Lexical Bundles in a Saudi General-Audience Podcast in English: A Corpus Analysis

Sahar Alkhelaiwi[1]

[1] Department of English Language and Translation, College of Sciences and Arts in Ar Rass, Qassim University, Saudi Arabia

Correspondence: Sahar Alkhelaiwi, Department of English Langauge and Translation, College of Sciences and Arts in Ar Rass, Qassim University, Saudi Arabia. E-mail: s.alkhelaiwi@qu.edu.sa

Received: January 22, 2023 Accepted: February 22, 2023 Online Published: February 22, 2023

doi:10.5430/wjel.v13n2p355 URL: https://doi.org/10.5430/wjel.v13n2p355

Abstract

Research finds that lexical bundles are vital to fluent language processing as they reduce cognitive load when memorized as chunked sequences of language, especially for second-language (L2) learners. Although lexical bundles in academic and political discourse have been studied, their use in podcasts is a less-researched domain in corpus linguistics even though podcasts are an increasingly popular medium that offers authentic public discourse for diverse audiences, including L2 learners and instructors. To address this gap, this study investigates the most frequently occurring and widely dispersed lexical bundles in an English-language, general-audience Saudi podcast. A specific corpus consisting of 10 podcast episodes (almost one hour each) was submitted to AntConc for the identification of lexical bundles based on three predefined parameters: length, frequency, and distribution. A lexical bundle was extracted if it consisted of a four-word sequence that occurred at least five times in at least five texts. From a podcast corpus of 111,174 words, 56 four-word lexical bundles were identified and ranked according to frequency, and their grammatical structures were analyzed. Results show that lexical bundles such as *thank you so much*, *a lot of people*, *be honest with you*, and *and I was like* are high on the list. Structurally, they consist of nominal, verb-based, and prepositional phrases among others, although verb-based bundles are the most common. The study concludes by providing a list of lexical bundles that may be used for podcast-based learning practice in a L2 listening class or for independent learning.

Keywords: AntConc, corpus linguistics, lexical bundles, n-grams, podcast

1. Introduction

In cognitive linguistics, *constructions* are considered the basic units of language. They are accepted by a speech community and established as part of a language learner's linguistic knowledge (Langacker, 1987; Simpson-Vlach & Ellis, 2010). Learners acquire constructions during meaningful communicative encounters and incorporate them into their structural inventory for that language (Pawley & Syder, 1983; Simpson-Vlach & Ellis, 2010). Studies in corpus linguistics have demonstrated the recurrent nature of these constructions (or *formulaic units* or *specific configurations*) in discourse (e.g., Biber et al., 1999; McEnery & Wilson, 1996). According to Gablasova, Brezina, and McEnery (2017), linguistic corpora provide "a rich source of information" about the frequency, regularity, and distribution of formulaic sequences in language, and formulaic language has enjoyed growing interest in corpus linguistics studies over the past decade (p. 156). As Martinez and Schmitt (2012) explain, one of the primary findings of corpus linguistics is that language consists not only of individual words, but also of a great deal of formulaic language. In fact, according to Durrant and Schmitt (2009), "language is largely formulaic in nature" (p. 157).

One impetus for recent research on formulaic language is, as Gablasova et al. (2017) point out (drawing on Ellis et al., 2015 and Wray, 2002), that research on both first and second language acquisition has found links between formulaic patterns and fundamental cognitive/mental processes during language learning and language use (e.g., storage, cognitive representation, and access to linguistic clusters in the mental lexicon). Researchers such as Simpson-Vlach and Ellis (2010), Shin and Nation (2008), and Pawley and Syder (1983) argue that collocations (another term for lexical bundles) help learners develop fluent language processing skills since they reduce cognitive demand and save processing time when they are learned as chunked expressions. Shin and Nation (2008) and Pawley and Syder (1983) also argue that collocations help learners sound "native-like," and, as noted by Ellis and Sinclair (1996), "[t]he attainment of fluency, in both native and foreign languages, involves the acquisition of memorized sequences of language" (p. 234). In a foreign-language context (which is the focus of the present study), Boers et al. (2006) find that second language (L2) speakers are judged to be more proficient when they use formulaic sequences. Ackermann and Chen (2013) and Laufer (2011) find that, although native speakers can easily recognize collocations, they are often challenging for L2 learners to acquire and use correctly. Ackermann and Chen (2013) attribute this to Nation's (2001) argument that collocations often involve "grammatical or lexical unpredictability or inflexibility" (p. 324).

