BILINGUAL PUBLISHING GROUP

Forum for Linguistic Studies

https://journals.bilpubgroup.com/index.php/fls

ARTICLE

Corpus-Based Analysis of Three-Word Lexical Bundles in Kazakh

Gulnar Adebietkyzy Sarseke [1*] **, *Aigerim Yerzhanovna Khopur*** [1] **, *Bibigul Yersainovna Akmagambetova*** [2] **, *Samal Yerbolovna Kaliyeva*** [3] **, *Amangul Muratkyzy Adilbek*** [4]

[1] *Kazakh Linguistics Department, L.N. Gumilyov Eurasian National University, Astana 000001, Kazakhstan*

[2] *Kazakh Language and Literature Department, M. Kozybayev North Kazakhstan University, Petropavlovsk 640000, Kazakhstan*

[3] *Foundation Department, L.N. Gumilyov Eurasian National University, Astana 000001, Kazakhstan*

[4] *Department of Practical Kazakh Language, L.N. Gumilyov Eurasian National University, Astana 000001, Kazakhstan*

ABSTRACT

Teaching Kazakh as a second language for academic purposes presents various challenges, including a lack of suitable teaching materials. There are currently no textbooks or syllabi that provide authentic and academic specific language for learners of Kazakh. One way to address this gap is through corpus-based studies, which help identify frequently used lexical bundles. Lexical bundles are recurrent word combinations that appear frequently in academic discourse, playing a crucial role in achieving fluency and mastering academic language conventions. Understanding these bundles can help learners develop language skills in academic Kazakh. This study identifies and analyzes the 100 most frequent three-word lexical bundles in academic Kazakh, focusing on their functional roles. The research is based on a corpus of 600 academic texts written by native speakers of Kazakh across six disciplines: 'Archeology and Ethnology', 'Translation Studies', 'Theology', 'Philology', 'Philosophy', and 'Oriental Studies'. The obtained lexical bundles are categorized into three main functional types: referential expressions, stance expressions, and discourse organizers. The findings reveal distinct patterns in the use of lexical bundles in academic Kazakh, highlighting how authors structure arguments, establish textual cohesion, and express their stance. This study contributes to the understanding of main lexical bundles in academic Kazakh, offering

*CORRESPONDING AUTHOR:

Gulnar Adebietkyzy Sarseke, Kazakh Linguistics Department, L.N. Gumilyov Eurasian National University, Astana 000001, Kazakhstan; Email: sarseke_ga@enu.kz

ARTICLE INFO

Received: 9 August 2025 | Revised: 1 September 2025 | Accepted: 9 September 2025 | Published Online: 27 October 2025
DOI: https://doi.org/10.30564/fls.v7i11.11567

CITATION

Sarseke, G.A., Khopur, A.Y., Akmagambetova, B.Y., et al., 2025. Corpus-Based Analysis of Three-Word Lexical Bundles in Kazakh. Forum for Linguistic Studies. 7(11): 1195–1210. DOI: https://doi.org/10.30564/fls.v7i11.11567

valuable insights for language instructors, curriculum developers, and researchers. The results can inform the development of academic writing resources, support syllabus design, and enhance instructional approaches in teaching Kazakh as a second language for academic purposes.

Keywords: Lexical Bundles; Academic Kazakh; Second Language; Language Teaching; Functional Category; Structural Type

1. Introduction

Numerous linguistic studies have focused on formulaic expressions since the second half of the 20th century, and this interest continues today[1–4]. Formulaic expressions have distinctive features in describing language registers. Among the range of formulaic expressions, lexical bundles have emerged as a prominent concept frequently utilised to characterise academic registers[5]. Lexical bundles distinguish between novice and expert use in both spoken and written contexts. They are a crucial aspect of fluency and ubiquitous in academic language use. It is widely asserted that experienced writers frequently use lexical bundles to convey specific meanings, whereas novice writers use them less frequently[6–8]. Lexical bundles are statistically the most common recurrent word sequences in academic texts compared to other registers[9]. These become apparent using corpus analysis software that extracts multi-word units with specified frequency and distribution criteria.

Corpus linguistic studies have contributed to the understanding of the distinctive linguistic characteristics of academic discourses. One of the central focus of the corpus linguistics is the notion 'lexical bundle'. The term 'lexical bundle' was first introduced in the Longman Grammar of Spoken and Written English[9]. Nevertheless, the notion of a lexical bundle can be traced back to Salem[10]. He conducted research on the analysis of a corpus of French government documents and texts. Altenberg was the first to examine recurrent word combinations in English using the London-Lund Corpus[11]. He developed a methodology for identifying recurrent word combinations that are specified by frequency and used both functional and grammatical analysis to classify them. Later Biber, Johansson, Leech, Conrad and Finegan extended the research that employs the corpus-based approach for the study of recurrent word sequences by comparing the most prevalent multi-word combinations in spoken and written English registers based on the data of the Longman Spoken and Written English Corpus (the LSWE Corpus)[9]. This approach has since been applied in subsequent studies that investigate lexical bundles in various registers in different languages[12–19].

Despite advancements in research of lexical bundles in English and other languages, there remains a lack of corpus-based studies in the Kazakh language, especially academic Kazakh. This study is pioneering work in the investigation of lexical bundles in Kazakh. No scientific research has been done on lexical bundles in Kazakh in any register or genre.

The field of academic Kazakh being a part of teaching Kazakh as a second language is not yet sufficiently developed as an independent discipline. Language for academic purposes is mostly a specific vocabulary which consists of words and phrases frequently used in academic texts. For the purpose of defining and creating the scope of necessary words and phrases to learn and teach language for academic purposes gains most tools and methods from corpora. According to Cotos, teaching language for academic purposes can benefit from specialized academic corpora as they contain the language of registers and genres which are of interest to the learners of the language for specific purposes[20]. Biber[16] and Hyland[21] implementing a corpus-based approach point out that academic language mostly consists of multi-word expressions or lexical bundles with an invariable set of their components.

In this research we aim to address the following research questions:

1. What are 100 most frequent three-word lexical bundles of the academic Kazakh language?
2. What are functional features of three-word lexical bundles in the academic Kazakh language?
3. What are structural types of three-word lexical bundles in the academic Kazakh?

Our study introduces a corpus-based approach for defining the most frequent three-word lexical bundles of academic

Kazakh and analyzing their functional categorization and structural composition in academic discourse. This methodology helps to determine the most frequent lexical bundles with maximum possible accuracy. Therefore, using the most frequent lexical bundles provided by the corpus we can be sure that we as teachers and educators teach a useful and relatable language to our learners. Furthermore, functional categorisation of the lexical bundles enables identification of their roles in academic discourse.

2. Literature Review

2.1. Definition and Classification of Lexical Bundles

The term 'lexical bundle' is defined in the Longman Grammar of Spoken and Written English as 'recurrent expressions, regardless of their idiomaticity, and regardless of their structural status' [9]. Lexical bundles are groups of three or more words that frequently recur in each register or genre. Cortes describes them as 'formulaic sequences strings of continuous or discontinuous words' or 'formulaic expressions' [22]. According to Vespignani et al., multiword expressions that frequently recurred in everyday language are defined as lexical chunks or lexical bundles and have a fixed nature [23]. Lexical bundles consist of three and more words. Some shorter bundles can be included into longer ones. It should also be mentioned that lexical bundles in academic texts do not represent a complete structural unit and are mostly used as important building blocks in a discourse.

