Lexical bundles in three oral corpora of university students

Purificación Sánchez Hernández, University of Murcia

Abstract

On the basis of previous lexical bundle studies, this paper examines the forms, structures and functions of 4-word bundles in three corpora of spoken English, one of them of native speakers of English and the remaining two of non-native speakers of English, corresponding to university students in their first year of an English Studies degree and to the same students after two years of university instruction. The study focuses on three major characteristics: the overall distribution of bundles, their typical structures and their functions. The findings show significant differences in the types of lexical bundles used by native and non-native students, as well as in their structure and function.

Our results support the idea that lexical bundles are important components in oral discourse. One of the pedagogical implications of this paper is that Spanish students should be exposed to more samples of spoken language.

1. Introduction

For years linguists have been interested in the study of frequent word combinations. "Phraseology" (Granger & Meunier 2008; Meunier & Granger 2007) and "formulaic sequences/language" (Schmitt 2004; Wray 2000, 2008) are two terms often used to refer to various types of multi-word units. In recent years, an increasing number of studies have made use of corpus data to add weight to the importance of multi-word units in language.

Recurrent word combinations, clusters, phrasicon, n-grams, or lexical bundles refer to word sequences frequently used and retrieved by means of a corpus-driven approach considering criteria of frequency and distribution across the corpus. A lexical bundle is a recurring sequence of three or more words that appears frequently in natural discourse, either oral or written (Biber et al., 1999). These chunks are fundamental parts of discourse whose research is becoming very important in EAP. Cortes (2004) and Hyland (2008b) have studied lexical bundles associated with disciplinary variation, and Biber, Conrad and Cortes (2004) have explored the role of lexical bundles in university teaching and textbooks.

Sánchez Hernández, Purificación. 2013. "Lexical bundles in three oral corpora of university students." *Nordic Journal of English Studies* 13(1): 187-209.

In a series of lexical bundle studies conducted by Biber and colleagues (Biber & Barbieri 2007; Biber & Conrad 1999; Biber, Conrad & Cortes 2003, 2004; Biber, Johansson, Leech, Conrad & Finegan 1999), it was found that conversation and academic prose present distinctive distribution patterns of lexical bundles. For example, most bundles in conversation are clausal, whereas in academic prose they are mainly phrasal. Much of the research published on lexical bundles has been carried out on written English texts. Comparatively, spoken English has not received sufficient attention so far. Biber et al. (2004) carried out a study on lexical bundles in classroom teaching and discussed the implications of their study for the theoretical status of lexical bundles.

This paper adopts an automated frequency-driven approach to identify frequently used word combinations (lexical bundles) in conversation. The study has been carried out on two spoken corpora of English of university students from the University of Murcia (Spain) and consequently compared with another corpus of spoken English collected at the Manchester Metropolitan University. The aim of this paper is to identify and analyse 4-word lexical bundles in the three oral corpora, applying a corpus-driven approach.

2. Background and state of the matter

As previously shown, a lexical bundle is a recurring sequence of three or more words that appears frequently in natural discourse, either oral or written (Biber et al. 1999). Research on these chunks as fundamental parts of discourse is becoming very important in EAP (Altenberg, 1987, Altenberg and Eeg-Olofsson, 1990, Butler, 1997, Biber and Tracy-Ventura, 2007). Lexical bundles associated with disciplinary variation have been studied by Cortes (2004), Hyland (2008b) and author (forthcoming). Biber, Conrad and Cortes (2004) have explored the role of lexical bundles in university teaching and textbooks.

To date, only a few studies have focused on lexical bundles in conversation. De Cock (1998) analysed highly recurring word combinations (HRWCs) in a corpus of spontaneous speak with native speakers and advanced learners of English. McCarthy and Carter (2004) researched multi-word strings in a large corpus of conversational English to identify the most common pragmatically integrated clusters. They discussed their functions and concluded that many clusters are more frequent than single words accepted as belonging to the core vocabulary of English. Biber, Conrad and Cortes (2004) compared the lexical bundles in classroom teaching and textbooks to those found in their previous research on conversation and academic prose, stating that lexical bundles serve as discourse framing devices. Nesi and Basturkmen (2006) focused on the cohesive role of lexical bundles in a corpus of 160 university lectures and reported that the majority of frequently occurring bundles were found to be used to signal discourse relations. Biber and Barbieri (2007) investigated the use of lexical bundles in a wide range of spoken and written university registers, and concluded that lexical bundles are very common in written discourse management in contrast to previous research which showed bundles as being much more common in speech than in writing. Furthermore, Tracy-Ventura, Cortes and Biber (2007) analysed lexical bundles in Spanish speech and writing, and concluded that although lexical bundles are more common in spoken registers than written registers in English, there is a much larger set of lexical bundles used in Spanish academic prose than in spoken interviews. Kim (2009) examined lexical bundles in a large corpus of Korean texts consisting of academic prose and conversation, stating their importance as building blocks in discourse. Csomay and Cortes (2010) investigated the relationship between the discourse functions of lexical bundles found in classroom teaching and their position and showed the existence of a strong relationship between intratextual linguistic variation and the corresponding shift in discourse. Ädel and Erman (2012) have investigated the use of lexical bundles in academic writing by native and non-native speakers, reporting that non-native speakers exhibit a more restricted repertoire of recurrent word combinations than native speakers do.

So far, however, lexical bundles have never been studied in conversation, taking into consideration oral corpora compiled with university students. De Cock (1990) carried out a study similar to ours from a methodological point of view but she focused on the methodology of the study rather than on the results.

