



Contents lists available at [SciVerse ScienceDirect](http://SciVerse.ScienceDirect.com)

# Journal of English for Academic Purposes

journal homepage: [www.elsevier.com/locate/jeap](http://www.elsevier.com/locate/jeap)



## The purpose of this study is to: Connecting lexical bundles and moves in research article introductions



Viviana Cortes\*

Department of Applied Linguistics and ESL, Georgia State University, United States

### A B S T R A C T

#### Keywords:

Lexical bundles  
Moves  
Academic writing  
Communicative functions

This article presents a group of lexical bundles identified in a corpus of research article introductions as the first step in the analysis of these expressions in the different sections of the research article. A one-million word corpus of research article introductions from various disciplines was compiled and the lexical bundles identified in it were classified grammatically and functionally. The findings of these analyses agreed with previous studies in the most frequent types of grammatical correlates for these bundles and the most frequent functions performed but showed several new qualities for these expressions (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2003, 2004). A further step in the analysis matched these lexical bundles to the moves and steps which are characteristic of research article introductions (Swales, 2004), discovering that a group of lexical bundles were exclusively linked to one move or step in a move while a second group occurred across several moves and steps. In addition, some of these expressions were used to trigger the steps that called for their use while others complemented other expressions and were used as comments.

© 2012 Elsevier Ltd. All rights reserved.

### 1. Introduction

Many studies conducted in the last decade have tried to investigate particular linguistic features that could be used to identify or specify the different stages of a certain genre. In the concrete case of research articles (RAs), for example, several researchers have been focusing in trying to identify a list of words or expressions that could characterize the different rhetorical moves which are considered part of the core organization of the different sections of the RA (Brett, 1994; Kanoksilapatham, 2003; Swales, 1981; Williams, 1999; Yang & Allison, 2003). As Kanoksilapatham (2007) explains, particular texts consist of a series of functional units called *moves* which help fulfill the communicative purposes of a genre. Moves vary in length and they can be the result of the combination of multiple steps. In his seminal study of moves in RA introductions, Swales (1981) presented four moves that help organize authors' communicative functions in these texts: *establishing the field*, *summarizing previous research*, *preparing present research*, and *introducing present research*. In spite of the fact that this was a small scale study based on a very limited number of RA introductions, Swales provided a brief list of words, expressions, and linguistic features that could be used to identify these four moves. Kanoksilapatham (2003) also looked for expressions that could be linked to Swales' move scheme. When analyzing a corpus of RAs in biochemistry and using multi-dimensional analysis, she found a number of linguistic characteristics that could be directly linked to the well-established moves identified in these articles. All in all, despite the strong research tradition that the study of rhetorical moves in RAs in general and

\* 34 Peachtree St. Suite 1200, Atlanta, GA 30303, United States. Tel.: +1 404 413 5195; fax: +1 404 413 5201.  
E-mail address: [vcortes@gsu.edu](mailto:vcortes@gsu.edu).

in RA introductions in particular has helped develop, empirical findings that linked particular linguistic features, single words, collocations, or longer word combinations to moves or steps in moves have been limited. These move-linguistic feature connections could bring about important developments in move-schema theory, because, once identified, these linguistic features could help better describe and illustrate the communicative functions of each move. Henry and Rosenberry (2001) explain that in order to teach a genre effectively, holistic knowledge of a genre needs “to be supplemented by a knowledge of the specific language associated with each move” (p. 155). Thus, linguistic features identified as frequently occurring in a particular move in RAs could become powerful tools that could be introduced to the curriculum of writing classes for students who are trying to get acquainted with the writing of this genre.

The objective of this article is to present a study that started identifying lexical bundles in RA introductions to later analyze the relationship between those bundles and the moves and steps that comprise the organization of these sections. By definition, lexical bundles are combinations of three or more words that frequently occur in a language or a given register (Biber, Johansson, Leech, Conrad, & Finegan, 1999). It is necessary to mention that frequency is the ultimate characteristic that defines lexical bundles. In RAs, frequent four-word lexical bundles are expressions such as *as a result of*, *in the context of*, *the extent to which*, *on the other hand*, and *at the beginning of*, to mention only a few. The number of lexical bundles identified in different corpora decreases with the length of the expressions: Biber et al. found a high number of three, four, and five-word bundles in their corpus study, but they found very few six-word bundles and no longer ones. Most of the literature that studied these expressions in context has focused on four-word bundles. The reason for this preference is that these bundles are much more frequent than five-word bundles and many of them include three-word bundles in their structure (as in the case of *as a result of* that includes *as a result*). In addition, several studies have analyzed the functions of lexical bundles in academic writing (Biber, Conrad, & Cortes, 2003; Cortes, 2004, 2008) explaining that these expressions perform three main functions: stance, discourse organizing, and referential functions, with several subcategories for each of these.

The aim of the study reported here was to identify the most frequent lexical bundles in a corpus of RA introductions from a variety of disciplines, focusing on four-word and longer bundles. After all the bundles were identified, they were analyzed in light of a move scheme adapted from those designed by Swales (1990, 2004) looking for structural and functional connections between the moves and the bundles frequently used in RA introductions. The following sub-sections will shed some light on the two main language features in the study, that is, lexical bundles and rhetorical moves, and a possible relationship between them.