There are different opinions concerning the classification of the lexical bundles. Commonly there are functional and structural classifications of the lexical bundles. Sidtis points out the problem of ambiguity for classifying some lexical bundles [24]. In terms of the function Biber, Conrad and Cortes define three main categories of the lexical bundles: 1. stance bundles that express attitude or assessment, 2. discourse organizers show connections between prior and coming discourse, 3. referential bundles indicate direct reference to physical or abstract entities [25]. This classification was applied in research by Ädel and Erman [26] and Wright [27]. However, Hyland refers to more research-focused genres, noting that Biber's taxonomy is a much broader corpus of spoken and written texts and, along with academic texts, covers a broader context [21]. According to Hyland's taxonomy, lexical bundles are classified into three main categories: research-oriented, text-oriented and participant-oriented. Hyland's functional classification was implemented in the study of Yin and Li [28]. Chen and Baker [29] also distinguish three major categories of the lexical bundles based on the taxonomy of Biber [16]. They are referential bundles, stance bundles, and discourse organizers. In this study we identify lexical bundles and analyse their functional features using categories clarified by Chen and Baker [29]. The subcategories of Chen and Baker are more complete and have finer distinctions [26].

2.2. Why are Lexical Bundles Important and Useful in Teaching a Second Language for Academic Purposes?

Lexical bundles are of great interest and importance in the field of second language learning [30,31]. They can help the learners in gaining mastery of the language [32]. Analyzing second language acquisition studies, we notice that native speakers make extensive use of lexical bundles and collocations [26]. In addition, the degree of the second language proficiency significantly depends on the proportion of the lexical bundles used in speech. Fillmore, Kempler and Wang note that formulaic expressions that are lexical bundles are 'memorized' rather than 'generated' [32]. Biber et al. claim that lexical bundles serve as a tool for developing fluency in spoken and written language. Therefore, we see the extreme necessity to teach and learn lexical bundles in second language [9].

Proficiency in academic language also includes the capacity to use lexical bundles. There are applications of lexical bundles as pedagogical tools for teaching articles to L2 English learners of different proficiencies [33]. Studies highlight that second language learners benefit from using lexical bundles in writing essays thus obtaining higher grades [24]. Lexical bundles have significant discourse functions, as they help structure ideas, establish temporal and spatial references, quantify information at multiple levels, express the speaker's stance, and convey implicit assumptions about the interlocutor [29].

Also it should be noted that learning and using lexical bundles (multiword phrases and chunks) are helpful in terms of memory and cognitive load. Highlighting the importance of multi-word phrases in second language learning

in a series of studies, Arnon and Snider conclude that more frequent phrases are processed faster[34]. Using and processing chunks reduce the load on working memory through the process of retrieving from long-term memory as a whole rather than individual elements of the chunk. Ellis states that our perceptional system is just tailored to percieve higher-frequency words and word combinations[35].

2.3. Three-Word Bundles in Kazakh

Kazakh belongs to the northwestern, or Kipchak, branch of the Turkic language family. Kazakh is an agglutinative language with suffixing morphology, sound harmony, and a head-final constituent order[36]. A variety of endings and suffixes are used in the Kazakh language. Kazakh shares all the characteristics of Turkic languages, including synharmonism, affix systems, and sentence structure.

In Kazakh, sentences are structured with a head-final, subject–object–verb order. Personal pronouns that indicate the subject do not always appear in a sentence. Kazakh allows pro-drop, which is a feature that allows subjects to be dropped when they are pragmatically or grammatically inferable[37]. If the subject is omitted, the verb suffix can be used to infer the person. For example, in the sentence *Балалармен ойнап отыр**мын** 'Balalarmen oinap otyr**myn**' ('I am playing with children')*, the subject is omitted and identified by the verb suffix -*мын 'myn'*, which indicates the first-person singular form.

Example: Балалармен ойнап отырмын. (I am playing with children.)

in Latin scripts	Balalarmen	oinap otyrmyn.
Literal translation	with children	I am playing
Grammatical function	object	verb

This syntactic feature is evident in the formation of lexical bundles in the Kazakh language. Many epistemic stance bundles in English are formed using personal pronouns, such as *I was going, I thought that, and if you want to*[25]. This type of structure is rarely found in Kazakh. Stance bundles in Kazakh are formed primarily from verb forms, without the involvement of personal pronouns. For example, *аман өткім келеді 'atap ótkim keledi' ('I would like to point out'), ерекше назар аудартамыз 'erekshe nazar audartamyz' ('we would like to draw special attention to'), ұсыныс ретінде айтамыз 'usynys retinde aitamyz' ('we would like to make a suggestion'),* etc. This language-specific feature is demonstrated by the fact that, of the 100 lexical bundles analysed in our study, only one includes the personal pronoun in its structure.

Another characteristic of the Kazakh language is the active involvement of verbs and auxiliary verbs in the formation of lexical bundles. Verb groups play an important role in the creation of lexical bundles in the Kazakh language (see the following tables). In Kazakh, lexical bundles that correspond to the English bundles structure *'that'* are mostly formed using different verb forms. Lexical bundles are often formed using participles (1), auxiliary verbs (2), and infinitives (3). For instance: (1) *Сондықтан олардың салыстырмалы түрде сирек **кездесетіні** таңқаларлық емес. – It is thus no surprise **that** these are relatively rare.* (2) *Мен бұл [оның] соңғы Олимпиадасы емес **деп** ойлағым келеді. – I would like to think **that** this is not [his] last Olympics.* (3) *Ақпараттық мақсаттарын ескере отырып, жаңалықтар мен академиялық прозаның басқа регистрлерге қарағанда есептік сандарды жиі **қолдануы** таңқаларлық емес. – Given their informational purposes, it is not surprising **that** news and academic prose use cardinals more frequently than the other registers.*

Of the auxiliary verbs, *де 'de' ('say')* was the most frequently encountered in the formation of lexical bundles in Kazakh academic texts. 'De' is an auxiliary verb used with various forms, such as *деді 'dedi' ('have said'), деген 'degen' ('said'), десе 'dese' ('if say'), демек 'demek' ('will say')* etc., in academic texts. They perform a connecting function in the discourse. In academic discourse, lexical bundles that include auxiliary verbs form one type of structure. For example, *деген атпен белгілі 'degen atpen belgili' ('known as'), деген сұрақ қояды 'degen suraq qoiady' ('asks the question'), деген қорытындыға келеді 'degen qorytyndyga keledi' ('comes to the conclusion'), деуге де болады 'deuge de bolady' ('it can be said').*

Although previous studies of lexical bundles in English have provided valuable insights into their structural and functional features, the question of whether these fea-

tures are universal across languages and contexts still requires in-depth investigation. Given that Kazakh belongs to a completely separate language family with different grammatical patterns and functional priorities, we assume that there may be more distinctive patterns of lexical bundles in Kazakh than those covered in the study. These patterns will be thoroughly studied in future research. The purpose of this study was to describe and illustrate the structure and major functions of lexical bundles in the Kazakh language and to compare the general patterns of their use in academic discourse with those found primarily in English. As a result, the study of lexical bundles in academic discourse in the Kazakh language will enrich the discourse on the structural and functional characteristics of lexical bundles in different languages.

3. Materials and Methods

The material of this study is a newly created corpus of academic written works of native speakers of Kazakh. The corpus includes 600 academic texts, 100 from each discipline as 'Archaeology and Ethnology', 'Translation Studies', 'Theology', 'Philology', 'Philosophy' and 'Oriental Studies'. These academic works include dissertations, textbooks, monographs, articles, abstracts, and conference papers. We created a corpus of academic written Kazakh. The total amount of tokens (words) is 3,519,626.

For this study we used the software AntConc (Version 4.3.1)[38]. AntConc is a corpus analysis toolkit designed by Laurence Anthony for classroom use[39]. It is a free corpus analysis program that reads plain text files and allows the user to find word patterns such as n-grams (3-, 4-, and 5-word sequences), concordances, and collocations. Because of its powerful multifunctional corpus-analysis tools and free use, AntConc is widely used in corpus linguistics, as well as other areas of linguistics and linguodidactics. In AntConc, multi-word sequences can be examined in two ways. They can be studied using the Word Clusters tool. An alternative way to search for multi-word sequences is to find lexical bundles, which are equivalent to n-grams where n usually varies between two and five words. AntConc includes the option to search for lexical bundles in the Word Clusters Tool[39].

