https://doi.org/10.22364/vnf.10.13

Lithuanian academic vocabulary: a corpus-based study[1]

Lietuviešu akadēmiskās valodas leksikas korpuslingvistiska izpēte

Agnė Lisauskaitė
Vilnius University, Faculty of Philology
Institute of Applied Linguistics
Department of Lithuanian Language
Universiteto St. 5, LT-01513 Vilnius, Lithuania
E-mail: *agne.lisauskaite@flf.stud.vu.lt*

The aim of this study is to explore the frequency of Lithuanian academic words and expressions, and to analyse their structure and semantics applying the corpus-based approach. This paper focuses on academic vocabulary used in Lithuanian research papers. The following objectives have been raised to achieve the aim: 1. to build a corpus of research papers published in the Lithuanian language over the past decade; 2. to identify the most frequent words and expressions (collocations) in Lithuanian academic writing applying statistical methods; 3. to explore the distribution, discourse functions, variability, semantic and pragmatic values of academic words. This paper first gives a brief overview of the key terms and theoretical background, then presents the research results and ends with the summary of the main findings.

Keywords: academic vocabulary; collocation; corpus; lexical bundle; Lithuanian; semantic (pragmatic) function.

1. Introduction

The purpose of this study is to explore academic vocabulary in Lithuanian research papers and to examine the frequency, structure and semantics of academic words and expressions applying the corpus-based approach. The following objectives have been raised to achieve the aim:

1) to compile a corpus of research papers published in the Lithuanian language over the past decade;
2) to identify the most frequent words and expressions (collocations) in Lithuanian academic writing applying statistical methods;
3) to compare some Lithuanian academic words and collocations with their English counterparts;
4) to explore the distribution, discourse functions, variability as well as semantic and pragmatic values of these academic words.

[1] The project was funded by Operational Programme for the European Union Funds' Investments (Contract No. 09.3.3-LMT-K-712-03-0042).

In the course of this study, a specialised corpus has been compiled and then used for the automated extraction of the most frequent words and expressions, which have then been analysed in terms of their function, structure (extensions, variability) and semantics (pragmatics).

Over the last decades, academic vocabulary, lexical bundles, collocations and their possible expansions have been studied extensively around the world. As a result, several corpus-driven lists of academic words and collocations have been compiled[2]. It has to be noted, however, that the vast part of this research has been confined to academic English.

This study investigates academic vocabulary, the definition of which remains the subject of some debate amongst foreign researchers. According to Paquot (2010, 9), academic vocabulary is often seen as a set of lexical items that are not core words[3], but which are relatively frequent in academic discourse, e.g., *chemical*, *colleague*, *consist*, *contrast*. Regardless of the discipline, these words tend to occur in a large proportion of academic texts. Summarising several available definitions, Paquot (2010, 28) offers the following definition of academic vocabulary: 'a set of options to refer to those activities that characterise academic work, organise scientific discourse and build the rhetoric of academic texts'.

Another key concept used in this paper is *collocation.* According to Halliday and Hasan's (1976, 287), collocation is a cover term that refers to 'cohesion that is achieved through the association of lexical items that regularly co-occur in similar environments'. In phraseology research, a collocation is generally seen as a continuum with varying degree of arbitrary restriction ranging from free combinations (*write an essay*), through restrictive collocations (*conduct / do research*) to frozen idioms (*generally speaking*). However, even free combinations are restricted by their semantic and/or syntactic environment. So, the boundaries between free combinations and restrictive collocations are sometimes blurred and are thus difficult to distinguish (Ackermann, Chen 2013, 2). It can be noted that collocations are generally recognisable without any difficulty by native speakers but often cause difficulties for language learners.

The study reported here also deals with lexical bundles. According to Salazar (2014, 13), lexical bundles were first introduced by Biber, Johansson, Leech, Conrad and Finegan in a chapter of the *Longman Grammar of Spoken and Written English* (1999, 989), where Biber and his colleagues offered a definition of lexical bundles as sequences of words that show a tendency to co-occur. The principal characteristic of lexical bundles is that the method of their identification relies mainly on frequency criteria. So lexical bundles can be described as frequently occurring lexical sequences. They can be classified according to their structure, meaning and function. The classification based on the number of constituents in

[2] Cf. ://www.victoria.ac.nz/lals/resources/academicwordlist, https://pearsonpte.com/wp-content/uploads/2014/07/AcademicCollocationList.pdf, http://www.phrasebank.manchester.ac.uk/

[3] Core words are the words that are frequently used in a language. A vocabulary of core words consists of function words (*and*, *about*, *be*, *to*, etc.) and content words like *lesson*, *person*, *suggest*, etc. Stubbs defines such words as pragmatically neutral words and adds that they give no indication about the field of discourse, in which they are used (Stubbs 1986, 104).

a lexical bundle is one of the most objective (*two-word*, *three-word*, etc.). The structural classification analyses lexical bundles in terms of their part-of-speech composition. The functional classification provided by Biber puts more focus on their semantic and pragmatic role and distinguishes among three functions: 1. referential expression; 2. discourse organising, or text-oriented; 3. stance expression, or participant-oriented (Juknevičienė 2011, 26–27).

