:*** корпусная лингвистика, корпусная лексикография, корпус, корпусные средства, англо-русский словарь высшего образования, корпусный испанско-русский глоссарий высшего образования, терминология высшего образования

***Для цитирования:*** Budykina V. G., Polyakova O. Corpus-driven lexicography: historical review and recent applications to gigher education terminology // Вестник Челябинского государственного университета. 2023. № 2 (472). Филологические науки. Вып. 131. С. 12–19.

## Introduction

John Sinclair believed that "the advent of the corpus was the most thrilling development in language study during his career" [12. P. 157] and, probably, ever.

The history of corpus linguistics dates back to the 1960s, when the Brown Corpus which is considered to be the first electronic corpus of the English language was published [3].

Patrick Hanks traced the development of corpus-based linguistics and lexicography. In one of his articles, he argues that there is an essential difference between corpus, generative and cognitive linguistics. He also states that corpus linguistics studies language as communicative behavior while the other two main streams of linguistics do not treat human language this way; they see it as an instrument for thoughts construction [5. P. 398].

Corpus linguistics has contributed to the development of computer-assisted language learning, translation studies and other spheres. Besides, lexicography is one of the areas driven by corpus linguistics. The term "corpus-driven" was developed by E. Tognini-Bonelli in 2001. Still, the first mentions of the term were made in her research paper published in IGL (the International Journal of Lexicography in 1996. She wrote: "In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence" [14. P. 84]. On the contrary, "corpus-based research seeks to support preconceived theories with judiciously selected examples from a corpus" [5. P. 417].

## Corpus linguistics timeline

The IJL has been published since 1988 but in the first issue there were no mentions of the term 'corpus'. The second issue of 1989 published three reviews discussing the pre-corpus approach used in the compilation process of the 1st and the 2nd editions of LDOCE (Longman Dictionary of Contemporary English)[1] and the first edition of Collins COBUILD English Language Dictionary[2] when a massive corpus was used for the selection of the dictionary's headwords.

In the first review, R. Carter, an applied linguist, valued the strength of the traditional approach which LDOCE used. Nevertheless, he made some remarks on the COBUILD and called it a remarkable and unique project. R. Carter pointed out that the very existence of corpora and its use as a base for the dictionary's entries are unique. He also stated that for the English language teachers, the dictionary publication is a great event, and the dictionary was highly evaluated as a "contribution to the theory and practice of communicative language teaching" [1. P. 41–42].

The second review was made by Ch. Fillmore, an American linguist who is famous for his three contributions to linguistic theory: construction grammar, case grammar, and frame semantics. The author concludes that he cannot make a choice between LDOCE and COBUILD and recommends buying both dictionaries. In his review, Ch. Fillmore contrasts explanatory techniques and differences of definitions in the two dictionaries under consideration. He also highlights the advantage of the authenticity of COUBUILD, which was compiled on the basis of corpus technologies [2. P. 81].

In the third review, F. Hausmann and A. Gorbahn mention that the corpus-based approach to selecting entries makes the COBUILD dictionary almost ideal learner's dictionary compared to the LDOCE. The entries of the latter contain too much extra information. However, the authors did not like the authentic examples in COBUILD and thus, criticize the use of corpora to select examples. They consider that in the dictionaries where learners are a target audience, illustrations should be made with pedagogical goals in mind. They formulated the arguments against in the following way:

1. Some illustrations taken out of context may look or sound strange and appeal to the learner's imagination.

2. Some authentic illustrations distract the reader's attention from the definition of a word because its illustrations may be very complicated.

3. Some illustrations reflect a very idiosyncratic use of English.

4. Some illustrations are highly abstract and take much time to decode the meaning and, therefore, divert the reader's attention from the real meaning of the word.

[1] *Summers D. et al.* (eds.). Longman Dictionary of Contemporary English. (Second edition; Third edition 1995.) London: Longman. (LDOCE). 1987.