As a first step we gathered 600 text files and converted each document into a text format. Then we removed all non-linguistic parts, such as titles, tables, references, and figure captions, that were not used for the count. Then we uploaded all 600 text files to AntConc software and obtained a target corpus of 600 files containing 3,519,626 tokens and consisting of different theses, research articles, books and coursebooks on disciplines mentioned above.

To generate a three-word bundle list using the built-in function N-Gram size 3 to get 500 hits to be refined and reduced to a number of 100 lexical bundles. Some research eliminated narrow discipline-specific bundles[40]. Therefore, during the process of refinement we also eliminated such words, as well as proper names and bundles containing some specific characters as brackets. In addition, there are some criteria for lexical bundles to be taken into account. These criteria concern cut-off frequency and dispersion. As noted by Biber et al. lexical bundles should recur in a register at least ten times per million words[9]. As well these occurences must be at least in five different texts of the register. Our register contains 3.5 million words, so in our study this occrurences should be more than 35 and take place in more than 18 texts. For this reason we eliminate those bundles that do not meet set criteria. So, for lexical bundles, the Clusters/N-grams tool extracted 3-word sequences with minimum frequency ≥35 and dispersion across ≥18 files. Word and keyword lists were generated. All settings and exports are archived for reproducibility. As Kazakh is an agglutinative language, whereby words can have several endings, we refined the phrases by removing endings that do not affect the meaning and putting the nouns in their initial form.

The retrieved and refined lexical bundles were manually categorized by functions implementing the description used by Chen and Baker[29] and also used by Ädel and Erman[26]. The main categories are:

Referential expressions are characterised by their functions in attribute specification.

- *Framing bundles* specify a given attribute or condition: *in the context of, the nature of the*
- *Quantifying bundles* relate to anything potentially measurable, suzh as size, number or amount: *a wide range of, in a number of*
- *Place/time/text-deictic bundles*: *are shown in figure, in the present study*

Stance expressions are frequently employed to articulate a writer's evaluation or attitude, the writer's judgement regarding the ability to execute an action.

- *Epistemic: are more likely to, may be due to*
- *Obligatory/directive: it is necessary to, that need to be, it has to be*
- *Ability: it is difficult to, to be able to*

Discourse organizers are employed for the purpose of text structuring.

- *Topic introduction*: *essay is going to, last but not least, in this essay*
- *Topic elaboration: in more detail in, on the other hand, can be used to*
- *Inferential: as a result of, in view of the, this is due to*
- *Identification/focusing: one of the most, there would be no, we can see that*

4. Results

4.1. Most Frequent Lexical Bundles in the Academic Kazakh

We have identified the most frequent lexical bundles in academic Kazakh based on the corpus we have created. **Table 1** shows a list of 100 most frequent lexical bundles refined according the criteria concerning cut-off frequency and dispersion.

As is seen in **Table 1** all the bundles meet the criteria of cut-off frequency, which is 47 and higher, the dispersion that is range is 20 and higher.

4.2. Functional Categories of the Identified Lexical Bundles

Table 2 shows the list of lexical bundles according to functional categories and sub-categories.

Table 1. 100 most frequent lexical bundles in the academic Kazakh.

No	Type	Translation into English	Rank	Freq	Range	Norm Freq	Norm Range
1	және т б	and so on/et cetera	1	1671	179	474.928	0.299
2	қандай да бір	any/some	4	464	119	131.877	0.199
3	және басқа да	and others	6	352	119	100.045	0.199
4	б з д	B.C. (Before Christ)	8	272	39	77.307	0.065
5	әлі күнге дейін	up to the present time	9	253	88	71.907	0.147
6	атап өтуге болады	it can be noted	16	178	52	50.591	0.087
7	б з б	before our era	19	164	18	46.612	0.030
8	уақыт өте келе	over time	21	156	83	44.338	0.139
9	т с с	and so on	23	153	56	43.485	0.093
10	былай деп жазады	writes as follows	24	147	49	41.780	0.082
11	және тағы басқа	and so on	25	146	45	41.496	0.075
12	бірі болып табылады	is one of	27	135	62	38.369	0.104
13	деп айтуға болады	it can be said that	30	131	66	37.233	0.110
14	деп айта аламыз	we can say that	31	128	42	36.380	0.070
15	негізге ала отырып	on the basis of	32	127	62	36.096	0.104
16	бүгінгі күнге дейін	to this day	35	123	64	34.959	0.107
17	осы тұрғыдан алғанда	from this perspective	36	122	42	34.675	0.070
18	қайта қалпына келтіру	to restore	39	115	24	32.685	0.040
19	айта кету керек	it should be noted	40	114	45	32.401	0.075
20	әлі де болса	still	43	113	48	32.117	0.080
21	осы уақытқа дейін	up to now	45	111	62	31.548	0.104
22	атап өткен жөн	it is worth noting	47	109	51	30.980	0.085
23	маңызды рөл атқарады	plays an important role	49	107	56	30.411	0.093
24	жылы жарық көрген	was published in (year)	51	105	41	29.843	0.068
25	белгілі бір дәрежеде	to some extent	53	103	49	29.274	0.082
26	күні бүгінге дейін	to this day	53	103	54	29.274	0.090
27	ұзақ уақыт бойы	for a long time	56	102	59	28.990	0.098
28	болуы мүмкін емес	is impossible	57	101	46	28.706	0.077
29	ішкі және сыртқы	internal and external	59	98	52	27.853	0.087
30	біз қарастырып отырған	which we are considering	62	94	26	26.716	0.043
31	хх ғасырдың басында	at the beginning of the 20th century	66	91	39	25.864	0.065
32	материалдық және рухани	material and spiritual	68	90	42	25.580	0.070
33	бұл өз кезегінде	this in turn	72	88	56	25.011	0.093
34	болып қала береді	continues to be	77	85	51	24.159	0.085
38	болуы да мүмкін	may also be	85	79	40	22.453	0.067
39	болуы мүмкін деген	said it was possible	85	79	26	22.453	0.043