### 1.1. Lexical bundles

Due to the advancements in computers and their use in the analysis of language corpora, there has been a strong shift in the study of formulaic expressions in the last two decades. In the past, the study of formulaic language was performed intuitively, with researchers making up lists of fixed expressions that they perceived as occurring frequently in the language. For example, Pawley and Syder (1983), in their seminal article on the importance of formulaic language to non-native speakers of English, explained that “A few minutes’ reflection produced the following sample of clauses that are familiar to the writers as habitually spoken sequences in Australian and New Zealand English.” (p. 206). The explanation of the authors’ methodology to identify habitual expressions was followed by a long list of short and long expressions such as *Need any help?* *Would you like some more?* *You shouldn’t have said that*, and *You’ve hurt his feelings*, which these authors perceived as frequent formulaic expressions in that geographical register. Lately, studies have favored more empirically-based research methods to identify recurrent expressions in particular registers. Advancements in corpus-based analysis make it possible to empirically identify fixed expressions that recur frequently in the language. These longer expressions, as well as two-word collocations, have often been researched within two traditions: the frequency-based tradition, which emphasizes the frequency of their co-occurrence, and the phraseological tradition, which focuses both on the grade of fixedness that these word combinations hold and on different ways to classify these expressions (Barfield & Gyllstad, 2010). Within the frequency-based tradition, some studies have reviewed the literature on formulaic expressions and checked their frequency in a corpus (Nattinger & DeCarrico, 1992). Other studies have used a strict frequency-driven approach to identify this type of expressions (Altenberg, 1993, 1998; Altenberg & Eeg-Olofsson, 1990). Using this frequency-driven methodology, Biber et al. (1999) coined the term *lexical bundles* to label “the most frequently occurring sequences of words” in a language or register (Biber, 2006, p. 134). Lexical bundles are identified using special software on a large corpus of language. These expressions are groups of three or more words that frequently recur in a language or in a particular register and meet arbitrarily-established cut-off points for frequency and range. For example, Biber et al. (1999) proposed a cut-off point of ten times per million words for a four-word expression to be considered a bundle. In addition, those expressions had to appear in five or more texts in the corpus under analysis, to avoid users’ idiosyncrasies. Succeeding studies have been more conservative and established higher cut-off points to ensure that these expressions are really frequent in a particular register (Cortes, 2004), with frequency cut-off points set at 20 or 40 occurrences per million words. As previously introduced, frequent lexical bundles in academic writing are expressions such as *as a result of*, *in the case of*, and *on the other hand*, among many others.

Lexical bundles have been studied in a wide variety of registers such as everyday conversation (Biber et al., 1999); research articles (Cortes, 2004); university textbooks and lectures (Biber, Conrad, & Cortes, 2004), doctoral dissertations and Master’s theses (Hyland, 2008), and English EU documents (Jablonkai, 2010). These studies tried to discover tendencies in the use of this type of formulaic expressions in different types of texts and some of these investigations introduced different taxonomies to classify these expressions structurally and functionally.

## 1.2. Move analysis

Move analysis was developed by Swales (1981) as part of genre analysis in order to identify and illustrate the rhetorical organization of particular texts (Kanoksilapatham, 2007). Texts are described in terms of their communicative purposes, categorizing the discourse units that make up those texts according to these purposes or rhetorical moves. "A move thus refers to a section of a text that performs a specific communicative function" (Kanoksilapatham, 2007, p. 23). In other words, a move is a stretch of written or spoken discourse, "which achieves a particular purpose within a text" (Henry & Rosenberry, 2001, p. 154). Moves are functional units in a text which, when working together, fulfill the overall communicative purpose of the genre (Connor, Davis, & De Rycker, 1995). The general organization of a text can be described as consisting of a number of moves, with some conventional moves occurring more often than other moves which are optional. Swales (1990) explains that moves often contain multiple elements or *steps* that, when combined, realize the move. The function of these steps is to achieve the purpose of the move to which they belong.

Many studies have proposed various move schema for the analysis of different genres such as book reviews (Motta-Roth, 1998), grant proposals (Connor & Mauranen, 1999), and job application letters (Bhatia, 1993; Henry & Rosenberry, 2001), to mention only some examples. The genre that has been studied through the analysis of move schema most frequently is, undoubtedly, the RA. Some of the studies that have used this methodology for the analysis of the organization of RAs have focused on the identification of moves in the whole RA in different disciplines such as medicine (Li & Ge, 2009; Nwogu, 1991), computer science (Posteguillo, 1999) or biochemistry (Kanoksilapatham, 2003). Some other studies have focused on specific sections of the RA in various disciplines or in a particular discipline (Anthony, 1999; Chu, 1996; Holmes, 1997; Swales, 1981; Yang & Allison, 2003).

The relationship between moves and their communicative functions in scientific texts can be clearly examined in Swales' move structure or framework for RA introductions (Swales, 1981). This scheme, commonly known as the Create a Research Space (CARS) model, showed a sequence of four moves that frequently occur in the introduction section of RAs. In a later study, Swales (1990) reduced the number of moves to three, creating the following framework: Move 1, *Establishing a territory*; Move 2, *Establishing a niche*; and Move 3, *Occupying a niche*. In addition, Swales introduced here a set of steps that can help characterize the internal organization of each move. Finally, Swales (2004) produced an even more comprehensive analysis of move three, identifying a higher number of steps in this move and classifying these steps as obligatory or optional, qualities that are in many cases connected to disciplinary variation.

Early studies that used move schema usually analyzed selected linguistic features by hand, trying to identify recurrent patterns and differences. In their studies of moves in letters of application, Henry and Rosenberry (2001) looked for linguistic markers of move boundaries. Their study analyzed only 40 articles and used a labor-intensive methodology to identify linguistic exponents. In the past five years, however, computers and corpus-based methodologies have been used to identify tendencies in the organization and linguistic conventions related to schema analyses. Kanoksilapatham (2007) reports that computers and language corpora help conduct more interesting analyses. She also explains that even though moves and steps are identified by their functional and semantic purposes, these moves will be realized through a variety of linguistic features. Durrant and Mathews-Aydinli (2011) conducted a study in which they compared formulaic language use and function between essays produced by novice writers and published research articles in similar disciplinary fields. They first annotated students' essays according to the communicative functions in them and then they manually looked for recurrent formulas associated with those functions. Their study made a valid attempt to find the connection between formulaic language and moves and even though they analyzed only 94 papers in each sub-corpus, their findings showed a short list of expression that reflected the move-formula connection. These observations have been part of the impetus for the study presented here, which tried to identify the tendencies in the use of recurrent multi-word units, lexical bundles, in the different moves of RA introductions.