Finally, this study seeks to explore the semantic (pragmatic) function of academic collocations, matching them to the rhetorical functions that are typical of certain sections of academic text. The analysis of the rhetorical functions of text parts started when several researchers tried to identify a list of words or expressions that could characterise different rhetorical moves which are considered part of the core organisation of different sections of the research article. The underlying assumption of this work was that a text is made up of functional units, or moves, that accomplish a communicative purpose of a given genre. Each move is comprised of several steps. Swales (1981) identified four moves typical of the research articles introductions: *establishing the field*, *summarising previous research*, *preparing present research*, and *introducing present research*. Since then, various researchers attempted to associate words or expressions with rhetorical moves in different sections of academic text (Brett 1994; Williams 1999; Ruiying, Allison 2003; Kanoksilapatham 2007; Flowerdew, Forest 2009; Cortes 2013) (Salazar 2014, 2). Having analysed the empirical material collected, this study has identified three principal moves, or rhetorical functions: 1. presentation of research (in the introductory part of the article); 2. methodology description and analysis (in the introduction and the main body of the paper); 3. presentation of research findings (in the conclusions).

There is a considerable body of research investigating various aspects of Academic Lithuanian. For example, Alaunienė (2005) examined the metalanguage and structure in academic discourse; Bitinienė' (2000, 2005 and 2013) analysed the syntactic features of academic texts and their intertextuality; Damošius (2007) explored the notion of evaluation/authorial stance. Despite all these efforts, corpus-based research of Academic Lithuanian is still lacking. In recent years, a number of studies have contributed to filling in this gap: Šinkūnienė (2014) examined the specific characteristics of Lithuanian humanities and social sciences discourse; several interlingual studies addressed such questions as authorial stance, hedging, modality, self-referencing and adverbalisation in scientific writing (Šinkūnienė 2010, 2011, 2015; Linkevičienė, Šinkūnienė 2012; Smetona, Usonienė 2012; Šinkūnienė, Van Olmen 2012; Mur Dueñas, Šinkūnienė 2016). All this research was based either on the corpora compiled by the researchers themselves or on the Lietuvių mokslo kalbos tekstynas [Corpus of Academic Lithuanian, available at: http://coralit.lt/].

However, only a small number of studies have so far targeted Lithuanian academic vocabulary. To date, there have been only two corpus-driven investigations of the use of specialised language and collocation in undergraduate student writing, one in social sciences and engineering (Gudavičienė 2018), and the other in humanities and engineering (Volungevičienė 2018). This work is being carried forward by another currently ongoing project [*Research of Phrasemes in Student*

Writing and an Interactive Phrase Bank, available at http://www.fraziskumas.flf.vu.lt/] funded by The State Commission of the Lithuanian Language (the project manager is Assoc. Prof. Vilma Zubaitienė) which seeks to explore academic phrases in undergraduate student writing in six main fields of study on the basis of a corpus of student papers that has been compiled specially for this purpose. The authors of the study analyse the most frequently used phrases, identify their types, describe phrase variety or lack thereof, note variations of phrase forms and register the frequency of their occurrence in discipline-specific texts and in different sections of text. The findings of these corpus-based studies will provide insights into linguistic choices of expert and novice academic writers.

2. Data and methods

For the purpose of this study, a specialised corpus of research articles published in Lithuanian between 2010 and 2016 has been compiled covering five main study areas (H000 – humanities, S000 – social sciences, P000 – physical sciences, B000 – biomedical sciences, T000 – engineering). The empirical research material was collected from the Lietuvos akademinė elektroninė biblioteka (eLABa) [The Lithuanian Academic Electronic Library, available at https://www.elaba.lt/elaba-portal/pradzia]. This corpus contains almost 2 million words and is divided onto five subcorpora according to study area, which in their turn are further subdivided into narrower disciplines.

Study Area	Number of texts	Disciplines
Humanities	126	Ethnology, philology, philosophy, history, theology, art history
Social sciences	168	Law, political science, economics, psychology, education, management, sociology, psychology, information and communication
Physical sciences	106	Mathematics, physics, chemistry, geography, geology, informatics, biochemistry, physical geography
Biomedical sciences	160	Biology, biophysics, ecology and environmental science, botany, zoology, medicine, dental surgery, pharmacy, public health, nursing
Engineering	203	Electrical engineering, civil engineering, transport engineering, environmental engineering, chemical engineering, power and thermal engineering, computer engineering, materials engineering, mechanics engineering, measurement engineering
Total:	**763**	

Table 1. The corpus composition and size

At the outset, the intention was to develop a corpus where each discipline was represented by at least three texts for each year between 2010 and 2016, but this turned out to be impossible because of technical constraints. So, paleontology and astronomy, for example, have been excluded from the physical sciences sub-corpus because most of the research papers available at the eLABa were either written in

English or did not correspond to the study period. Some other disciplines could not meet the three-article criterion: the search returned fewer than three articles, and for some disciplines, none. This could be explained by the fact that some earlier papers, mostly those dating back to 2010–2011, had invalid weblinks and could no longer be found. It should also be noted that the number of Lithuanian scientists choosing to publish their work in English is constantly increasing, especially in the area of biomedical sciences. Due to all these factors, some disciplines have been excluded from the corpus, while others are represented by fewer than three research papers.