Table 1. *Cont.*

No	Type	Translation into English	Rank	Freq	Range	Norm Freq	Norm Range
40	деп атап көрсетеді	points out that	85	79	25	22.453	0.042
41	деген пікір айтады	states the view that	92	76	23	21.601	0.038
42	сонымен қатар бұл	in addition, this	92	76	45	21.601	0.075
38	болуы да мүмкін	may also be	85	79	40	22.453	0.067
39	болуы мүмкін деген	said it was possible	85	79	26	22.453	0.043
40	деп атап көрсетеді	points out that	85	79	25	22.453	0.042
41	деген пікір айтады	states the view that	92	76	23	21.601	0.038
42	сонымен қатар бұл	in addition, this	92	76	45	21.601	0.075
43	алғашқылардың бірі болып	being among the first	98	74	41	21.032	0.068
44	дамып келе жатқан	developing	98	74	46	21.032	0.077
45	ғасырлар бойы қалыптасқан	formed over centuries	98	74	33	21.032	0.055
46	деген мағынаны білдіреді	conveys the meaning that	101	73	45	20.748	0.075
47	өмір сүріп жатқан	currently living	101	73	43	20.748	0.072
48	осы күнге дейін	up to this day	106	72	46	20.464	0.077
49	бөлігі болып табылады	is a part of	109	71	44	20.179	0.073
50	қамтамасыз ету үшін	in order to ensure	109	71	41	20.179	0.068
51	қол жеткізу үшін	in order to achieve	109	71	46	20.179	0.077
52	ұзақ жылдар бойы	for many years	115	68	34	19.327	0.057
53	теориялық және практикалық	theoretical and practical	116	67	40	19.043	0.067
54	ғасырдың екінші жартысында	in the second half of the century	116	67	39	19.043	0.065
55	қазақ халқы үшін	for the Kazakh people	116	67	29	19.043	0.048
56	саяси және экономикалық	political and economic	122	66	34	18.758	0.057
57	ғасырда өмір сүрген	lived in the (century)	125	65	32	18.474	0.053
58	әлеуметтік және мәдени	social and cultural	129	64	35	18.190	0.058
59	өмір мен өлім	life and death	129	64	26	18.190	0.043
60	арқылы жүзеге асады	is carried out through	132	63	32	17.906	0.053
61	былай деп жазды	wrote as follows	135	62	26	17.622	0.043
62	бір сөзбен айтқанда	in a word	135	62	40	17.622	0.067
63	тәуелсіздік алғаннан кейін	after gaining independence	135	62	31	17.622	0.052
64	анықтауға мүмкіндік береді	allows to determine	140	61	35	17.337	0.058
65	ерекше орын алады	plays a special role	140	61	37	17.337	0.062
66	жүзеге асыру үшін	in order to implement	140	61	35	17.337	0.058
67	сол себепті де	for this reason	140	61	33	17.337	0.055
68	хх ғасырдың басындағы	after beginning of the 20th century	140	61	24	17.337	0.040
69	қалыптасуы мен дамуы	formation and development	149	60	33	17.053	0.055
70	дәлел бола алады	can serve as evidence	152	59	35	16.769	0.058
71	деп атауға болады	can be referred to as	159	58	40	16.485	0.067
72	деректерге сүйене отырып	based on the data	162	57	19	16.200	0.032
73	мақсаты мен міндеттері	objectives and tasks	164	56	44	15.916	0.073
74	маңызды болып табылады	is of importance	164	56	42	15.916	0.070
75	тағы да бір	another	164	56	29	15.916	0.048
76	заман талабына сай	in accordance with the demands of the time	169	55	35	15.632	0.058
77	негізі болып табылады	is the basis of	169	55	32	15.632	0.053
78	жақсылық пен жамандық	good and evil	172	54	25	15.348	0.042
79	бұрын соңды болмаған	unprecedented	177	53	32	15.064	0.053
80	бұқаралық ақпарат құралдары	mass media	177	53	28	15.064	0.047
81	қорытынды жасауға болады	it can be concluded	177	53	41	15.064	0.068
82	деп те атайды	is also referred to as	184	52	37	14.779	0.062
83	кеңістік пен уақыт	space and time	184	52	20	14.779	0.033
84	мыңдаған жылдар бойы	for thousands of years	184	52	29	14.779	0.048
85	ол белгілі бір	it is a certain	184	52	37	14.779	0.062
86	ол өз кезегінде	this in turn	184	52	26	14.779	0.043
87	өзіне ғана тән	peculiar only to it	184	52	34	14.779	0.057
88	және өзге де	as well as others	197	51	30	14.495	0.050
89	кез келген адам	any person	197	51	33	14.495	0.055
90	экономикалық және саяси	economic and political	197	51	29	14.495	0.048
91	жасауға мүмкіндік береді	makes it possible to	210	50	34	14.211	0.057
92	көрінісі болып табылады	is a manifestation of	210	50	32	14.211	0.053
93	т б сияқты	like et cetera	210	50	35	14.211	0.058
94	деп атап өтеді	notes that	223	49	20	13.927	0.033
95	шығыс пен батыс	East and West	223	49	21	13.927	0.035
96	адам мен қоғамның	of the human and society	232	48	24	13.642	0.040
97	түсінуге мүмкіндік береді	enables to understand	232	48	23	13.642	0.038
98	қалай болғанда да	in any case	232	48	33	13.642	0.055
99	білім және ғылым	education and science	245	47	31	13.358	0.052
100	бұл ең алдымен	this is first of all	245	47	26	13.358	0.043

Table 2. Lexical bundles according to functional categories and sub-categories.

No	Type	Translation into English	Function Category	Function Sub-Category
1	және т б	and so on/et cetera	referential	framing
2	қандай да бір	any/some	referential	framing
3	және басқа да	and others	referential	framing
4	б з д	B.C. (Before Christ)	referential	time
5	әлі күнге дейін	up to the present time	referential	time
6	атап өтуге болады	it can be noted	stance	epistemic
7	б з б	before our era	referential	time
8	уақыт өте келе	over time	referential	time
9	т с с	and so on	referential	framing
10	былай деп жазады	writes as follows	discourse organizer	identification/focusing
11	және тағы басқа	and so on	referential	framing
12	бірі болып табылады	is one of	discourse organizer	identification/focusing
13	деп айтуға болады	it can be said that	stance	epistemic
14	деп айта аламыз	we can say that	stance	epistemic
15	негізге ала отырып	on the basis of	discourse organizer	inferential
16	бүгінгі күнге дейін	to this day	referential	framing
17	осы тұрғыдан алғанда	from this perspective	discourse organizer	inferential
18	қайта қалпына келтіру	to restore	discourse organizer	identification/focusing
19	айта кету керек	it should be noted	stance	obligatory/directive
20	әлі де болса	still	discourse organizer	topic elaboration
21	осы уақытқа дейін	up to now	referential	framing
22	атап өткен жөн	it is worth noting	stance	obligatory/directive
23	маңызды рөл атқарады	plays an important role	discourse organizer	identification/focusing
24	жылы жарық көрген	was published in (year)	referential	framing
25	белгілі бір дәрежеде	to some extent	referential	quantifying
26	күні бүгінге дейін	to this day	referential	time
27	ұзақ уақыт бойы	for a long time	referential	time
28	болуы мүмкін емес	is impossible	stance	epistemic
29	ішкі және сыртқы	internal and external	referential	framing
30	біз қарастырып отырған	which we are considering	discourse organizer	identification/focusing
31	хх ғасырдың басында	at the beginning of the 20th century	referential	time
32	материалдық және рухани	material and spiritual	referential	framing
33	бұл өз кезегінде	this in turn	discourse organizer	topic elaboration
34	болып қала береді	continues to be	discourse organizer	identification/focusing
35	белгілі бір деңгейде	to some level	referential	quantifying
36	десе де болады	it can also be said	stance	epistemic
37	өмір сүріп отырған	living at present	discourse organizer	identification/focusing
38	болуы да мүмкін	may also be	stance	epistemic
39	болуы мүмкін деген	said it was possible	discourse organizer	identification/focusing
40	деп атап көрсетеді	points out that	discourse organizer	identification/focusing
41	деген пікір айтады	states the view that	discourse organizer	identification/focusing
42	сонымен қатар бұл	in addition, this	referential	framing
43	алғашқылардың бірі болып	being among the first	discourse organizer	identification/focusing
44	дамып келе жатқан	developing	discourse organizer	identification/focusing
45	ғасырлар бойы қалыптасқан	formed over centuries	discourse organizer	identification/focusing
46	деген мағынаны білдіреді	conveys the meaning that	discourse organizer	identification/focusing
47	өмір сүріп жатқан	currently living	discourse organizer	identification/focusing
48	осы күнге дейін	up to this day	referential	time
49	бөлігі болып табылады	is a part of	discourse organizer	identification/focusing
50	қамтамасыз ету үшін	in order to ensure	discourse organizer	topic elaboration
51	қол жеткізу үшін	in order to achieve	discourse organizer	topic elaboration
52	ұзақ жылдар бойы	for many years	referential	time
53	теориялық және практикалық	theoretical and practical	referential	framing
54	ғасырдың екінші жартысында	in the second half of the century	referential	time
55	қазақ халқы үшін	for the Kazakh people	discourse organizer	topic elaboration
56	саяси және экономикалық	political and economic	referential	framing
57	ғасырда өмір сүрген	lived in the (century)	discourse organizer	identification/focusing
58	әлеуметтік және мәдени	social and cultural	referential	framing
59	өмір мен өлім	life and death	discourse organizer	topic elaboration
60	арқылы жүзеге асады	is carried out through	discourse organizer	topic elaboration
61	былай деп жазды	wrote as follows	discourse organizer	identification/focusing
62	бір сөзбен айтқанда	in a word	discourse organizer	inferential
63	тәуелсіздік алғаннан кейін	after gaining independence	referential	framing
64	анықтауға мүмкіндік береді	allows to determine	stance	epistemic
65	ерекше орын алады	plays a special role	discourse organizer	identification/focusing
66	жүзеге асыру үшін	in order to implement	discourse organizer	topic elaboration
67	сол себепті де	for this reason	discourse organizer	inferential

