LEXICAL BUNDLES IN MEDICAL RESEARCH ARTICLES: STRUCTURES AND FUNCTIONS

By

Ndeye Bineta Mbodj-Diop

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Teaching English to Speakers of Other Languages – Master of Arts

2016

ABSTRACT

LEXICAL BUNDLES IN MEDICAL RESEARCH ARTICLES: STRUCTURES AND FUNCTIONS

By

Ndeye Bineta Mbodj-Diop

Lexical bundles, also referred to as "multi-word sequences" (Biber et al., 2004); "formulaic language" (Wray & Perkins, 2000); prefabricated routines and patterns (prefabs) (Erman & Warren, 2000); or chunks elsewhere; have been defined by Hyland (2012) as "statistically the most frequent recurring sequences of words in any collection of texts" (p. 150). Such *sequences of words* have received considerable attention in corpus studies in English for Academic Purposes (EAP) and English for Specific Purposes (ESP), with the underlying assumptions being that (a) experts in different discourse communities combine words in different ways to convey field-specific meanings and perform a variety of rhetorical functions; and (b) control of field-specific bundles is a key component of language production – be it written or spoken. The present study looked at the use of lexical bundles in medical research articles. Using a corpus of 1.1 million words, it investigated the frequency, structures, and functions of 4-word bundles in this particular genre of academic writing. Over 200 bundles were identified and the analysis indicated (1) a predominance in medical articles of lexical bundles beginning with noun phrases or prepositional phrases; (2) a more frequent use of research-oriented bundles compared to participant-oriented and text-oriented bundles; and (3) an extremely low frequency of specialized lexical items in the identified bundles. These results, as well as their pedagogical implications, are discussed in the present paper.

The present work is dedicated to my mother for her constant prayers and to my belated father. I know he would be proud of me…

ACKNOWLEDGEMENTS

I would like to express my gratitude to the Fulbright Program for giving me the opportunity to complete my Master's Degree in TESOL at the Department of Linguistics & Germanic, Slavic, Asian, and African Languages at Michigan State University. This has been a rewarding experience for me and my thanks go to all my professors for their highly valuable teachings and advice.

I am forever grateful to Dr. Charlene Polio, my thesis Director, and Dr. Amanda Temples, my second reader, without whom this work would not have been possible. I have learnt a lot throughout the writing process of this thesis and I could never thank them enough for their time, patience, and guidance. I feel fortunate to have had the opportunity to work with them.

I would also would like to thank my field-expert advisor, Dr. Mamadou Mourtalla Ka, Head of the Health Department at Thies University for his encouragements and insightful advice. I am also thankful to Dr. Paula Winke and Dr. Debra Hardison for their availability and valuable advice throughout these two years.

My heartfelt thanks go also to my cohort for being good, supportive friends. You were much more than just classmates. It was a blessing to meet each of you.

Last but not least, many thanks to my husband for his everlasting understanding, patience, and support; and for making it all the way to Michigan to attend my defense. It means a lot to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

BASE: British Academic Spoken English

LLC: London Lund Corpus of Spoken English

LOB: Lancaster-Oslo-Bergen corpus

LSWE: Longman Spoken and Written English

MICASE: Michigan Corpus of Academic Spoken English

MRA: Medical Research Articles

MRAC: Medical Research Article Corpus (created for the present study)

T2K-SWAL: TOEFL 2000 Spoken and Written Academic Language

INTRODUCTION

Lexical bundles, also referred to as *multi-word sequences* (Biber, Conrad, & Cortes 2004); *formulaic language* (Wray & Perkins, 2000); *prefabricated routines and patterns* (prefabs) (Erman & Warren, 2000); or chunks elsewhere; have been defined by Hyland (2012) as "statistically the most frequent recurring sequences of words in any collection of texts" (p. 150) The important role of lexical bundles in discourse – be it written or spoken – has been widely demonstrated in corpus linguistics research and literature (e.g., Biber & Barbieri, 2007; Biber et al., 2004; Biber, Johansson, Leech, Conrad, & Finegan; 1999; Cortes, 2004, 2006; Hyland, 2008a, 2008b; Wray & Perkins, 2000).

The large proportion of these studies underscored the frequent use of lexical bundles in both oral and written texts, in general, and in academic discourse, in particular. For example, Warren and Erman (2000) investigated, among other points, the "proportion of prefabs" in oral and written texts. Using a collection of texts made of extracts from the *London Lund Corpus of Spoken English* (LLC), the *Lancaster-Oslo-Bergen corpus* (LOB), and the *Goldilocks*, they found that lexical chunks accounted for 58.6% of the spoken texts, and 52.3% of the written texts. As for the pervasiveness of lexical bundles in academic discourse, Biber et al., (1999) found the extremely high frequency per million words of over 60,000 times for 3-word lexical bundles and more than 5,000 times for 4-word bundles in academic prose.