[2] *Sinclair J. M., Hanks P. et al*. (eds.). Collins COBUILD English Language Dictionary. (First edition.) London: Collins. (COBUILD). 1987.

5. Some illustrations may contain words of the same stem as the headword.

So, F. Hausmann and A. Gorbahn conclude that the illustration taken from the corpora should be adapted to the purpose of the learner's dictionary or basic needs of the target audience [6. P. 46]. Presumably, intuition can help in making examples.

G. Sampson, in his turn, disagrees with the practice of making illustrations based on intuition. He calls such examples abundant and unreliable [10. P. 2]. Since Johnson (1755), lexicographers believe that to avoid or at least reduce the number of inadvertent errors, every lexicographic job should be initiated on a solid basis of textual evidence. A simple collection of literature, citations, and dictionaries are now substituted by a corpus containing thousands of examples from the real speech, mass media, literature, and etc. Sometimes, the intuition evidence coincides with the corpus evidence, and it was proved by P. Hanks that the corpus size matters when analyzing examples and their selection. One could explain many properties of lemmas only by employing significant contexts.

Another conclusion of Patrick Hanks is that corpus-driven lexicography mostly deals with everyday language and phraseology. It can illustrate how words and phrases are used in real communication, but it does not provide any information about the boundaries of its possible usage. Studying the word usage as it is reflected in the corpus, corpus-driven lexicographers decide about the word meaning. So, statements about the word meanings are made by interpretations of data obtained by close examinations and analyses made by lexicographers [5. P. 404]. Therefore, there is a great deal of evidence to support the view that the use of corpus in headwords and examples selection is an intrinsic part of the dictionary compilation process, but the blind selection is not appropriate in our opinion as the intuition and knowledge of the lexicographer is a must. "The value of the corpus should not be underestimated or overestimated. It cannot replace the lexicographer; nor should it be regarded as inferior to the knowledge of the lexicographer in any respect".

Example selection is not the only way how corpus can be used in lexicography. This is how U. Stuttgart explains the ways to use corpora and other technological advancements. He states that lexicographers can rely on corpus data concerning the selection of raw material from which he/she gets evidence based on both quantitative and qualitative criteria; the selection of linguistic data put in the dictionary, and corpus-linguistic tools used in lexicographic work) [12. P. 131–132].

Lemma selection is a frequently used argument among lexicographers to motivate the use of corpus data. A lemma candidate list with the indicators of frequency could be extracted from the corpus. Frequency counts may contain figures for text types, registers, domains, regions or time periods. Word frequency is the starting point for determining the dictionary nomenclature. It is also used to determine inclusion/exclusion headwords while updating the lemma inventory. The words which have become frequent in the corpora will be included, and the words which are not frequently used any more will be removed from the lemma list [7; 4]. However, these lists are not sensitive to polysemy, and polysemy-related problems may arise.

Another widespread corpus tool in lexicography is called concordancer. This is a generator of KWIC indices. The concept of Keyword in Context indexing had been first proposed and implemented manually by librarian A. Crestadoro in 1856–1864. KWIC is a system which sorts and aligns the words and forms an index of most high frequency words contained in the texts to be searchable alphabetically in the index. There are custom-made tools [15], tools offered by large national corpora (BNC and the Sara/Xaira tool)[1], query tools of the Bank of English, the COSMAS tool[2], Frantext[3]. There are also open source tools, for example, The IMS Open Corpus Workbench[4] and some commercial products.

Specialized corpus tools have been created to meet the needs of computer-driven lexicography, i. e. to find evidence for typical syntagmatic properties of words and word combinations (collocations, sets of contexts illustrating the frequent uses, etc.) and to condense similar corpus instances into a sort of "type" as for most frequently used words, thousands of contexts may be extracted from the corpus tools [13. P. 144]. Here are some corpus query tools to mention: the WASPS[5] or the Sketch Engine[6] among others.