Table 2. *Cont.*

No	Type	Translation into English	Function Category	Function Sub-Category
68	хх ғасырдың басындағы	after beginning of the 20th century	referential	time
69	қалыптасуы мен дамуы	formation and development	discourse organizer	topic elaboration
70	дәлел бола алады	can serve as evidence	stance	epistemic
71	деп атауға болады	can be referred to as	stance	epistemic
72	деректерге сүйене отырып	based on the data	discourse organizer	inferential
73	мақсаты мен міндеттері	objectives and tasks	discourse organizer	topic elaboration
74	маңызды болып табылады	is of importance	discourse organizer	identification/focusing
75	тағы да бір	another	referential	framing
76	заман талабына сай	in accordance with the demands of the time	referential	framing
77	негізі болып табылады	is the basis of	discourse organizer	identification/focusing
78	жақсылық пен жамандық	good and evil	discourse organizer	topic elaboration
79	бұрын соңды болмаған	unprecedented	referential	framing
80	бұқаралық ақпарат құралдары	mass media	discourse organizer	topic elaboration
81	қорытынды жасауға болады	it can be concluded	stance	epistemic
82	деп те атайды	is also referred to as	discourse organizer	identification/focusing
83	кеңістік пен уақыт	space and time	discourse organizer	topic elaboration
84	мыңдаған жылдар бойы	for thousands of years	referential	time
85	ол белгілі бір	it is a certain	referential	framing
86	ол өз кезегінде	this in turn	discourse organizer	identification/focusing
87	өзіне ғана тән	peculiar only to it	referential	framing
88	және өзге де	as well as others	referential	framing
89	кез келген адам	any person	referential	framing
90	экономикалық және саяси	economic and political	referential	framing
91	жасауға мүмкіндік береді	makes it possible to	stance	epistemic
92	көрінісі болып табылады	is a manifestation of	discourse organizer	identification/focusing
93	т б сияқты	like et cetera	referential	framing
94	деп атап өтеді	notes that	discourse organizer	identification/focusing
95	шығыс пен батыс	East and West	discourse organizer	topic elaboration
96	адам мен қоғам	of the human and society	discourse organizer	topic elaboration
97	түсінуге мүмкіндік береді	enables to understand	stance	epistemic
98	қалай болғанда да	in any case	discourse organizer	inferential
99	білім және ғылым	education and science	discourse organizer	topic elaboration
100	бұл ең алдымен	this is first of all	discourse organizer	topic elaboration

In **Figure 1** it is clearly shown that majority of the lexical bundles are of discourse organizers being 48% and referential category being 38%, while stance bundles are only 14%.

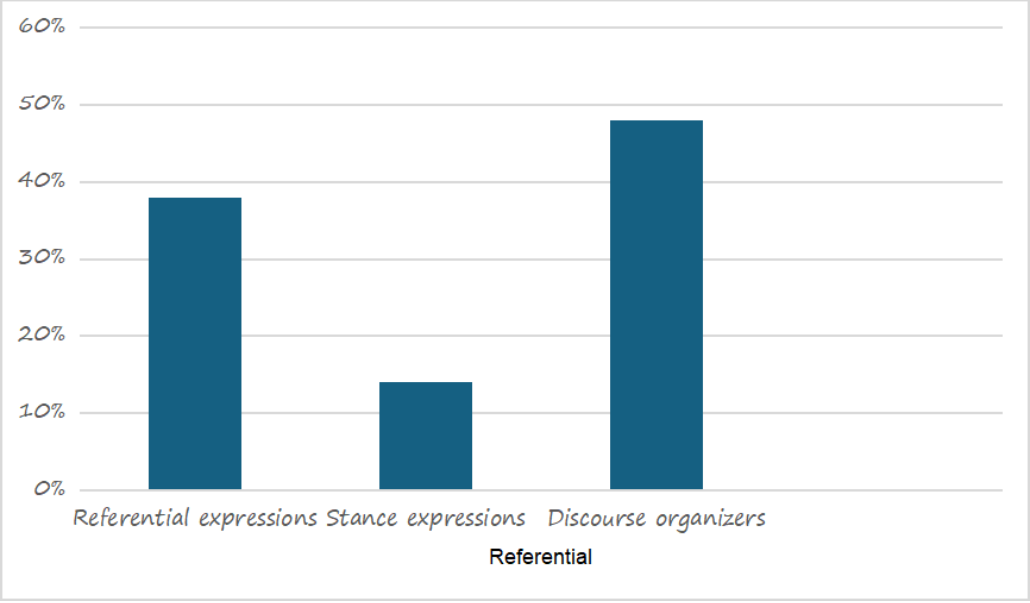

Figure 1. Functional distribution of lexical bundles by categories.

Here are the examples for functional types of lexical bundles from the corpus:

Referential expressions:

Бұл мәселені зерттеу кезінде, осы тараптан Сепир-Уорф теориясына, В. фон Гумбольдтың философиялық-лингвистикалық теориясына, Г.Г. Гадамердің ой-тұжырымдамасына, американдық философ У. Куайнның көзқарасына *және т.б.* жүгінеміз.

Translation into English:

In studying this issue, we will turn to the Sapir-Whorf's theory, the philosophical and linguistic theory of B. von Humboldt, the thought of G.G. Gadamer, the views of the American philosopher W. Quine, *etc.*

Таяу Шығыстағы сопылық мектептері Еуропаға бұрындары ене бастаған және бұл құбылыс *күні бүгінге дейін* жалғасуда.

Translation into English:

Sufi schools from the Middle East began to penetrate Europe long ago, and this phenomenon continues *to this day*.

Stance expressions:

Өз кезегінде, ғалымдар қолданған анықтамаларды сыни талдау одан әрі пәнаралық зерттеулер үшін қажетті негізгі жұмыс аппаратын *анықтауға мүмкіндік береді.*

Translation into English:

In turn, a critical analysis of the definitions used by scientists *allows* us *to identify* the basic working apparatus necessary for further interdisciplinary research.

Қай тіл болмасын өзінің жалғызіліктілігінде атрофияға ұшырайды, яғни семіп қалады *деп айтуға болады.*

Translation into English:

It can be said that any language suffers from atrophy, that is, it withers away, in its solitude.

Discourse organizers:

Ғылыми шығармашылықтың этикалық аспектісі зерттеушілердің мінез-құлқын реттейтін және ғылыми жұмыстың барлық кезеңдеріне әсер ететін нормалар мен принциптерді анықтау арқылы ғылыми шығармашылықта *маңызды рөл атқарады*.

Translation into English:

The ethical aspect of scientific creativity *plays an important role* in scientific creativity by defining norms and principles that regulate the behavior of researchers and affect all stages of scientific work.

Адам әлеммен және басқа адамдармен қарым-қатынасқа түсе бастаған жерде адам сөзсіз, суық, жансыз, өлі объективтілікке тап болады, ол «сыртқы» нәрсені субъективтіліктің «жауына» айналдырады, *бұл өз кезегінде* оның бөтенденуіне әкеледі.