2. Background of the Study

2.1 Lexical Bundles

In previous literature, this phenomenon—a series of words that frequently occur together in a given register—has often been referred to by different terms (Siyanova-Chanturia & Pellicer-Sánchez, 2018), including lexical bundles, multiword constructions, multiword expressions, n-grams, clusters, academic formulas, formulaic language, collocations, chunks, chains, duads/tryads, prefabricated patterns, and word combinations. Majuddin, Siyanova-Chanturia, and Boers (2021) use *multiword expressions* (MWEs) as an umbrella term for a wide range of linguistic patterns above the single-word level. They argue that MWEs differ from formulaic language, which may include single-word items as well as longer constructions. MWEs may be collocations (*strong tea*), binomials (*bride and groom*), idioms (*spill the beans*), lexical bundles (*on the other hand*) (which are the focus of this study), or proverbs (*better late than never*). The present study, however, focuses only on lexical bundles, which Biber, Conrad, and Cortes (2004) define as follows: "Lexical bundles are usually not complete grammatical structures nor are they idiomatic, but they function as basic building blocks of discourse" (p. 371). O'Keeffe, McCarthy, and Carter (2007) describe n-grams or lexical bundles as "fragmentary strings" and "meaningful strings", such as *a number of* and *thank you* (p. 61). Although they are fragments, they provide "important clues as to how interaction unfolds" (O'Keeffe et al., 2007, p. 70).

2.2 Podcasts

According to Panagiotidis (2021), the term *podcast*, first introduced in 2004, is a neologism taken from "pod" and "broadcast"; the term *pod* became particularly popular after the global release of the Apple iPod in 2001. Other scholars suggest that *podcast* is derived from "personal on demand (POD)" and "broadCAST'' (Panagiotidis, 2021). In fact, both views are valid; as Panagiotidis (2021) points out, podcasts are available "on demand" and are usually transmitted to users' phones or other mobile devices. Nurmukhamedov and Sharakhimov (2021) define a podcast "as a multimedia file, such as a radio program or video, that can be downloaded or streamed from the Internet onto a computer or mobile device" (p. 2). Nurmukhamedov and Sharakhimov (2021) also point to the growing popularity of podcasts, especially since many professional groups (e.g., TED Talks), periodicals (e.g., *The Economist*), and universities (e.g., Harvard and Massachusetts Institute of Technology) are now offering free podcasts, allowing listeners to easily access many podcasts on different topics. Many listeners find podcasts attractive because, as Nurmukhamedov and Sharakhimov (2021) and Thorne and Payne (2005) point out, they do not have to sit and listen; rather, they can listen to their favorite podcasts while riding a bus or subway, driving a car, or walking around a university campus or shopping mall.

It seems worthwhile to analyze lexical bundles in podcasts since listeners, who may include L2 learners, might be exposed to these bundles. Since Biber et al. (2004) argue that "different registers tend to rely on different sets of lexical bundles" (p. 377), and Durrant (2018) confirms that such recurrent formulas tend to be highly context-dependent and associated, not only with particular communities, but also with certain topics, registers, and genres—the present study will thus analyze and characterize such bundles in podcasts as a form of spoken discourse. Previous corpus linguistic studies comparing spoken and written discourse have found that lexical bundles and collocations are more common in spoken discourse (e.g., Biber et al., 1999; Biber et al., 2004; Shin, 2007; Shin & Nation, 2008; Leech, 2000). Shin and Nation (2008) argue that such formulaic units are more important to spoken discourse than to written register; this may be because speech is constructed in real time and thus makes greater demands on working memory than writing, increasing the need to depend on formulas, as it is easier to retrieve something from long-term memory than to construct it anew (e.g., Kuiper, 1996).

2.3 Previous Research on Lists of Lexical Bundles

Many previous studies of lexical bundles have developed lists of these bundles (e.g., Biber et al., 2004; Alasmary, 2019; Shin & Nation, 2008; Simpson-Vlach & Ellis, 2010; Nesi & Basturkmen, 2006; DeCarrico & Nattinger, 1988; Ackerman & Chen, 2013; Hyland, 2008; Crawford-Camiciottoli, 2007; Wood & Apple, 2014). However, nearly most of these studies focus on academic discourse and make recommendations for teaching and testing in the area of English for Academic Purposes (EAP). Although the present study deals with podcasts (a spoken non-academic register), it is useful to review some previous studies of lexical bundles used in academic discourse as this study will follow some of their methodologies for identifying lexical bundles from a corpus. To the best of the researcher's knowledge, to date, no study has examined recurrent lexical bundles in podcasts.

