Pakistan Social Sciences Review
www.pssr.org.pk

RESEARCH PAPER

Formulaic Language in Social Sciences: A Functional Analysis of Lexical Bundles in Native and Non-Native Academic Discourse

Moazzam Ali Malik [1] Hafiz Zeeshan Fazal [2] Hassan Moavia [3]

1. Assistant Professor,Department of English, University of Gujrat, Punjab, Pakistan
2. M. Phil Scholar, Department of English, University of Lahore, Chenab Campus, Gujrat, Punjab, Pakistan
3. M. Phil Scholar, Department of English, University of Lahore, Chenab Campus, Gujrat, Punjab, Pakistan

PAPER INFO	ABSTRACT
Corresponding Author: moazzam.ali@uog.edu.pk	The current study explicates how differentially the native and non-native writers functionally use the lexical bundles in the academic discourse of PhD theses. The corpus data comprised 200 PhD theses produced by non-native Pakistani scholars and the native scholars in the five different disciplines of the social sciences. To conduct the analysis, the study employs a combination of functional taxonomies of the lexical bundles previously proposed and used by Biber and his associates (see, Biber, et al., 2004 & 2003); Biber & Barbieri, 2007). For data analysis, a corpus tool, AntConc 3.3.5, was used to generate and enlist the 4-word units of lexical bundles found in the collected corpus. These identified lexical bundles were, later, categorized into three functional categories as *referential*, *discourse organizing*, and *stance*. The findings of the study reveal that the native and non-native writers make significantly more use of *referential* lexical bundles and a minimum use of *discourse organizing* and *stance* bundles with certain qualitative differences. It is expected that the findings of the study can help EAP scholars in developing teaching materials to assist non-native writers in improving their academic skills

Introduction

The term formulaic sequence is a broader term that refers to the frequently occurring word-units ranging from sentence-stems to the lengthy clauses. So, an example of a short formulaic sequence can be as *"in conclusion"* and that of a longer formulation can be as *"In this section of the chapter"*. Formulaic language is a very common feature of any academic discourse. According to Wray (2002), these expressions are "prefabricated" in the minds of the author who store and retrieve them from their memory (p. 9). These pre-constructed phrases are cognitively

accessible to the scholars who systematically use them as building blocks of the academic discourse. There are many functions that are performed by the formulaic sequences in the academic discourse. For instance, the formulaic sequences give coherence to the text, make it interactive, contextualize the meaning, establish a writer's rapport and mark his/her identity with a particular discourse community (see Siyanova-Chanturia & Martinez, 2015). Nonetheless, formulaic language is a baggy term that, in discourse studies, is employed to refer to the various kinds of word structures generally termed as phraseological Units, lexical bundles, social routines, sentence-stems, idiomatic expressions, collocations, etc (see, Wood, 2010). Lexical bundles, the focus of study, are regarded as the lexical units that "often co-occur in longer sequences" or "show a statistical tendency to co-occur" (Biber et al., 1999). Lexical bundles, according to Biber et al., (1999) are so much recurrent that they cover approximately 21% of written academic discourse.

Academic discourse is a broad term which refers to any form of discourse produced either *by* the academic community or *for* the academic community. A number of registers and text-types produced by the academic community, e.g., assignments, presentations, lectures, study notes, academic speeches, theses, term-papers, etc., fall into the category of academic discourse. Interest in the structure of the academic discourse is not a recent phenomenon. Hyland (2009) explains that since the late 60s, scholars are taking interest in exploring the textual nuances of the academic discourse. In recent times, with its empirical focus on generic and rhetorical features, metadiscursive aspects, rhetorical structures, syntactic patterning, etc., the interest in academic discourse has expanded tremendously (see Hyland, 2000: Hyland, 2009; Swales, 2004; Yang & Allison, 2003). Likewise, studies exploring the academic discourse, for the use of lexical bundles have got their impetus due to the empirical contributions of textual linguists and the practical orientations of ESP scholars.