3. Research objectives

The main objective of this study is to identify and analyse lexical bundles in conversation across three different corpora of students at university level, so that the findings of this work can be a starting point for establishing their pedagogical implications. I aim to answer the following research questions:

1) What are the most frequent 4-word lexical bundles in conversation in the three corpora involved?
2) Are there important differences from the point of view of their structure between the 4-word lexical bundles used in corpora of native English speakers and Spanish students studying an English degree?
3) What are the functions of the bundles in the three corpora?
4) What are the pedagogical implications of these findings?

Our final goal is to highlight the importance of exposing students of foreign languages to real samples of spoken language.

4. Methodology

4.1 Corpora used for this study

The present study is based on three oral corpora compiled between 2005 and 2007. Our learner data (C1) was collected during 2005 (N= 59, average age = 19.6). The average number of years that these learners had spent studying English before starting university was 8.8. Almost half of the informants had travelled to English-speaking countries, 45.8% spending an average of 1.9 months abroad. All of them (9 male/19 female) were enrolled in the English Studies degree offered by the University of Murcia. The corpus of English speaker language (C2) was compiled using the same structure at the Manchester Metropolitan University, UK[1]. The number of informants (12 male/16 female) in C2 was 28, all of them native speakers of English (average age = 22.25). This corpus was collected in 2006. Corpus 3 (C3) was collected using the same structure as C1 and C2 at Murcia University during 2007 (N= 18, average age = 21.6). The 18 informants (5 male/13 female) were some of the students who had started their degree in 2005 and had been

[1] Further details can be found at http://cecl.fltr.ucl.ac.be/CeclProjects/Lindsei/lindsei.htm#data

interviewed for corpus 1, now repeating the exercise after having completed two years of their English Studies degree.

4.2 Data

Native speakers of English led the interviews for all three corpora. The interviews followed the OPI format of the *Louvain International Database of Spoken English Interlanguage* (LINDSEI) corpus and were divided into three parts. First, speakers were given three topics to choose from: an experience that has taught them an important lesson, a country that has impressed them or a film or play which they particularly enjoyed or disliked. This was the personal narrative component of the interview. A small part of the interview was then devoted to interpersonal communication. Finally, students were given four pictures which told a story and were asked to describe them and offer an account of what was going on. This was the picture description component of the interview.

For this study, and in order to have samples somewhat comparable, we selected 28 interviews[2] from the 59 we had in C1 and made a new C1 with the same number of interviews as C2 in order to maintain an equal number of interviews. Thus, the total number of words in the new C1 (considering only the informants' production) was 24390, and the mean word count was 871.03 per contributor. In the British speakers' corpus (C2), the total number of words considered was 21509, and the mean word count was 796.62 per contributor. In Corpus 3, the total number of words was 18094 and the mean word count per contributor was 1005.22.

After transcription by qualified native speakers of English, the three corpora were tagged at the University of Northern Arizona under the supervision of Prof. Douglas Biber. It was impossible to obtain the exact number of words from each group of speakers or each subject; hence in the final word count we had a few words more in Corpora 1 and 2 than in Corpus 3 (Table 1).

Table 1. Corpora word counts

Corpus C1	Corpus C2	Corpus C3	Total number of words
24390	21509	18094	63983

[2] 18 informants were the students who had been interviewed for corpus 3 after having completed two years of their English Studies degree. The remaining 10 were selected at random.

4.3 Categorisation of lexical bundles

Biber et al. (1999) considered lexical bundles all those word combinations that recurred over 10 times in a million words and were repeated in five or more texts in the Longman Corpus. Later, Cortes (2004), Biber et al. (2004) and Hyland (2008b) established the cut-off point of 20 times per million words for large written corpora, whereas for relatively small spoken corpora a raw cut-off frequency is often used ranging from 2-10 (Altenberg 1998; De Cock 1998). However, the actual cut-off frequency used to identify lexical bundles is somewhat arbitrary.

Our study focused on 4-word bundles because they are more common than 5 or 6-word bundles and offer a wider range of structures and functions than 3-word bundles which, on the other hand, are much more frequent in academic prose (Biber et al., 1999). Moreover, working with 4-word bundles allows us to establish comparisons with other studies of a similar type (Biber & Barbieri 2007; Biber et al. 2004; Cortes 2004; Hyland 2008b). A sequence must be used in at least 3 to 5 different texts to be counted as a lexical bundle (Cortes 2004; Biber & Barbieri 2007). In this context, in the present study the 4-word lexical items must recur in at least 3 texts to be considered a lexical bundle. A smaller number of occurrences could be considered idiosyncratic of the speakers.

A free to use software tool (http://conc.lextutor.ca/tuples/eng/) was used to generate 4-word bundle lists for the texts of each corpus. Some word sequences containing words identifying the students (e.g. *English*, *United Kingdom*) or any other repeated context-dependent bundles were manually excluded from the extracted bundle lists.

The number of types and tokens across the three corpora are shown in Table 2. It is worth mentioning that the lowest number of lexical bundles, both in terms of types and tokens, was registered in corpus 2, collected from British students. The most obvious deduction looking at the types (40) and the tokens (131) in C2 is that native speakers tend to repeat lexical bundles less than non-native speakers of the language.