One of the earliest attempts to compile a list of recurrent sequences was made by Biber et al. (2004). They compared lexical bundles in large corpora of spoken (e.g., classroom lessons and university conversations) and written (e.g., academic prose and textbooks) registers. They analyzed data from the TOEFL 2000 Spoken and Written Academic Language Corpus, which they compared to the non-academic Longman Grammar of Spoken and Written English corpus. They then conducted a frequency analysis (to be included, a lexical bundle had to recur at least 40 times per one million utterances and consist of four words) and found that lexical bundles occurred more frequently in spoken discourse than in written discourse. Some lexical bundles they found in spoken registers included *I don't know*, *I think it was*, *you have to be*, and *it's going to be*. In a study that did not focus exclusively on academic language, Shin and Nation (2008) analyzed data from the ten million spoken English section of the British National Corpus (BNC), and compiled a long list of high-frequency collocations of different lengths. They identified 4,698 collocations, but only the 100 most frequent are included in their published list. Of these, 77 per cent are two-word collocations (e.g., *you know* or *a bit*), while the rest are three- (e.g., *at the moment*, *I think that*) and four-word sequences (e.g., *thank you very much*).

A key lexical bundle list is published in what is probably one of the most significant and methodologically robust studies on this topic

(according to Alasmary, 2019). In this study, Simpson-Vlach and Ellis (2010) compiled an Academic Formulas List (AFL), which they aggregated from academic speech corpora, the Michigan Corpus of Academic Spoken English (MICASE) and the BNC, as well as an academic writing corpus consisting of research articles designed by Hyland (2004). These corpora were then compared to two other non-academic corpora. They included formulaic sequences occurring at least ten times per one million words and appearing in at least four out of five academic disciplines studied. Simpson-Vlach and Ellis (2010) used a series of analyses to confirm the importance of combining quantitative methods, such as frequency, n-grams, log likelihood statistical measures, and mutual information (MI). MI is a statistical measure that is commonly employed in the field of information science to assess the degree to which the words in a phrase occur together more frequently than would be expected by chance. A higher MI score suggests a stronger association between the words, while a lower MI score indicates that their co-occurrence is more likely due to chance. Simpson-Vlach and Ellis (2010) also asked some experts (experienced EAP instructors and language testers) to rate a stratified random sample of formulaic sequences using specific criteria to predict formulas teaching worth (FTW)—that is, which academic formulaic sequences are worth teaching to L2 learners (p. 488). This was done based on the reasoning that it is better to avoid long lists of formulaic sequences. Their list includes recurrent patterns of different lengths, including three-word (e.g., *blah blah blah*), four-word (e.g., *does that make sense*), and five-word sequences (e.g., *you know what I mean*).

Using the BNC, Martinez and Schmitt (2012) built a phrasal expressions list that includes 505 lexical phrases (e.g., *have to*, *of course*, *turn into*, and *something like that*). To be included on this list, a given n-gram had to meet six criteria. Included n-grams range from two to four words in length, occur at least 787 times in the corpus, have a one-word equivalent in English (e.g., *put up with* has the synonym *tolerate*), are not semantically transparent (i.e., they would be expected to cause potential difficulties for L2 learners), and are deceptively transparent (i.e., they use words that learners think they know, but knowing those words does not necessarily convey understanding of the n-gram). Ackermann and Chen (2013) used a set of statistical measures to analyze entries that met the quantitative parameters for normed frequency, MI score, and t-score. They also conducted a computational corpus analysis by creating lists of content words, nod words, and stop lists; the list was then manually vetted by the researchers and expert judges (a linguistics professor and a dictionary consultant). They eventually produced a list called the Academic Collocation List (ACL), a cross-disciplinary list based on the Pearson International Corpus of Academic English (PICAE) that encompasses 2,468 entries that are mainly two-word sequences using the following patterns: (1) noun + noun (e.g., *anecdotal evidence*) (nearly three-quarters of all entries); (2) verb + noun/adjective (e.g., *seem plausible*); (3) adverb + verb (e.g., *explicitly state*); and (4) adverb + adjective (e.g., *highly controversial*).

To conclude, using existing corpora, the aforementioned studies underscore the importance of lexical bundles, as seen in the development of multiword lists. However, little research has examined lexical bundles outside academic (or political) contexts. To fill this gap, it is the aim of this paper to identify lexical bundles that appear in general-audience podcasts and to examine their structures. The findings will have some pedagogical implications for language instructors' podcast-based learning practices.

2.4 Research Questions

This study seeks to answer the following questions:

1. Which lexical bundles appear most frequently in the context of English-language podcasts targeting a general audience of Saudi listeners?
2. What are the structural patterns of common lexical bundles that appear in English-language podcasts targeting a general audience of Saudi listeners?