In terms of size, the largest sub-corpus is that of engineering (26,61 %), followed by social sciences (22,02 %), biomedical sciences (20,97 %), humanities (16,51 %) and physical sciences (13,50 %).

The present study is based on the corpus-driven methodology. The corpus is comprised of research papers extracted from the eLABa. Once the relevant texts had been selected, the .pdf files were converted to .txt files and then transferred to .doc files. All the information that is not relevant to the current study, such as abstracts, keywords, examples, footnotes, references, summary and page numbers, was deleted. Then every paper was assigned a name according to its field of science, discipline, publication year and number (e.g., B000_01B_2010_1). Academic vocabulary (both words and expressions) was retrieved from the corpus using *AntConc 3.4.4* software programme (Anthony 2014). The same *AntConc* software was then used to extract frequent word lists (*Word List* function) from both the general corpus (763 research articles) and subject-specific sub-corpora (each containing a 100 research papers) as well as to generate lists of lexical bundles, which in their turn served as a basis for the manual extraction of collocations. Finally, collocation extensions were retrieved by the *AntConc* software.

3. Results

3.1. Academic vocabulary lists

First, the most frequent words were extracted from the corpus applying the cut-off frequency of 100 occurrences. As explained above, the corpus has been divided into five subcorpora of 100 scientific papers each. The classification of research material by subcorpora and the subsequent development of discipline-specific word lists has allowed to distinguish between subject-specific and cross-disciplinary, that is more general, lexical items.

Table 2 shows the number of academic words per subcorpus.

Study Area (subcorpus)	Number of Academic Words
Humanities	51
Social sciences	175
Physical sciences	49
Biomedical sciences	44
Technological sciences	51
Total:	**370**

Table 2. The total number of academic words per subcorpus

Then lexical items have been lemmatised: in other words, each entry includes all the inflected forms of a word. Table 3 shows the lemma of an academic word and its frequency of occurrence in research articles.

No.	Frequency	Academic word
1.	8770	*Tyrimas* ‘research’
2.	7647	*Metas* ‘time’
3.	7519	*Turėti* ‘to have’
4.	5138	*Darbas* ‘work’
5.	4945	*Nustatyti* ‘to establish’
6.	4917	*Sistema* ‘system’
7.	4404	*Duomuo* ‘data’
8.	4164	*Naudoti* ‘to use’
9.	3878	*Rezultatas* ‘result’
10.	3725	*Socialinis* ‘social’

Table 3. 10 most frequent academic words

As can be seen in Table 3[4], the most frequent word in Lithuanian academic writing is *tyrimas* ‘research/study’. The rest of the words shown in this Table, such as *metas* ‘time’, *darbas* ‘work’, *naudoti* ‘to use’, *turėti* ‘to have’, are not specific to the academic register as they occur frequently in other types of text. However, the fact that these words appear at the very top of the list shows that they are indispensable in academic writing. From a linguistic point of view, the noun is the most prevalent part of speech: there are only three verbs and one adjective. This linguistic distribution of academic words seems to corroborate the assumption that academic texts are characterized by the high prevalence of nouns. It is also interesting to note that the only adjective among the first 10 most frequent academic words, *socialinis* ‘social’, is the only discipline-specific word: it has a particularly high frequency in social sciences papers. Just three Lithuanian words, *tyrimas (research)*, *nustatyti (to establish)* and *duomuo (data)*, have their counterparts in the *Academic Word List* (AWL)[5]. Other Lithuanian academic words do not have equivalents in the AWL.

Turning now to the subcorpus lists, it has been found that some sublists are dominated by discipline-specific lexical items, while others comprise mostly

[4] Cf. The most frequent academic nouns in Bachelor’s theses in social sciences and engineering are *analizė* ‘analysis’, *duomenys* ‘data’, *įtaka* ‘influence’, *klausimas* ‘question’, *lentelė* ‘table’, *metodas* ‘method’, *paveikslas* ‘picture’, *problema* ‘problem’, *tikslas* ‘aim’ and *tyrimas* ‘research’ (Gudavičienė 2018, 7).

[5] AWL was developed by Averil Coxhead. First of all, a written corpus of academic English was compiled for the purpose of finding out which words occurred in a wide range of academic texts from a variety of subject areas. It includes four faculty sections: Arts, Commerce, Law and Science. The Academic Corpus contained academic journal articles, book chapters, course workbooks, laboratory manuals, and course notes (Coxhead 2000, 219). The word list has been divided into sublists. For example, the words in Sublist 1 occur more frequently in the corpus than the other words in the list.

general, or cross-disciplinary, academic words. A good example of the former would the humanities sublist presented in the table below.