## Higher education lexicography and terminology projects

Education plays a significant role in the modern society as it creates new knowledge attuned to the fast-changing conditions and needs. Globalization is

[1] http://www.natcorp.ox.ac.uk/tools/

[2] http:// www.ids-mannheim.de/cosmas2/

[3] https://www.frantext.fr/

[4] http://cwb.sourceforge.net

[5] https://archaea.i2bc.paris-saclay.fr/wasps/

[6] https://www.sketchengine.eu/

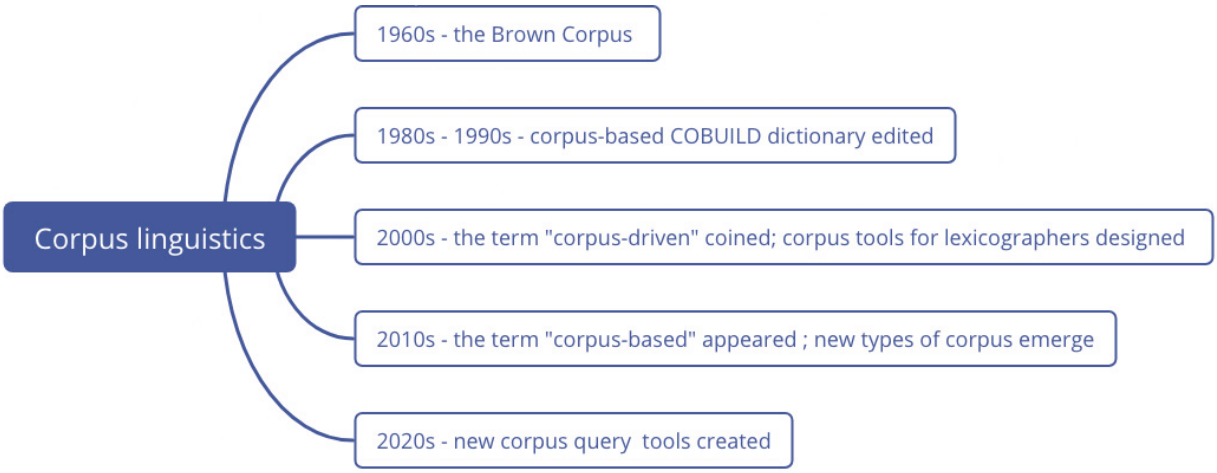


*Fig. 1. Corpus linguistics timeline (own elaboration)*

spreading in different spheres such as politics, culture, and research. Thus, a great demand is placed upon highly qualified employees to be competitive within the international market. A number of projects have been launched worldwide to promote cooperation in the education sphere. The number of students studying abroad is astounding, and the academic mobility of staff and students is on the rise before the start of the pandemic. As educational systems vary significantly from country to country, participants of projects, students, and education counterparts face the problem of misunderstanding terms used in education. No corpus on global education has been ever built. If any, the number of source languages is limited, no extra language information is given that does not make education systems of different countries transparent, and does not recognize qualifications.

Some efforts have been made in different countries to harmonize the education terminology. In Russia, the American-Focused Dictionary of Higher Education was published in Moscow in 2017[1]. The target audience of the dictionary are learners, translators, and teachers who are interested in the terminology of elementary, secondary and higher education; special education, psychology, philosophy, knowledge assessment, and such practical issues as accommodation, tuition fees, and etc. Various disciplines are integrated in a single reference guide which can be relied on by educators, translators, and students in their careers and endeavors. It is not a corpus-based dictionary, but it should be included in the corpus as well as other lexicographic works.

Spanish higher education institutions make a lot of efforts to harmonize Spanish, English, and Catalan terminology to attune to the needs of new requirements aimed at internalization of programmes, approaches, and policies. UGRTERM, TERMCAT and UPVTERM are the three terminology projects conducted by the University of Granada, Government of Catalonia and Polytechnic University of Valencia respectively. The research on different language pairs is conducted in collaboration with language normalization institutions.

[1] Budykina V. American-focused English-Russian Dictionary of Higher Education. Moscow. 2017. 400 p.