3. Methodology

3.1 Corpus

The current study collects data from the podcast *The Mo Show*. As described on its website, this podcast offers a front-row seat to the life and culture of Saudi Arabia, including personal stories about the accomplishments and experiences of residents of Saudi Arabia (Islam, 2022–present). According to the podcast categories defined by Nurmukhamedov and Sadler (2011), *The Mo Show* is a general-audience podcast. This podcast was selected for the present study because it is the first English-language podcast targeting a Saudi audience. Ten episodes were randomly selected to create the corpus (focusing on episodes published in 2022 for which transcripts are available). The included podcast episodes cover a wide range of themes, including women's empowerment, wellness, fashion, sports, business, and space. To create a corpus, episode transcripts were copied from YouTube and pasted into a Microsoft Word file. The podcasts are also available on Apple Podcasts, Spotify, and Google Podcasts; thus, it may be assumed that people may listen to these podcasts more than watching them on YouTube. Titles were manually removed from the transcripts, and each episode transcript was numbered to facilitate concordance checks and example extraction. This process resulted in a corpus of 111,174 words (see Table 1). Written permission to use *The Mo Show podcast* in this study was obtained.

Table 1. The study's corpus

No.	Episode title	Interviewee	Length in time	Length in words	Link
1	HRH Princess Reema bint Bandar AlSaud | Mo Show 70	HRH Princess Reema bint Bandar AlSaud	1:05:36	10,811	https://www.youtube.com/watch?v=UqYkZd0FM-A&t=206s
2	Houssam Abiad | The Mo Show 63 | Planning for Cities, Love for Makkah, Celebrating our Identity	Houssam Abiad	1:10:46	14,041	https://www.youtube.com/watch?v=WCZZVYUNJzk
3	Emon Shakoor | The Mo Show 67 | Scaling Blossom, Seizures to Superpowers & Inclusive Innovation	Emon Shakoor	1:01:14	11,395	https://www.youtube.com/watch?v=0qyvE8W0NPs
4	Marloes Knippenberg | The Mo Show 69 | The Hospitality Space, Localisation & Sustainability	Marloes Knippenberg	1:07:53	10,706	https://www.youtube.com/watch?v=EBZ8Eqfxy7g
5	HRH Prince Abdulaziz Bin Turki Al-Faisal 55	HRH Prince Abdulaziz Bin Turki Al-Faisal	56:55	9,943	https://www.youtube.com/watch?v=4T23VbA6xyU
6	Mishaal Ashemimry 49 | The Mo Show Podcast | Rocket Science, Aerospace Engineer & Space Exploration	Mishaal Ashemimry	57:22	10,488	https://www.youtube.com/watch?v=UMa0Ycs20EY
7	Lina Hashem 48 | The Mo Show Podcast | Mental Health Therapist	Lina Hashem	1:13:17	12,036	https://www.youtube.com/watch?v=6lCx7n5v8qI
8	Daneh Buahmad | The Mo Show 73 | Fashion Design, Growing up in ARAMCO and Vivienne Westwood	Daneh Buahmad	1:09:13	11,175	https://www.youtube.com/watch?v=Ag6P7fsvd6E
9	Noura Alturki | The Mo Show 64 | Nesma, Corporate Governance, Social impact	Noura Alturki	59:53	9,824	https://www.youtube.com/watch?v=9TAlyLd8ZVM
10	Mohamed Mostafa | The Mo Show 61 | Wellbeing, Emotions and Weight-loss	Mohamed Mostafa	1:02:07	10,899	https://www.youtube.com/watch?v=0T0nF00WoAY

