New Perspectives in Science Education

International Conference NEW PERSPECTIVES in SCIENCE EDUCATION

Teaching Phraseology in Engineering ESP Writing: From learner corpus analysis to Data-Driven Learning

Andreea Dinca

West University of Timisoara, Romania

Abstract

Corpus-based research has demonstrated that academic writing is inherently phraseological, with multi-word sequences such as lexical bundles functioning as "building blocks of discourse" [2, p. 400]. At the same time, research shows that non-native university-level students enrolled in English for Specific Purposes (ESP) courses often have a solid command of English grammar and vocabulary, yet many continue to face challenges in the use of recurrent word combinations [9, p. 469]. Using a corpus linguistics methodology, this study analyzes a corpus of Romanian university student writing in English from the field of Engineering to identify recurrent phraseological patterns that characterize non-native Engineering writing. Findings indicate that Romanian Engineering students' writing relies on engagement-oriented and procedure-describing phraseology, foregrounding writer presence and reader guidance rather than the impersonal stance and hedging strategies typical of expert academic argumentation. Drawing on these findings, the paper proposes a set of data-driven learning (DDL) activities specifically designed to support the development of phraseological competence in Engineering ESP writing by engaging learners in the direct exploration of authentic language use captured in corpora [7]. The proposed activities are designed for use with a corpus of expert multidisciplinary academic writing, EXPRES [6], available through a user-friendly interface that supports classroom implementation. The paper argues that combining learner corpus analysis with structured DDL tasks offers a practical approach to teaching effective phrase use in specialized academic writing.

Keywords: *English for Specific Purposes (ESP), Engineering, academic writing, Data-Driven learning (DDL)*

1. Introduction

Research in corpus linguistics has consistently demonstrated that academic writing is fundamentally phraseological in nature (see e.g., [2]). Broadly defined, phraseology in corpus linguistics refers to the systematic study of recurrent multi-word sequences or lexical phrases that reflect conventionalized patterns of language use [11]. From this perspective, academic discourse is not constructed word by word, but through recurring combinations that fulfil rhetorical and disciplinary functions.

As a pedagogically motivated movement, English for Specific Purposes (ESP) corpus-supported research has investigated various types of phraseological units in student writing, including collocations (e.g., [14]) and lexical bundles (e.g., [8]). Studies have shown that non-native university-level students often display a solid command of English grammar and vocabulary, yet encounter difficulties in the use of recurrent word combinations [9, p. 469]. More specifically, findings indicate that non-native university students writing in English tend to rely on a limited repertoire of phraseological units, overusing highly frequent, general-purpose expressions while underusing or misusing more discipline-sensitive and expert-like phrases [19]. Such a restricted phraseological range may result in texts that are grammatically accurate but rhetorically limited.

In response to these challenges, corpus-informed ESP pedagogy has increasingly proposed the Data-Driven Learning (DDL) approach, which enables learners to explore authentic corpus data and inductively identify patterns of usage. A growing body of research suggests that DDL can promote deeper noticing of lexico-grammatical patterns and contribute to the refinement of academic writing skills, particularly when activities are guided and pedagogically structured (see e.g., [4]; [1]).

However, although previous studies have examined the use of phraseological units in learner writing (e.g., [12]) and others have explored the pedagogical benefits of DDL in ESP contexts (e.g., [1]), fewer investigations have systematically combined learner corpus diagnosis with guided corpus-based exploration in a discipline-specific setting. In other words, there remains a need for pedagogical models that move explicitly from empirical analysis of student writing to the design of targeted, phraseology-focused DDL activities grounded in authentic expert discourse.

The present study addresses this gap by using a learner corpus of Engineering student writing as a diagnostic tool to identify novice phraseological patterns and by transforming these findings into structured DDL tasks based on an expert academic corpus. The research questions guiding this study are:

RQ1. What phraseological patterns characterize non-native Engineering student writing in English compared to proficient models?

RQ2. How can corpus-based findings be transformed into data-driven learning activities for Engineering ESP writing instruction?

2. Literature Review

The study of multi-word units has its origins several decades ago, as illustrated by a 1983 study by Pawley and Syder, which states that that, in addition to stored, ready-made clauses and sentences, native speakers have a vast repertoire of “phraseological expressions, each of which is something less than a completely specified clause” (p. 205). Since then, the field of phraseology has seen a rapid expansion, aided by the advent of computers and by corpus linguistics methodology, which has shown that meaning is often conveyed through words co-occurring in preferred syntagmatic structures, rather than functioning as isolated units [18, p. 460]. As a result, the study of language corpora has helped identify new categories of multi-word units, such as n-grams (or lexical bundles), co-occurrences, collocational frameworks, colligations, etc. [10].

