DIGITAL TRANSLATION METHODS IN THE CONTEXT OF FOREIGN LANGUAGE EDUCATION: CORPORA APPROACH

Mikhaylova A.G.[1],
e-mail: steba1971@mail.ru,

Kobzeva N.A.[2],
candidate of pedagogical sciences, associate professor,
e-mail: nadiatom@mail.ru,
[1]*Sevastopol State University, Sevastopol, Russia*
[2]*National Research Tomsk Polytechnic University, Tomsk, Russia*

The article describes the use of corpora as promising translation methods in the context of foreign language education. Corpus-driven approaches is known to have become popular in lexicography following digital technological advances. This method have made it possible for computer programs to process vast amounts of data very rapidly. The purpose of the study: to identify the main role of corpora in teaching a foreign language. The following research methods are: content analysis of scientific literature on the research topic, interpretation, systematization, structuring. The following results are obtained: currently the problem of corpus approach use as digital means to solve foreign language education issues are relevant due to Foreign language study enhances human's opportunities in technologies. It is necessary to create a link between the digital society and humanitarian education, that is, digital culture. The use of corpus technologies can provide significant assistance in the professional activities of a foreign language teacher. The corpus approach is an important tool in organizing the learning process, finding the necessary content of educational tools, and its application can ensure the quality of teaching a foreign language. In conclusion it is noted that expanding the corpus capacity to translate texts contributes to the diversification of learning processes through digital technologies.

Keywords: digital technologies, corpora, lexicon, translation methods, foreign language, culture, genre, foreign language skills

ЦИФРОВЫЕ МЕТОДЫ ПЕРЕВОДА В КОНТЕКСТЕ ИНОЯЗЫЧНОГО ОБРАЗОВАНИЯ: КОРПУСНЫЙ ПОДХОД

Михайлова Алла Григорьевна[1],
e-mail: steba1971@mail.ru,

Кобзева Надежда Александровна[2],
канд. пед. наук, доцент,
e-mail: nadiatom@mail.ru,
[1]*Севастопольский государственный университет, г. Севастополь, Россия*
[2]*Национальный исследовательский Томский политехнический университет, г. Томск, Россия*

В данной статье исследуется использование корпусов как перспективных методов перевода в контексте обучения иностранным языкам. Известно, что подходы, основанные на корпусе, стали популярными в лексикографии после достижений цифровых технологий. Этот метод позволил компьютерным программам очень быстро обрабатывать огромные объемы данных. Цель исследования: выявить основную роль корпусов в обучении иностранному языку. Методами исследования являются: контент-анализ научной литературы по теме исследования, интерпретация, систематизация, структурирование. Результаты: в настоящее время проблема применения корпусного подхода как цифрового средства решения вопросов обучения иностранному языку актуальна, поскольку иноязычные умения расширяют возможности человека в сфере различных технологий. Необходимо создать связь между цифровым обществом и гуманитарным образованием, то есть цифровой культурой. Применение корпусных технологий может оказать существенную помощь в профессиональной деятельности преподавателя

иностранных языков. Корпусный подход – важный инструмент в организации процесса обучения, нахождении необходимого содержательного наполнения учебных средств, а применение может повысить качество преподавания иностранного языка. В заключении отмечено, что расширение возможностей корпуса по переводу текстов способствует диверсификации процессов обучения посредством цифровых технологий.

Ключевые слова: цифровые технологии, корпус, лексика, способы перевода, иностранный язык, культура, жанр, иноязычные умения

Introduction

The attention has been paid to corpus-based study in the sphere of applied linguistics in translation studies. Corpus-driven approaches is known to have become prevalent "in lexicography following recent technological advances, which have made it possible for computer programs to process vast amounts of data very rapidly"[1]. Currently the issues of corpus approach use as digital means to solve problems of foreign language education are relevant due to Foreign language study enhances one's opportunities in technology, military, business, government, medicine, law, industry, marketing, etc. [1, p. 462]. There are possibilities of using corpus technologies in the pedagogical process.

Foreign languages translation is founded on an interdisciplinary basis [2]. It is aimed at "communicative, cognitive, sociocultural, informational, professional and general cultural competencies forming" [3, p. 127]. The ability to translate professional texts is valuable for specialists [4]. The corpus approach is empirical and analyzes real word usage in the natural language environment; uses a fairly extensive, typical selection of texts. The task of educators is to teach students to translate texts taking into account the factor of cultural differences [5].