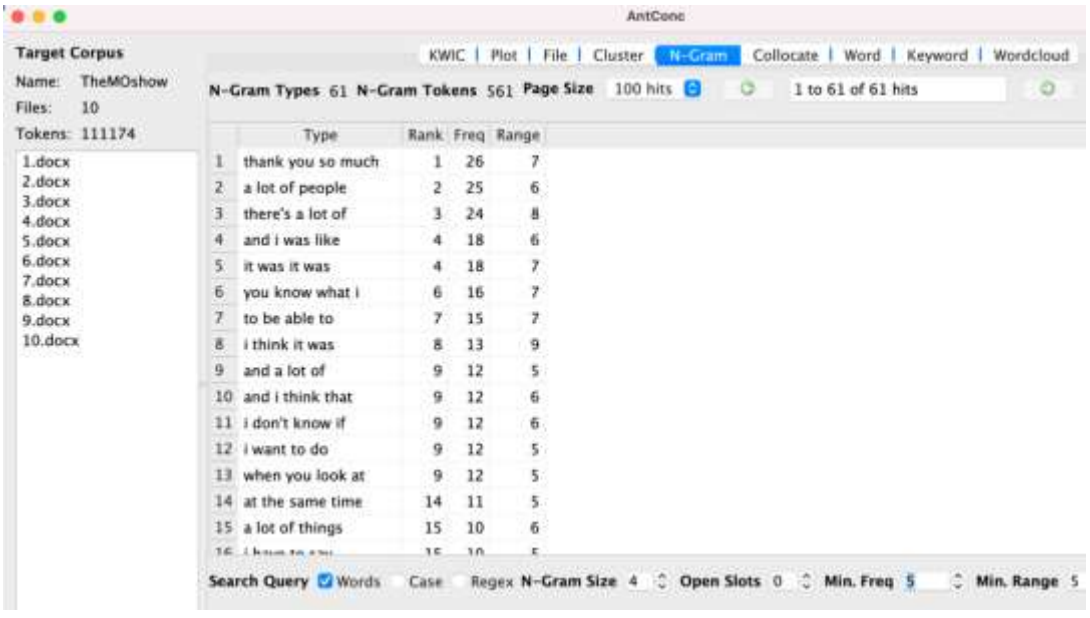

Figure 1. AntConc's list implementing cluster/n-gram features, displaying four-word bundles

3.2 Lexical Bundle Extraction and Analyses

All 10 text files were uploaded to AntConc (Anthony, 2022; version 4.1.4), and the n-gram feature was used to extract four-word bundles (this bundle length is favored by many scholars, such as Biber et al., 2004; Hyland, 2008). Previous studies have arbitrarily set cut-offs for the minimum frequency and ranges of included lexical bundles; the cut-off frequency in previous studies ranges from 10 to 40

occurrences per million words. Since this corpus is small, I followed Jablonkai (2009), who recommends that a lexical cluster should appear at least five times in five different texts of a small corpus to be included. However, it should be noted that the cluster feature of AntConc was executed, because when running the n-gram feature alone, AntConc produced bundles such *I am able to* with no contractions. Researchers such Biber et al. (2004) recommend that the orthographic word unit should be dealt with as one single word (e.g., *I'd*, *It's*) to precisely identify clusters in a corpus—a method that was also followed in this study. Therefore, the following parameters were used in AntConc: frequency = 5; n-gram = 4; range = 5 (see Figure 1). This search returned 61 four-word bundles.

Then, some bundles were excluded. Following Chen and Baker (2016), bundles that include proper nouns were excluded as they are highly context-dependent (*The Mo Show Podcast*), or those referencing the specific context of the podcast (*coming on the show*, and *in Saudi Arabia and*). In addition, overlapping bundles (that is, bundles that are parts of longer lexical bundles) (*you so much for* and *you for having me*) were excluded. Next, the structures of the extracted bundles were classified and counted using Alasmary's taxonomy (2019), which is based on traditional frameworks such as Biber et al. (2004). This was considered useful because Biber et al.'s (2004) framework includes both oral and written registers.

4. Results

4.1 Lexical Bundles in the Mo Show Podcast Corpus

The final list includes 56 bundles (see Table 2), listed according to frequency and distribution amongst texts. The most common sequence was *thank you so much* (appearing 26 times in seven texts), followed by *a lot of people* (appearing 25 times in six texts), *there's a lot of* (appearing 24 times in eight texts), *and I was like* (appearing 18 times in six texts), and *it was, it was* (appearing 18 times in six texts).

Table 2. All lexical bundles from the podcast corpus

No.	Lexical bundle	Rank	Freq.*	Range	No.	Lexical bundle	Rank	Freq.*	Range
1	thank you so much	1	26	7	32	**you so much for	24	8	5
2	a lot of people	2	25	6	33	a little bit about	33	7	6
3	there's a lot of	3	24	8	34	and and you know	33	7	5
4	and I was like	4	18	6	35	and you know what	33	7	5
5	it was it was	4	18	7	36	as much as I	33	7	5
6	you know what I	6	16	7	37	**coming on the show	33	7	6
7	to be able to	7	15	7	38	know what I mean	33	7	5
8	I think it was	8	13	9	39	of the reasons why	33	7	5
9	and a lot of	9	12	5	40	one of the reasons	33	7	5
10	and I think that	9	12	6	41	one of the things	33	7	6
11	I don't know if	9	12	6	42	thank you for having	33	7	7
12	I want to do	9	12	5	43	**you for having me	33	7	7
13	when you look at	9	12	5	44	and I don't know	44	6	6
14	at the same time	14	11	5	45	and I think it's	44	6	5
15	a lot of things	15	10	6	46	and you have to	44	6	5
16	I have to say	15	10	5	47	at the end of	44	6	5
17	I want to be	15	10	5	48	back in the day	44	6	5
18	I would like to	15	10	7	49	have to deal with	44	6	6
19	to be part of	15	10	5	50	I don't know how	44	6	5
20	be honest with you	20	9	5	51	**the mo show podcast	44	6	5
21	in the world and	20	9	6	52	to do it and	44	6	6
22	is one of the	20	9	8	53	was like you know	44	6	5
23	that you want to	20	9	5	54	and there was a	54	5	5
24	a little bit more	24	8	5	55	and you know you	54	5	5
25	and I want to	24	8	5	56	I think it's a	54	5	5
26	at the time I	24	8	6	57	if you want to	54	5	5
27	do you want to	24	8	5	58	**in Saudi Arabia and	54	5	5
28	in the in the	24	8	6	59	in the middle of	54	5	5
29	so I was like	24	8	5	60	know you have to	54	5	5
30	you know the the	24	8	5	61	you don't want to	54	5	5
31	you know you have	24	8	6					