Table 2. Number of lexical bundles in the three corpora

Corpus	Number of lexical bundles (types)	Number of lexical bundles (tokens)	Type/token ratio
1	44	178	0,24
2	40	131	0,30
3	59	233	0,25

After comparing the frequencies and patterns across the disciplinary corpora, all the bundles were categorised structurally, in terms of their grammatical types, and functionally, according to their meaning in the texts. In this study Biber et al's (2004) classification has been used for the structural and functional analysis, since their study was carried out considering oral and written samples. According to this classification, there are three main structural types: a) lexical bundles that incorporate verb phrase fragments; b) lexical bundles that incorporate dependent clause fragments; and c) lexical bundles that incorporate noun phrase and prepositional phrase fragments. The different types and subtypes are listed in Tables 5 and 6.

5. Results and discussion

5.1 Lexical bundles in our corpora

As shown in Table 2, we found 44 different lexical bundles in C1, 40 in C2 and 59 in C3, totalling 178, 131 and 233 individual cases respectively, which accounts for 0.72% of the total words in C1, 0.61% in C2 and 1.28% in C3. Notably, C1 and C3 have 9 lexical bundles in common; however, none of them can be found in C2.

I like very much, *In the first picture* and *I I don't know* were the most frequent lexical bundles in C1, C2 and C3 respectively. Surprisingly, there are no lexical bundles common to all three corpora. However, as shown in Table 3, some coincidences exist between C1 and C3 (both corpora of non-native speakers of English), which share 9 lexical bundles (in bold), and between C2 and C3 which share just 1 (underlined). Our results do not coincide with those reported by Chen and Baker (2010) who found several bundles common to three corpora of native and non-native academic writing.

Table 3. 40 of the most common lexical bundles in the three corpora

Corpus 1	Freq.	**Corpus 2**	Freq.	**Corpus 3**	Freq.
i like very much	28	in the first picture	6	i i don't know	9
are a lot of	8	the third picture	4	**i would like to**	9
there are a lot	8	very happy with it	4	**or something like**	8
in the in the	7	my mum and dad	4	**that**	
and i don't know	7	i've been to france	4	don't know how to	7
i don't know i	6	in the morning and	4	**i don't know how**	7
like it very much	6	i thought it was	4	in the first picture	7
i don't know what	6	in the fourth picture	4	i think it was	6
or something like	6	and i would say	4	**i don't know i**	5
that		country that i've	4	**i don't know the**	5
i want to go	6	visited		in the second one	5
how do you say	5	it was a bit	3	**it's not the same**	5
a lot of things	5	she seems to be	3	i don't know if	5
i i want to	5	i think in the	3	**how do you say**	4
i like it very	5	to go to the	3	i don't know and	4
i don't know the	4	it was really good	3	i don't know it's	4
go to the cinema	4	i'd like to go	3	it's a it's a	4
no i don't know	4	was a bit strange	3	in in in the	4
don't know what to	4	don't think i could	3	the second time i	4
i don't know how	4	i don't think i	3	when i was there	4
mm i don't know	4	in the u k	3	know how to say	4
i go to the	4	met a lot of	3	mm i don't know	4
a lot of english	4	quite a few times	3	in the in the	4
know what to do	3	doesn't look very	3	would like to to	3
do you say that	3	happy		i was there i	3
it's very beautiful	3	look very happy with	3	don't know if i	3
and	3	happy with what she	3	i i really like	3
. and the last one	3	it's a lot more	3	to go to the	3
i was in a	3	the the the the	3	i think it's a	3
want to go there	3	it looks like he's	3	**a lot of things**	3
was going to be	3	and then in the	3	a lot of people	3
i don't know but	3	and things like that	3	i don't know	3
for me it was	3	she's showing her	3	because	
it's not the same	3	friends		a portrait of a	3
similar to spanish		o'clock in the	3	in the second	3
people	3	morning		picture	
the next city we	3	with what she sees	3	she is showing the	3
in the third one	3	it to her friends	3	so we had to	3
i would like to	3	showing it to her	3	**i don't know what**	3
doesn't want to		it was very different	3	there is a a	3
continue	3	i want to go	3	i don't know mm	3
with a lot of	3	and it was really	3	in in the first	3
i went to england	3	she's showing it to	3	i i would like	3
he doesn't want to		begins to draw her	3		

5.2 Structure of bundles

As shown in Tables 4 and 6, the bundles were categorised according to their structure and function (Biber 2004: 381, 384) From the structural point of view the three corpora offer significant differences in terms of the types of bundles used and in terms of percentages (Tables 4 and 5).

Table 4. Raw percentages of structural types in C1, C2 and C3

Structural types	**C1**		**C2**		**C3**	
	Tokens	*%*	Tokens	*%*	Tokens	*%*
1. Lexical bundles that incorporate verb phrase fragments	27	*61.31*	21	*52.50*	32	*54.23*
2. Lexical bundles that incorporate dependent clause fragments	6	*13.63*	4	*10.00*	11	*18.64*
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments	11	*25.00*	15	*37.50*	16	*27.13*
Total	44	*100*	40	*100*	59	*100*

As the results also indicate, there are important differences in the structural types of bundles used in the three corpora. The figures reveal that, in conversation, the highest percentages of lexical bundles incorporate verb phrase fragments (61.31%, 52.50% and 54.24 % respectively in C1, C2 and C3), whereas Biber et al. (2004: 380) report that 90% of the lexical bundles incorporated verb phrase fragments in their study. The informants in Corpus 1 use a much higher percentage of “Lexical bundles that incorporate verb phrase fragments” than those in corpora 2 and 3 which share similar rates of use. However, this trend changes in the use of “Lexical bundles that incorporate dependent clause fragments”, because the percentages of the Spanish speakers of C1 are closer to native speakers of English (C2) than to C3. Similar percentages of use are shared by C1 and C3 in the employment of “Lexical bundles that incorporate noun phrase and prepositional phrase fragments”.