Today, "the concept of correspondence and/or equivalence has been a complex issue among translation experts and theorists due to the cultural dimension involved in the languages involved (source language and target language)" [6, p. 50]. A corpus gives lexicographers access to a wider context of use for a term compared to paper slips. A new corpora-based approach is used to cluster short texts, since they give rise to the problem of insufficient word matching and lack of contextual information.

Being a computer-based means, corpus linguistics is based on providing a ready resource of authentic texts. One can search for examples from authentic sources not only with whole words and phrases for study and vocabulary elaboration, but also examples of certain morphemes, articles, suffixes and endings for composing grammatical tasks by means of use of concordance.

Analysis of recent achievements and publications shows that advances in the area of both translation technology have highly facilitated the interpreters' role and "enabled them to produce target texts that are clearer and more nuanced" [7, p. 2]. A recent phenomenon is corpora using in language teaching that mostly focused on English language. "When people communicate linguistically, they "talk to convey informational content – and we do this by means of words" [8, p. 417]. Concordance provides different possibilities for developing tasks for all levels of learners. That is, the corpus can be composed of texts that correspond to the level of students' knowledge, which means that tasks developed on their basis can be effective. The corpus is a complete resource, and any tasks based on it have answers in the corpus itself.

Statements can be classified as various genres. The growth of corpus evidence have resulted in numerous kinds of dictionaries – corpus-based reference publications. Genres are infrequently investigated and highly influential such as corpus tools. They proved to be useful.

Methodological and theoretical basis

Corpus linguistics have been investigated by scientists S. Voelkel, F. Kretzschmar. A corpus as a cross-cultural textual analysis was studied by I. Mel'cuk and A. Polguère. They consider theory and practical appli-

---

[1] The role of corpora in dictionary making. Future Learn. – URL: https://clck.ru/38Dvjd (date of application: 02.03.2024). – Text: electronic.

cation of lexicographic definition and present a vision on how to simulate term meanings in dictionaries or in lexicographic models [8].

Some scientists believed that the possible corpora were not universal solutions that could meet all the needs of a translator and translators would "be required to create their own corpora if they embark on a project for which few materials are available" [9, p. 2]. Tawffeek A.S. Mohammed proposed the main ways of word translation by means of Corpora approach [7]. He highlights the significance of applying various corpora in translation context and expresses the possibility of creating corpora of one's own. A. Saricoban believes that content interpretation claims "awareness of native language (L1) and target language (L2) in both language structures and cultures exploring similarities and differences between them" [10, p. 4].

X. Li suggests that "the most in-demand competences are linguistic competence in working languages, psycho-physiological competence, interpersonal competence, extra-linguistic knowledge, and instrumental competence" [11, p. 4]. A. Negoescu writes about discourse aspects. According to his opinion the terms and expressions should be "very carefully chosen to create something unique, surprising and memorable in order to achieve the most impact upon the audience" [12, p. 3]. However, when considering language in use in real situations these boundaries between the levels of the language are blurred [12].

We perceive the world through our feelings, memories, the culture we belong to and the language we speak. Therefore, languages are considered to be a major part of our understanding of the world as they provide the interaction between representatives of different cultures. "Culture, religion and science depend on language" [13, p. 367] and language carries them all. Language is thus a prerequisite for the differentiation of worlds, and barriers between languages must be reduced by translation or interpretation.

A.C. Parlog proves that receptor response is usually challenged by the content, form and style of the ST and that it loses its importance if the translation has not previously been evaluated. The TT will be assessed in relation to the lexis used, phonetic transformations and translators' personal interventions in the literary text. Translationese is under the pressure of the "language of Europe", i.e. English; the translation of the EU documents in European languages has increasingly become a process of utter "translationese" [14].

In general, the problem can be defined as the following: what are the effective digital means of text interpretation? Its connection with main scientific and practical issues is in the fact that digital educational technologies are an innovative way of the educational process organization. It is based on the electronic systems application that provides visibility. The purpose of using digital technologies is to improve the quality and efficiency of the educational process.

Translators should identify factors which can affect the quality of interpreted text. The role of corpora approach is "to provide the most correct examples of the use of language units…" [13, p. 367]. A corpus as a set of texts processed is applied as a basis for the study of a language (confirmation of linguistic rules in a given language; testing of statistical hypotheses; statistical analysis).