Translation into English:

Where a person begins to enter into relations with the world and other people, a person inevitably encounters a cold, lifeless, dead objectivity that turns the 'external' into the 'enemy' of subjectivity, *which in turn* leads to its alienation.

In the present study we identified three major structural types of lexical bundles, shown in **Table 3**. Type 1 bundles incorporate verb phrase fragments. For example, Types 1a and 1b begin with a noun or adjective followed by a verb phrase (e.g., *қорытынды жасауға болады 'conclusion can be made', негізге ала отырып 'the basis can be taken', негізі болып табылады 'the basis is', маңызды болып табылады 'important is')*. Type 1c begins with a noun in the dative case followed by the verb phrase *мүмкіндік береді* 'allow' (e.g., *түсінуге мүмкіндік береді 'to undestanding allow', жасауға мүмкіндік береді 'to... allow')*. Type 1d begins with the auxiliary verb *де* followed by a verb phrase (*e.g., деп аталған өтеді 'is called', десе де болады 'can be said'*). Type 1e begins with a verb followed by a verb fragment with modality (*e.g., айта кету керек 'noted it should be', аталған өтуге болады 'noted it can be'*). Type 1f begins with an adjective followed by a phrasal compound verb (*e.g.,*

маңызды рөл атқарады 'important plays role', ерекше орын алады 'special place occupies').

Type 2 bundles incorporate noun phrase fragments. For example, Type 2a begins with a pronoun followed by noun phrase components *(e.g., бұл өз кезегінде 'this in turn', осы тұрғыдан алғанда 'this from perspective').* Type 2b begins with the phrase *белгілі бір* 'certain' followed by a noun *(e.g., белгілі бір деңгейде 'certain to an extent', белгілі бір дәрежеде 'certain to an extent').* Type 2c bundles consist of nouns of temporality such as *күн 'day', уақыт 'time', жыл 'year', ғасыр 'century' (e.g., күні бүгінге дейін 'day to this day', уақыт өте келе 'time over', ғасырдың екінші жартысында 'of the century in the second half', ғасырлар бойы қалыптасқан 'over centuries formed', ғасырда өмір сүрген 'in the century lived').*

Type 3 bundles incorporate conjunction fragments. Type 3a begins with a verb phrase followed by the conjunction *үшін* (*e.g.,* қамтамасыз ету үшін 'to provide', қол жеткізу үшін 'to achieve'). Type 3b begins with a conjunction *және* followed by post-modifier fragments (*e.g., және т. б. and so on, және басқа да 'and so on').* Type 3c begins with a conjunction followed by a verb phrase (e.g., арқылы жүзеге асады '*through carried out*', *арқылы ықпал етеді 'by influenced').*

Table 3. Structural types of lexical bundles.

1	**Lexical bundles with verb phrase fragments**
1a	Noun + verb phrase fragment Example bundles: *қорытынды жасауға болады, дәлел бола алады, деректерге сүйене отырып*
1b	Noun + verb phrase fragment *болып табылады* 'is' Example bundles: *көрінісі болып табылады, бөлігі болып табылады, мәселе болып табылады*
1c	Noun + verb phrase with passive verb Example bundles: *түсінуге мүмкіндік береді, жасауға мүмкіндік береді, табуға мүмкіндік береді*
1d	Auxiliary verb *де* + verb phrase fragment Example bundles: *деп атап өтеді, десе де болады, деп атауға болады*
1e	Verb + verb phrase with modal verb Example bundles: *айта кету керек, атап өткен жөн, атап өтуге болады, болып қала береді*
1f	Adjective + verb phrase fragment Example bundles: *маңызды рөл атқарады, ерекше орын алады, маңызды болып табылады*
2	**Lexical bundles with noun phrase fragments**
2a	Pronoun + noun phrase fragments Example bundles: *бұл өз кезегінде, осы тұрғыдан алғанда, ол өз кезегінде*
2b	Noun *phrase of dimensions* Example bundles: *белгілі бір деңгейде, белгілі бір дәрежеде*
2c	Nouns phrase of temporality Example bundles: *күні бүгінге дейін, уақыт өте келе, әлі күнге дейін, ұзақ уақыт бойы, ұзақ жылдар бойы, ғасырдың екінші жартысында, ғасырлар бойы қалыптасқан*
3	**Lexical bundles with conjunction phrase fragments**
3a	Noun phrase + conjunction Example bundles: *қамтамасыз ету үшін, қол жеткізу үшін*
3b	Conjunctions with other post-modifier fragments Example bundles: *және басқа да, және тағы басқа, және өзге де*
3c	Conjunctions with verb phrase fragments Example bundles: *арқылы жүзеге асады, арқылы ықпал етеді*

5. Discussion

In this study we aimed to define the most frequent one hundred three-word lexical bundles of the academic Kazakh language. Also, we tried to meet the criteria set for lexical bundles noted by Biber et al.[9]. As a result, we obtained a cut-off frequency of 47 and higher and the dispersion or range of 20 and higher.

To address the second research question, we made attempt to specify functional categories and sub-categories of the obtained lexical bundles. We have defined the functions, but the distribution turned out to be very complex and ambiguous. This is due to the difference in the structure of lexical bundles of the Kazakh and English languages. In the Kazakh language there are significant number of lexical bundles containing a verb, adjective, participial expressions, and nouns, which was quite interesting and surprising to us. Lexical bundles in English, based on the definitions of categories and sub-categories, and the lexical bundles of Kazakh, appear very different in structure. This fact made the process of functional categorization and sub-categorization quite difficult. Ädel and Erman [26], implementing the functional categorization of Chen and Baker [29], also state that there is some vagueness of criteria to decide which (sub)category a given bundle should belong to, which has led to some inconsistencies in previous studies.

Using the definitions of categories and sub-categories proposed by Chen and Baker [29] and our speculations we categorized lexical bundles containing verbs as discourse organizers, and further to identification. Obtained lexical bundles consisting of adjectives and conjunctions were identified by us as referential category and framing sub-category due to characteristics of attribute specification. However, the lexical bundles containing participial expressions were identified by us as discourse organizers category and identification sub-category because they provide a more complete focus and identification rather than in case of adjectives in referential category and framing sub-category. Those lexical bundles consisting of nouns and conjunctions were attributed by us to discourse organizers categories and topic elaboration sub-category as we reflect that the next information in the discourse will be elaborated around these nouns. We categorized lexical bundles containing adverbial participial expressions as discourse organizers with inferential sub-category as they implement some conclusion or reasoning. We identified the referential categories with time and quantifying sub-category and stance categories with epistemic and obligatory/directive sub-category since these signs are obvious. Despite considerable difficulty we have defined the functions of the obtained lexical bundles with distributions 48%, 38% and 14% for discourse organizers, referential bundles and stance bundles, respectively.

To address the third research question, we have looked at the structural composition of lexical bundles in the academic Kazakh discourse. Verbs, including auxiliary verbs and participles, play a significant part in the construction of lexical bundles in Kazakh. The majority of lexical bundles are formed with verb phrase fragments. The stance bundles beginning with a personal pronoun followed by a verb phrase are rarely seen in the academic discourse in Kazakh. In Kazakh academic discourse, it is quite uncommon to find a lexical bundle that combines a pronoun and a verb phrase to convey the author's position. As an illustration, among the 100 lexical bundles we found, just one—*біз қарастырып отырған* 'we are considering'—contains the pronoun *біз* 'we'. It can be explained by the Kazakh language's grammatical characteristics. A pronoun is not required in a sentence in Kazakh since verbs are the primary means of expressing personality; the personal endings that are appended to the verb identify the speaker or writer. The stance bundle *деп айта аламыз* 'we can say' (№14 in **Table 1**), for instance, identifies the author of the text by using the verb (*айта аламыз* where -мыз indicates personality) without the pronoun ('біз'). It is also worth noting that the pronoun *біз* 'we' is used more frequently than *мен* 'I' in Kazakh academic discourse and that the pronoun 'I' is hardly ever employed in academic writing. In the Kazakh academic setting, using the pronoun 'we' to convey an author's position has become standard writing practice. Even when a single individual is the text's author, scholarly writing in Kazakh tends to use the pronoun "we". This is one of the main characteristics of Kazakh stance bundles. In contrast, the pronoun "I" forms a large number of lexical bundles in English [9].