Most of the research on the use of lexical bundles focuses on their structural configurations or the functional use. In structure, arguably, lexical bundles are not considered having well-defined structural units. It is more appropriate to view them as bearing certain structural associations and phrasal combinations than the fixed lexical combinations. For instance, lexical bundles like "*a large number of*" and "*in the conclusion of*" are, respectively, the noun and prepositional phrases that may come with different constituting components but their structural (phrasal) properties remain relatively fixed. The most widely acknowledged framework for the functional use of lexical bundles has been proposed by Biber et al., (1999, pp. 1014-1024), in which lexical bundles are divided into four different categories based on prepositional, nominal, verbal and clausal structures.

Rationale of Study

Over the last two decades, research in lexical bundles has evolved into two different strands. Scholars like Cortes (2004), Hyland (2008), Kashiha & Heng, (2013) have explored how academic discourses have different disciplinary

orientations, primarily attributable to their subject matter and the empirical inclination. These scholars have demonstrated that there exists a qualitative and quantitative difference in the use of the lexical bundles across the knowledge territories of the disciplines (see Hyland, 2008; Strunkyt & Jurkūnait, 2008). Their research has inspired many ESP and EAP scholars who are convinced that the appropriate use of multi-word expressions like lexical bundles, idiomatic expressions, collocations, etc, can improve the language competence of the learners in specific domains of the second language use. Nonetheless, there is a dearth of research that aims to explicate how the lexical bundles are used to functionally contextualize meaning, interact with the reader, build textual coherence, and mark disciplinary boundaries in the academic discourse. Second, there are also studies (e.g., Adel and Erman, 2012; Hernandez 2013; Crowley, 1991) that focus on the cross-linguistic differences in the use of functional lexical bundles in different types of academic discourses. The findings of these studies are contradicting each other in explaining the cross-linguistic use of the functional categories of lexical bundles. One of the possible reasons for the conflicting results of the previous studies could be the variation in the corpus-texts (e.g. spoken and written) that they use for the comparisons. As different texts are produced under the influence of various social, cultural and contextual contexts, so these extra-textual influences affect the structural and functional configuration of the text-design. We believe that to truly understand the differences in the use of academic language among native and non-native writers, ideally, a study must use a compatible corpus for the comparison with the only single cross-linguistic difference of the mother tongue.

This current study, from a cross-linguistic perspective, unearths the functional aspects of 4-word lexical bundles in the academic discourse of social sciences. The objective of this endevour is to document how the formulaic language of lexical bundles functionally structure the the PhD theses produced by the English native and the Pakistani non-native writers. So, the research objectives of the current study are:

1. What is the functional distribution of the lexical bundles in the PhD theses written by the native and non-native writers in the domain of social sciences?

2. How does the use of *stance*, *discourse organizing*, and *referential* lexical bundles mark the communicative functions of interaction, textuality, and refrentiality in the academic discourse of social sciences?

Literature Review

Biber and his associates (e.g., Biber et al, 1999; 2004) conducted some influential studies at the turn of this century. Through their influential work these scholars have proposed structural and functional taxonomies of lexical bundles have given rise to a bulk of literature that investigates lexical bundles in academic discourse. In their study, Biber and Barbieri (2007) have elaborated upon how the

lexical bundles are different from other formulaic sequences like collocations, idiomatic expressions, phraseological units, etc. They explain that lexical units do not carry idiomatic meaning and have little perceptual salience. Furthermore, lexical bundles are not the typical linguistic structures; rather they conjoin certain structural units (clause/phrase) of a text.

Different studies like Biber (2006), Hyland (2008), Biber and Barbieri (2007) have proposed different taxonomies explicating the structural configuration of lexical bundles. These taxonomies are quite similar as they identify four major types of structural lexical bundles. These structural units are generally regarded as composed of a) verb phrase elements, b) noun phrase elements, c) prepositional phrase elements, and d) clausal structures. Each of these four types of structural units has multiple manifestations and types in the form of clause fragments and pragmatic clues. As the objective of the current study is to explore the functional distribution of the lexical bundles, so in the following we will discuss them in detail.