No statistically significant differences were found when analyzing the results shown in Tables 4 and 5. In Table 4 three chi-squared tests were performed, juxtaposing the results of C1 and C2 (*p-value = 0.454*); C2 and C3 (*p-value = 0.722*); C1 and C3 (*p-value = 0,366*). In the case of Table 5, the three structural types were subjected to individual statistical analysis: a chi-squared test (*p-value = 0.532*) was used for “Lexical bundles that incorporate verb phrase fragment”. Another chi-squared test revealed no statistically significant differences for “Lexical bundles that incorporate dependent clause fragments" (*p-value = 0.829*). Finally, a last chi-squared test was used for "Lexical bundles that incorporate noun phrase and prepositional phrase fragments" (*p-value = 0.196*).

For comparative purposes, we will consider Corpus 2 as the control corpus, since it was collected from native speakers of English and lexical bundles are considered expressions "universally presented as typically native-like"[3] (Granger, 1998).

A more detailed analysis of these results reveals that, as shown in Table 5, the first structural category, "Lexical bundles that incorporate verb phrase fragments**"**, registers the highest number of occurrences with respect to the other categories. The percentage values reveal that there are no occurrences in the categories Discourse marker+VP fragment", Verb phrase (with passive verb) and Yes-no question fragments in any of the three corpora. This finding seems consistent with the type of interviews carried out where students had to speak about personal experiences and tell a story by describing a series of pictures. However, there are important differences in the use of 1st/2nd person pronoun + VP fragment since 28.57% of the bundles used by native speakers of English fall in this category, whereas in the case of the native speakers of C1 and C3 the bundles amount to 48.92% and 43.75% respectively. These data in corpora 1 and 3 are similar to those described by Biber et al. (2004: 380) who report that approximately 50% of these lexical bundles begin with a personal pronoun+verb phrase. Surprisingly, in our corpora, the Spanish informants followed this trend, whereas the native speakers differed strikingly from such finding. Hence, it seems more likely possible that the less instruction there is on language, the greater the use of personal pronouns in oral discourse. As for the category 3rd person pronoun +VP fragment, the informants of C2 exhibit the highest percentage of use (38.09%), followed by those interviewed in C3 (18.75%) and the informants of C1 (14.81%). This could be an indicator that the use of the 3rd person pronoun + VP fragment increases after instruction and resembles the way native informants use this grammatical category.

[3] I am aware of the implications of ELF paradigm for EAP research (Björkman, 2011). However, one of the most important issues for EAP instruction is the needs and expectations of the specific group. C1 and C3 informants are enrolled in the English Studies degree.

Table 5. Detailed percentages of structural types in C1, C2 and C3 (actual numbers of occurrences in brackets)

Structural types	**Subtypes**	**C1**	**C2**	**C3**
1. Lexical bundles that incorporate verb phrase fragments	1a. 1st/2nd person pronoun + VP fragment	(13) 48.92%	(6) 28.57%	(14) 43.75%
	1b. 3rd person pronoun + VP fragment	(4) 14.81%	(8) 38.09%	(6) 18.75%
	1c. Discourse marker + VP fragment	–	–	–
	1d. Verb phrase (with non-passive verb)	(9) 33.33%	(7) 33.33%	(11) 34.37%
	1e. Verb phrase (with passive verb)	–	–	–
	1f. Yes-no question fragments	–	–	–
	1g. WH-question fragments	(1) 3.70%	–	(1) 3.12%
2. Lexical bundles that incorporate dependent clause fragments	2a. 1st/2nd person pronoun+dependent clause/fragment	(5) 83.33%	(3) 75%	(9) 81.81%
	2b. WH-clause fragments	–	–	–
	2c. If-clause fragments	–	–	(1) 9.09%
	2d. To-clause fragment	(1) 16.67%	(1) 25%	(1) 9.09%
	2e. That-clause fragment	–	–	–
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments	3a. Noun phrase with of-phrase fragment	(3) 27.23%	–	(4) 25%
	3b. Noun phrase with other post-modifier fragment	–	(1) 6.66%	–
	3c. Other noun phrase expressions	(4) 36.40%	(6) 40%	(3) 18.75%
	3d. Prepositional phrase expressions	(3) 27.23%	(8) 53.33%	(9) 56.25%
	3e. Comparative expressions	(1) 9.04%	–	–

It should be highlighted that the addition of the categories 1st/2nd person pronoun + VP fragment and 3rd person pronoun + VP fragment of the three corpora show similar results: non-native students concentrate on the 1st/2nd person pronoun +VP fragment, whereas native students do not rely as much on their personal experiences, as illustrated below:

> .. the ... how to change lives and how your like and .. *and I don't know* and make you .. well the poverty to .. to make us richer ... and it's quite unfair .. (C1)
>
> .. *I don't think I* have a .. specific type of film I like .. like a different range .. of films erm (¿?) ... like probably more: .. action films are my favourite and then like ... (C2)
>
> your .. letter writing you know at your home erm .. *I don't know I* don't know what to do very long (C3)

As expected, and in good agreement with the results described in the previous category, in the second group, including "Lexical bundles that incorporate dependent clause fragments**"**, some subcategories were absent, namely WH.clause fragments and That-clause fragment. The subcategory If-clause fragments registered only 1 occurrence, and To-clause fragment showed only 1 occurrence in C2 and C3 respectively. In contrast, 1st/2nd person pronoun+dependent clause/fragment showed the highest percentages of use in the three corpora, with 5, 3 and 9 occurrences respectively, suggesting a trend of relying on the use of more personal pronouns, as the students are less proficient in the use of the language. These results are consistent with the nature of the interviews and also with the idea that speakers in general and especially non-proficient speakers tend to use personal pronouns focusing the information on their own world and experiences as previously stated.