The field of English for Specific Purposes and its sub-branch, English for Academic Purposes (EAP), have extensively researched phraseology with two main aims. First, researchers have sought to identify those phraseological units that characterize proficient academic writing and are worth teaching to students. For this purpose, language corpora containing expert academic writing were studied. Second, another body of research has focused on analyzing learner corpora, which are language corpora containing non-native student academic writing. The aim was to identify possible gaps or limitations in students’ writing that could be addressed by ESP/EAP instruction (see e.g., [12]). The literature shows that most learners tend to overuse a limited repertoire of phraseological units, alongside the misuse and underuse of expert-like phraseological units [19].

Scholars have also examined the discourse functions performed by certain phraseological units, such as lexical bundles. A widely used functional classification is proposed by Hyland [13], who distinguishes three broad categories: research-oriented, text-oriented and participant-oriented bundles. Research-oriented phraseological units assist writers in structuring their activities and representing real-world experiences, text-oriented expressions relate to the organisation of the text and the construction of meaning as a message or argument, while participant-oriented bundles focus on the writer and the reader of the text (pp. 13–14).

Practitioners in ESP/EAP are concerned with writing instruction, and various writing support resources and teaching strategies have therefore been proposed. As a result of this research, writing support materials have been developed, such as The Academic Formulas List (AFL) [22] and the Academic Phrasebank [15]. Subsequently, digital corpora began to be used as part of ESP/EAP courses, where students were taught how to explore authentic language use through Data-Driven Learning (DDL) [7, p. 252]. Corpora are accessed through corpus linguistics tools, or more recently through user-friendly corpus interfaces specifically designed for teaching purposes (see e.g., [17]; [6]). Recent research has shown that this approach is effective in improving students’ writing skills and ESP-related competencies, including discipline-specific genre awareness and professional communicative competence ([4]; [23]; [3]).

3. Methods

The primary corpus under analysis in this paper is a dataset from the Romanian Genre Corpus (ROGER) [5]. ROGER is a bilingual, comparable learner corpus containing academic writing produced by students enrolled in several Romanian universities. The full corpus totals 3.11 million words and includes various text genres from five disciplines, written either in the students’ native language, Romanian (referred to as "L1 Romanian"), or in English as a Foreign Language (referred to as "L2 English").

The sub-corpus selected for this study consists of short texts (e.g., essays) written in L2 English by native Romanian undergraduate students. These texts were produced within ESP courses by students in the field of Engineering (Information Technology, Mathematics, etc). Details concerning the corpus composition are presented in Table 1 below.

Discipline sub-corpus	Text Genre	Number of texts	Number of words
Engineering	Case Study, Documentation, Essay, Research, Report	126	229,078

Table 1. ROGER-Engineering corpus dataset

The secondary dataset used in this investigation is the *British Academic Written English (BAWE) Corpus* [16], a collection of proficient student writing from British higher education totaling approximately 6.5 million words. This corpus was accessed through the corpus query platform SketchEngine and serves as the reference corpus for comparative analysis.

The corpus linguistics analysis was conducted using SketchEngine[1], which offers several analytical affordances, including N-gram extraction, Keyword Analysis, Key Word in Context (KWIC) concordances, and the calculation of relative frequency per million words.

The extraction of phraseological units followed a semi-automated procedure enabled by SketchEngine. Lexical bundles (or N-grams) are continuous sequences of words identified according to predefined extraction parameters. For this study, 3- to 5-word N-grams were extracted with a minimum frequency threshold of 20 occurrences per one million words. To prevent bundle overlap, shorter N-grams embedded within longer sequences were grouped using the N-gram nesting function. Additionally, all bundles were manually checked to ensure functional coherence and to exclude incomplete or structurally fragmented sequences. Importantly, no pre-defined phraseological structures were searched for in advance. Instead, the identified patterns emerged inductively from the corpus data.

4. Results and Discussion

Table 2 lists the top 20 lexical bundles specific to the ROGER-Engineering corpus and the top 20 lexical bundles specific to BAWE. These key bundles were identified using the Keyword Analysis function in SketchEngine, which allows the identification of n-grams that are significantly more frequent in the corpus under investigation (ROGER-Engineering) than in the reference corpus (BAWE). As such, these bundles may be taken to reflect recurrent rhetorical and discourse-functional preferences in each corpus.