*Frequency

**Bundles were excluded.

4.2 Structural Analyses of Lexical Bundles in The Mo Show Podcast Corpus

The first structural category is nominal bundles. Only three of these were found in the corpus, and two of them begin with the indefinite pronoun *one* followed by *of* as embedded constructions, as shown in Table 3. (Due to space limitations, each category is illustrated with two quotes.)

Table 3. Nominal lexical bundles

Structural category	Grammatical pattern	Lexical bundle
NP	NP + of	• One of the reasons
	Other NPs	• One of the things
		• Of the reasons why

(1) you know **one of the things** was we need to start thinking in a different way like business plans you know legal department compliance governance risk where we weren't really focused on that in the earlier days (episode 9)

(2) it is uh it's it's a miracle how it all came together like one **of the reasons why** people are so curious in astronomy (episode 6)

Six prepositional sequences were found in the corpus (see Table 4). All of these begin with *in* or *at*; one sequence begins with a preposition and concludes with either a marker for the next clause (e.g., *at the time I*) or a conjunction (e.g., *in the world and*). Two of these sequences end with *of* (e.g., *at the end of*), and one involves a prepositional repetition (*in the in the*).

Table 4. Prepositional lexical bundles

Structural category	Grammatical pattern	Lexical bundle
PP	PP + of	• at the same time
	Other PPs	• in the world and
		• at the time I
		• in the in the
		• at the end of
		• in the middle of

(3) I feel we are very connected but **at the same time** extremely disconnected yeah how funny the irony yeah so does mental health problems do they develop or stem from a young age (episode 7)

(4) it's an extension of the education system where you take employees and you appraise them **at the end of** the year (episode 9)

Table 5 shows the 26 identified lexical bundles that contain verbs; these use several different structures. Some include the copular be + an adjective (e.g., *be honest with you*); some include a copula + a noun (e.g., *to be part of*); some are existential phrases (e.g., *there's a lot of*); some start with a $1^{st}/2^{nd}$ personal pronoun (e.g., *I want to be*; *you know the the*); some are sentence fragments that include a verb (e.g., *was like you know*), and a single question (*Do you want to?).*

Table 5. Verb-based lexical bundles

Structural category	Grammatical pattern	Lexical bundle
VP	Copular be + NP + AP	• was like you know
	Existential phrases	• to be able to
	It clauses	• to be part of
	A question	• there's a lot of
	Fragment with a verb	• I want to be
		• be honest with you
		• I have to say
		• I want to do
		• I don't know if
		• you know what I
		• I think it was
		• I would like to
		• thank you so much
		• do you want to
		• you know the the
		• you know you have
		• know what I mean
		• thank you for having
		• have to deal with
		• to do it and
		• I think it's a
		• know you have to
		• you don't want to
		• it was it was
		• is one of the
		• back in the day

(5) that we are doing is we're launching a community platform a mini neighborhood platform which really becomes a tool for our channel managers and properties **to be able to** almost curate their neighborhood get the suppliers and if that's retail or FMB or you know (episode 4)

(6) I am so grateful for my wife for my for my son for my job for you know we we lose sight of that we take things for granted and **I have to say** that's something that this job makes me do reminds me to count my blessings (episode 6)

Table 6 lists the rest of the bundle sequences (21 clusters), which are mainly sequences (12) that start with a conjunction followed by a 1st/2nd personal pronoun (*and I think that, and you know what, so I was like*) or a conjunction followed by an indefinite article (*and a lot of*); of these, the most common structures use *and.* This category also includes a single conditional *if* (*if you want to*), and adverbial clusters (*a little bit more, as much as I, a lot of people*).