A new corpora-based approach is used to cluster short texts, since they give rise to the problem of insufficient word matching and lack of contextual information. The features of the corpus are collection of texts available for translated material analysis.

We should identify unsolved parts of the general problem that the article is devoted to: the role of corpora in lexicographic practice to solve translation problem as well as corpus linguistic approaches it the quality of interpreted text.

The purpose

In this article we study the issues of the role of corpora in lexicographic practice to solve translation problem. Language can be considered as an essential part of understanding of the world. The barriers occurred have to be reduced by means of the translations "instrument". The following tasks are to be solved: to define the role of electronic corpora in appropriate selection of vocabulary; to find a type of language corpora to process text information.

Results

Electronic corpora layout allows translators to find all occurrences of a particular word or expression and find examples of verbal intricacies [15; 16]. This provides the word selection depending on the text register (informal, formal, slang, etc), the author, the subject matter, the geographical location and so on"[2].

The corpus as a source of term examples can be taken as the starting point of the process the 'corpus-based' approach. Corpus-driven approaches is known to have become very essential in lexicography. The classification of language corpora is given in the table 1.

Undoubtedly, corpus technologies application can provide significant assistance in the professional activities of foreign language teachers, become an important tool in organizing the learning process, finding the necessary content for the course and learning aids. Corpus technologies influence the improvement of foreign language teaching quality.

Table 1 – Classification of Language Corpora[3]

Attribute	Types of Corpus
Type of language data	Oral, written, multimedia, mixed
Language of presentation of texts	Monolingual, bilingual, multilingual, mixed
Language norm	Literary, dialectal, mixed
Discursive and stylistic affiliation of texts	Literary, scientific (terminological), colloquial, journalistic, folklore, mixed
Genre-generic affiliation of texts	Prose, poetic, dramatic, mixed
Purpose of use Multi-purpose, specialized	Access method Freely available, commercial, closed, mixed
Purpose Research, illustrative, mixed	Chronological feature Synchronic, diachronic
The way of existence of the case is dynamic (replenished), static	Indexing marked up (annotated), unmarked
The nature of the markup is morphological, syntactic, semantic, prosodic, mixed	The volume of texts is full-text, fragmentary, mixed

There following possibilities of using corpus technologies in the pedagogical process are:

– the corpus approach is focused on the applied learning of language, its functioning in real situations and texts, which is important for a foreign language teacher;

– one of the main advantages of materials developed on the basis of corpora is the ability to increase the autonomy of teachers and students;

– the use of corpus technologies can accelerate the educational process significantly within the framework of autonomous educational activities of students;

– students can do various corpus studies in the field of morphology, phraseology, syntax that provide formation of their independence;

– using corpus technologies is a factor in improving the quality of teaching foreign languages.

Processing allows us to search for any necessary information in the case. In corpora, as a rule, the markup of texts is carried out by parts of speech, this allows us to find words in specified grammatical forms and in specified structures and combinations (collocates).

Special concordance programs are used to process information. They search for information in an array of texts and form a concordance – a list of all contexts in which a word or phrase occurs in the considered text. Here, for example, is a part of the concordancer for the verb crack (color markers show parts of speech in the immediate environment of the searched word) [13].

The size of corpora is a few billion words. One example of typical corpora is the Oxford English Corpus. Containing 21bn words, it is "used by lexicographers working on the OED and other Oxford dictionaries"[4].

The presented example is from the Corpus of Contemporary American English (COCA) (figure 1, 2). Particular corpora contain different types of texts (written only, oral only, both types) and different sets of genres. COCA presents a wide range of genres.