## Corpus-based methodology

One of the initiatives to standardize higher education terminology through corpus technologies has resulted in a corpus-based glossary of Spanish-Russian higher education that has been recently compiled [8; 9]. In this case, the researchers advocated for a corpus-driven approach based on documentation collection, corpus construction and terminological database design. Despite the limitations of a bilingual study, its lexical scope deployed samples of scientific, institutional, administrative and academic documentation in both languages. The corpus of 2,127,457 words represents a wide variety of academic areas via a 700 bilingual units glossary.

In this research, the scholars followed *ad-hoc* corpus design guidelines based on the general and specific criteria [8; 9; 10]. Therefore, the architecture for analyzing professional terminology consisted of four components:

– Academic and administrative general corpus (AAG): the monolingual corpus of 19 files in Spanish (90% of words) and 17 files in Russian (10% of words).

– Specialized university texts corpus (SUT): the monolingual corpus comprised of 10 files in Spanish (78% of words) and 8 files in Russian (22% of words).

– Legal texts corpus (LT): the monolingual corpus included 25 files in Spanish (69% of words) and

23 files in Russian (31% of words). Legal texts corpus (LT): the only bilingual parallel corpus covered 7 files in Spanish (67% of words) and 7 files in Russian (33% of words).

After processing the texts mentioned above, a considerable wordlist of glossary candidates was obtained and classified within a specialized database of relevant specialized terms. Spanish and Russian equivalents, context examples, collocations, information sources, translation types, synonyms or abbreviations were categorized into ten conceptual domains in the following manner [10] (table 1).

The application of a translation perspective to the research of the EHEA environment allowed for the collection of quantitative data on translation approaches in each of the indicated translation groups. Furthermore, the above-mentioned information analyzed directs us toward presenting the results. Fig. 2 depicts correspondences that might be useful in the contrastive analysis process.

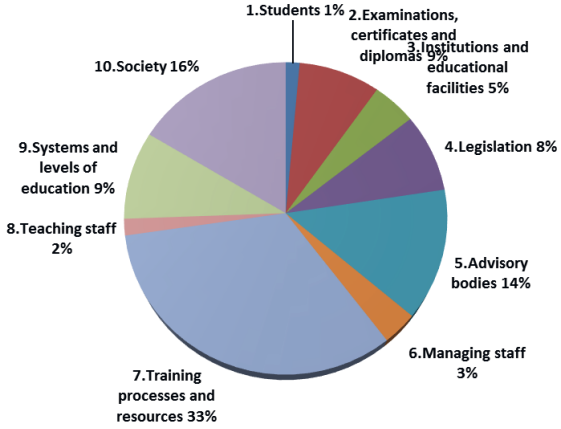


*Fig. 2. Visual representation of the thematic density of 10 corpus domains (own elaboration).*

*Table 1*

**Examples of bilingual terminology arranged by domains (own elaboration)**

| Domain | Bilingual terms | Examples |
|---|---|---|
| 1. Students | 10 | Student body [*alumnado — студенческий контингент*], scholarchip holder [*becario — стипендиат*], student [*estudiante — студент*] |
| 2. Examinations, certificates and diplomas | 59 | Accreditation [*acreditación — аккредитация*], transcript [*certificado académico академическая справка*], graduate [*licenciado/a — специалист*] |
| 3. Institutions and educational facilities | 32 | Building [*edificio — корпус*], Higher Education Institution [*institución de educación superior высшее учебное заведение*], vice-chancellor's office [*vicerrectorado — управление*] |
| 4. Legislation | 57 | Authorisation [*autorización — разрешение*], communiqué [*comunicado — коммюнике*], declaration [*declaración — заявление*] |
| 5. Advisory bodies | 95 | Assembly [*asamblea — собрание*], academic council [*claustro universitario — Ученый совет*], forum [*foro — форум*] |
| 6. Managing staff | 24 | Director [*director/a — директор*], deputy chairman [*vicepresidente — заместитель председателя*], member of committee [*vocal — член комиссии*] |
| 7. Training processes and resources | 232 | Payment [*abono — оплата*], learning [*aprendizaje — обучение*], course [*curso — курс*] |
| 8. Teaching staff | 12 | Professor [*catedrático — профессор*], faculty member [*docente — преподаватель],* research staff [*personal dedicado a la investigación — исследовательский персонал*] |
| 9. Systems and levels of education | 64 | Admission [*admisión — прием*], doctorate [doctorado -аспирантура], higher education [*enseñanza superior — система высшего образования*] |
| 10. Society | 115 | Academic community [*comunidad académica — академическое сообщество*], CV [*Curriculum Vitae — резюме*], strategy [*estrategia — стратегия*] |