Table 6. Other types of structures

Structural category	Grammatical pattern	Lexical bundle
Other	Conditional if	• if you want to
	Conjunctions	• and and you know
	Adverbial phrase	• as much as I
		• of the reasons why
		• and I don't know
		• and I think it's
		• and I was like
		• and a lot of
		• and I think that
		• when you look at
		• a lot of people
		• a lot of things
		• that you want to
		• a little bit more
		• and you know what
		• and I want to
		• so I was like
		• a little bit about
		• and you have to
		• and there was a
		• and you know you

(7) um you're not going to be driven **and and you know** you're not going to be able to dedicate that part of your life for others um so that for me is always that dichotomy and that conversation existence what's in it for you what's in (episode 2)

(8) and um it's it's tied or associated with the world economic forum **so I was like** you know this is a great you know way to meet other motivated talented young people work with them, so I joined them in uh and actually it was (episode 3)

5. Discussion

After the AntConc parameters for extracting lexical bundles were set, 56 four-word clusters were identified. Although this corpus is smaller than those used in previous studies, 56 lexical bundles were identified and this can be enough for a lexical bundle list; in terms of size, this is quite similar to those of, for example, Alasmary (2019) 65-list, and Biber et al. (2004) who reports 43 lexical bundles in spoken registers and 84 that appear during classroom teaching. In addition, this finding (the size of this study's list) is consistent with previous findings that spoken registers contain perhaps more multiword expressions than written registers (e.g., Biber et al., 2004; Shin & Nation, 2008). Some bundles in the list developed here also appear in Biber et al.'s (2004) list, including *I would like to*, *you know what I*, *I don't want to*, *to be able to*, and *there's a lot of*, even though Biber et al. (2004) examined spoken academic discourse while the present study looks at spoken non-academic discourse. However, as Shin and Nation (2008) point out to explain the frequency of bundles such as *thank you*, *you know*, and *a bit* (similar to clusters appeared in this analysis), such bundles reflect the personal, interactional, and here-and-now nature of spoken discourse. The structural analyses in this study show that four grammatical categories of lexical bundles, as identified in previous research (e.g., Alasmary, 2019; Biber et al., 2004), also appeared in the corpus (verb-based, nominal, prepositional, and other); verb-based lexical bundles, especially those that begin with the personal pronouns *I* or *you*, occur more frequently than other verb-based structures or nominal and prepositional-based bundles. This aligns with Bieber et al.'s (2004) findings regarding an important variation between spoken and written registers; they find that verbs and pronouns are more common in spoken registers and that nouns are often dominant in written registers. Biber et al. (2004) attribute this difference to the subjective interpersonal purposes of spoken discourse—such as podcasts.

6. Conclusion

In general, the current study has attempted to take lexical bundle research into a less-researched area by examining lexical bundle use in a podcast (spoken non-academic English) produced in a foreign-language context. This study has also identified a list of lexical sequences that may be useful for English-language learners who want to listen to such podcasts to improve their L2 listening comprehension skills. These findings may also be useful for English-language teachers, as this study provides a list of the most common multiword sequences in a general-audience podcast. Nurmukhamedov and Sadler (2011) emphasize that podcasts, in particular, can be very useful during language learning, especially if this is accompanied by handouts containing linguistic information generated from a podcast episode (e.g., vocabulary), whether this is integrated during a listening class or independent learning. Further, in L2 listening research, researchers such as Wakamoto and Rose (2021) suggest that L2 listeners need to practice listening and to look for devices and materials outside the classroom to improve their listening skills (as part of the development of their self-regulation listening strategies)—and podcasts could be

one form of such materials and tools that L2 listeners need to use. In addition, recent research in the field of Global Englishes finds that L2 listeners increasingly need to be exposed to a wider variety of Englishes so that L2 learners can communicate with a wide range of native and non-native English inerlocuotrs (Wakamoto & Rose, 2021)—thus listening to a podcast like the one used in the current study might offer such an opportunity to L2 listeners in the Saudi context. Therefore, there are some pedagogical implications that emerge from this study.

It should be noted, however, that the list included here is not comprehensive. Future studies should examine larger more extensive podcast corpora, which might lead to more representative results. In addition, obtaining expert judgments on the included lexical bundles and more complex statistical measures, as recommended in previous literature, might result in a list that is more useful for teachers and material writers. This study was exploratory in nature and collected data from only one type of a podcast (general audience) and from a podcast produced in a single geographical location. Future studies including other types of podcasts with different contexts may paint a fuller picture of the use of lexical bundles in podcasts. Another possible limitation is that only one podcast was used, and that podcast has one host who might do a significant portion of the talking on it, and there could be some linguistic idiosyncrasies that might also have impacted on the results.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the academic collocation list (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. https://doi.org/10.1016/j.jeap.2013.08.002

Alasmary, A. (2019). Academic lexical bundles in graduate-level math texts: A corpus-based expert-approved list. *Language Teaching Research, 26*(1), 99-123. https://doi.org/10.1177/1362168819877306