---

[2] The role of corpora in dictionary making. Future Learn. – URL: https://clck.ru/38Dvjd (date of application: 02.03.2024). – Text: electronic.
[3] Compiled by the authors.
[4] The role of corpora in dictionary making. Future Learn. – URL: https://clck.ru/38Dvjd (date of application: 02.03.2024). – Text: electronic.
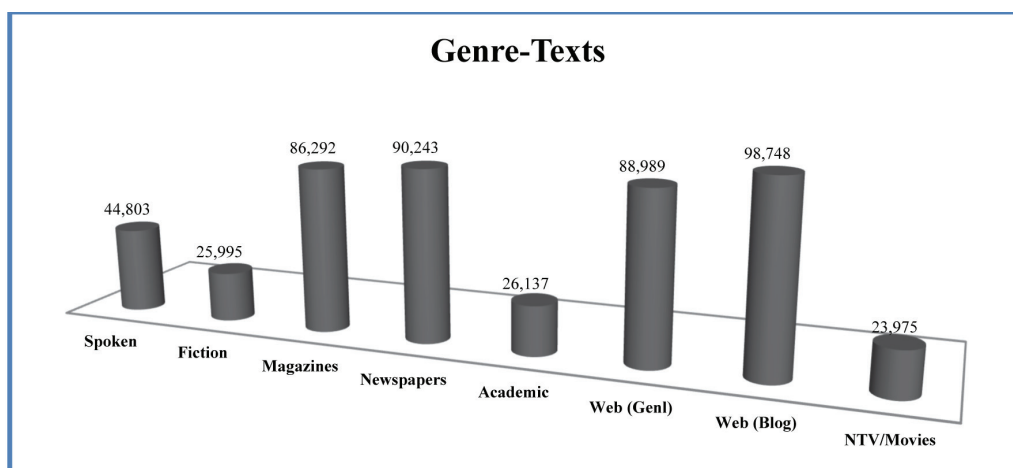
Figure 1 – Corpora contain different types of texts and sets of genres[5]

The table below present different types of genres and their explanation (figure 2).

Genre	# texts	# words	Explanation
Spoken	44,803	127,396,932	Transcripts of unscripted conversation from more than 150 different TV and radio programs (examples: All Things Considered (NPR), Newshour (PBS), Good Morning America (ABC), Oprah)
Fiction	25,992	119,505,305	Short stories and plays from literary magazines, children's magazines, popular magazines, first chapters of first edition books 1990-present, and fan fiction.
Magazines	86,292	127,352,030	Nearly 100 different magazines, with a good mix between specific domains like news, health, home and gardening, women, financial, religion, sports, etc.
Newspapers	90,243	122,958,016	Newspapers from across the US, including: USA Today, New York Times, Atlanta Journal Constitution, San Francisco Chronicle, etc. Good mix between different sections of the newspaper, such as local news, opinion, sports, financial, etc.
Academic	26,137	120,988,361	More than 200 different peer-reviewed journals. These cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business
Web (Genl)	88,989	129,899,427	Classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages (by Serge Sharoff). Taken from the US portion of the GloWbE corpus.
Web (Blog)	98,748	125,496,216	Texts that were classified by Google as being blogs. Further classified into the web genres of academic, argument, fiction, info, instruction, legal, news, personal, promotion, review web pages. Taken from the US portion of the GloWbE corpus.
TV/Movies	23,975	129,293,467	Subtitles from OpenSubtitles.org, and later the TV and Movies corpora. Studies have shown that the language from these shows and movies is even more colloquial / core than the data in actual "spoken corpora".
	485,179	1,002,889,754	

Figure 2 – Summary of Data on COCA[6]

The frequency of collocation "adj. + problem" in the SOS case corpus is shown in the figure 3.

A Specialized 8 Million Civil Engineering Research Paper Corpus in Tokyo is presented. "National linguistic corpora includes a variety of genres, such as Bank of English (200 million word usages), British National Corpus (100 million word usages)" [14, p. 23]. An English translation corpus is a linguistic one, which carries out transformation to text, filtering, and normalization. It can then be processed by usual tools of corpus linguistics and implemented in a search corpus system. Corpus-based lexicographic analysis clearly helps to reveal the contextual use of certain words, especially synonymous ones. The translator presents some options for certain word application in the original texts.

Scholars could use corpus data to analyze English as a Foreign Language (TEFL) curricula and facts. The textural organization of the corpus consists of a traditional content with reference to translations and a tabular form, which is more appropriate to comparison.

The following tasks of the corpus method can be highlighted:

---

[5] Compiled by the authors.
[6] Source: https://www.english-corpora.org/coca/.