Throughout the research, the scholars tested some working hypotheses. They worked on the documentary typology detection and organization as well as proposed its joint classification, grounded on the singularities of national education systems. Furthermore, the bilingual terminology database facilitated the systematization of different types of equivalence within the specialized translation samples. In the analytical section, the research helped observe a quantitative variation of equivalence according to the thematic domain and translation strategy employed (see Table 2 below).

The project brought together both countries' university management documents and approached the sustainable construction of a bilingual database transformable into a corpus-driven lexicography product. A wide range of its applications varies from the professional environment of translation, mediation, standardization and information management to the teaching field and use by the interested public.

Motivated by the recent technological advancements and the development of corpus-based lexicography, it is suggested that a new lexicographic project will be initiated. It will involve linguistics, cognitive linguistics, lexicography, education, IT, and neuroscience. The project will result in the creation of a unique corpus on Global Education, including international education terminology, samples of curricula, syllabi, and diplomas, descriptions of the best practices and teaching techniques, school and university management in pandemic and post-pandemic times. It will allow educational systems and their terminologies to be transparent that will contribute to the development of international collaboration, cooperation of education communities worldwide, as well as provide more opportunities for student and academic mobility, etc.

The project will contribute to developing internationally harmonized terminology of education and a database containing basic information on education systems worldwide, description of qualifications, samples of diplomas, nostrification procedure, main Education acts, etc. Moreover, the corpus will contain information on new education concepts and programs resulting from the global pandemic and describe the best practices that work in this new environment. This is in great demand within an international education and research community. Linguists and lexicographers from different countries, domain and technical specialists should join their forces to conduct this interdisciplinary project that will make a difference in many fields related to education, contribute to mutually beneficial international cooperation, and enrich the science and research.

## Conclusions

As corpus-driven lexicography is being improving, a lot of new corpus tools have been developed that contribute to the advancement in generative, cognitive and corpus linguistics, let alone corpus lexicog-

*Table 2*

**Translation equivalence by domain and translation group [10]**

| Domain | Calque | Trans-literation | Descriptive translation | Approximate translation | Total amount of bilingual records |
|---|---|---|---|---|---|
| 1. Students | 8 | 1 | 1 | 0 | 10 |
| 2. Examinations, certificates, and diplomas | 29 | 11 | 3 | 16 | 59 |
| 3. Institutions and educational facilities | 16 | 12 | 0 | 4 | 32 |
| 4. Legislation | 47 | 7 | 0 | 3 | 57 |
| 5. Advisory bodies | 62 | 11 | 5 | 17 | 95 |
| 6. Managing staff | 10 | 8 | 0 | 6 | 24 |
| 7. Training processes and resources | 155 | 51 | 8 | 18 | 232 |
| 8. Teaching staff | 4 | 0 | 1 | 7 | 12 |
| 9. Systems and levels of education | 42 | 5 | 8 | 9 | 64 |
| 10. Society | 78 | 24 | 4 | 9 | 115 |
| **TOTAL AMOUNT** | **451** | **130** | **30** | **89** | **700** |

raphy. The data from National corpora are used in different areas of research. The scientific community should pay particular attention to creating specialized corpora that are in great demand at the moment. As education terminology is in the area of the authors' research interest, this sphere was more closely examined and analyzed in terms of the existence of such corpora and the efforts made to harmonize the cross-language terminologies. A new lexicographic project on the Global Education corpus is announced to contribute to the theory and practice of corpus linguistics, development of education, global cooperation, and mutual understanding.