No.	Frequency	Academic Word
1.	1258	*Kalba* ‘language’
2.	744	*Kalbėti* ‘to talk’
3.	700	*Žodis* ‘word’
4.	683	*Tekstas* ‘text’
5.	639	*Tapti* ‘to become’
6.	617	*Istorija* ‘history’
7.	571	*Kultūra* ‘culture’
8.	570	*Reikšmė* ‘meaning’
9.	551	*Kūrinys* ‘work’
10.	499	*Dalis* ‘part’

Table 4. The most frequent academic words in the humanities papers

Nearly all the words in Table 4 are closely associated with various branches of humanities: for example, words like *kalba* ‘language’, *kalbėti* ‘to talk’, *žodis* ‘word’, *tekstas* ‘text’, *reikšmė* ‘meaning’, *kūrinys* ‘work’ are clearly linked to philology and language studies; the lexical item *kultūra* ‘culture’ has a high frequency of occurrence in academic writing on philosophy, philology and history of art. Perhaps it comes as no surprise that the word *istorija* ‘history’ is frequently found in research papers on history. Only two items on the list, *tapti* ‘to become’ and *dalis* ‘part’, have a general meaning and are not discipline-specific. Just one academic word from the humanities list, i.e., *text* appears also in the AWL sublists.

In contrast to the humanities list, the physical sciences list is dominated by lexical items with a general, or cross-disciplinary, meaning.

No.	Frequency	Academic Word
1.	940	*Duomuo* ‘data’
2.	881	*Sistema* ‘system’
3.	876	*Tyrimas* ‘research/study’
4.	816	*Naudoti* ‘to use’
5.	760	*Nustatyti* ‘to establish’
6.	686	*Metodas* ‘method’
7.	631	*Gauti* ‘to obtain’
8.	604	*Atlikti* ‘to do/perform/carry out’
9.	603	*Darbas* ‘work/study’
10.	597	*Modelis* ‘model’

Table 5. The most frequent academic words of physical sciences

Table 5 reveals that all the most frequents words on the physical sciences list convey abstract meaning: in other words, they could potentially be used in scientific writing across all study areas. These academic words are not subject-specific, and their use is not restricted to physical sciences. Several lexical items

on this list, such as *data*, *obtained*, *research*, *established*, *method*, are also typical of English academic discourse as evidenced by the AWL.

It is interesting to note that the lexical item *tyrimas* 'research/study' features among the first 10 most frequent academic words in all the sublists, except for the humanities one. This could be explained by the fact that this word is highly cross-disciplinary and, as such, is inevitable in academic writing. A similar status is enjoyed by the word *nustatyti* 'to establish', which also appears on almost every sublist. This lemma could be understood as a contextual academic word that has a high frequently of occurrence in other types of text.

3.2. Academic lexical bundles

A list of three- and four-word bundles was extracted from the corpus in order to examine the lexical environment in which academic words function, and to identify regularly re-occurring word sequences[6]. *Cluster / N-Grams* tool returned 200 lexical bundles based on the following criteria: the raw minimum frequency in research articles was set to 7 and the same type of lexical bundle had to occur in at least 5 different texts.

No.	Frequency	Occurrence in text	Lexical bundle
1.	368	217	*galima teigti kad* 'it can be assumed that'
2.	165	109	*tai reiškia kad* 'it means that'
3.	137	137	*mokslas lietuvos ateitis* 'science Lithuania's future'
4.	127	133	*taip pat yra* 'is also'
5.	126	111	*taip pat buvo* 'was also'
6.	112	72	*rezultatai parodė kad* 'the results/ findings showed that'
7.	109	78	*ir tai kad* 'and the fact that'
8.	101	68	*rezultatai rodo kad* 'the results/findings show that'
9.	99	69	*daryti išvadą kad* 'to draw a conclusion'
10.	98	65	*dėl šios priežasties* 'for this reason'

Table 6. The most frequent lexical bundles in academic discourse

As can be seen in Table 6, three-word bundles turned out to be the dominant structure[7] throughout the list. Some of these bundles, such as *ir tai kad* 'and the fact that', *taip pat yra* 'is also', *taip pat buvo* 'was also' are constituted only of function words, while other bundles contain content words (i.e., *galima teigti kad* 'it can be assumed that', *rezultatai rodo kad* 'the results/findings show that', *daryti*

[6] Given the considerable size of the list of lexical bundles (200 lexical bundles), only the most significant examples are presented in the current paper.

[7] Cf. Lexical bundles in Bachelor's theses containing a noun as their member (Gudavičienė 2018, 8).

išvadą kad ‘to draw a conclusion that’). The presence of a content word in a bundle has been interpreted as a signal of a possible collocation[8].

3.3. Academic collocation list

Following the assumption that lexical bundles may contain collocations within their structure, the 200-word sequences extracted during the previous stage of research were analysed to compile two collocation lists: one general (20 items) and the other discipline-specific (12 items). The table below presents the overview of these collocations according to the part of speech of their constituents.