Romanian students frequently employ argument-structuring phrases such as *when it comes to* and *on the other hand*. Based on Hyland's [13] functional classification, many of the bundles identified in the ROGER-Engineering corpus perform topic-referential or procedure-describing functions (e.g*. the running time is, in which the user, array is already sorted*). In addition, students tend to rely on engagement-oriented frames such as *we can see that, as you can see*, and *if you want to*, which explicitly foreground writer presence and reader involvement. Rather than realising stance or discourse-organising functions, these bundles are typically used to introduce explanations or guide the reader through computational procedures. This pattern may indicate a stronger emphasis on technical specificity and process description, as opposed to the adoption of a more conventional academic writing style characterised by impersonal stance and cautious evaluation.

The concordance evidence further illustrates this tendency. In example (a), the repeated use of *we can see* frames the interpretation of results in an explicitly instructional manner, while in (c), the use of *if you want to* introduces a procedural explanation through a conversational structure that resembles directive or user-oriented discourse. Such constructions are comparatively rare in expert academic writing, where claims are more frequently presented through impersonal stance markers.

In contrast, the BAWE corpus contains bundles that contribute to the construction of rhetorical strategies, particularly hedging and evidentiality, through phrases such *as it is possible to, are more likely to,* and *is likely to be*. Many of these bundles realise stance and discourse-organising functions, typically through anticipatory it constructions and passive reporting structures (e.g. *it can be seen that, it is clear that*). These forms allow writers to present claims in a depersonalised manner and to signal

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

degrees of epistemic certainty. As illustrated in examples (d)–(f), interpretation and generalisation are framed through evidential reference (*it can be seen that; from the above table, it can be seen that*) or cautious inference (*is likely to be*), thereby contributing to a more formal and impersonal academic style.

Taken together, the functional distribution of lexical bundles in the two corpora suggests that while ROGER-Engineering students demonstrate familiarity with discipline-specific terminology and computational description, their writing shows a tendency to rely on engagement-oriented and procedure-focused phraseology. By comparison, the BAWE corpus reflects a greater use of stance and discourse-organising bundles that support argumentation, hedging, and the presentation of claims in accordance with established academic conventions.

ROGER-Engineering			**BAWE**		
Rank	**Bundle**	**Relative freq. pmw**	**Rank**	**Bundle**	**Relative freq. pmw**
1	one of the most	158.95	1	can be seen in	43.54
2	the running time is	117.32	2	it is possible to	39.23
3	the end of the	117.32	3	the way in which	32.15
4	the name of the	83.26	4	can be seen that	28.91
5	at the same time	83.26	5	the extent to which	27.35
6	to be able to	79.47	6	in terms of the	26.99
7	when it comes to	75.69	7	It is important to	26.27
8	array is already sorted	68.12	8	it is clear that	26.03
9	can be used to	60.55	9	it can be seen	25.43
10	is one of the most	60.55	10	is due to the	25.07
11	has the purpose to	52.98	11	the World Wide Web	24.23
12	of the most popular	52.98	12	it is difficult to	23.75
13	on the other hand	52.98	13	are more likely to	23.03
14	of the most important	49.20	14	in relation to the	22.55
15	As you can see	37.84	15	in the same way	21.35
16	we can see that	34.06	16	the role of the	21.23
17	if you want to	34.06	17	it can be seen that	20.99
18	in which the user	30.28	18	as a result of the	20.75
19	the running time of	30.28	19	the case of the	17.75
20	Sort and Quick Sort	30.28	20	is likely to be	17.75

Table 2. Lexical bundles in ROGER-Engineering and in BAWE

Concordances from the ROGER-Engineering corpus

(a) In order to see the impact of these results better, **we can** check the graph below: **We can see** in the graph, how the running time of insertion sort increases
(b) Randomly sorted arrays: for **the running time is** 730s;
(c) This operation is used **if you want to** list your account to the sale.

Concordances from the BAWE-Physical Sciences corpus

(d) From equation 6 **it can be seen that** a plot of logb against logA (see Figure 1) will produce a straight line graph […]
(e) It is already known that hot dark matter **is likely to be** made up of particles that can travel close to the speed of light
(f) From the above table, **it can be seen** that WACC increased every year from 7.15 in 2001 to 9.06 in 2005.