## 1.3. Lexical bundles and moves as building blocks

The identification of lexical items and lexical and grammatical patterns used to signal the onset of rhetorical moves has always been a topic of interest for researchers studying discourse organization. Swales (1981) already identified some lexical items as key markers for moves. For example, he stated that when asserting centrality, a function often expressed in the first move of introductions, authors use expressions that reflect that the issues to be discussed in those articles raise questions of interest using words that denote *interest* or *importance* (p.26). A wide variety of expressions have often been intuitively identified as frequently initiating or defining a move. That is the case of expressions such as *it is the purpose of this paper*, or *the aim of this investigation*, which have been found to introduce the arrival of Move 3 (introducing the present work). Most of these expressions resulted from the analysis of a reduced number of texts or have been perceived as frequently used by authors to convey that communicative function. Identifying lexical bundles in the rhetorical moves of RA introductions would empirically demonstrate which expressions are frequently used to initiate rhetorical moves in this section of RAs.

There is another important concept that brings lexical bundles and rhetorical moves together. Each of these linguistic features has been considered building blocks to be used in the construction of discourse. Lexical bundles have been defined as recurrent expressions that can be retrieved from our memory to be used as "text building blocks" (Biber, Conrad, & Leech, 2002, p. 443). Hyland (2008) also maintains that bundles have been increasingly seen as "important building blocks of coherent discourse and characteristic features of language use in particular settings" (p. 8). In a similar way, rhetorical moves

have also been considered building blocks. Biber, Connor, and Upton (2007) explain that move types can be seen as the “main building blocks” of a genre (p. 53). Dudley-Evans (1995) considered rhetorical moves such an inherent part of a genre that they could be used to teach novice writers how to produce successful texts in that particular genre.

It seems both lexical bundles and moves show similar characteristics: while lexical bundles are seen as lexico-grammatical building blocks associated with basic functions used to bind the text together, rhetorical moves are seen as “segments of discourse that provide the building blocks of texts” (Biber et al., 2007, p.9). A description of the relationship between lexical bundles and moves in a particular register could provide more evidence towards a complete picture of the tendencies used in the organizational and lexico-grammatical patterns used to build discourse by different speech communities. There is undoubtedly a need to further investigate the relationship between these two types of building blocks in different registers.

## 2. Corpus description and methodology

The corpus used for this study was compiled using the introduction sections of RAs from the Published Research Article Corpus (PRAC) (Gray & Cortes, 2010). Even though the scope of this study did not focus on disciplinary differences in the use of lexical bundles in RA introductions, the corpus includes texts that represent writing from various academic disciplines, as shown in Table 1. The introduction sections were identified in the articles and stored separately, after having been cleaned out of any marks which did not belong to the prose of these sections (such as footnotes, superscripted numbers, and page numbers, for example). As introductions tend to be short sections in the RAs of many disciplines, more than a thousand texts extracted from RAs in 56 journals were necessary to reach the one-million word corpus which should be the basic corpus size threshold for the identification of lexical bundles. The average length for the introductions in this corpus was 736 words with a range that went from about 300 to 1300 words.

Lexical bundles were identified in the corpus of introductions using a variation of the Lexical Bundle Program (LBP) (Cortes, 2004). The LBP was modified to look for recurring expressions of four+ words. Different cut-off points were set out for the program to identify those recurrent word combinations as lexical bundles. The conservative cut-off point of 20 times per million words (pmw) was established for four-word bundles and lower cut-off points were established for longer bundles: 10 times pmw for five-word bundles, eight times pmw for six and seven-word bundles, and six times pmw for any longer bundle that could be identified. In addition, the cut-off point for range (the number of texts in which a bundle has to occur) was kept at 5 or more texts in the corpus for all bundle lengths following the trend for range in the literature.

The bundles identified by the LBP were analyzed structurally and functionally using taxonomies previously developed in the literature (Biber et al., 1999; Biber et al., 2003, 2004). Finally, move analysis was conducted to identify the moves in which the bundles occurred. For the purpose of the present study and in order to use all the steps included in the frameworks created by Swales (1990, 2004), both schemes were merged: for Move 1, Establishing a territory, the steps introduced by Swales in 1990 were used, while for Moves 2 and 3, the steps introduced in his 2004 work were selected. The moves and steps used in the present study are shown in Table 2.

The methodology used in this study was bottom-up, that is, my observation was the lexical bundle, which once identified was analyzed in context to discover the communicative function that the lexical bundle was conveying or helping convey. Each of the 3849 tokens of the 135 lexical bundle types identified in this study were analyzed in their context by this researcher in order to establish the rhetorical move and step in which the bundles occurred, carefully studying the communicative purpose of each of these expressions. In addition, and for reliability measures, a second well-trained rater classified 50% of randomly selected occurrences of the bundles identified by the LBP. After this procedure was completed, the percentage of agreement for communicative purpose identification (both moves and steps) between both researchers was

**Table 1**  
Description of the corpus of research article introductions.

Disciplines	Words	Texts
Agronomy (Ag)	58,565	92
Applied Linguistics (Al)	99,097	87
Animal Science (As)	40,210	101
Biology (Bi)	98,962	147
Business (Bu)	86,753	122
Chemistry (Ch)	53,008	94
Civil and Materials Engineering (Ce)	96,585	159
Communication Studies (Co)	76,105	57
Computer Science (Cs)	98,394	144
Economics (Ec)	95,265	92
Physics and astronomy (Pa)	89,051	112
Statistics (St)	63,840	91
Urban and Regional Planning (Up)	46,913	74
Total number of words and texts	1,002,748	1372

**Table 2**

Moves and steps in research article introductions (adapted from Swales, 1990, 2004).