With such high presence in academic discourse, lexical bundles are often considered as markers of proficiency. For example, Cortes (2004) suggested that the frequent use of lexical bundles appears to indicate "competent language use within a register to the point that learning conventions of register use may in part consist of learning how to use certain fixed phrases" (p. 398). Such a claim clearly hints at the high pedagogical value of corpus-based studies that look at the use of lexical bundles in specialized academic texts. Perhaps such pedagogical value is better summarized by Hyland (2008a) who stated that:

> [G]aining control of a new language or register requires a sensitivity to expert users' preferences for certain sequences of words over others that might seem equally possible. So, if learning to use the more frequent fixed phrases of a discipline can contribute to gaining communicative competence in a field of study, there are advantages to identifying these clusters to better help learners acquire the specific rhetorical practices of their communities. (Hyland, 2008a; p.5)

Several corpus-based studies have thus been conducted in an aim to better understand and describe how discourse is constructed in different fields. Many of them have focused on the use of lexical bundles in a variety of academic disciplines (e.g., Cortes, 2004, 2006; Hyland, 2008a, 2008b). The present study adds to such a tradition by investigating the use of lexical bundles in published medical research articles (MRAs).

The motivations underlying the focus on MRAs in the present study can be summarized in two main points. First, there is a need to address some pedagogical concerns and fulfill a professional responsibility. As an English for specific purposes (ESP) instructor in charge of designing the English curriculum at the Health Department of Thies University, Senegal, and being given the opportunity to complete a Master's degree in TESOL in the USA as a Fulbright grantee, I naturally decided to take advantage of this great opportunity by conducting this study that hopefully will contribute to a better fulfillment of my professional duties both as a course designer and a language instructor. It should be added that Cortes' (2004) and Hyland's (2008a) statements about the pedagogical value of corpus-based studies are all the more relevant to the Senegalese context where ESP teachers have to design their instructional materials basically from scratch, in the absence of appropriate teaching resources.

The second point of the rationale behind the choice to focus on MRAs is that this genre is highly representative of the learners' target language use situations. Because Senegal is an EFL context where instruction in content areas is provided in French, one of the rare pieces of academic writing – other than ESP writing assignments – that students are to produce in English is the MRA. As future medical professionals, they will necessarily have to publish articles in English to both gain recognition from their peers and advance in their careers. Given the context described above and the earlier mentioned importance of lexical bundles in academic prose, it is pure logic that the present study sets out to investigate the use of lexical bundles in MRAs.

Several other studies have used MRA corpora to investigate language resources used in this particular genre of academic writing. While many of the work in this domain focused on field-specific vocabulary in MRAs (e.g., Chen & Ge, 2007; Mungra & Canziani, 2013; Wang, Liang &Ge, 2008), to my knowledge, only two studies (Gledhill, 2000; Marco, 2000). addressed the use of multiword expressions in MRAs. Gledhill examined the discourse function of collocations in the Introduction sections of cancer research articles. His study focused mainly on collocations including the verbs *has, have, been,* and *is* and the prepositions *of* and *to*. Marco's study investigated collocational frameworks in medical articles focusing on intermediate words that commonly serve as fillers of the following three frameworks: *the ... of, a ... of,* and *be ... to.*

As can be seen, these studies were limited to either one section of MRAs or a pre-determined set of frameworks. The present study takes a more exploratory approach by investigating the frequency, structures, and functions of lexical bundles in all sections of medical articles. Using a corpus of over 1 million words made up of 250 MRAs from five different journals, this study simply sets out to answer the following research questions:

1- What lexical bundles are frequently used in MRAs?

2- What are the structural patterns of lexical bundles in MRAs?

3- What rhetorical functions do such bundles have in MRAs?

CHAPTER 1. LITERATURE REVIEW

2.1 Lexical Bundles

As mentioned, different terms have been used in studies that looked at the use of word combinations that frequently occur together in language production. The term *lexical bundle* was first used by Biber et al. (1999) and was defined as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (p. 990). This definitional framework has been adopted in many subsequent studies (e.g., Biber et al., 2004; Cortes, 2004; 2006; Neely & Cortes, 2009; Hyland, 2008a, 2008b; Grabowski, 2015) that have added to our understanding of lexical bundles and contributed to the setting of identification criteria for such word combinations.