For this study, three major studies by Biber and his associates (e.g., Biber et al 2003& 2004; Biber & Barbieri, 2007) provided a functional taxonomy of lexical bundles According to this taxonomy, lexical bundles perform three major functions of taking stance, organizing discourse and referring to the textual and extra textual entities. These three categories and their sub-categories are given in table 1.

Table 1
Functional Taxonomy of Lexical Bundles

S. No	Stance Bundles	Discourse Organizers	Referential Bundles
1	Epistemic Stance Bundles	Topic introduction bundles	Framing
2	Desire Bundles	Topic elaboration, clarification bundles	Quantifying
3	Obligation (directive) bundles	Identification/focus bundles	Time/place/text-deixis bundles
4	Intension/prediction bundles		
5	Ability bundles		

This table gives the functional taxonomy of lexical bundles with three major functional categories along with their eleven sub-categories.

Stance Bundles

These lexical bundles are used to expresses the attitudes of the speaker or the writer about the propositional content of the statements. This major category has five other sub-categories as well.

a. Epistemic stance bundles

These lexical bundles express writer/speaker's knowledge status on the propositional content of the statement. For example, *It is quite probable that, I am certain that.* Epistemic *stance* bundles can have two types in structure,i.e., personal (involving personal pronouns) or impersonal (without personal pronouns).

b. Desire bundles

These bundles express writer/spearker's wish or desire for something, e.g., *I would prefer to, I wish to.*

c. Obligation/ directive bundles

These lexical bundles impose some obligation on the addressee or they are used to direct someone. Generally, a second person pronoun is used in these bundles, e.g., *you will have to, I might ask you to.*

d. *Intention/Prediction bundles*

These lexical bundles express the writer/speaker's intention, e.g., *I would like to, we intend to*

e. Ability bundles

These *stance* bundles express the ability of the writer/speaker to do something. Generally, this ability is expressed by using modals. There examples are---- *to be able to, to manage the*

Discourse Organizers

These are the cohesive lexical bundles as they connect t different parts of the discourse. They can be used to take turns, introduce a topic, and draw attention of the addressee. They have three sub-categories as given below.

a. Topic introduction/focus

These lexical bundles mark the beginning of a new topic, e.g., *I would like to explain, I have to say*

b. Topic elaboration/clarification

These bundles give clarification or add additional information about what is already known or has been said, e.g., *as you know, in the same way*

c. Identification/focus bundles

These lexical bundles organize a discourse. Their examples are *as has been said* and *one of the most.*

Referential bundles

Referential bundles are used to make references. These references can make textual (endophoric) and extra-textual (exophoric) references. These bundles have three following sub-categories:

a. *Framing*

These lexical bundles, in academic writings, are used to frame a discourse. The examples of *framing* are: *the presence of the*, *in the wake of*

b. *Quantifying*

These lexical bundles attempt to quantify an entity. e.g., *a great variety of, a huge number of*

c. *Time/place/text-deixis bundles*

These lexical bundles have deixis which refer to some time, place or the part of a text, e.g., *in the beginning of the, above the given picture.*

Research in Lexical Bundles

With an exploratory orientation, early studies on lexical bundles aimed at exploring the recurrent word-units in registers and text-types drawn from different origin. Later, the scholar became interested in conducting cross-disciplinary analysis of the different texts to comparatively highlight the differential use of lexical bundles in the target texts. For instance, Biber et al., (2004) comparatively analyzed the corpus of classroom language and text-books for the presence of lexical bundles. Their findings reveal that the corpus of classroom language had twice more lexical bundles than in the textbooks. In another study, Biber (2006) conducted a comprehensive analysis of the Longman Spoken & Written English Corpus (LSWE) to explore the patterns of recurrent word-units and their differential occurrences in the written (academic prose) and the spoken (conversation) registers. The foundational research by Biber and his associates provided the base-line data for the future researchers and the English for Academic Purposes (EAP) scholars to investigate more in this direction.