## References

1. Carter R. Review Articles: LDOCE and COBUILD. I*nternational Journal of Lexicography.* 1989;2(1):30-43.

2. Fillmore CJ. Review Article: Two Dictionaries. *International Journal of Lexicography.* 1989;2(1):57-83.

3. Francis WN., Kučera H. Manual to accompany a standard sample of present-day edited American English, for use with digital computers. Original edition 1964, revised 1971, revised and augmented 1979. Department of Linguistics, Brown University, Providence, RI. 1979. Available from: http://icame.uib.no/brown/bcm.html.

4. Geyken A. Korpora als Korrektiv für einsprachige Wörterbücher: Philologie auf neuen Wegen. *LiLi. Zeitschrift für Literaturwissenschaft und Linguistik*. 2004;34(136):72-100.

5. Hanks P. The Corpus Revolution in Lexicography. *International Journal of Lexicography.* 2012;25(4):398-436.

6. Hausmann FJ., Gorbahn A. Review Article: COBUILD and LDOCE II — A Comparative Review. *International Journal of Lexicography.* 1989;2(1):44-56.

7. Heid U., Säuberlich B., Debus-Gregor E., Scholze-Stubenrecht W. Tools for Upgrading Printed Dictionaries by Means of Corpus-based Lexical Acquisition. Proceedings of the Fourth Language Resources and Evaluation Conference. Lisboa, ELRA; 2004. Pp. 419–423.

8. Polyakova O. An integrated approach to the higher education terminology in Spanish-Russian university texts. *Onomázein: Revista de lingüística, filología y traducción de la Pontificia Universidad Católica de Chile*. 2021;51:57-82.

9. Polyakova O., Candel-Mora M.Á. Building a corpus-based glossary of Spanish-Russian higher education for specialised translation. *Sendebar*;2019:(30):141-162.

10. Polyakova Nesterenko O. Estudio de las peculiaridades de la terminología en el entorno académico del EEES en ruso y en español [Tesis doctoral no publicada]. *Universitat Politècnica de València*. 2013. Available from: https://riunet.upv.es/handle/10251/34509

11. Sampson G. Empirical Linguistics. London, Continuum; 2001. 226 p.

12. Sinclair JM. Preface. *International Journal of Corpus Linguistics*. 2007;2(2):155-157.

13. Stuttgart UH. Corpus Linguistics and lexicography.Corpus linguistics: an international handbook, edited by Anke Lüdeling and Merja Kytö. Berlin/New York, W. de Gruyter; 2008. Pp. 131–153.

14. Tognini-Bonelli E. Corpus linguistics at work. Amsterdam/Philadelphia: John Benjamins Publishing, 2001. 223 p.

15. Walter E, Harley A. The role of corpus and collocation tools in practical lexicography. Proceedings of the Tenth EURALEX International Congress, EURALEX; 2002. Pp. 851–857.

## Information about the authors

**Vera G. Budykina** — Candidate of Philological Sciences, Associate Professor, Dean, Faculty of Eurasian and Oriental Studies.

**Oksana Polyakova** — Candidate of Philological Sciences (PhD), Lecturer, Applied Linguistics Department.

## Информация об авторах

**В. Г. Будыкина** — кандидат филологических наук, доцент, декан факультета Евразии и Востока ЧелГУ.

**О. Полякова** — кандидат филологических наук (доктор филологии), преподаватель кафедры прикладной лингвистики.

*Статья поступила в редакцию 29.01.2022; одобрена после рецензирования 10.04.2022; принята к публикации 26.12.2022.*

*The article was submitted 29.01.2022; approved after reviewing 10.04.2022; accepted for publication 26.12.2022.*

Вклад авторов: оба автора сделали эквивалентный вклад в подготовку публикации.
Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article.
The authors declare no conflicts of interests.