> the first year you go out all night .. and then you .. you are bored your well *I want to be* pained (C1)
>
> come back to Manchester for your final year and your like oh god *I want to go* abroad again .. but yeah no definately definately go back to both of those places ... (C2)
>
> .. er (que es er?) he told me okay *if you want to* be .. a lecturer I know how to English is that saying like have you (C3)

The last structural category, comprising "Lexical bundles that incorporate noun phrase and prepositional phrase fragments**"**, registers the highest number of occurrences with respect to the other categories (Table 5). The subcategory Noun phrase with of-phrase fragment reveals high percentages in C1 (27.23%), 25% in C3 and none in C2, which means that only non-native speakers of the language use it. However, the most common structure for C2 and C3 is Prepositional phrase expressions (53% and 56.25% respectively), which registers only 27.23% of the occurrences in C1. This structure is commonly used to show logical relationships between prepositional elements. It should be noted that the structure Noun phrase with of-phrase fragments is one of the most commonly used in academic prose as reported by Biber et al. (2004: 282) who state that "this structure accounts for 70% of the common bundles in academic prose" and also by Hyland (2008b: 10) who informs that "this expression comprises about a quarter of all forms in his corpora of academic texts". However, in our study, this structure accounts for almost 10% of bundles in C1 and a bit less in C3, whereas native speakers of English in C2 do not use it. This fact lends support to the idea that the spoken production of Spanish speakers shares some characteristics of written language. It would seem that their foreign language instruction may have been based on grammar rules rather than colloquial speech.

> also you you know *a lot of people* and that area a good thing .. you know people from .. (C1)

> or ... that's what makes .. them funny *and things like that* and then that builds up and ... (C2)

> she isn´t interested in the media *or something like that* .. and all .. her classmates .. er makes (C3)

The remaining categories show no occurrences or minimal ones, as in the case of Noun phrase with other post-modifier fragment with only 6.66% in C2.

> at the picture. .and I don't think she's.. maybe not very *happy with what she* sees in it (C2)

5.3 Functions of bundles

As Table 6 indicates, no important differences were found among the functional categories across corpora. However, one prominent feature was the greater concentration of stance expressions in the three corpora, amounting to 62%, 53.9% and 66.1% in corpus 1, 2 and 3 respectively. Such results are in good agreement with the findings of Biber et al. (2004) and Biber and Barbieri (2007) who reported that stance bundles account for over 60% in conversation. There were no occurrences in "Special conversational functions" in any of the three corpora and the category "Discourse organizers" exhibited almost three times more occurrences in C1 than in C2 or C3. Regarding "Referential expressions", the results seem to indicate that the tendency to use them increases with instruction, since the informants of C3 use them more than those of C1, although they fail to reach the percentages which correspond to the native speakers of English.

Discourse organizers and referential expressions are considerably less common than stance bundles, which is consistent with other studies on these types of expressions (Biber & Barbieri 2007).

Table 6. Percentages of use of functional types of lexical bundles across corpora (Biber 2004)

Functional types of lexical bundles	**C1**		**C2**		**C3**	
	Tokens	*%*	Tokens	*%*	Tokens	*%*
I. Stance expressions	28	*63.56*	23	*57.50*	38	*64.28*
II. Discourse organizers	5	*11.36*	2	*5.00*	3	*5.07*
III. Referential expressions	11	*24.97*	15	*37.50*	18	*30.41*
IV. Special conversational functions	–	–	–	–	–	–
Total	44	*100*	40	*100*	59	*100*

If we compare our results with those reported by Biber, Conrad and Cortes (2004) we can see that our percentages of Stance expressions (63.56, 57.50 and 64.28% for C1, C2 and C3 respectively) are similar to their results in conversation (69.05%). However, regarding Discourse organizers, the results of C1 (11.36%) are closer to those described by these authors in textbooks (11.11%) and in C2 and C3, our percentages (5.00% and 5.07% respectively) are similar to those reported for academic prose (5,26%).

With respect to the third category, Referential expressions, the results of C2 (37.50%) are similar to those found by Biber et al (2004) in classroom teaching (38.09%) while the results of C1 and C3 follow the same trend (24.97% for C1 and 30.41% for C3) after two years of instruction at University.

Summarizing the results shown in Table 4 we could conclude that the use of Stance expressions in the 3 corpora is characteristic of the conversation register. The use of Referential expressions in C2 is distinctive of the classroom teaching register, whereas in C3, the percentages approximate those of C2. It seems that the use of referential expressions increases with instruction.

No statistically significant differences were found in Table 6; a chi-squared test was applied, obtaining a *p-value = 0.557*. In Table 7, the three functional types were subjected to individual statistical analysis: a chi-squared test performed for “Stance expressions”, associating results in subtypes A and B (gathering, therefore, sub-subtypes “B1” to “B6” in “B”), showed statistically significant differences (*p-value = 0.0039*).

Nevertheless, no statistically significant differences were found in the two other functional types, either when performing a Z-Test for proportions (confidence level 95%) for “Discourse organizers”, or when applying a chi-square test (*p-value* = 0.152) for “Referential expressions”.