Hyland (2008) has analysed a large-sized academic corpus of 3.5 million words comprising texts taken from research articles, masters and doctoral theses from four different disciplines. Hyland (2008), due to some methodological reasons, focused on the 4-word lexical bundles for analyzing the distribution of lexical bundles in academic discourse. This was a significant study highlighted lexical

bundles as an important building blocks of the academic discourse and the defining feature of the disciplinary boundaries.

In their study, Herbel-Eisenmann, Wanger and Cortes (2010) delimited their focus to a single category of lexical bundles, i.e., *stance* bundles. Their corpus comprised of secondary mathematics classroom corpus, conversation corpus, and university classroom corpus. By comparing these three different corpora, they concluded that *stance* bundles were frequently employed by the teachers and the students in their secondary level mathematics classroom.

In a cross-linguistic study, Hernandez (2013) has studied the spoken corpus of L1 and L2 speakers across three types of corpus. It was an important study as it, for the first time, compared the spoken academic corpus. The findings of the study show that L2 speakers made more frequent use of lexical bundles than L1 speakers. Nonetheless, the use of stance bundles was found more pronounced for L1 speakers. Additionally, among the functional types, it was found that *stance* bundles were the most frequently used across all three spoken corpora. Moreover, *referential* expressions were more frequently used in the corpus of L1 speakers.

Chen and Baker (2010) have explored the corpus comprising L1 and L2 written academic discourse to comparatively study the frequency occurrences of the lexical bundles. Their corpus comprised a) published academic texts (FLOB-J) and b) academic writings produced by L1 (BAWE-EN) and L2 (BAWE-CH) speakers. Qualitative and quantitative methods were used for the study. The findings of the study show that the L1 and the L2 corpora were functionally similar for the use of lexical bundles. Nonetheless, structurally, the noun phrase bundles are significantly different between all three corpora. Moreover, it was found that the L2 corpus had more frequent use of idiomatic expressions and connectors than the native corpus.

In a similar study by Byshokocsa and Lee (2017), the undergraduate students' corpus of English argumentative essays was compared for the English and Chinese speakers. By using Biber et al.'s (1999) and (2004) taxonomies the researchers explored the lexical bundles in these corpora for their structural and functional features. For analysis, 4-word lexical units were identified and analyzed to meet the research objectives. The results show that L1-Chinese students make more use of lexical bundle types than L1-English students. Among structural lexical bundles, the L1- English students were found to make more use of noun and preposition phrases but the L1-Chinese students mostly used the verb phrases. Moreover, L1-Chinese students also frequently made grammatically erroneous structures and committed mistakes in the use of prepositions and articles.

Materials and Methods

For this study the quantitative approaches were employed to analyze the functional types of 4-word lexical bundles in the corpus of PhD theses written by in the native and non-native writers. These lexical bundles were identified and

quantitatively explored by using a corpus tool (AntConc 3.3.5). The qualitative findings required a close analysis of the contextual use of the identified lexical bundles before grouping them into functional categories.

Developing Corpus

Two different corpora (native & non-native) were compiled for the study to compare the functional distribution of the lexical bundles the PhD theses of social sciences.

Non-native corpus data comprised 100 Ph.D. theses downloaded from Pakistan Research Repository (http://eprints.hec.gov.pk) maintained by Pakistan Higher Education Commission (HEC), Pakistan. These theses were randomly chosen from five different fields of social sciences (see table 2). The selected fields of social sciences were Education, English, History, Political Science and Psychology. All the theses were downloaded in PDF format and then converted into the plain text and cleaned before the files were analyzed.