---

Move 1 Establishing a territory
Step 1 Claiming centrality
Step 2 Making topic generalization/s
Step 3 Reviewing items of previous literature
Move 2 Establishing a niche
Step 1A Indicating a gap or
Step 1B Adding to what is known
Step 2 Presenting positive justification
Move 3 Presenting the present work
Step 1 Announcing present research descriptively and/or purposively
Step 2 Presenting research questions or hypotheses
Step 3 Definitional clarifications
Step 4 Summarizing methods
Step 5 Announcing principal outcomes
Step 6 stating the value of the present research
Step 7 Outlining the structure of the paper

---

computed. It is necessary to explain that the corpus used in this study was not tagged for moves or steps. It is undeniable that such a corpus would facilitate the analysis of any type of linguistic or organizational feature. For the purpose of the study presented here, however, it was not necessary to use a corpus tagged for moves as there were many moves in the corpus that did not use lexical bundles. Only those stretches of discourse related by proximity or communicative function to the lexical bundle under analysis were studied for move identification.

### 3. Lexical bundles in RA introductions: structures and functions

This section will introduce the results of the identification and analysis of lexical bundles in RA introductions. Lexical bundles of different lengths will be presented in frequency order followed by their structural and general function classification.

#### 3.1. From four-word to nine-word lexical bundles

The longest bundles identified by Biber et al. (1999) in the academic writing section of the Longman Corpus were six-word expressions such as *from the point of view of* and *in such a way as to*. When analyzing the corpus of introductions used in this study, the LBP yielded a wide variety of bundles longer than six words, recurrent expressions of seven, eight, and nine words that met the pre-established frequency cut-off point criteria to be considered lexical bundles. Appendix A shows all the bundles identified by the LBP grouped by the number of words in the bundle together with their frequency and range. As shown in Appendix A, the corpus analysis yielded five nine-word, one eight-word, six seven-word, four six-word, twenty-nine five-word, and ninety four-word lexical bundles. As expected, as the number of words in the expression increases, the frequency of the bundles decreases. Furthermore, the relationship between the number of four and five-word bundles identified here agreed with previous studies that claimed four-word bundles are much more frequent than five-word and longer bundles (Biber & Conrad, 1999; Cortes, 2004).

It is necessary to point out that even though in many cases a longer bundle contains a shorter bundle, in this study the shorter expression was only included as a lexical bundle when it met the frequency cut-off point as a shorter bundle. For example, the lexical bundle *the paper is organized as follows* occurred 30 times by itself and not as part of longer lexical bundles such as *the rest of the paper is organized as follows* or *the remainder of the paper is organized as follows*. That is why these expressions are considered two different lexical bundles. When the shorter expressions did not meet the cut-off point to be considered lexical bundles, only the longer expressions were identified, counted, and analyzed as bundles. The corpus of introductions yielded longer expressions as lexical bundles, which provides new opportunities to enrich the use of the structural and functional taxonomies that have been frequently used in the literature to classify mostly four-word bundles.

#### 3.2. Classifying bundles structurally and functionally

Structurally, the bundles identified in this study mostly belonged to the three main groups previously defined in the literature (Biber et al., 2004):

- a. Lexical bundles that incorporate noun phrases or prepositional phrase fragments, such as *in the present study*, *the objective of this paper*, *in the next section*, *one of the most important*



- b. Lexical bundles that incorporate verb phrase fragments, in expressions such as *little is known about, is related to the, it has been shown that, it is necessary to, it has been suggested that*
- c. Lexical bundles that incorporate dependent clause fragments, as in *that there is a, was to determine the, we show that the*

Previous studies have emphasized that lexical bundles are often incomplete units (Biber et al., 1999; Biber & Conrad, 1999). The longer bundles identified here, however, were in some cases complete structures, complete clauses, and sometimes even sentences. That is why the following new category was created:

- d. Lexical bundles that include noun phrases and verb phrases (fragments or whole phrases or clauses) in bundles such as *the rest of the paper is organized as follows, and the objective of this study was to evaluate.*

The lexical bundles *the rest of the paper is organized as follows* and *the remainder of the paper is organized as follows* occurred as whole sentences (followed by a period) fourteen and eleven times respectively in the corpus.

The taxonomy created by Cortes (2002) and improved by Biber et al. (2003, 2004) was used to complete a general functional classification of the lexical bundles identified in the corpus of introductions. These bundles could be grouped in three major functional groups:

- a. Stance markers:
  - a.1. Impersonal attitudinal/modality stance bundles in expressions like *it is important to, it is necessary to, there is a need to, can be used to*; and
  - a.2. Impersonal epistemic stance markers in bundles such as *are likely to be, the fact that the*
- b. Discourse organizers:
  - b.1. Topic introduction/focus bundles as in the next section, the rest of the paper is organized as follows, in this study the and
  - b.2. Topic elaboration/clarification in bundles such as *on the other hand, as well as the*
- c. Referential expressions:
  - c.1. Identification/focus bundles like *one of the most, one of the major* and
  - c.2. Specification of attributes referential bundles such as:
    - c.2.1. Framing attributes in expressions such as *the nature of the, in the context of, and*
    - c.2.2. Time, place, or multifunctional referential expressions like *at the end of, or at the same time.*

As with previous uses of this functional taxonomy, the classification of a bundle into a certain category does not mean that the bundle was used exclusively in that function. Lexical bundles were categorized by the function they most often occurred in (Biber et al., 2004).

#### 4. Bundles and moves: from general functions to communicative purposes

After they had been classified structurally and functionally, the bundles identified in the corpus of introductions were analyzed in their contexts, taking into account the communicative purpose of the surrounding discourse to establish the relationship between the bundle and the move and step in which the lexical bundle was used. Table 3 shows all the bundles identified in the corpus, presented in the moves in which they appeared most frequently. The first group of bundles in each step presented in Table 3 (in **bold**) shows the bundles that were identified in their contexts that occurred in only one step in each move. When a lexical bundle appeared in more than one move or step, it was recorded in each of the moves/steps the bundle occurred in. The superscript number next to certain lexical bundles in the next columns for each step in the table indicates the number of steps the bundle was identified in. As shown in Table 3, bundles longer than four words occurred in only one move and step; expressions such as *the remainder of the paper is organized as follows* occurred only in Move 3 Step 7, *outlining the structure of the paper*, as shown in the following example:

Ex. 1. *The remainder of the paper is organized as follows: Section 2 contains a brief description of the class of disparities in general followed by a discussion of the proposed divergences. In Section 3 we discuss ...* (Statistics)

Shorter bundles, on the other hand, occurred in several moves or steps. That is the case of, for example, *in the context of*, which could be identified in four steps across three different moves.