According to Cortes (2004) lexical bundles differ from other word combinations such as pure idioms in that their meanings are typically transparent from the words comprised in them. For instance, the meaning of the bundle *in the presence of* would appear much more salient to a language learner than that of the idiom *to kick the bucket* (to die). Cortes further suggested that like other word combinations, lexical bundles are fixed expressions because computer programs used to identify them only recognize one form of a bundle at a time. For that reason, the bundle *there was no difference* and what apparently stands for its plural form *there were no differences,* both from the list of four-word bundles in the present study, were identified by the software as two separate types of bundles; each fixed form representing one type of bundle. However, Cortes pointed out that the fixedness of lexical bundles is different from that of other word combinations because it is primarily based on frequency; meaning that an expression (e.g., *were similar in the*) can be identified as a lexical bundle but another form of the same expression (e.g., *was similar in the*) may not qualify as a lexical bundle because it does not occur frequently enough in the corpus under study.

Indeed, frequency, according to Biber et al. (1999) is the defining characteristic of lexical bundles. They suggested that for a word combination to qualify as a lexical bundle, it has to occur at least 10 times per million words and across a range of at least five texts. This frequency can be much higher for the most common bundles in a corpus and can reach over 100 times per million words (Cortes, 2004). However, it should be noted that although there seems to be a consensus about frequency as the main criterion for identifying lexical bundles, the cutoff used in studies of lexical bundles is quite arbitrary and varies from one study to another. For instance, while some studies such as Hyland (2008a, 2008b) or Cortes (2004) used a higher frequency than Biber et al. (1999), setting their cutoff point at 20 times per million words, other studies (e.g. Biber et al., 2004; Biber & Barbieri, 2007) used an even higher frequency cutoff of 40 times per million words; the idea being that the higher the frequency, the more representative of a given register the identified bundles are.

In addition to occurring frequently in a selection of texts, multiword combinations must appear across multiple texts to qualify as lexical bundles (Biber et al., 1999). Such a requirement contributes to a better representation of linguistic choices typical of a given register or discipline by avoiding "idiosyncratic uses of individual writers or speakers" (Biber & Barbieri, 2007; p. 268). As suggested by Hyland (2012), the recurrence of lexical bundles in several texts by different writers or speakers indicates "some perceptual salience among users and conventionalization within a particular discourse community" (p. 152). As in the case of the frequency of occurrence, the range of texts at which word combinations should occur to qualify as lexical bundles varies from one study to another, the minimum range being five texts as suggested by Biber et al. (1999).

Another characteristic of lexical bundles is that although they function as a whole, they do not represent complete structural units. Biber et al. (1999) found that over 95 % of the lexical bundles identified in their corpus of academic prose did not constitute complete structural units. Indeed, as suggested by Biber et al. (2004), most lexical bundles bridge two structural units, meaning that they start at a clause or phrase boundary and end with the first words of another structural unit. They further suggested that most bundles in academic prose typically bridge two phrases as in these examples from the list of identified four-word bundles in the present study: *in view of the, with respect to the, the basis of a.* Cortes (2004) nicely summarized the characteristics of lexical bundles discussed above by stating that:

> Lexical bundles are identified empirically, rather than intuitively, as word combinations that recur most commonly in a register, and therefore, lexical bundles are usually not complete structural units, but rather fragmented phrases or clauses with new fragments embedded. (p. 400)

However, Cortes further pointed out that although they do not constitute complete structural units, lexical bundles have strong grammatical correlates that can serve as the basis of their structural classification. Referring to Biber et al.'s (1999) classification of lexical bundles – which is the model of analysis used for the lexical bundles identified in the present study and is discussed in the next chapter of this paper – she suggested that while most bundles in conversation are clausal *(e.g., I want you to, it's going to be),* the majority of lexical bundles in academic prose are phrasal; that is, they are parts of noun phrases or prepositional phrases such as in the following examples from the present study: *on the basis of, in the absence of, the baseline characteristics of, a reduction in the,* to name but a few. These bundles, according to Biber and Barbieri (2007) express functions and meanings that differ dramatically across registers and academic disciplines. Therefore, research on the use of lexical bundles in academic contexts would be better served if conducted from an ESP perspective that considers each register or discipline separately. Such has been the approach in many studies of lexical bundles, some of which are presented below.

1.2 Studies of Lexical Bundles in EAP/ESP

Lexical bundles have been shown as being particularly frequent and constituting important building blocks in academic discourse (Hyland, 2008b). This certainly can explain the growing body of corpus-based studies, over the last two decades, focusing on the use of lexical bundles in academic contexts (e.g., Biber et al., 1999; Biber & Barbieri, 2007; Biber et al., 2004; Conrad, 1999; Cortes, 2004; 2006; Grabowski, 2015; Hyland, 2008a; 2008b; Marco, 2000; Neely & Cortes, 2009; Nesi & Basturkmen, 2006).