Native corpus data was downloaded from British online library (www.ethos.bl.uk). In the native corpus, 100 Ph.D. theses were randomly selected from the same disciplines as selected for compiling non-native corpus. The detail of both the corpora is as below:

Table 2
Scheme for the Corpus for PhD Theses in Social Sciences

Type of Corpus	Non-native Social Sciences		Native Social Sciences	
	Subjects	**No. of theses**	**Subjects**	**No. of theses**
	Education	20	Education	20
	English	20	English	20
	History	20	History	20
Disciplines	Political Science	20	Political Science	20
	Psychology	20	Psychology	20
Total No of Theses	100		100	
Total no. of Words	6350130		13,026,919	

Criteria for the Analysis

There were three criteria which were met for the data analysis. First, lexical bundles can range from the length of 2 words to 6 words or even more. However, the lexical bundles which are less or more than 4-word length are not considered suitable of the study by many scholars (e.g., Hyland, 2008; Cortes, 2004; Chen, 2008). It is believed that a shorter or a longer word string of the lexical bundles will not be able to provide rich insights about the functional use of lexical bundles.

Second, there were many 4-unit words that were found less than 5 times per million words in both the corpus. By definition, as the lexical bundles are known for their frequent occurrence in a texit, so the frequency occurrence of 5 lexical bundles per million words was considered a cut-off value and any lexical item that occurred less than this cut-off value was not considered to put under any functional category of the lexical bundles.

Third criterion for the selection of lexical bundles was their functional salience. Not all 4-word units can have functional orientation in the text. So, the random combinations of the 4-word units were ignored.

Data Analysis Tool

For quantitative analysis of the present study, two softwares i.e., PDF file converter and AntConc 3.3.5 were used.

a) AntConc Corpus Analysis Tool

This software consists of options such as concordance, file view, wordlist, N grams, and clusters. In order to investigate the corpus, this program scans the corpus word by word and concludes the results with their rank, range, and frequencies. It identifies the lexical bundles within the corpora according to the range and frequency size set by the researcher.

b) PDF Files Converter

All the files for the corpus compilation were downloaded in the PDF format. Then, these files were converted by using PDF file converter into the plain text to execute them in AntConc.

Framework

To achieve the research objectives, the current study used the functional categorization of lexical bundles derived from Biber et al., (2004 & 2003) and Biber & Barbieri (2007). This functional framework is divided into 3 major (*stance*, *discourse*, and *referential*) and multiple sub-categories of lexical bundles (see section Literature Review).

Data Analysis

Identification of 4-word lexical units

At the first stage of data analysis, AntConc corpus tool was employed to calculate the number of words in the target corpus of PhD theses. It was noted that, the corpus of PhD theses by native speakers had 13 million words and that of non-native speakers had 6.3 million words. Moreover, the n-gram option was employed to produce the lists of the 4-word units present in the corpus of PhD theses. It is

important to recall here that according to the data analysis criterion, only those 4-word lexical bundles were analyzed which had their frequency occurrence of more than 5 times per million words for each of the corpus.

In the following the most frequently occurring top-20 four-word units present in the PhD thesis of the native and non-native writers are compared. Table 3 shows that the lexical bundles of "*on the other hand*" and "*on the bases of*" have the frequency occurrences 210 and 155 per million words. Moreover, for the native corpus, the most frequently occurring lexical bundle per million words is *at the same time* with 26 occurrences followed by "*on the other hand*" with 23 occurrences. Moreover, the data shows that "*in the light of*" and "*in relation to the*", are the least occurring lexical patterns in both the corpus with their respective occurrences of 40 and 7 in the corpus. Another lexical bundle which is found common in the corpus is, "*in the form of*". In the following, there is a comparative list of the first twenty lexical bundles with their frequency.