Move 1 step 2, making topic generalizations

Ex. 2. *The effect of the microbiota and its metabolic activities require special consideration when viewed in the context of pig production in which efficient animal growth is a primary objective.* (Animal Science)

Move 2 step 1, indicating a gap/adding to what is known

Ex. 3. *Although third-person perception is a robust phenomenon, it has never been examined in the context of a CD-ROM.* (Communication Studies)

**Table 3**

Lexical bundles in research article introduction moves. The superscript numbers indicate the number of steps in which the bundle was identified.

Move 1	Step 1: Claiming relevance of field	<b>a great deal of,</b> <b>As one of the,</b> <b>In a variety of,</b> <b>is one of the most,</b> <b>one of the major,</b> <b>one of the most important,</b> <b>one of the most,</b> <b>Play an important role in the,</b> <b>The importance of the</b>	a wide range of <sup>3</sup> , a wide variety of <sup>2</sup> , An analysis of the <sup>2</sup> , In the field of <sup>3</sup> , The relationship between the <sup>3</sup>	on the other hand <sup>2</sup> , on the use of the <sup>2</sup> , play a role in <sup>2</sup> , that there is a, the degree to which <sup>3</sup> , the effects of the <sup>2</sup> , the fact that the <sup>2</sup> , the nature of the <sup>2</sup> , the use of the <sup>2</sup> , the ways in which <sup>2</sup> , to the use of <sup>3</sup>
	Step 2: making topic generalizations	<b>a function of the,</b> <b>an understanding of the,</b> <b>as a measure of,</b> <b>as well as in,</b> <b>as well as the,</b> <b>for a variety of,</b> <b>has been shown to,</b> <b>has been shown to be,</b> <b>have been shown to,</b> <b>in the use of,</b> <b>in terms of the,</b> <b>it has been shown that,</b> <b>it is well known that,</b> <b>the development of the,</b> <b>the presence of the,</b> <b>the use of a,</b> <b>there has been a,</b> <b>to the development of,</b> <b>to the extent that</b>	a great deal of <sup>2</sup> , are more likely to <sup>2</sup> , as a result of <sup>2</sup> , at the same time <sup>2</sup> , at the time of <sup>2</sup> , have been shown to be <sup>2</sup> , in a number of ways <sup>2</sup> , in relation to the <sup>2</sup> , in the absence of <sup>2</sup> , in the case of <sup>5</sup> , in the context of <sup>4</sup> , in the field of <sup>3</sup> , in the form of <sup>2</sup> , in the presence of <sup>3</sup> , is based on the <sup>3</sup> , it is difficult to <sup>2</sup> , it is possible to <sup>2</sup> , on the basis of <sup>2</sup> ,	
	Step 3: reviewing items of previous literature	<b>are likely to be,</b> <b>as a result of the,</b> <b>at the end of,</b> <b>in a number of,</b> <b>by the presence of the,</b> <b>in the development of,</b> <b>in the use of the,</b> <b>it has been suggested that,</b> <b>it was found that the,</b> <b>referred to as the,</b> <b>studies have shown that,</b> <b>the impact of the,</b> <b>the size of the,</b> <b>the structure of the,</b> <b>with the use of</b>	a wide variety of <sup>2</sup> , are more likely to <sup>2</sup> , as a function of the <sup>2</sup> , as a result of <sup>2</sup> , as well as in the <sup>3</sup> , at the time of <sup>2</sup> , can be used to <sup>2</sup> , for the first time <sup>3</sup> , have been shown to be <sup>2</sup> , in addition to the <sup>2</sup> , in the absence of <sup>3</sup> , in the case of <sup>5</sup> , in the case of the <sup>2</sup> , in the context of the <sup>2</sup> , in the field of <sup>3</sup> ,	in the form of <sup>2</sup> , in the presence of <sup>3</sup> , is related to the <sup>2</sup> , on the basis of the <sup>2</sup> , play a role in <sup>2</sup> , the fact that the <sup>2</sup> , the value of the <sup>2</sup> , to the use of <sup>3</sup> ,
Move 2	Step 1A: indicating a gap Step 1B: adding to what is known	<b>it is necessary to,</b> <b>it should be noted that,</b> <b>the effect of the,</b>	is known about the <sup>2</sup> , little is known about the <sup>2</sup> , the degree to which <sup>3</sup> , in the context of <sup>4</sup> , it is difficult to <sup>2</sup> ,	the relationship between the <sup>3</sup> , there is a need to <sup>2</sup> , there are a number of <sup>2</sup> , with respect to the <sup>2</sup>
	Step 2: presenting positive justification	<b>a better understanding of,</b> <b>a better understanding of the,</b>	is known about the <sup>2</sup> , little is known about the <sup>2</sup> , there is a need to <sup>2</sup>	
Move 3	Step 1: announcing present research descriptively and/or purposefully	<b>of this study was to,</b> <b>the purpose of the present study was to,</b> <b>the aim of this paper is to,</b> <b>the aim of this study,</b> <b>in this paper we,</b> <b>the objective of this paper is to,</b> <b>the objectives of this study were,</b> <b>the objective of this study was to evaluate the,</b> <b>the purpose of this study is to,</b> <b>the purpose of this paper is to,</b> <b>the purpose of this study was to,</b> <b>to determine the effects of,</b> <b>used in this study,</b> <b>we show that the</b>	an analysis of the <sup>2</sup> , for the first time <sup>3</sup> , in addition to the <sup>2</sup> , in the context of <sup>4</sup> , in the context of the <sup>2</sup> , in the present study <sup>2</sup> , it is possible to <sup>2</sup> , is based on the <sup>3</sup> , is related to the <sup>2</sup> , on the basis of the <sup>2</sup> , on the use of the <sup>2</sup> , the effects of the <sup>2</sup> , the nature of the <sup>2</sup> , the performance of the <sup>2</sup> , the use of the <sup>2</sup> , the ways in which <sup>2</sup> , to the use of <sup>3</sup> ,	