No.	Component I	Part of speech	Component II	Part of speech	Translation of collocation
1.	*analizė*	N	*rodo*	V	‘the analysis shows’
2.	*atkreipti*	V	*dėmesį*	N	‘to draw attention to/to highlight’
3.	*atsakyti į*	V + PREP	*klausimą*	N	‘to answer a question’
4.	*darbo*	N	*tikslas*	N	‘aim of the study/ paper’
5.	*daryti*	V	*išvadą*	N	‘to draw a conclusion’
6.	*daryti*	V	*prielaidą*	N	‘to assume’
7.	*dėmesys*	N	*skiriamas*	V	‘focuses on’
8.	*gauti*	V	*duomenys*	N	‘data obtained’
9.	*glaudžiai*	ADV	*susijęs*	V	‘closely linked’
10.	*pateikti*	V	*lentelėje*	N	‘are presented/ shown in the table’

Table 7. 10 most frequent general collocations in academic discourse

Structurally, the general collocation list is made up almost entirely of two-word collocations, with only two three-word exceptions (*statistinė duomenų analizė* ‘statistical data analysis’ and *statistiškai reikšmingas skirtumas* ‘statistically significant difference’). In contrast, discipline-specific collocations tend to have a three-word structure. Another interesting result of the structural analysis has been the identification of a recurrent pattern (*modal verb + infinitive)* among two-word sequences, which has not been included into the collocation list, but which could, nonetheless, be treated as a separate group. Some examples of this pattern would be the combinations of the modal verbs *galima* ‘can’, *leidžia* ‘let’, *reikia* ‘should / ought to’ and such infinitive forms of a verbs as *manyti* ‘to think (imperfective)’, *pasakyti* ‘to tell (perfective)’, *pastebėti* ‘to observe (perfective)’,

[8] Although noun is a content word, the lexical bundle *mokslas lietuvos ateitis* ‘science Lithuania’s future’ referring to the journal title was excluded from the further analysis of collocations.

pažymėti ‘to indicate (perfective)’, *sakyti* ‘to tell (imperfective)’, *suskirstyti* ‘to divide (perfective)’, *teigti* ‘to affirm (imperfective)’.

The table below presents the list of discipline-specific collocations.

No.	Component I	Part of speech	Component II	Part of speech	Component III	Part of speech	Translation of collocation
1.	*Antrasis*	NUM	*pasaulinis*	ADJ	*karas*	N	‘Second World War’
2.	*asmens*	N	*sveikatos*	N	*priežiūra*	N	‘human healthcare’
3.	*Lietuvos*	N	*Didžioji*	ADJ	*Kunigaikštystė*	N	‘Grand Duchy of Lithuania’
4.	*lietuvių*	N	*kalba*	N			‘Lithuanian language’
5.	*Lietuvos*	N	*Respublikos*	N	*Konstitucija*	N	‘Constitution of Lithuania’
6.	*Lietuvos*	N	*Respublikos*	N	*Vyriausybė*	N	‘Government of Lithuania’
7.	*socialinių*	ADJ	*mokslų*	N	*studijos*	N	‘studies of social sciences’
8.	*struktūriniai*	ADJ	*fondai*	N			‘structural funds’
9.	*sveikatos*	N	*priežiūros*	N	*specialistai*	N	‘healthcare professionals’
10.	*viešasis*	ADJ	*valdymas*	N			‘public governance’
11.	*viešoji*	ADJ	*vadyba*	N			‘public management’
12.	*žmogaus*	N	*teisės*	N			‘human rights’

Table 8. The most frequent discipline-specific collocations in academic writing

As shown in the examples above, discipline-specific collocations are commonly composed of three words. These collocations cover several fields of study: for example, *Antrasis pasaulinis karas* ‘Second World War’, *Lietuvos Didžioji Kunigaikštystė* ‘Grand Duchy of Lithuania’ belong to the history domain; the expressions *asmens sveikatos priežiūra* ‘human healthcare’, *sveikatos priežiūros specialistai* ‘healthcare professionals’ are typical of biomedical research; the collocation *lietuvių kalba* ‘Lithuanian language’ occurs frequently in humanities papers, especially in those written by philologists; expressions like *Lietuvos Respublikos konstitucija* ‘Constitution of Lithuania’, *Lietuvos Respublikos Vyriausybė* ‘Government of Lithuania’, *socialinių mokslų studijos* ‘studies of social sciences’[9], *struktūriniai fondai* ‘structural funds’, *viešasis valdymas* ‘public governance’, *viešoji*

[9] In some contexts, it may refer to the journal title.

vadyba ‘public management’, *žmogaus teisės* ‘human rights’ are part of the discourse of social sciences.

Finally, it has also been found that while some of the Lithuanian general academic collocations, such as *to draw attention to*, *to draw a conclusion*, *to make an assumption* and *closely linked,* have counterparts in the *Academic Collocation List* (ACL)[10], others were not common in English academic prose.

3.3.1. Variants of collocation components

In order to further our understanding of the functioning of collocations in an academic text, this study has analysed their structural variability. The table below presents examples of the variants of several cross-disciplinary collocations discussed above.