1) the task of the student is to discover something new in a foreign language;

2) the task of the teacher is to provide a context in which these strategies can be implemented;

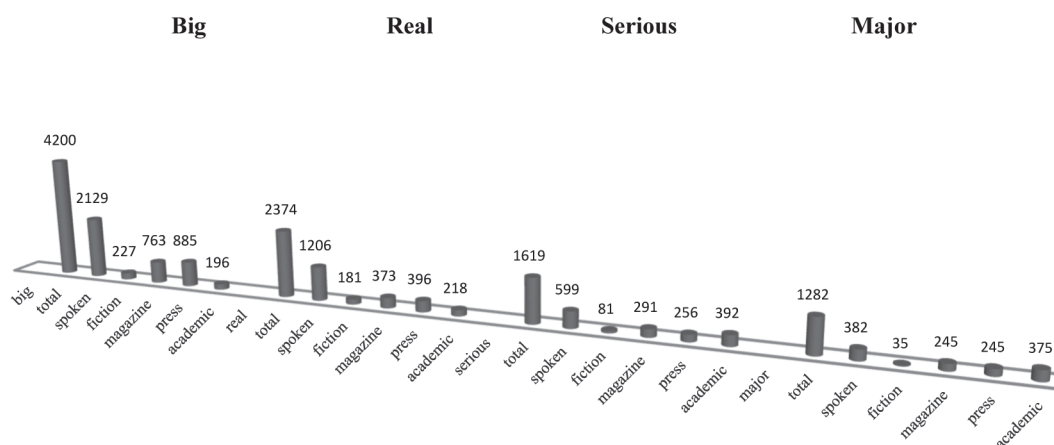3) the corpus approach provides relevant aspects realization of foreign language proficiency.

Figure 3 – The frequency of collocation "adj. + problem" in SOS case corpus[7]

Thus, the effective digital means of text interpretation is corpus-based lexicographic analysis. We based on the assumption that the purpose of using digital technologies is to improve the quality and efficiency of the educational process. A corpora-based approach is used to cluster short texts, since they give rise to the problem of insufficient word matching and lack of contextual information. It is applied as a basis for linguistic rules in a given language.

Translators can determine factors which affect the quality of interpreted text. The role of corpora approach is to provide the most correct examples of the use of language units. According to mentioned example the frequency of collocation "adj. + problem" may be over 4 thousand patterns that proved a great efficiency and quality of text interpretation and translation.

Conclusion

Language education as a process and result of cognitive activity offers a modern concept of teaching foreign languages. It is based on the forming of the ability and readiness to participate in oral, written, virtual, intercultural, organizational, corporate, scientific communication using foreign languages as a means. In the context of foreign language education high requirements for teaching practical skills of translating texts in the specialty and special terminology are put in. The analysis of different corpora types are the subjects of papers in machine translation, in which corpora are used to create hidden models.

Today, the issue of correct translation in the professional areas is the main one for higher education, since a modern specialist must be ready to work with professional foreign language literature. Hence, the relevance of special training of future specialists, regardless of the training profile, in the skills of translating special texts is obvious.

Corpora can be useful in teaching foreign languages. One of the major features of corpus-based learner dictionaries is their inclusion of quantitative data extracted from a corpus. The essential point is that dictionaries when defining the entry for a vocabulary select the terms and expressions used from a controlled set.

The following possible ways to use concordances in foreign language education: a) authentic examples demonstration of the use of certain lexical units, set phrases, grammatical structures, text structure; b) grammatical rules formulation and ones for the lexical units application; c) expanding the vocabulary (searching for new meanings of known words, correlating words and phrases with the context, compiling a dictionary entry inde-

---

[7] Compiled by the authors.

pendently or with the help of an educator); d) texts comparison created by native and non-native speakers. These forms of work are quite traditional for the methodology of teaching foreign languages. The use of corpus technologies can significantly accelerate the educational process in the context of autonomous students' activities

There are no diagnostic and experimental data to confirm effectiveness of corpora approach, but our task is to prove that it is feasible and productive. Our universities should practise this technology.

References

1. *Mikhaylova A.G.* Corpus Approach to Professional Texts Translation in the Context of Digital Culture Possibilities for Foreign Language Education // Digital Transformation of Social and Economic Systems: Proceedings of the International Scientific and Practical Conference, Moscow, January 26, 2024. – Moscow: Moscow Witte University, 2024. – P. 462–474.

2. *Chroma M.* The Role of Translation in Professional Communication // The Routledge Handbook of language and Professional Communication. – Publisher: Routledge, 2014. – Chapter 11. – Edition 1. – P.147–164.