Table 3
Comparison of the 4-word units in PhD Corpus

Non-native Corpus				Native Corpus		
Sr.	Lexical Bundles	F	Normalized Frequency Per million	Sr.	Lexical Bundles	F	Normalized Frequency Per million
1	on the other hand	1316	210.56	1	at the same time	337	25.949
2	on the bases of	974	155.84	2	on the other hand	297	22.869
3	is no significant difference	734	117.44	3	the end of the	252	19.404
4	there is no significant	666	106.56	4	in the case of	243	18.711
5	with the help of	452	72.32	5	in the light of	195	15.015
6	in the field of	404	64.64	6	at the end of	173	13.321
7	the performance of the	348	55.68	7	on the one hand	182	14.014
8	at the same time	336	53.76	8	the role of the	164	12.628
9	at the time of	336	53.76	9	as a result of	162	12.474
10	as far as the	328	52.48	10	as well as the	162	12.474
11	in the form of	318	50.88	11	the fact that the	156	12.012

12	in the same way	298	47.68	12	the fact that the	143	11.011
13	as well as the	292	46.72	13	the nature of the	133	10.241
14	as a result of	290	46.4	14	**on the basis of**	132	10.164
15	in the case of	290	46.4	15	**in the form of**	129	9.933
16	at the end of	276	44.16	16	in the context of	113	8.701
17	a large number of	266	42.56	17	on the part of	104	8.008
18	the end of the	260	41.6	18	for the sake of	100	7.7
19	of the opinion that	250	40	19	a member of the	89	6.853
20	in the light of	248	39.68	20	in relation to the	89	6.853

Analysis of Lexical Bundles across Main Categories

Here, we will discuss the functional distribution of the lexical bundles across their three major categories. After applying the criterion of data analysis mentioned earlier, it was noted that native speakers used lower number of different combinations of the 4-word units (327) as compared with non-native speakers (558). Later, the identified lexical bundles were grouped according to their three functional categories (see Fig.1).

In the following figure, the functional distribution of the lexical bundles across three major categories is given in total numbers and percentages. The data shows that the most prominent type of function is referential function with the highest number of lexical bundles for both the corpus. There are 75% (246) and 64% (358) *referential* bundles in the native and non-native corpus, respectively. Furthermore, the second most prominent type of lexical bundle is that of *stance*. This functional category has 20% (65) and 22% (124) lexical bundles in the native and non-native corpus, respectively. At the end, the least frequently occurring lexical bundles fall into the category of *discourse organizers*. Its occurrences are 5% (16) in the native corpus and 13% (76) in non-native corpus.

Figure 1
Functional Distribution of LBs for Main Categories

Functional Distribution of LBs

	No. of Bundles	Referential (%)	Discourse (%)	Stance (%)
Native Corpus	377	75	5	70
Non-native Corpus	558	64	13	22
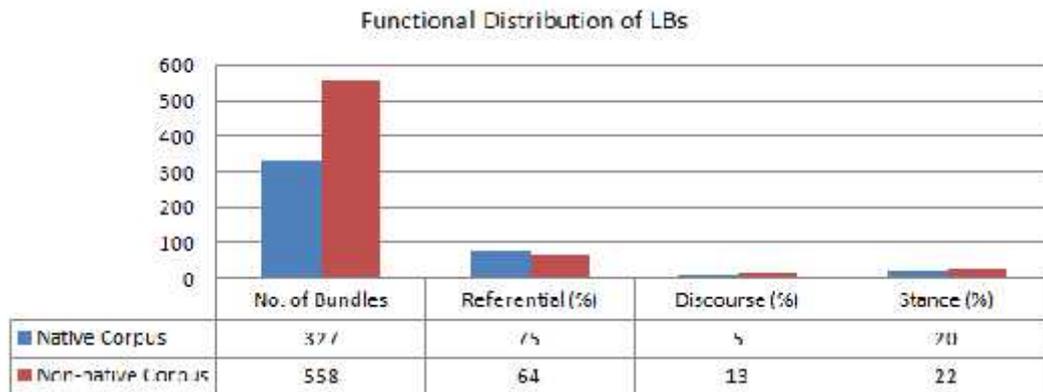
The findings of the study show that non-native writers make more use of lexical bundles than the native speakers. This finding is in-line with Hernandez (2013) and Byshokocsa and Lee (2017) but contradict with the findings of Adel and Erman (2012) and that of Chen and Baker (2010).