Step 2: presenting R Q's or hypotheses		there are a number of <sup>2</sup> , in the case of <sup>5</sup> ,	
Step 3: definitional clarifications		as well as in the <sup>3</sup> , on the other hand <sup>2</sup> , the degree to which <sup>3</sup> ,	
Step 4: summarizing methods		a wide range of <sup>3</sup> , as a function of the <sup>2</sup> , as well as in the <sup>3</sup> , at the same time <sup>2</sup> , in the absence of <sup>3</sup> , in the case of <sup>5</sup> , in the case of the <sup>2</sup> , in the context of <sup>4</sup> ,	in the presence of <sup>3</sup> , in the present study <sup>2</sup> , on the basis of <sup>2</sup> , is based on the <sup>3</sup> , the relationship between the <sup>3</sup> , the rest of the <sup>2</sup> ,
Step 5: announcing principal outcomes	<b>an increase in the, the results of the,</b>	in the case of <sup>5</sup> , in relation to the <sup>2</sup> , with respect to the <sup>2</sup>	
Step 6: stating the value of the present research	<b>in the sense that the,</b>	a wide range of <sup>3</sup> , can be used to <sup>2</sup> , for the first time <sup>3</sup> , in a number of ways <sup>2</sup> , the value of the <sup>2</sup>	
Step 7: outlining the structure of the paper	<b>are presented in section, at the end of the, in the next section, the paper is organized as follows, this paper is organized as follows, the remainder of the paper is organized as follows, the remainder of this paper is organized as follows, the rest of the paper is organized as follows, the rest of this paper is organized as follow</b>	the performance of the <sup>2</sup> , the rest of the <sup>2</sup> ,	

Move 3 step 1, announcing present research descriptively and/or purposefully

Ex. 4. *In this paper, we examine the same question in the context of arbitrary hierarchies that are consistent with the technology.* (Economics)

Move 3 step 4 summarizing methods

Ex. 5. *The present research tests the two explanations in the context of a real-life case.* (Communication studies)

In order to have more reliable bundle-move connection identification, 50% of the occurrences of the bundles were analyzed and classified into moves and steps by a second researcher. The comparison showed 99% of agreement between researchers for moves and 92% for steps in moves. The main cause of disagreement lied in the fact that some lexical bundles appeared in sentences that may represent more than one step and this was recorded differently by each analyst. When this occurred, the use of the bundle under analysis was revised and the bundle was re-classified as occurring in the two steps that may have caused the mismatch.

The reason for longer bundles (lexical bundles with five or more words) to be used in only one move is simple: longer bundles contain more lexical words within their composition which help them be more informational than shorter bundles, which are mainly made up of function words that surround a central lexical word, such as a noun. The nouns in shorter bundles, however, are in most cases *shell nouns* (Aktas & Cortes, 2008; Schmid, 2000), a type of abstract noun that has little or no meaning in itself, particularly when used in academic discourse. The content of these nouns can often be found in the surrounding discourse, in preceding clauses or in succeeding phrases or noun complement clauses. That is the case of the nouns in lexical bundles such as *the fact that the*, in which the noun *fact* is used as a shell noun, as shown in the following example:

Ex. 6. *The three-fold distinction in the types of publication reflects the fact that the readers of the journals are researchers and/or practitioners.* (Applied Linguistics)

In example 6, the *fact* is realized in the noun complement clause following the noun (the readers of the journals are researchers and/or practitioners). The shell noun *fact* is encapsulating the meaning that is explained in the succeeding discourse.



#### 4.1. The bundle-move connection extended

A new characteristic of lexical bundles was also identified while conducting the bundle–move relationship analysis. Some bundles, particularly those bundles with more than five words, were considered *triggers* in that they started the move and were used at the beginning of the clause that triggered the move or step. Some other bundles, however, appeared in clauses related to a particular move, but they were *complements*, used in the second part of the clauses or phrases identified as belonging to a particular step. Some examples of *triggers* are bundles such as *the purpose of the present study*, *the objective of this study*, *it has been suggested that*, *it has been shown that*, and *little is known about the*, among others, as shown in the following example:

Ex. 7. *Hispanic children who are developing both their first (L1) and second language skills (L2) may also exhibit differences in performance on narrative production and comprehension tasks, which might partially explain their performance on recall and comprehension of written narratives in school. However, little is known about the narrative skills of Spanish-speaking children who are in the process of becoming bilingual in English.* (Applied Linguistics)

The first sentence in this example is summarizing the findings of previous studies that show the state of the field, differences in performances on narrative production and comprehension tasks for children developing their first and second language. The sentence that starts with *however* + the lexical bundle introduces the gap, showing that the field has not yet paid attention to the narrative skills of Spanish-speaking children (which was in fact the objective of the study being reported in that article).

Other lexical bundles were used as *complements*, in sentences that completed the clauses triggering the move, adding a commentary to the main communicative purpose of the discourse used in that particular section. This can be shown in the following example with the bundle *in the sense of*.

Ex. 8. *The objectives here are to determine if hedge funds exhibit performance persistence in the sense that some funds consistently have higher returns than others and to test hypotheses about the source of returns for hedge funds. In addition, the impact of size on performance will be studied.* (Business)

Ex. 9. *Essayist literacy has been widely depicted as decontextualized in the sense that it stands apart from its context of production and in theory supplies the contexts necessary for interpretation within the text itself.* (Applied Linguistics)

In example 8, Move 3 Step 1 (announcing present research descriptively and/or purposefully) is triggered by the expression *the objectives here are to determine*. The bundle *in the sense that* is used in the sentence to introduce a commentary on those objectives. In example 9, the sentence is part of Move 1 Step 2, making topic generalizations, but the lexical bundle *in the sense that* is used as a comment for the trigger of the generalization, indicated here by the topic and the use of the present perfect in the verb phrase.