Details pertaining to the percentages allocated to the different subcategories of functional types of bundles are shown in Table 7.

Table 7 is based on the categories proposed by Biber et al. (2004). However, in the functional type “Stance expressions” and under the category Attitudinal/modality stance, we have identified three more subcategories that were not present in Biber et al’s classifications, namely *opinion, like/dislike* and *description.*

“Stance Expressions” provide a framework for the interpretation of the following proposition. Epistemic stance bundles focus on the knowledge status of the information and attitudinal bundles express speaker attitudes (Biber 2004: 389). When considering the percentages of “Stance expressions”, one of the most striking differences among the corpora is the high percentage of opinion bundles (both personal and impersonal) in C2 (56.64%), and the low percentages in C1 and C3 (3.56% and 5.30% respectively). This may be due to the fact that giving opinions requires a more elaborate use of language.

Table 7. Detailed percentages of functional types in C1, C2 and C3 (adapted from Biber et al. 2004). (Actual numbers of occurrences in brackets)

Functional types	**Subtypes**	**C1**	**C2**	**C3**
I. Stance expressions	A. Epistemic stance			
	- Personal	(11) 39.65%	(4) 16.94%	(23) 60.52%
	- Impersonal	–	–	–
	B. Attitudinal/modality stance			
	B1) Desire			
	- Personal	(6) 21.46%	(2) 8.60%	(5) 13.65%
	- Impersonal	–	–	–
	B2) Obligation/directive			
	- Personal	–	–	(1) 2.65%
	- Impersonal	–	–	–
	B3) Intention/prediction			
	- Personal	(3) 10.71%	(2) 8.60%	(2) 5.30%
	- Impersonal	–	–	–
	B4) Opinion			
	- Personal	–	(10) 43.60%	(1) 2.65%
	- Impersonal	(1) 3.56%	(3) 13.04%	(1) 2.65%
	B5) Like/dislike	(3) 10.71%	–	(2) 5.30%
	B6) Description	(4) 14.28%	(2) 8.60%	(3) 7.95%
II. Discourse organizers	A. Topic introduction/ focus	–	–	–
	B. Topic elaboration/ clarification	(5) 100%	(2) 100%	(3) 100%
III. Referential expressions	A. Identification/focus		(1) 6.66%	(2) 11.11%
	B. Imprecision	(3) 27.7%	(1) 6.66%	(1) 5.55%
	C. Specification of attributes			
	C1) Quantity specification	(6) 54.54%	(5) 33.30%	(4) 22.20%
	C2) Tangible framing attributes			
	C3) Intangible framing attributes			
	D. Time/place/text reference			
	D1) Place reference	(1) 9.09%	(2) 13.32%	(3) 16.65%
	D2) Time reference		(3) 19.98%	(1) 5.55%
	D3) Multi-functional reference	(1) 9.09%	(3) 19.98%	(7) 38.85%

The remaining subcategories also show some differences. The most relevant is the low percentage of personal involvement of C2 (16.94%) in comparison with C1 (39.65%) and C3 (60.52%). Another feature of C2 is the even distribution of percentages in the remaining subcategories of Attitudinal modality stance (8.60% *desire*, *intention* and *description*, respectively) in contrast with the irregular percentages allocated to the subcategories in C1 and C3, where *desire* totals 21.46% and 13.25% respectively, *intention* 10.71 and 5.30, and *description* 14.28 and 7.95. Surprisingly, there are no occurrences in the subcategory *like/dislike* in C2, whereas this category registers 10.71 and 5.30 in C1 and C3 respectively.

> erm one book .. erm ... erm .. Shakespeare ... poesía poem yeah poems yeah *I don’t know the name* (C1)

> Em..it was really hot when I went to Paris..I think it was the hottest day they’d had for about.. twenty years (C2)

> which she doesn't *I don't know why* cos she with the effect of the reality I don't know (C3)

“Discourse organizing bundles” serve the functions of Topic introduction/focus and Topic elaboration/clarification. In our corpora, with respect to the “Discourse Organizers**”,** there were no occurrences in Topic introduction/focus. All 100% of the bundles take place in Topic elaboration/clarification, albeit with few occurrences.

“Referential bundles” usually identify an entity or highlight some particular attribute as especially important (Biber 2004: 393). In the percentages allocated to **“**Referential expressions**”**, a noteworthy feature is the high presence of Imprecision in C1 (27.7%) in comparison with 6.66% and 5.55% in corpora 2 and 3.

The occurrences of C1 take place in Specification of attributes: Quantity specification (54.54%), Imprecision (27.7%) and Time/place/text reference with Place and Multi-functional reference (9.09%). However, in C2, the percentages are distributed among Identification (6.66%), Imprecision (6.66%), Quantity specification (33.30%), Place reference (13.32%), Time reference (19.98%) and Multi-functional reference (19.98%). In C3 there are occurrences in all categories and the percentages of C3 are more similar to C2 than to C1, which means that the use of referential expressions has improved with instruction.

> and h=here there is a lot of .. th=*there are a lot of* .. tourists only tourists and museums .. em .. (C1)
>
> take your own lessons but here we're told what we're gonna study over there *it's a lot more* relaxed like I said (C2)
>
> erm .. er for example in London er there are *a lot of things* around London and I think (C3)

Perhaps the most important finding resulting from the analysis and comparison of the functional bundles in the three corpora is the evolution that can be seen in the use of "Discourse organizers" and "Referential expressions" by C3 informants. This finding reflects the importance of instruction in the use of bundles.