Functional Analysis of Lexical Bundles across Sub-categories

In the following, we provide the frequency occurrences of the lexical bundles along with the sub-categories of *referential*, *discourse* and *stance* bundles. The table 4 and fig 2 show the total number of frequency occurrences of the lexical bundles in each category along with their normalized occurrences per million words.

The analysis shows that there is the highest frequency use of lexical bundles for the sub-category of framing (see fig 2). This means that the most of the lexical bundles used in PhD theses have the referential orientation. *Framing* bundles are used to frame the discourse, quantify the entities or indicate time, text or place. According to table 3, the percentage occurrence of the *framing* bundles is 35 % and 13% for non-native and the native corpus respectively.

Table 3
**Functional Distribution of Lexical Bundles for Sub-categories**

	Native Discourse			Non-native Discourse		
Functions	**Sub Category**	**Total F**	**Normalized Frequency per million**	**Sub Category**	**Total F**	**Normalized Frequency per million**
Stance	Epistemic Stance	32	2.464	Epistemic Stance	60	9.6
	Obligation	7	0.539	Obligation	18	2.88

	Desire/int ention /attitudin al	19	1.463	Desire/int ention/atti tudinal	32	5.12
	Ability	7	0.539	Ability	14	2.24
Discourse Organizer	Topic introducti on	0	0	Topic introductio n	4	0.64
	Topic Elaboratio n	8	0.616	Topic Elaboratio n	54	8.64
	Identificati on	8	0.616	Identificati on	18	2.88
Referenti al	Framing	171	13.167	Framing	216	34.56
	Quantifica tion	22	1.694	Quantificat ion	62	9.92
	Time/text /place	53	4.081	Time/text /place	80	12.8

The data also shows that the least use of the 4-unit words in both the corpus is present for the lexical bundles of *discourse* (see fig. 2). Among the sub-categories of *discourse* bundles, the lexical bundles used for *topic introduction* were not found in the native corpus and had a very limited occurrence in non-native PhD corpus. Moreover, among the *stance* bundles, according to the findings of the study, non-native speakers comparatively more frequently use the *epistemic stance* bundles (see table 3). This shows that the Pakistani PhD scholars were more interactive in their academic writings than that of native speakers.

Fig. 2
Functional Distribution of Lexical Bundles for Sub-categories
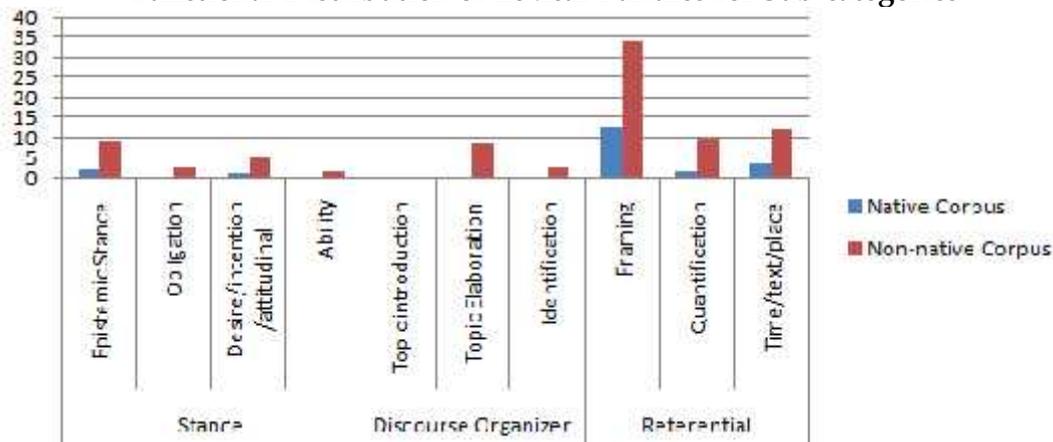
Discussion

The functional use of lexical bundles in the selected academic corpus shows that there are quantitative differences among the native and non-native writers in

terms of their preferences for the use of lexical bundles. It was found that non-native writers use higher frequency of lexical bundles as compared with the native writers. This finding is similar to the studies conducted by Hernandez (2013) and Byshokocsa and Lee (2017) but it challenges the findings of the two other studies conducted by Adel and Erman (2012) and Chen and Baker (2010). Apparently, the frequent use of lexical bundles can be belived as associated with the respective competence of the native and non-native writers in English. Nonetheless, such associations can be complex in nature and need further probe. One tentative interpretation of the current study can be that non-native writers, through their input of the academic discourse, develop an academic jargon composed of lexical bundles as the key building blocks of their linguistic repertoire. Later, these lexical bundles are overused when they produce any academic discourse.