A preliminary analysis of *triggers* and *complements* showed that all bundles made up of more than five words (six+-word bundles) always performed as *triggers*, initiating a particular step in a move (See Appendix A for trigger/complement classification of five+-word bundles). As previously discussed, these bundles have in many cases various lexical words loaded with information essential for accomplishing a particular communicative purpose, as in the case of bundles such as *the rest of the paper is organized as follows*, or *the objective of this study was to*. Within the five-word bundle group, 20 bundles always acted as *triggers*, 8 acted as *complements* and 1, *on the basis of the*, acted as *trigger* and *complement* (47% of the time as a *trigger* and 53% as a *complement*). A more detailed analysis of this new development in the study of lexical bundle functions, particularly of those bundles used to perform both functions and of shorter bundles, will help complete the functional spectrum performed by these expressions.

It should also be noted that there is a strong relationship between certain bundles and the move/step in which they occur. As shown in Table 3, in Move 1, all steps show a wide variety of lexical bundles that were used in only one step, with many of these bundles used as *triggers* (as in *one of the most important*, and *play an important role in*). Only a few bundles were used exclusively in the Steps within Move 2, with some acting as *triggers* (*there is a need to*, *little is known about the*) and some others acting as *complements* (*with respect to the*, *the effect of the*), which reflects that writers may have used more creative rather than formulaic language to convey the communicative purposes related to this move. In Move 3, steps 1, 5, 6, and 7, showed exclusive bundles, with most of the bundles in this group triggering the corresponding steps. This finding emphasizes the existence of language pervasiveness in this genre. The writers of these papers made frequent use of formulas that convey particular communicative purposes. Flowerdew and Li (2007) discussed the language re-use phenomenon, explaining that it is common practice in scientific discourse. The direct extraction of common phrases from published work is normal in some fields such as biomedical science. These authors commented that “scientific writing is given to formulaicity” and that this phenomenon can apply from one word to a phrase and from a single step or move to the whole rhetorical organization of a piece (p. 460).

Those bundles that are functioning as *triggers* for moves and steps could be used to construct a bundle-move/step scheme that could work as a skeleton to describe particular moves and their communicative functions. This scheme could become a useful tool to introduce moves and steps together with frequent linguistic exponents used to trigger those moves/steps in genre-based writing classes.

## 5. Conclusion

The purpose of the study reported in this article was to identify lexical bundles of different lengths in a corpus of RA introductions and to try to connect those bundles to the moves and steps that represent the communicative functions conveyed in those texts. The findings previously reported show a strong relationship between lexical bundles and the moves and steps in the Introduction framework. This relationship was better illustrated in the case of longer bundles (expressions of six or more words) which tend to strictly relate to only one step in a particular move.

Two specific findings need to be emphasized. First, the fact that the analysis of RA introductions yielded extremely long lexical bundles (such as *the rest of the paper is organized as follows* or *the purpose of the present study was to*), phenomenon that had never been reported before in studies of this type of formulaic language. Recurrent expressions made up of more than six words had never been identified before as lexical bundles. Structurally, some of these expressions are complete units, which slightly deviate from one of the constant characteristics of lexical bundles (i.e. they are not complete units). Second, the general functions that these bundles performed are functions previously found to be conveyed by bundles in academic texts, such as referential functions or discourse organizing functions (Biber et al., 2003, 2004). When further analyzing the communicative purposes of lexical bundles in relation to the moves in which these bundles often occur, some bundles were found to be used to trigger the communicative function of the move/step, while other lexical bundles were used as complements, accompanying the language that marked the beginning of a particular move or step. The former were mostly longer lexical bundles, while 4 and 5-word lexical bundles were also used as complements.

Many expressions that have been found to be extremely frequent in this very specialized corpus of introductory sections of RAs had never been identified before as recurrent expressions in corpora described as made up of academic texts, or even in corpora made up of whole RAs. It is necessary to point out that the findings of this study show that the frequency of individual lexical bundles becomes higher as the corpus becomes more focused or restricted. For example, we can now predict that an expression such as *studies have shown that* would have an extremely high frequency in a corpus consisting of Move 1 Step 3 texts exclusively. This prediction needs to be taken into consideration when creating corpora for the identification of lexical bundles in particular registers or sub-registers and for avoiding making generalizations of recurrent expressions in broader registers, such as academic writing.

The original purpose of the present study did not include a comparison in the use of lexical bundles in RA moves across disciplines mainly because the differences in disciplinary sub-corpora size may not make the comparison reliable. However, Appendix A shows certain tendencies that can be easily observed. All the disciplines represented in this corpus of RA introductions used lexical bundles longer than five words for Move 3 Step 1 (Introducing present research) but only a few disciplines (Computer Science, Economics, Statistics, and Urban and Regional Planning) used these longer bundles for Move 3, step 7 (Outlining the structure of the paper). Moreover, all disciplines made use of five and four-word bundles to convey the communicative purposes closely tied to the introduction sections of RAs. Appendix A shows that very few four-word bundles were never used in a particular discipline. Even disciplines that are represented by the smallest corpora in this study, as in the case of Statistics or Agronomy, made frequent use of bundles in all functions.

The relationship between lexical bundles and moves needs to be further developed not only in Introductions but also in other sections of RAs. The use of certain lexical bundles as triggers and/or complements calls for further analysis, particularly in the case of four-word bundles that tend to occur in several moves and steps and perform more than one function. In addition, the use of bundles in RA introductions across disciplines deserves further investigation, as lexical bundles seem to be discipline-bound (Cortes, 2004; Hyland, 2008). A better-balanced corpus (one which has equivalent number of words and number of texts in each of its disciplines) could yield more reliable findings for a discipline specific study of lexical bundles in introductions and other sections of RAs.