5. Pedagogical Applications

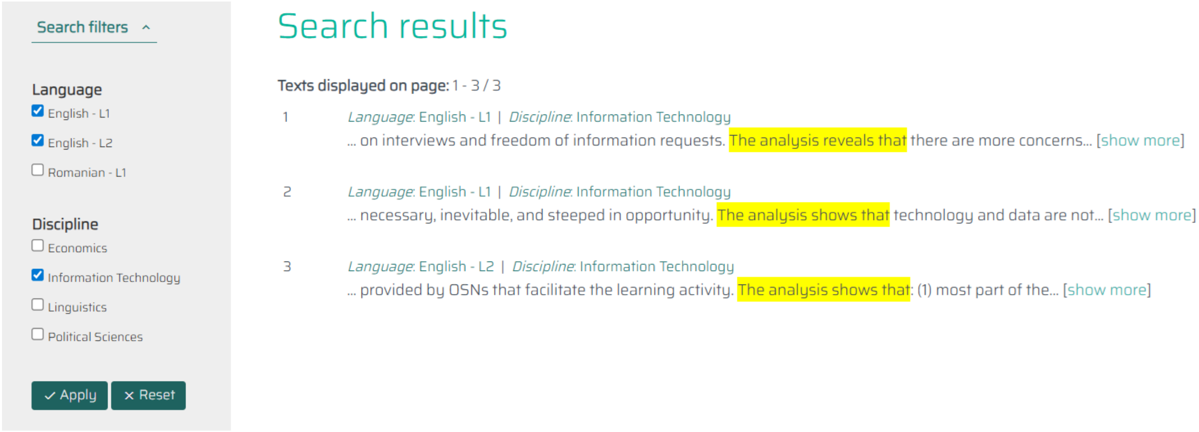
5.1. The EXPRES Corpus Query Platform

The Corpus of Expert Writing in Romanian and English (EXPRES) [6] is a bilingual corpus of expert academic writing in English and Romanian. It comprises peer-reviewed research articles published in high-impact scientific journals across four disciplines: Linguistics, Political Science, Economics, and Information Technology. The corpus includes three language varieties: English L1, English L2, and Romanian L1. Articles in the English L1 sub-corpus were authored by scholars whose native language is English, whereas the English L2 sub-corpus contains articles written in English by native Romanian scholars. The third component consists of research articles written in Romanian by Romanian native scholars.

The EXPRES online platform provides free and user-friendly access to the corpus and offers multiple data-filtering options. Texts can be selected according to discipline and language variety, as illustrated in Figure 1. The possibility of filtering by discipline is particularly relevant in ESP settings, as it enables students to investigate discipline-specific vocabulary, recurrent phraseological patterns, and rhetorical conventions. At the same time, cross-disciplinary comparison can facilitate greater awareness of how academic discourse varies across fields.

The platform supports simple corpus consultation searches where users can enter a word or phrase into the search field, and the platform retrieves all corresponding occurrences in the corpus. Each concordance line is accompanied by metadata indicating the discipline and language variety of the source text. The surrounding textual context is displayed, allowing learners to observe how the phrase functions in authentic academic discourse. This type of guided corpus exploration is central to DDL, as it encourages students to notice recurring patterns and identify typical word combinations in a given discipline.

In addition, the platform supports wildcard searches, where a word within a phrase can be replaced by the character "*", enabling the retrieval of multiple structural variations. For example, a common expression in student writing is "the results show that". By searching for the string "the analysis * that", learners can identify alternative formulations such as "the analysis reveals that". Such searches can be incorporated into DDL tasks targeting lexical variety while reducing formulaic repetition.



Search filters ^

Language
☑ English - L1
☑ English - L2
☐ Romanian - L1

Discipline
☐ Economics
☑ Information Technology
☐ Linguistics
☐ Political Sciences

✓ Apply ✕ Reset

Search results

Texts displayed on page: 1 - 3 / 3

1 *Language*: English - L1 | *Discipline*: Information Technology
... on interviews and freedom of information requests. The analysis reveals that there are more concerns... [show more]

2 *Language*: English - L1 | *Discipline*: Information Technology
... necessary, inevitable, and steeped in opportunity. The analysis shows that technology and data are not... [show more]

3 *Language*: English - L2 | *Discipline*: Information Technology
... provided by OSNs that facilitate the learning activity. The analysis shows that: (1) most part of the... [show more]

Fig. 1. Screenshot EXPRES platform

5.2. Teaching Discipline-specific Phraseology through DDL

The ROGER analysis showed that Romanian engineering students frequently rely on personal stance markers (e.g. *we can say that*) rather than the impersonal constructions typical of academic writing. The following DDL activities aim to expand students' repertoire of impersonal stance expressions and promote more objective, discipline-appropriate language use.