Another interesting finding of the study is the high frequency presence of *referential* bundles in the PhD theses of both the corpus. One explanation for the high frequency use of referential bundles could be the referential orientation of language in general and the academic discourse in particular. It means that the scholars of social sciences quite frequently refer to the textual and extra-textual entities in their academic discourse. The frequency occurrences of all three sub-categories of *referring* lexical bundles are high, particularly for *framing*. *Framing* is an important aspect of any formal discourse—academic or non-academic. The study also shows that, for both the corpus, the number of the 4-word units used for *framing* is close to twice the number of the lexical bundles used for the rest of the functional categories in the PhD corpora.

Conclusion

Previous research on the use of lexical bundles in academic discourse has produced conflicting results regarding their frequency use by the native and non-native writers (e.g., Adel and Erman, 2012; Byshokocsa and Lee, 2017). Anticipating that one of the reasons for the contradictory findings can be the variation in the corpus used for the studies; the researchers for this study decided to delimit the current study to the academic discourse of PhD theses in the domain of social sciences. A large corpus of 200 PhD theses by native and non-native writers was analyzed to identify and categorize functional lexical bundles according to the taxonomy based on Biber et al., (2004 & 2003) and Biber & Barbieri (2007). The study shows a significantly high frequency use of *referential* bundles for both types of corpus. Furthermore, the study highlights that within the *referential* bundles, most of the lexical bundles are used by the writers for *framing* and contextualizing their research discourse. It is envisaged that the current study will help the EAP scholars and teachers in developing material for teaching academic writing skills to non-native learners.

References

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.

Biber, D. (2006). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97-116.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for specific purposes*, *26*(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey* Leech (pp. 71–93). Frankfurt: Peter Lang.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, *25*(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English* (Vol. 2). London: longman.

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes*, *30*, 38-52.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, *5*(5), 31-64.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language learning & technology*, *14*(2), 30-49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23, 397–423.

Crowley, S. (1991). A Personal Essay on Freshman English. *Pre-Text: A Journal of Rhetorical Theory*, *12*, 155-76.

Herbel-Eisenmann, B., Wagner, D., & Cortes, V. (2010). Lexical bundle analysis in mathematics classroom discourse: The significance of stance. *Educational Studies in Mathematics, 75*(1), 23–42.

Hernández, P. S. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, *12*(1), 187-209.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4–21.

Hyland, K. (2009). *Academic discourse: English in a global context*. A&C Black.

Kashiha, H., & Heng, C. S. (2013). *The Effect of Role-play on EFL Learners' Retention and Recall of Idiomatic Expressions*.

Marshall, S. L., & While, A. E. (1994). Interviewing respondents who have English as a second language: challenges encountered and suggestions for other researchers. *Journal of Advanced Nursing*, *19*(3), 566-571.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Rafiee, M., Tavakoli, M., & Amirian, Z. (2011). Structural analysis of lexical bundles across two types of English newspapers edited by native and non-native speakers. *Modern Journal of Applied Linguistics*, *3*(2), 218-236.

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549-569.

Strunkyt, G., & Jurkūnait, E. (2008). *Written academic discourse: Lexical bundles in humanities and natural sciences* (Unpublished BA thesis. Vilnius: University Faculty of Philology, Department of English Philology).

Wood, D. (2010). Lexical clusters in an EAP textbook corpus. *Perspectives on formulaic language: Acquisition and communication*, *88*.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.