3. *Kobzeva N.A., Mikhaylova A.G.* Translation Peculiarities in Professional Communication of Non-Linguistic Universities // Yazyk. Obshchestvo. Obrazovanie [Language. Society. Education]. Collection of Scientific Papers of the III International Scientific and Practical Conference. Edited by Yu.V. Kobenko. – Tomsk, 2022. – P. 127–131.

4. *Mikhaylova A.G.* Organizational and Pedagogical Conditions for Increasing the Motivation of Students with Disabilities in the Context of Inclusive Education // Voprosy Metodiki Prepodavaniya v Vuze [Teaching Methodology in Higher Education]. – 2021. – Vol. 10, No. 39. – P. 67–77.

5. *Chu Tao Zheng, Cheng Liu, Hau San Wong.* Corpus-Based Topic Diffusion for Short Text Clustering // Neurocomputing. – 2018. – Vol. 27531. – P. 2444–2458. – DOI 10.1016/j.neucom.2017.11.019.

6. *Anthony N., Zayyanu Z.M.* Achieving Correspondence and/or Equivalence in Translation, An Evaluation of the Translation Ekwensi's Burning Grass into French as La Brousseardente by Francoise Balogun // Journal of Research in Humanities and Social Science. – 2021. – Vol. 9, Issue 5. – P. 50–55.

7. *Tawffeek A.S.* Mohammed. The Use of Corpora in Translation into the Second Language: A Project-Based Approach // Front. Educ. Sec. Higher Education. – 2022. – Vol. 7. – URL: https://doi.org/10.3389/feduc.2022.849056 (date of application: 12.03.2024). – Text: electronic.

8. *Mel'cuk I., Polguère A.* Theory and Practice of Lexicographic Definition // Journal of Cognitive Science. – 2018. – No. 19 (4). – P. 417–470. – DOI 10.17791/jcs.2018.19.4.417.

9. *Johansson S.* On the Role of Corpora in Cross-Linguistic Research. Corpora and Crosslinguistic Research. – Amsterdam, 1999. – P. 3–24.

10. *Saricoban A.* Translation to Test Language Areas and Skills // Professional Communication and Translation Studies. International Conference. – Politehnica University of Timisoara, Romania, 2011. – 7 edition. – URL: https://docplayer.org/60395657-Professional-communication-and-translation-studies.html (date of application: 12.03.2024). – Text: electronic.

11. *Li X.* Identifying In-Demand Qualifications and Competences for Translation Curriculum Renewal: A Content Analysis of Translation Job Ads // The Interpreter and Translator Trainer. – 2022. – Vol. 16, Issue 2. – URL: https://www.tandfonline.com/doi/full/10.1080/1750399X.2021.2017706 (date of application: 12.03.2024). – Text: electronic.

12. *Negoescu A.* Aspects of Advertising Discourse // Professional Communication and Translation Studies. International Conference. – Politehnica University of Timisoara, Romania, 2011. – 7 edition. – URL: https://docplayer.org/60395657-Professional-communication-and-translation-studies.html (date of application: 12.03.2024). – Text: electronic.

13. *Mikhaylova A.G.* Korpusnyj Podhod k Perevodu Professional'nyh Tekstov [Corpus Approach to the Translation of Professional Texts] // Cifrovaya Gumanitaristika i Tekhnologii v Obrazovanii [Digital Humanities and Technologies in Education (DHTE 2023)]: Collection. Articles of the IV International Scientific and Practical Conference. November 16–17, 2023 / ed. V.V. Rubtsova, M.G. Sorokova, N.P. Radchikova. – Moscow: Publishing house FSBEI HE MGPPU, 2023. – P. 367–380.

14. *Parlog A.C.* Recreating Meaning in Translation // Professional Communication and Translation Studies. International Conference. – Politehnica University of Timisoara, Romania, 2011. – 7 edition. – URL: htt-

ps://docplayer.org/60395657-Professional-communication-and-translation-studies.html (date of application: 12.03.2024). – Text: electronic.

15. *Gilmore A., Millar N.* The language of Civil Engineering Research Articles: A Corpus Based Approach // English for Specific Purposes. – 2018. – Vol. 51. – P. 1–17. – DOI 10.1016/j.esp.2018.02.002.

16. *Staples S.* Using Corpus-Based Discourse Analysis for Curriculum Development: Creating and Evaluating a Pronunciation Course for Internationally Educated Nurses // English for Specific Purposes. – 2019. – Vol. 53. – P. 13–29. – DOI 10.1016/j.esp.2018.08.005.