There is a strong relationship between structural types and functional categories for lexical bundles. This is also mentioned by Biber, Conrad and Cortes [25]. Most stance bundles are composed of verb phrase fragments, while most referential bundles are composed of noun phrase fragments in Kazakh too. Discourse organizers are the only functional category used in all three structural types.

As we noted in the introduction, there are no studies on lexical bundles in Kazakh. As well, there are not many recent studies concerning lexical bundles of a certain language except English. Gong, Le and Buckingham investigate the cross-sectional distribution of four-word lexical bundles across IMRD sections of articles in medicine in English and

conclude that sets of lexical bundles and their functions vary significantly across sections [41]. Samraj investigates lexical bundles in different disciplines written in English and shows that lexical bundles depend on the discipline and methodological settings of extraction (length, frequency, variance), which is why research results may diverge between disciplines and corpora [42]. A study similar to ours in its objectives is research provided by Shirazizadeh and Amirfazlian on the basis of a large corpus of ≈5.7 million words [43]. They investigate forms and functions of 4-word bundles and discuss previous pedagogical applications of lexical bundle research in academic discourse.

Most of the studies on lexical bundles are devoted to their usage by learners of English as a second language. Appel presents a corpus study of L2 essays containing 3–5-word bundles used by students and how this relates to holistic assessment of writing quality; also, statistical relationships between bundle profile and writing scores are shown [44]. Shin and Won investigate parallel corpora of written essays and oral presentations by the same second language authors and state that genre significantly influences the choice of bundles, while oral and written products demonstrate different sets of lexical bundles [45]. Li and Lei compare master's theses by native English and L1 Chinese authors and reveal similarities and differences in bundle frequencies/types and their genre functions in academic writing [46]. In our research we analysed the three most frequent three-word lexical bundles peculiar to the Kazakh language.

In this research we also highlight the necessity of investigating forms and functions of lexical bundles of different length, frequency and variance, and their subsequent application for the learning and teaching of a second language for academic purposes. Puimège emphasises that knowing lexical bundles, i.e., formulaic sequences, helps learners to fulfil a wide range of discourse functions and proposes to acquire them through meaning-focused activities [47]. Different studies have demonstrated a strong association between second language proficiency and fluency and lexical bundle knowledge. Learners who use lexical bundles more frequently in their second language speech tend to be perceived as more proficient and fluent [48]. As well as texts written by the second language learners that incorporate a greater density of lexical bundles tend to earn higher evaluations [49]. All these results indicate that formulaic language, i.e., lexical bundles, is a core component of second language proficiency.

6. Conclusions

This study contributing the necessity of defining lexical bundles in teaching purposes, especially in teaching second language in academic purposes, has identified 100 most frequent three-word lexical bundles of the academic Kazakh language and their functions using software AntConc on the basis of corpus compiled of 600 academic written texts from disciplines as 'Archaeology and Ethnology', 'Translation Studies', 'Theology', 'Philology', 'Philosophy' and 'Oriental studies'. The total amount of tokens is 3,519,626. As demonstrated in this study, 100 most frequent lexical bundles of academic Kazakh language primarily serve as discourse organizers and of referential expressions. These categories are employed to structure ideas and texts, and are characterised by their functions in attribute specification.

We consider that the results presented in this study are of high value. Pedagogically it would be useful in gaining mastery in learning Kazakh for academic purposes. Teachers of Kazakh for academic purposes can include these lexical bundles in teaching syllabuses and materials as a language input. Functional categorization will help the learners use the lexical bundles appropriately for academic purposes for organizing ideas, structuring texts, specifying information and giving the author's stance.

As it was mentioned before academic Kazakh for nonnatives as a new field needed development of different materials, textbooks, and syllabi in large quantities in a short time. As well we suggest future research for identifying the most frequent four and five-word lexical bundles based on the methods implemented in this study. In addition, it is essential to deeper examine lexical bundles with respect to their structures specifically in the Kazakh language. Therefore, this study is an early step in the development of academic Kazakh teaching and learning.

Author Contributions

Conceptualization, G.A.S.; methodology, G.A.S.; software, G.A.S. and A.Y.K.; validation, G.A.S., B.Y.A. and S.Y.K.; formal analysis, G.A.S. and A.Y.K.; investigation, G.A.S., A.M.A. and S.Y.K.; resources, G.A.S.; writing—original draft preparation, G.A.S. and A.Y.K.; writing—

review and editing, G.A.S.; supervision, G.A.S.; project administration, G.A.S.; funding acquisition, G.A.S. All authors have read and agreed to the published version of the manuscript.

Funding

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23488585 'Digital Humanities: Creating a Corpus of Academic Kazakh').

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data utilised in this research are available upon request.

Acknowledgments

Special thanks are extended to Professor Hilary Nesi at the University of Coventry for her insightful feedback on the draft version of the paper and support on the development of the academic corpus in Kazakh. Additionally, we would like to express our gratitude to the anonymous reviewers whose valuable comments enabled us to improve this paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Ellis, N.C., 1996. Sequencing in SLA: Phonological memory, chunking, and points of order. Studies in Second Language Acquisition. 18(1), 91–126. DOI: https://doi.org/10.1017/S0272263100014698

[2] Moon, R., 1997. Vocabulary connections: Multi-word items in English. In: Schmitt, N., McCarthy, M. (eds.). Vocabulary: Description, Acquisition and Pedagogy. Cambridge University Press: Cambridge, UK. pp. 40–63.

[3] Howarth, P., 1998. Phraseology and second language proficiency. Applied Linguistics. 19(1), 24–44. DOI: https://doi.org/10.1093/applin/19.1.24

[4] Wray, A., 2002. Formulaic Language and the Lexicon. Cambridge University Press: Cambridge, UK.

[5] Hyland, K., 2018. Academic lexical bundles. International Journal of Corpus Linguistics. 23(3), 383–407. DOI: https://doi.org/10.1075/ijcl.17080.hyl

[6] Bamberg, B., 1983. What makes a text coherent? College Composition and Communication. 34(4), 417–429. DOI: https://doi.org/10.2307/357898

[7] McCulley, G., 1985. Writing quality, coherence, and cohesion. Research in the Teaching of English. 19(3), 269–282.

[8] Haswell, R., 1991. Gaining Ground in College Writing: Tales of Development and Interpretation. Southern Methodist University Press: Dallas, TX, USA. p. 236.

[9] Biber, D., Johansson, S., Leech, G., et al., 1999. Longman Grammar of Spoken and Written English. Pearson Education: Harlow, UK. p. 1204.

[10] Salem, A., 1987. Practice Repeated Segments. L'Institut National de la Langue Française: Paris, France. p. 334. (in French)

[11] Altenberg, B., 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In: Cowie, A.P. (ed.). Phraseology: Theory, Analysis, and Applications. Oxford University Press: Oxford, UK. DOI: https://doi.org/10.1093/oso/9780198294252.003.0005

[12] Partington, A., Morley, J., 2004. From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In: Lewandowska-Tomaszczyk, B. (ed.). Practical Applications in Language and Computers – PALC 2003. Peter Lang: Frankfurt am Main, Germany. pp. 179–192.