Using the EXPRES corpus, students explore authentic examples of anticipatory it-constructions and related impersonal patterns, analyse differences in degrees of certainty and caution, and contrast these forms with personal expressions found in learner writing. The final task requires students to reformulate excerpts from engineering papers using appropriate impersonal constructions.

Activity 1: Understanding impersonal expressions
Task: Open the EXPRES corpus, select English L1 and English L2 and the Information Technology discipline. Search for the phrases below and observe the context in which they appear. Write down one example for each phrase. What do they all have in common when compared to personal expressions like “we can say that” or “From my point of view”?

- It can be argued that
- It can be stated that

- The evidence suggests that
- One could argue that
- It might be proposed that
- It is plausible to suggest that
- It is evident that

Activity 2: Ordering task (adapted from [21])
Order the following phrases in a scale, from most certain to least certain. Search the EXPRES corpus and write down one a sentence containing each phrase.
it is possible; it is likely; it is unlikely; it is clear; it seems that;

Activity 3: Exploring anticipatory “it” constructions

a) Open the EXPRES corpus and search for the phrases: **it can be * that** and **it * be that**. Write down a few examples for each phrase.
b) Search the EXPRES corpus and write down one sentence for each of the following words: *may; could; might; is likely to; it seems that; it is possible*. Why do you think these words and phrases are used in the author's writing?
c) Using the expressions you retrieved, try and categorize them into the following groups:

Certainty (e.g. *it can be concluded that*)	Speculation/ Being cautious (e.g. *it can be suggested that*)	Evidence-based (e.g. *it can be observed that*)

d) Below are excerpts from student papers that use a variety of personal expressions. Rewrite the excerpts using impersonal expressions encountered in the previous exercises.
   1) “In order to see the impact of these results better, **we can check** the graph below**: We can see** in the graph, how the running time of insertion sort increases.” (Engineering paper)

   2) “Based on these information **we can compare** sorting algorithms based on complexity. **We can say tha**t the first three algorithms with are slow because they require scanning the whole array “even it's almost or already sorted.” (Engineering paper)
   3) **We can observe that** the winner for overall performance and utility is the Quicksort algorithm.” (Engineering paper)

6. Conclusion

This study has shown that Romanian Engineering students’ academic writing is characterised by a preference for engagement-oriented and procedure-focused phraseological patterns, which reflect a more personal and instructional style than that typically found in proficient academic discourse. By contrast, expert writing demonstrates a greater reliance on impersonal stance expressions that enable writers to present claims cautiously and objectively. Drawing on these findings, the paper has illustrated how learner corpus analysis can be used to inform the design of discipline-specific DDL activities aimed at developing students’ phraseological competence. The proposed tasks encourage learners to explore authentic corpus data and to notice functional differences between personal and impersonal constructions in academic writing. More broadly, the study suggests that integrating learner corpus diagnosis with guided corpus consultation may represent a viable pedagogical approach to supporting the development of effective phrase use in Engineering ESP contexts.

REFERENCES

[1] Ackerley K., “Students’ preferences and strategies”, in Charles M., Frankenberg-Garcia A. (Eds.), Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis, London, Routledge, 2021, pp. 78–99.

[2] Biber D., Conrad S., Cortes V., “If you look at …: Lexical bundles in university teaching and textbooks”, Applied Linguistics, Oxford, Oxford University Press, 2004, 25(3), pp. 371–405, doi:10.1093/applin/25.3.371.

[3] Boumediene H., “Enhancing ESP Language Learning with Corpus Tools and Data-Driven Approaches”, Journal of Studies in Language, Culture, and Society (JSLCS), 2025, 8(3), pp. 37–48.

[4] Chitez M., Bercuci L., “Data-driven learning in ESP university settings in Romania: Multiple corpus consultation approaches for academic writing support”, in Meunier F., Van de Vyver J., Bradley L., Thouësny S. (Eds.), CALL and complexity – Short papers from EUROCALL 2019, Louvain-la-Neuve, Research-publishing.net, 2019, pp. 75–81, doi:10.14705/rpnet.2019.38.989.