Anthony, L. (2022). *AntConc* (Version 4.1.4) [computer software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/ software/antconc/. Accessed 1 December 2022.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English.* Edinburgh: Pearson Education Limited.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245-261. https://doi.org/10.1191/1362168806lr195oa

Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880. https://doi.org/10.1093/applin/amu065

Crawford-Camiciottoli, B. (2007). *The language of business studies lectures*. Amsterdam: John Benjamins. https://doi.org/10.1075/pbns.157

DeCarrico, J., & Nattinger, J. R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes, 7*(2), 91-102. https://doi.org/10.1016/0889–4906(88)90027–0

Durrant, P. (2018). Formulaic language in English for academic purposes. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language* (pp. 211-227). Routledge. https://doi.org/10.4324/9781315206615-12

Durrant, P., & Schmitt, N. (2009). To what extent do native and non–native writers make use of collocations? *IRAL–International Review of Applied Linguistics in Language Teaching, 47*(2), 157-177. https://doi.org/10.1515/iral.2009.007

Ellis, N. C., & Sinclair, S. G. (1996). Working memory in the acquisition of vocabulary and syntax: Putting language in good order. *The Quarterly Journal of Experimental Psychology Section A, 49*(1), 234-250. https://doi.org/10.1080/713755604

Ellis, N. C., Simpson-Vlach, R., Römer, U., O'Donnell, M., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 357–378). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.016

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning, 67*(1), 155-179. https://doi.org/10.1111/lang.12225

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*. Ann Arbor: University of Michigan Press.

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

Islam, M. (Host). (2022–present). *The Mo Show Podcast* [audio and video podcast]. Retrieved from https://www.themopodcast.com/

Jablonkai, R. (2009). "In the light of": a corpus-based analysis of lexical bundles in two EU-related registers. *Corvinus University of Budapest: WopaLP, 3*, 1-26.

Kuiper, K. (1996). *Smooth talkers: The linguistic performance of auctioneers and sportscasters*. Mahwah, New Jersey: Lawrence Erlbaum. http://doi.org/10.1017/S0047404500020054

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites*. Stanford: Stanford University Press.

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography, 24*(1), 29-49. https://doi.org/10.1093/ijl/ecq039

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*(4), 675-724. https://doi.org/10.1111/0023-8333.00143

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition, 43*(5), 985-1008. https://doi.org/10.1017/S0272263121000036

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320. https://doi.org/10.1093/applin/ams010

McEnery, T., & Wilson, A. (1996). *Corpus linguistics: An introduction*. Edinburgh: Edinburgh University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-304. https://doi.org/10.1075/ijcl.11.3.04nes

Nurmukhamedov, U., & Sadler, R. (2011). Podcasts in four categories: Applications to language learning. In B. R. Facer & M. Abdous (Eds.), *Academic podcasting and mobile assisted language learning: Applications and outcomes* (pp. 176-195). Hershey, PA: IGI Global. https://doi.org/10.4018/978-1-60960-141-6.ch011

Nurmukhamedov, U., & Sharakhimov, S. (2021). Corpus-based vocabulary analysis of English podcasts. *RELC Journal, 0*(0), 1-15. https://doi.org/10.1177/0033688220979315

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511497650

Panagiotidis, P. (2021). Podcasts in language learning research review and future perspective. In *Proceedings of EDULEARN21 Conference, 1*(6). https://doi.org/10.21125/edulearn.2021.2227

Pawley, A., & F. Syder (1983). Two puzzles for linguistic theory: Native-like selection and native-like fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication (pp. 191-225)*. London, UK: Longman.

Shin, D. (2007). The high frequency collocations of spoken and written English. *English Teaching, 62*(1), 199-218. https://doi.org/10.15858/engtea.62.1.200703.199

Shin, D., & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal, 62*(4), 339-348. https://doi.org/10.1093/elt/ccm091

Simpson–Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512. https://doi.org/10.1093/applin/amp058

Siyanova-Chanturia, A., & Pellicer-Sánchez, A. (2018). *Understanding formulaic language: A second language acquisition perspective*. London, New York: Routledge. https://doi.org/10.4324/9781315206615

Thorne, S., & Payne, J. (2005). Evolutionary trajectories, internet-mediated expression, and language education. *CALICO Journal, 22*(3), 371-397. https://doi.org/10.1558/cj.v22i3.371-397

Wakamoto, N., & Rose, H. (2021). Learning to listen strategically: Developing a listening comprehension strategies questionnaire for learning English as a global language. *System, 103*, (102670), 1-11. https://doi.org/10.1016/j.system.2021.102670

Wood, D. C., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes, 15*, 1-13. https://doi.org/10.1016/j.jeap.2014.03.002

Wray, A. (2002). *Formulaic Language and the lexicon.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772