Nr.	Variants (I)	Component I	Component II	Variants (II)
1.		*analizė*	*rodo*	*parodė* ‘show-PST.3’
2.	*atkreipia* ‘turn-PRS.3’ *atkreipę* ‘turn-PTCP. ACT. PST-NOM.PL’ *atkreipus* ‘turn-ADV.PTCP. ACT. PST’ *atkreipėme* ‘turn-PRS.1PL’ *atkreipė* ‘turn-PST.3’	*atkreipti*	*dėmesį*	
3.	*Atsakoma* ‘answer-PTCP.PASS. PRS-NOM’ *atsakant* ‘answer-ADV.PTCP. ACT. PRS’ *atsakė* ‘anwer-PST.3’	*atsakyti į*	*klausimą*	
4.	*daro* ‘draw-PRS.3’ *darome* ‘draw-PST.1PL’	*daryti*	*išvadą*	
5.	*Darant* ‘make-ADV.PTCP.ACT. PRS’ *padarytą* ‘make- PTCP.PASS. PST.-ACC.SG.F’ *sudaro* ‘make- PRS.3’ *darome* ‘make-PRS.1PL’ *darė* ‘make-PST.3’ *darėme* ‘make-PST.1PL’	*daryti*	*prielaidą*	
6.		*dėmesys*	*skiriamas*	*skirtas* ‘turn-PTCP. PASS.PST-NOM.SG.M’
7.		*glaudžiai*	*susijęs*	*susiję* ‘link-PTCP.ACT. PST-NOM.PL’ *susijusi* ‘link-PTCP. ACT.PST-NOM.SG.F’

Table 9. The possible variants of collocation components

[10] ACL contains 2469 most frequent and pedagogically relevant lexical collocations in written academic English. It was compiled from the written curricular component of the Pearson International Corpus of Academic English.

The analysis of collocations has revealed that it is the verb constituent that tends to have variants, which can be either finite verb forms inflected for person, number and tense (*atkreipia*, *atsakė*, *darėme*, etc.) or nonfinite verbals (*atkreipę*, *atsakant*, *padarytą*, etc.). Participles constituents can be inflected for gender (*susiję*, *susijusi*).

No variants have been identified among the discipline-specific collocations, which could be explained by the absence of the verb in their composition.

3.3.2. Academic collocation extensions

In order to establish the lexical environment in which the collocations in question function, their extensions were extracted and then analysed to find out how collocations expand on either side and within their own boundaries.

The analysis showed that on the left, collocations can be extended by different noun forms (*duomenų* 'data' (GEN), *kolegija* 'college' (NOM), *politikoje* 'politics' (LOC), etc.), verb forms (*atlikta* 'performed-PTCP.PASS.PST-NOM. SG.F', *siekiant* 'seeking to', *yra* 'is', *linkęs* 'inclined', etc.), adjectives (*išskirtinis* 'exclusive', *pagrindinis* 'basic', *didžiausias* 'biggest', etc.), e.g., *duomenų analizė rodo* 'the data analysis shows', *didžiausias dėmesys skiriamas* 'the main focus of attention', etc. There are also several examples of extensions by the pronoun (*mes* 'we', *mūsų* 'our', *visas* 'all', *šis* 'this') and the adverb (*adekvačiai* 'adequately', *iš anksto* 'in advance', *teigiamai* 'positively', *pagaliau* 'finally'): e.g., *adekvačiai atsakyti į klausimą* 'to provide an adequate answer to the question', *pagaliau atkreipsiu dėmesį* 'I will finally draw your attention to', etc.

On the right side, collocations are often extended by the conjunction *kad* 'that'. These collocations may behave like independent clauses of a complex sentence, for example *autoriai daro išvadą kad* 'the authors concluded that'. Besides, right-side extensions may include prepositions (*į* 'to', *su* 'with'), adverbs (*kodėl* 'why', *kaip* 'how'), pronouns (*kas* 'who / which') or verb forms (*įkeliami* 'have been uploaded-PTCP.PASS.PRS-NOM.PL.M', *laikyti* 'have been considered to be', *parodė* 'has shown', etc.), for example *gauti duomenys laikyti* 'the data obtained have been considered to be'.

There has been only one case found where the extension occurred within the collocation: *pateikti lentelėje* 'are given in the table'. In this example, the collocation is extended by a numeral indicating the number of the table, e.g., *pateikti X lentelėje* 'are given in Table X'.

3.4. Collocations' semantic (pragmatic) functions

For the purpose of the semantic analysis of academic collocations, the *AntConc* software (Collocation Function) was used to extract collocations from specific sections of text on the basis of a set of key words such as, *prielaida* 'assumption', *uždaviniai* 'objectives', *klausimai* 'questions', *hipotezė* 'hypothesis', *tikslas* 'aim', *metodas* 'method', *duomenys* 'data', *informacija* 'information', *lentelė* 'table', *tyrimas* 'research', *išvados* 'conclusions' .