[5] Chitez M., Bercuci L., Dincă A., Rogobete R., Csürös K., Corpus of Romanian Academic Genres (ROGER) [Data set], Timișoara, West University of Timișoara, 2021.

[6] Chitez M., Mureșan V., Rogobete R., Dincă A., Corpus of Expert Writing in Romanian and English (EXPRES) [Data set], Timișoara, West University of Timișoara, 2022.

[7] Cotos E., “Language for specific purposes and corpus-based pedagogy”, in Chapelle C.A., Sauro S. (Eds.), The handbook of technology and second language teaching and learning, Hoboken, John Wiley & Sons, 2017, pp. 248–264, doi:10.1002/9781118914069.ch17.

[8] Ebeling S.O., Hasselgård H., “Learners’ and native speakers’ use of recurrent word-combinations across disciplines”, Bergen Language and Linguistics Studies, Bergen, University of Bergen, 2015, 6, pp. 87–106, doi:10.15845/bells.v6i0.810.

[9] Flowerdew L., “Learner corpora and language for academic and specific purposes”, in Meunier F., Gilquin G., Granger S. (Eds.), The Cambridge handbook of learner corpus research, Cambridge, Cambridge University Press, 2015, pp. 465–484, doi:10.1017/CBO9781139649414.021.

[10] Granger S., Paquot M., “Disentangling the phraseological web”, in Meunier F., Granger S. (Eds.), Phraseology: An interdisciplinary perspective, Amsterdam, John Benjamins, 2008, pp. 27–49.

[11] Gray B., Biber D., “Phraseology”, in Biber D., Reppen R. (Eds.), The Cambridge handbook of English corpus linguistics, Cambridge, Cambridge University Press, 2015, pp. 160–181.

[12] Hasselgård H., “Phraseological teddy bears: Frequent lexical bundles in academic writing by Norwegian learners and native speakers of English”, in Wiegand V., Mahlberg M. (Eds.), Corpus linguistics, context and culture, Berlin, De Gruyter, 2019, pp. 339–362, doi:10.1515/9783110489071-013.

[13] Hyland K., “Academic vocabulary and lexical bundles in disciplinary discourses”, Applied Linguistics, Oxford, Oxford University Press, 2008, 29(1), pp. 1–24.

[14] Laufer B., Waldman T., “Verb-noun collocations in second language writing: A corpus analysis of learners' English”, Language Learning, Hoboken, Wiley, 2011, 61(2), pp. 647–672, doi:10.1111/j.1467-9922.2010.00621.x.

[15] Morley J., Academic phrasebank: A compendium of commonly used phrasal elements in academic English, Manchester, The University of Manchester, 2018.

[16] Nesi H., Gardner S., The British Academic Written English (BAWE) corpus [Data set], Warwick, University of Warwick, 2012.

[17] O’Flynn J., “Lexical bundles in the academic writing of the Arts and Humanities: from corpus to CALL”, Yearbook of Phraseology, Berlin, De Gruyter, 2022, 13(1), pp. 81–108, doi:10.1515/phras-2022-0006.

[18] Paquot M., “Lexicography and phraseology”, in Biber D., Reppen R. (Eds.), The Cambridge handbook of English corpus linguistics, Cambridge, Cambridge University Press, 2015, pp. 460–477, doi:10.1017/CBO9781139764377.026.

[19] Paquot M., Granger S., “Formulaic language in learner corpora”, Annual Review of Applied Linguistics, Cambridge, Cambridge University Press, 2012, 32, pp. 130–149, doi:10.1017/S0267190512000098.

[20] Pawley A., Syder F.H., “Two puzzles for linguistic theory: Nativelike selection and nativelike fluency”, in Richards J.C., Schmidt R.W. (Eds.), Language and communication, London, Longman, 1983, pp. 191–225.

[21] Salazar D., Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching, Amsterdam, John Benjamins Publishing Company, 2014, doi:10.1075/scl.65.

[22] Simpson-Vlach R., Ellis N.C., “An academic formulas list: New methods in phraseology research”, Applied Linguistics, Oxford, Oxford University Press, 2010, 31(4), pp. 487–512, doi:10.1093/applin/amp058.

[23] Yan H., “A blended system for data-driven learning of English for specific purposes”, International Journal of Emerging Technologies in Learning (iJET), Kassel, iJET, 2022, 17(12), pp. 121–134, doi:10.3991/ijet.v17i12.29653.