[13] Nesi, H., Basturkmen, H., 2006. Lexical bundles and discourse signaling in academic lectures. International Journal of Corpus Linguistics. 11(3), 283–304. DOI: https://doi.org/10.1075/ijcl.11.3.04nes

[14] Biber, D., Barbieri, F., 2007. Lexical bundles in university spoken and written registers. English for Specific Purposes. 26(3), 263–286. DOI: https://doi.org/10.1016/j.esp.2006.08.003

[15] Cortes, V., 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. Corpora. 3(1). DOI: https://doi.org/10.3366/E1749503 20800006

[16] Biber, D., 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. International Journal of Corpus Linguistics. 14(3), 275–311. DOI: https://doi.org/10.1075/ijcl.14.3.08bib

[17] Li, J., Schmitt, N., 2009. The acquisition of lexical

phrases in academic writing: A longitudinal case study. Journal of Second Language Writing. 18(2), 85–102. DOI: https://doi.org/10.1016/j.jslw.2009.02.001

[18] Lee, J., Chen, S.X., 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. Journal of Second Language Writing. 18(3), 281–296. DOI: https://doi.org/10.1016/j.jslw.2009.05.004

[19] Kim, Y., 2009. Korean lexical bundles in conversation and academic texts. Corpora. 4(2), 135–165. DOI: https://doi.org/10.3366/E1749503209000288

[20] Cotos, E., 2017. 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. John Wiley & Sons: Oxford, UK. pp. 248–264. DOI: https://doi.org/10.1002/9781118914069.ch17

[21] Hyland, K., 2008. As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes. 27(1), 4–21. DOI: https://doi.org/10.1016/j.esp.2007.06.001

[22] Cortes, V., 2024. Lexical bundles in academic writing. Reference Module in Social Sciences. Elsevier: Amsterdam, Netherlands.

[23] Vespignani, F., Canal, P., Molinaro, N., et al., 2010. Predictive mechanisms in idiom comprehension. Journal of Cognitive Neuroscience. 22(8), 1682–1700. DOI: https://doi.org/10.1162/jocn.2009.21293

[24] Sidtis, D., 2021. Foundations of Familiar Language: Formulaic Expressions, Lexical Bundles, and Collocations at Work and Play. John Wiley & Sons: Hoboken, NJ, USA. p. 464.

[25] Biber, D., Conrad, S., Cortes, V., 2004. If you look at …: Lexical bundles in university teaching and textbooks. Applied Linguistics. 25(3), 371–405. DOI: https://doi.org/10.1093/applin/25.3.371

[26] Ädel, A., Erman, B., 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. English for Specific Purposes. 31(2), 81–92. DOI: https://doi.org/10.1016/j.esp.2011.08.004

[27] Wright, H.R., 2019. Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. English for Specific Purposes. 54, 1–14. DOI: https://doi.org/10.1016/j.esp.2018.09.001

[28] Yin, X., Li, S., 2021. Lexical bundles as an intradisciplinary and interdisciplinary mark: A corpus-based study of research articles from business, biology, and applied linguistics. Applied Corpus Linguistics. 1, 100006. DOI: https://doi.org/10.1016/j.acorp.2021.100006

[29] Chen, Y.-H., Baker, P., 2010. Lexical bundles in L1 and L2 academic writing. Language Learning & Technology. 14(2), 30–49. DOI: https://doi.org/10.64152/10125/44213

[30] Pérez-Llantada, C., 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. Journal of English for Academic Purposes. 14, 84–94. DOI: https://doi.org/10.1016/j.jeap.2014.01.002

[31] Salazar, D., 2014. Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching. John Benjamins: Amsterdam, Netherlands; Philadelphia, PA, USA. p. 212.

[32] Fillmore, C.J., Kempler, D., Wang, W.S., 2014. Introduction. In: Fillmore, C.J., Kempler, D., Wang, W.S. (eds.). Individual Differences in Language Ability and Language Behavior. Academic Press: New York, NY, USA. pp. 1–10.

[33] Shin, Y.K., Kim, Y., 2017. Using lexical bundles to teach articles to L2 English learners of different proficiencies. System. 69, 79–91. DOI: https://doi.org/10.1016/j.system.2017.08.002

[34] Arnon, I., Snider, N., 2010. More than words: Frequency effects for multi-word phrases. Journal of Memory and Language. 62(1), 67–82. DOI: https://doi.org/10.1016/j.jml.2009.09.005

[35] Ellis, N.C., 2017. Chunking: The Changing English Language – Psycholinguistic Perspectives. Cambridge University Press: Cambridge, UK. pp. 113–147.

[36] Johanson, L., Csató, É.A., 2015. The Turkic Languages. Routledge: London, UK.

[37] Dotton, Z., Wagner, J.D., 2018. A Grammar of Kazakh. Duke University, Duke Center for Slavic, Eurasian, and East European Studies: Durham, NC, USA. p. 69.

[38] Anthony, L., 2024. AntConc (Version 4.3.1) [computer software]. Waseda University: Tokyo, Japan. Available from: https://www.laurenceanthony.net/software/antconc/

[39] Anthony, L., 2005. AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom. In Proceedings of the 2005 IEEE International Professional Communication Conference, New York, NY, USA, 10–13 July 2005; pp. 729–737. DOI: https://doi.org/10.1109/IPCC.2005.1494244

[40] Pan, F., Reppen, R., Biber, D., 2020. Methodological issues in contrastive lexical bundle research. International Journal of Corpus Linguistics. 25(2), 215–229. DOI: https://doi.org/10.1075/ijcl.19063.pan

[41] Gong, H., Le, T.N.P., Buckingham, L., 2025. Lexical bundles across IMRD-structured Medicine research article sections: A within-register perspective. Journal of English for Academic Purposes. 74, 101487. DOI: https://doi.org/10.1016/j.jeap.2025.101487

[42] Samraj, B., 2024. Disciplinary differences in lexical bundles use: A cautionary tale from methodological variations. Journal of English for Academic Purposes. 70, 101399. DOI: https://doi.org/10.1016/j.jeap.2024.101399

[43] Shirazizadeh, M., Amirfazlian, R., 2021. Lexical bun-

dles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. Journal of English for Academic Purposes. 49, 100946. DOI: https://doi.org/10.1016/j.jeap.2020.100946

[44] Appel, R., 2022. Lexical bundles in L2 English academic texts: Relationships with holistic assessments of writing quality. System. 110, 102899. DOI: https://doi.org/10.1016/j.system.2022.102899

[45] Shin, Y.K., Won, D.O., 2024. To what extent do L2 learners produce genre-appropriate language? A comparative analysis of lexical bundles in argumentative essays and speeches. Journal of English for Academic Purposes. 69, 101389. DOI: https://doi.org/10.1016/j.jeap.2024.101389

[46] Li, Y., Lei, H., 2025. Lexical bundles in L1 and L2 English academic writing: Convergent and divergent usage. SAGE Open. 15(2). DOI: https://doi.org/10.1177/21582440251333850

[47] Puimège, E., 2024. Learning L2 formulaic sequences from meaning-focused activities. ITL – International Journal of Applied Linguistics. 175(2), 163–186. DOI: https://doi.org/10.1075/itl.23014.pui

[48] Stengers, H., Boers, F., Housen, A., et al., 2011. Formulaic sequences and L2 oral proficiency: Does the type of target language influence the association? International Review of Applied Linguistics in Language Teaching. 49(4), 321–343. DOI: https://doi.org/10.1515/iral.2011.017

[49] Ohlrogge, A., 2009. Formulaic expressions in intermediate EFL writing assessment. In: Corrigan, R., Moravcsik, E.A., Ouali, H., et al. (eds.). Formulaic Language. Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations. John Benjamins: Amsterdam, Netherlands; Philadelphia, PA, USA. pp. 375–386.