6. Conclusions and pedagogical implications

The main objective of this paper was to identify and analyse 4-word lexical bundles in three oral corpora, applying a corpus-driven approach. We have shown the overall distribution of such lexical bundles and their typical structures and functions in the three corpora from native English speakers and students of English of a similar age range and education.

As has been shown in this paper C1 and C3 (corpora of non-native speakers of English) offer a larger number of lexical bundles than C2 (native speakers of English) contrary to the results reported by Chen & Baker (2010), and Ädel and Erman (2012) for academic writing. There are important differences in the structural types of bundles used in the three corpora, the lexical bundles which incorporate verb phrase being those which register the highest percentages in the three corpora.

Regarding the functional types of lexical bundles, one of the most prominent features is the greater concentration of stance expressions in the three corpora, which coincides with the results of other researchers. As we have shown, the Discourse organizers bundles share more features with written than with oral registers as described by Biber et al (2004). C1 exhibits bundles similar to those likely to appear in textbooks, whereas the bundles analysed in C2 and C3 are more similar to those found in academic prose by the same researchers. With respect to the

Referential expressions, the results show that their use more resembles classroom teaching than conversation in C2; as can be seen, use of referential expressions increases with instruction, so that the percentages found in C3 are more similar to those registered in C2 than those described in C1. The type of interview carried out may explain the results. There was no proper conversation in the sense that there was no dialogue, since the interviewer was only allowed to ask a few questions and elicit conversation. This could be the reason why the informants made use referential expressions in a way similar to that described in classroom teaching, which is an intermediate register between oral and written.

Building on previous studies of lexical bundles (Biber et al. 2004; Cortes 2004; Hyland 2008b), the aim was to highlight the pedagogical implications of teaching lexical bundles to students of English by showing the differences between the samples collected from native students of English and learners of English. Biber and Barbieri (2007) suggest that, as these formulaic expressions are so frequent, we might assume that students will naturally acquire them and, consequently, that there is no need for them to be overtly taught. However, it is necessary to expose the students to more samples of spoken language in all environments and not only to instructional approaches. The findings of this study show that even though students might have frequently encounter these expressions in their classes, simple exposure to the frequent use of lexical bundles does not result in the acquisition and mastery of these expressions by university students.

I am aware of the difficulty in introducing lexical bundles effectively in L2 teaching curricula. Lewis (1993), Nattinger and DeCarrico (1992) and Willis (1990) proposed three major pedagogical frameworks, reviewed by Wray (2000) who found them all inadequate to some extent. Following Nation (2009), Byrd and Coxhead (2010) suggest that teachers should draw attention to bundles in class readings/class materials and propose that some explicit instruction should be provided. Then, after the instruction, keeping track of the bundles presented and studied in the classroom is also of paramount importance. Coxhead (2004) proposes the use of vocabulary boxes, Nation (2001) and Schmitt (2000) recommend vocabulary notebooks. Revisiting bundles regularly and creating opportunities for feedback (Webb, 2007) are also important techniques for the students to acquire them.

However, there are still two key issues in the teaching and learning of lexical bundles: the selection of the bundles to be taught and the activities to be used. More research should be done on the criterion for the selection of the bundles; most studies adopt the criterion of frequency when selecting the bundles to be taught; however, their function in discourse could also be a good factor to be taken into account. The sequencing of activities used to teach lexical in another point to be considered. Further attention should be drawn on these key points.

Acknowledgements
I am deeply indebted to the two anonymous reviewers for the valuable comments and suggestions on the manuscript.

References

Ädel, Annelie and Britt Erman. 2012. "Recurrent Word Combinations in Academic Writing by Native and Non-native Speakers of English: A Lexical Bundles Approach." *English for Specific Purposes* 31: 81-92.

Altenberg**,** Bengt. 1987. "Causal Ordering Strategies in English Conversation." *Grammar in the Construction of Texts*. Ed. James Monaghan. London: Frances Pinter. 50-64.

Altenberg, Bengt. 1998. "On the Phraseology of Spoken English: The Evidence of Recurrent Word-combinations." *Phraseology: Theory, Analysis and Applications*. Ed. Anthony Paul Cowie. Oxford: Oxford University Press. 101-122.

Altenberg, Bengt and Mats Eeg-Olofsson. 1990. "Phraseology in Spoken English." *Theory and Practice in Corpus Linguistics*. Eds. Jan Aarts and Willem Meijs. Amsterdam: Rodopi. 1-26.

Biber, Douglas and Federica Barbieri. 2007. "Lexical Bundles in University Spoken and Written Registers." *English for Specific Purposes* 26: 263-286.

Biber, Douglas and Susan Conrad. 1999. "Lexical Bundles in Conversation and Academic Prose." *Out of Corpora: Studies in Honor of Stig Johansson.* Eds. Hilde Hasselgard and Signe Oksfjell. Amsterdam: Rodopi. 181-9.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2003. "Lexical Bundles in Speech and Writing: An Initial Taxonomy." *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*. Eds. Andrew Wilson, Paul Rayson, and Tony McEnery. Frankfurt/Main: Peter Lang. 71-92.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. "If you look at… Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics* 25(3): 371-405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas and Nicole Tracy-Ventura. 2007. "Dimensions of Register Variation in Spanish." *Working with Spanish Corpora*. Eds. Giovanni Parodi. London: Continuum. 87-152.