Three main semantic (pragmatic) functions linked to the rhetorical text functions have been distinguished: *presentation of research, description of methodology and analysis, presentation of research findings.* The *presentation*

of research function is performed by the following collocations: *kelti prielaidą, kad* 'to make an assumption that', *iškelti tokie uždaviniai* 'the following objectives have been raised', *kelia daug klausimų* 'raises a lot of questions', *vertėtų iškelti hipotezę* 'it is worth formulating the hypothesis', *šio darbo tikslas* 'the aim of this study', etc. The *description of methodology and analysis* function is fulfilled by the collocations *taikant X metodą* 'applying the X method', *iš pateiktų duomenų matyti* 'the data provided show', *pateikta informacija apie* 'provides information about', *atsakyti į šį klausimą* 'to answer this question', *pateikta X lentelėje* 'shown in Table X', etc. The collocations *apibendrinant tyrimo rezultatus* 'summing up the results/findings of the study', *apibendrinus tyrimo duomenis* 'having summed up the results/findings of the study', *leido daryti išvadą* 'allowed to draw a conclusion' can be matched with the *presentation of research findings* function. The observation that has emerged from this analysis is that not all key words contain an unambiguous rhetorical function in their semantics. For example, the word *rezultatai* 'results/findings', which has an extensive semantic range, may appear in the collocations that perform both the *description of methodology and analysis* and *presentation of research findings* functions.

Conclusions

1. This study has shown that the most frequent academic word in Lithuanian academic writing is *tyrimas* 'research/study'. From a linguistic point of view, the noun is the most prevalent part of speech in the Lithuanian academic word list. The analysis of the subject-specific sublists has revealed that depending on the study area, the most frequent words, that is those that feature at the top of the list, can be either subject-specific, as is the case of the humanities sublist, or more general in meaning, as exemplified by the physical science sublist.
2. The counterparts of the Lithuanian academic words *tyrimas* 'research', *nustatyti* 'to establish', *duomuo* 'data', *tekstas* 'text', *gauti* 'to obtain', *metodas* 'method' are commonly used in English academic prose.
3. The study has found that academic discourse shows a clear preference for three-word lexical bundles. Some lexical bundles contain only function words, whereas others include content words.
4. The collocations extracted from lexical bundles have been classified into two categories: cross-disciplinary and subject-specific. The cross-disciplinary collocation list comprised 20 items of mostly two-words collocations, whereas most subject-specific collocations had a three-word structure. This investigation has also identified the word string *modal verb + infinitive* as a distinct type of collocation.
    4.1. Some collocations (*atkreipti dėmesį* 'to draw attention to', *daryti išvadą* 'to come to a conclusion', *daryti prielaidą* 'to make an assumption' and *glaudžiai susijęs* 'closely linked') typical of Lithuanian academic writing are also found in the English ACL.
    4.2. The verb constituent of collocation is prone to have variants, i.e., finite (inflected for person) or infinite verbs (inflected for gender).

4.3. The analysis of the expansion patterns has demonstrated that on the left side, Lithuanian collocations are regularly extended by different cases of the noun, verb forms, adjectives, pronouns and adverbs. On the right, the most frequent extension is the conjunction *kad* 'that'. Other right-side extensions may include prepositions, adverbs, pronouns and verbal forms. The study has found only one example of expansion within a collocation: a numeral qualifying a noun (*pateikti X lentelėje* 'are presented in Table X').

5. The semantic analysis of the key word collocations extracted from the corpus has revealed that these collocations carry out three main semantic (pragmatic) functions, which are closely associated with the rhetorical functions of a scientific text: *presentation of research* (*kelti prielaidą, kad* 'to make an assumption that', etc.), *description of methodology and analysis* (*taikant X metodą* 'applying X method', etc.), *presentation of research findings* (*apibendrinant tyrimo rezultatus* 'summing up the results/findings', etc.). It has been noted that some key words have an extensive semantic range and can thus exercise several semantic (pragmatic) functions.

Abbreviations

1, 2, 3	person
ACC	accusative
ACT	active
ADJ	adjective
ADV	adverbial (participle)
F	feminine
GEN	genitive
LOC	locative
M	masculine
N	noun
NOM	nominative
NUM	numeral
PASS	passive
PL	plural
PREP	preposition
PRS	present
PST	past
PTCP	participle
SG	singular
V	verb

Sources

1. *Lietuvos akademinė elektroninė biblioteka* (eLABa). (The Lithuanian Academic Electronic Library.) Available: https://www.elaba.lt/elaba-portal/pradzia
2. *Lietuvių mokslo kalbos tekstynas.* (Corpus of Academic Lithuanian.) Available: http://coralit.lt/
3. *Academic Word List*. Available: https://www.victoria.ac.nz/lals/resources/academicwordlist/sublists