Finally, the pedagogical application of the findings reported here needs to be carefully considered. There are currently many genre-based classes that use computers and corpora to help students analyze the tendencies in the use of linguistic conventions and discourse organization of different academic genres (Cortes, 2011). Those courses that focus on the analysis of RAs, for example, could benefit from findings like those presented in this article. While showing students the characteristics of these texts and the moves in these sections, academic writing instructors could also introduce a set of lexical bundles identified as frequent in a particular move/step for students to investigate in their own corpora (ideally made up of texts from journals in their disciplines). That way, students could compare and contrast the use or lack of use of bundles to convey particular communicative functions in the RAs of different disciplines, adopting those bundles that they discover to be frequent in a particular move and making use of both moves and bundles as real building blocks in the construction of these particular texts.

## Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.jeap.2012.11.002>.

## References

- Aktas, N., & Cortes, V. (2008). A comparative corpus-based study of shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes*, 7, 3–14.
- Altenberg, B., & Eeg-Olofsson, M. (1990). Phraseology in spoken English: presentation of a project. In J. Aarts, & W. Meijs (Eds.), *Theory and practice in corpus linguistics* (pp. 1–26). Amsterdam: Rodopi.

- Altenberg, B. (1993). Recurrent word combinations in spoken English. In J. D'Arcy (Ed.), *Proceedings of the fifth Nordic Association for English Studies conference* (pp. 17–27). Reykjavik: University of Iceland.
- Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. In A. H. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 101–122). Oxford: Clarendon Press.
- Anthony, L. (1999). Writing research article introductions in software engineering: how accurate is the standard model? *IEEE Transactions in Professional Communication*, 42, 38–46.
- Barfield, A., & Gyllstad, H. (2010). *Researching collocations in another language*. Houndmills, Basingstoke, Hampshire: Macmillan Publishers Limited.
- Bhatia, V. (1993). *Analysing genre: Language use in professional settings*. London: Longman.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard, & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181–189). Amsterdam: Rodopi.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Pearson.
- Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Harlow: Pearson.
- Biber, D., Conrad, S., & Cortes, V. (2003). Towards a taxonomy of lexical bundles in speech and writing. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the Lune: A Festschrift for Geoffrey Leech* (pp. 71–92). Frankfurt/Maine: Peter Lang.
- Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25, 371–405.
- Biber, D., Connor, U., & Upton, T. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: Benjamin.
- Brett, P. (1994). A genre analysis of the result sections of sociology articles. *English for Specific Purposes*, 13, 47–59.
- Chu, B. K. S. (1996). *Introductions in state-of-the-art, argumentative, and teaching tips TESL journal articles: Three possible sub-genres of introductions?* City University of Hong Kong, Research Monograph No. 12.
- Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals: European Union research grants. *English for Specific Purposes*, 18, 47–62.
- Connor, U., Davis, J., & De Rycker, T. (1995). Correctness and clarity in applying for overseas jobs: a cross-cultural analysis of US and Flemish applications. *Text*, 15, 457–475.
- Cortes, V. (2002). Lexical bundles in academic writing in history and biology. Unpublished doctoral dissertation. Northern Arizona University, Flagstaff, Arizona.
- Cortes, V. (2004). Lexical bundles in published and student writing in history and biology. *English for Specific Purposes*, 23, 397–423.
- Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3, 43–58.
- Cortes, V. (2011). Genre analysis in the academic writing class: with or without corpora? *Cuadernos de Filologia. Estudios Linguísticos*, XVI, 41–64.
- Dudley-Evans, T. (1995). Genre: how far can we, should we go? *World Englishes*, 16, 353–358.
- Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, 30, 58–72.
- Flowerdew, J., & Li, Y. (2007). Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics*, 28, 440–465.
- Gray, B., & Cortes, V. (2010). Perception vs. evidence: an analysis of *this* and *these* in academic prose. *English for Specific Purposes*, 30, 31–43.
- Henry, A., & Rosenberry, R. L. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: 'Letter of Application'. *English for Specific Purposes*, 20, 153–167.
- Holmes, R. (1997). Genre Analysis and the social sciences: an investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes*, 16, 321–337.
- Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27, 4–21.
- Jablonskai, R. (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.
- Kanoksilapatham, B. (2003). A Corpus-based investigation of scientific research articles: linking move analysis with multidimensional analysis. Unpublished doctoral dissertation, Georgetown University, Washington, DC.
- Kanoksilapatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor, & T. A. Upton (Eds.), *Discourse on the move: Using corpus analysis to describe discourse structure* (pp. 73–119). Amsterdam: John Benjamins.
- Li, L., & Ge, G. (2009). Genre analysis: structural and linguistic evolution of the English-medium medical research article (1985–2004). *English for Specific Purposes*, 28, 93–104.
- Motta-Roth, D. (1998). Discourse analysis and academic book reviews: a study of text and disciplinary cultures. In I. Fontanet, S. Posteguillo, J. Palmer, & J. Coll (Eds.), *Genre studies in English for academic purposes* (pp. 29–57). Castellon, Spain: Universita Jaume I.
- Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
- Nwogu, K. (1991). Structure of science popularizations: a genre-analysis approach to the schema of popularized medical texts. *English for Specific Purposes*, 10, 111–123.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: native like selection and native like fluency. In J. Richards, & R. Schmidt (Eds.), *Language and communication* (pp. 191–227). London: Longman.
- Posteguillo, S. (1999). The schematic structure of computer science research articles. *English for Specific Purposes*, 18, 139–160.
- Schmid, H. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. Berlin: Walter de Gruyter.
- Swales, J. (1981). *Aspects of article introductions*. Birmingham, UK: The University of Aston, Language Studies Unit.
- Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.
- Swales, J. (2004). *Research genres: Exploration and applications*. Cambridge: Cambridge University Press.
- Williams, I. A. (1999). Results sections of medical research articles: analysis of rhetorical categories for pedagogical purposes. *English for Specific Purposes*, 18, 347–366.
- Yang, R., & Allison, D. (2003). Research articles in applied linguistics: moving from results to conclusions. *English for Specific Purposes*, 22, 365–385.