Björkman, Beyza. 2011. "English as a Lingua Franca in Higher Education: Implications for EAP." *Ibérica* 22: 79-100.

Butler, Chris. 1997. "Repeated Word Combinations in Spoken and Written Text: Some Implications for Functional Grammar." *A Fund of Ideas: Recent Development in Functional Grammar*. Eds. C. Butler, J.Connolly, R. Gatwards, & M. Wismans. Amsterdam: Institute for Functional Research into Language and Language Use. 60-77.

Byrd, Pat and Averil Coxhead. 2010. "*On the other hand*: Lexical Bundles in Academic Writing and in the Teaching of EAP." *University of Sydney Papers in TESOL* 5: 31-64.

Cortes, Viviana. 2004. "Lexical Bundles in Published and Student Disciplinary Writing: Examples From History and Biology." *English for Specific Purposes* 23: 397-423.

Coxhead, Averil. 2004. "Using a Class Vocabulary Box: How, Why, When, Where, and Who." *RELC Guidelines* 26(2): 19-23.

Chen, Yu-Hua and Paul Baker. 2010. "Lexical Bundles in L1 and L2 Academic Writing." *Language Learning and Technology* 14(2): 30-49.

Csomay, Enico and Viviana Cortes. 2010. "Lexical Bundle Distribution in University Classroom Talk." *Corpus Linguistic Applications: Current Studies, New Directions*. Eds. Gries, Stefan, Stefanie Wulff and Mark Davies. Amsterdam/New York: Rodopi V.B. 153-168.

De Cock, Sylvie. 1998. "A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-native Speakers of English." *International Journal of Corpus Linguistics* 3(1*)*: 59-80.

Granger Sylviane. 1998. "Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae." *Phraseology: Theory, Analysis and Applications*. Ed. Anthony P. Cowie. Oxford: OUP. 145-160.

Granger, Sylviane and Fanny. Meunier. 2008. "Phraseology in Language Learning and Teaching. Where to From Here?" *Phraseology in Foreign Language Learning and Teaching*. Eds. Sylviane Granger and Fanny Meunier: John Benjamins. 247-252.

Hyland, Ken. 2005. "Stance and Engagement: A Model of Interaction in Academic Discourse." *Discourse Studies* 7(2): 173-191.

Hyland, Ken. 2008a. "Academic Clusters: Text Patterning in Published and Postgraduate Writing." *International Journal of Applied Linguistics* 18(1): 41-62.

Hyland, Ken. 2008b. "As can be seen: Lexical Bundles and Disciplinary Variation." *English for Specific Purposes* 27(1): 4-21.

Jablonkai, Reka. 2010. "English in the Context of European Integration: A Corpus-driven Analysis of Lexical Bundles in English EU Documents." *English for Specific Purposes* 29: 253-267.

Kim, YouJin**.** 2009. "Korean Lexical Bundles in Conversation and Academic Texts." *Corpora* 4:135-165.

Lewis, Michael. 1993. *The Lexical Approach.* Hove UK: Teacher Training Publications.

McCarthy, Michael and Ronald Carter. 2004. "This that and the other: Multi-word Clusters in Spoken English as Visible Patterns of Interaction." *Teanga: The Irish Yearbook of Applied Linguistics* 21: 30-53.

Meunier, Fanny and Sylviane Granger. 2007. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins Publishing.

Nation, Paul. 2001. *Learning Vocabulary in Another Language.* Cambridge: CUP.

Nation, Paul. 2008. *Teaching Vocabulary: Strategies and Techniques*. Boston: Heinle, Cengage Learning.

Nattinger, James R. and Jeanette S. DeCarrico. 1992. *Lexical Phrases and Language Teaching*. Oxford: OUP.

Nesi, Hilary and Helen Basturkmen. 2006. “Lexical Bundles and Discourse Signalling in Academic Lectures.” *International Journal of Corpus Linguistics* (Special issue on Cohesion) Eds. Michaela Mahlberg and John Flowerdew. 11(3): 147-168.

Reeves, Carol. 2005. *The Language of Science*. New York: Routledge.

Salem, André. 1987. *Pratique des Segments Répetés*. Paris: Institut National de la Langue Francaise.

Schmitt, Norbert. 2000. *Vocabulary in Language Teaching*. Cambridge: CUP.

Schmitt, Norbert. 2004. *Formulaic Sequences: Acquisition, Processing, and Use*. Amsterdam: John Benjamins.

Scott, Michael. 1996. *Wordsmith Tools 4*. Oxford: Oxford University Press.

Sinclair, John McH. 1987. “The Nature of the Evidence.” *Looking Up: An Account of the COBUILD Project in Lexical Computing*. Ed. John McH. Sinclair. London: Collins. 150-159.

Thompson, Geoff. 2001. “Interaction in Academic Writing: Learning to Argue with the Reader.” *Applied Linguistics* 22(1): 58-78.

Tracy-Ventura, Nicole, Viviana Cortes and Douglas Biber. 2007. “Lexical Bundles in Spanish Speech and Writing.” *Working with Spanish corpora. Ed. Giovanni*. Parodi London: Continuum. 354-375.

Webb, Stuart. 2007. “The Effects of Repetition on Vocabulary Knowledge.” *Applied Linguistics* 28(1): 46-65.

Willis, Dave. 1990. *The Lexical Syllabus*. London: Harper Collins.

Wray, Alison. 2000. “Formulaic Sequences in Second Language Teaching: Principle and Practice.” *Applied Linguistics* 21(4): 463-489.

Wray, Alison. 2008. *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.