References

1. Ackermann, Kirsten, Chen, Yu-Hua. 2013. Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*. 12 (4), 235–247.
2. Alaunienė, Zita. 2005. Akademinių tekstų struktūra ir jos raiška. *Žmogus ir žodis.* 7(1), 63–67.
3. Biber, Douglas et al. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.
4. Bitinienė, Audronė. 2000. Vientisiniai mokslinio stiliaus sakiniai. *Kalbotyra*. 48(1)–49(1), 19–27.
5. Bitinienė, Audronė. 2005. Mokslinis stilius ir jo intertekstualumas. *Žmogus ir žodis*. 7(1), 68–72.
6. Bitinienė, Audronė. 2013. *Mokslinio teksto stilistika: monografija*. Vilnius: Edukologija.
7. Brett, Paul. 1994. A genre analysis of the result sections of sociology articles. *English for Specific Purposes*. 13(1), 47–59.
8. Cortes, Viviana. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*. 12(1), 33–43.
9. Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly*. 34(2), 213–238.
10. Damošius, Saulius. 2007. Vertinimo raiška mokslinio stiliaus tekstuose. *Lituanistica*. 53, 4 (72), 51–62.
11. Flowerdew, John, Forest, Richard W. 2009. Schematic structure and lexico-grammatical realization in corpus-based genre analysis: the case of Research in the PhD literature review. *Academic Writing: At the Interface of Corpus and Discourse*. Charles, Maggie, Pecorari, Diane, Hunston, Susan (eds.). London: Continuum, 15–36.
12. Gudavičienė, Eglė. 2018. Socialinių ir technologijos mokslų bakalauro darbų kalbos ypatybės. *Žmogus ir žodis / Didaktinė lingvistika*. 20 (1), 4–13.
13. Halliday, Michael, Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.
14. Juknevičienė, Rita. 2011. *Leksinės samplaikos svetimkalbių ir gimtakalbių vartotojų rašytinėje anglų kalboje*. PhD dissertation. Vilnius: Vilnius University.
15. Kanoksilapatham, Budsaba. 2007. Rhetorical moves in biochemistry research articles. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 73–119.
16. Linkevičienė, Nijolė, Šinkūnienė, Jolanta. 2012. Asmeniniai įvardžiai mokslo kalboje. *Kalbotyra*. 64 (3), 78–102.

17. Mur-Dueñas, Pilar, Šinkūnienė, Jolanta. 2016. Self-reference in research articles across Europe and Asia: A review of studies. *Brno Studies in English*. 42 (1), 71–92.
18. Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing*. New York: Continuum.
19. Ruiying, Yang, Allison, Desmond. 2003. Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes.* 22(4), 365–385.
20. Salaz, Danica. 2014. *Lexical Bundles in Native and Non-native Scientific Writing*. Amsterdam / Philadelphia: John Benjamins Publishing Company.
21. Smetona, Antanas, Usonienė, Aurelija. 2012. Autoriaus pozicijos adverbialai ir adverbializacija lietuvių mokslo kalboje. *Kalbotyra*. 64 (3), 124–139.
22. Stubbs, Michael. 1986. Language development, lexical competence and nuclear vocabulary. *Educational Linguistics*. Stubbs, Michael (ed.). Oxford and New York: Blackwell, 98–115.
23. Swales, John. 1981. *Aspects of Article Introductions*. Birmingham: The University of Aston, Language Studies Unit.
24. Šinkūnienė, Jolanta. 2010. Autoriaus pozicijos raiška asmeniniais įvardžiais rašytiniame akademiniame diskurse. *Filologija*. 15, 124–141.
25. Šinkūnienė, Jolanta. 2011. *Autoriaus pozicijos švelninimas rašytiniame moksliniame diskurse: gretinamasis tyrimas*. PhD dissertation. Vilnius: Vilniaus universitetas.
26. Šinkūnienė, Jolanta, Van Olmen, Daniël. 2012. Modal verbs of necessity in academic English, Dutch and Lithuanian: Epistemicity and / or evidentiality? *Darbai ir Dienos.* 58, 153–181.
27. Šinkūnienė, Jolanta. 2014. *Lietuviškojo humanitarinių ir socialinių mokslų diskurso ypatybės*. Vilnius: Vilniaus universiteto leidykla.
28. Šinkūnienė, Jolanta. 2015. Neepisteminis modalumas lietuvių ir anglų mokslo kalboje: kiekybiniai ir kokybiniai vartosenos ypatumai. *Kalbotyra*. 67, 131–154.
29. Volungevičienė, Skaistė. 2018. Humanitarinių ir technologijos mokslų studentų darbų fraziškumas. *Lietuvių kalba*. 12, 38–54.
30. Williams, Ian A. 1999. Results sections of medical research articles: Analysis of rhetorical categories for pedagogical purposes. *English for Specific Purposes*. 18(4), 347–366.

Kopsavilkums

Pētījumā mēģināts izpētīt lietuviešu akadēmiskās valodas vārdu un frāžu biežumu, struktūru un semantiku, izmantojot korpuslingvistikas metodes. Pētījuma objekts ir lietuviešu akadēmiskās valodas leksika zinātniskos rakstos. Mērķa sasniegšanai izvirzīti šādi uzdevumi: 1) izmantojot pēdējās desmitgades dažādu nozaru zinātniskos tekstus, izveidot tekstu korpusu; 2) izmantojot statistiskās metodes, noteikt biežākos akadēmiskā diskursa vārdus un frāzes (kolokācijas); 3) izpētīt akadēmiskās leksikas (vārdu un frāžu) sadalījumu un funkcijas tekstā, variantumu, semantisko un pragmatisko vērtību. Šo pētījumu veido vairākas daļas: vispirms skaidroti pamattermini un darba teorētiskā pieeja, sniegti pētījuma rezultāti, nobeigumā formulēti secinājumi.

Atslēgvārdi: akadēmiskā leksika; kolokācija; vārdu savienojumi; lietuviešu valoda; semantiskā (pragmatiskā) funkcija; tekstu korpuss.