These studies differed in their purposes and the registers they focused on. While some of them compared lexical bundles in spoken vs. written academic registers (e.g., Biber et al., 2004), others investigated the use of lexical bundles in academic vs. non-academic university registers (e.g., Biber & Barbieri, 2007). Still, some of these studies compared the use of lexical bundles across disciplines and/or by novice and expert writers within the same discipline (e.g., Cortes, 2004; Hyland, 2008a, 2008b). Finally, a few of these studies focused on one specific genre within academic discourse. For example, Nesi and Basturkmen (2006) investigated the role of lexical bundles in university lectures while Gabroski (2015) and Marco (2000) focused on the use of multi-word expressions in specific genres: patient information leaflets and summaries of product characteristics, and the medical research article, respectively.

A consistent finding in most of these studies remains the high variation of lexical bundles across registers, disciplines, and genres. For example, in their study that compared lexical bundles in written (academic prose and textbooks) and spoken (conversation and classroom teaching) registers, using texts from the TOEFL 2000 Spoken and Written Academic Language Corpus (T2K SWAL) and the Longman Spoken and Written English Corpus (LSWE), Biber et al. (2004) found that lexical bundles occurred much more frequently in teaching sessions and that classroom discourse comprised the widest variety of lexical bundle types. They attributed such a high density of lexical bundles in classroom discourse to the fact that this register used both *oral* bundles – meaning bundles specific to spoken discourse – and *literate* bundles, specific to written discourse.

Similar results were reported by Nesi and Basturkmen (2006) who found that lexical bundles in their corpus of 160 monologic lectures from the British Academic Spoken English (BASE) corpus and the Michigan Corpus of Academic Spoken English (MICASE) included both *oral* and *literate* bundles. They also found that the majority of lexical bundles in academic lectures were used for discourse signaling purposes and suggested that there is a necessity to raise learners' awareness of such use of lexical bundles to facilitate their comprehension when attending academic lectures.

In addition to the variation of bundles types across registers, Biber et al. (2004) also observed dramatic differences in the way different registers relied on particular functional types of bundles; with (1) *stance* bundles being extremely frequent in both conversation and classroom teaching; (2) *discourse organizing* bundles being most frequently used in classroom teaching and a little less in conversation; and (3) *referential* bundles being extremely common in classroom teaching and slightly less in academic prose and textbooks. Furthermore, the authors also found that there was a relationship between the structures of the bundles and the two different registers. While *oral* bundles consisted mostly of declarative and interrogative clause fragments (e.g.

that's one of the, what do you think), *literate* bundles used in academic prose, for instance, contained mainly prepositional phrases and noun phrases (e.g., *the end of the, at the same time*).

Using a different approach, Cortes (2004) compared lexical bundles used by expert writers in two different disciplines (history and biology); and then looked at the use of such bundles by novice writers in these two disciplines at three different university levels: undergraduate lower division, undergraduate upper division, and graduate level. The data of her corpora revealed great variations in the use of lexical bundles both across the two disciplines and by expert and novice writers.

Drawing from the structural classification of lexical bundles in previous studies such as Biber et al. (1999) – which is also used in the present study and discussed in the next chapter – she found that bundles used by experts in history consisted mainly of expressions beginning with noun phrases and prepositional phrases. On the other hand, bundles used by experts in biology followed a wider variety of structural patterns including: prepositional phrases, noun phrases, it + Vbe +adjective clause fragments (e.g., *it is important to*), Vbe + complement clause fragments (e.g., *is consistent with the*), and noun phrase +V +complement clause fragments (e.g., *studies have shown that*).

Cortes also found that the differences between the two disciplines expanded to the functions performed by lexical bundles in her two sub-corpora. Her data indicated that bundles identified in both history and biology as time markers were used for different purposes from one discipline to the other. For instance, while in history time markers were mainly used to indicate years or time periods in which historical events happened, in biology, bundles in this same category mainly served the purpose of indicating "stages in the evolutionary or developmental processes of different biological phenomena" (p. 411). Another interesting observation regarding the functions of lexical bundles in the two disciplines was the prominence in biology of bundles used as *epistemic-impersonal* / *probable-possible* stance markers (e.g., *it is possible that, are likely to be);* while such bundles were nonexistent in her history corpus.

Similar variations from one discipline to another were also found by Hyland (2008a) who compared the use of lexical bundles in two distinct fields: pure sciences, represented by electrical engineering and microbiology; and social sciences, represented by applied linguistics and business studies. His findings indicated variations not only between the two fields – as could be expected – but also between disciplines within the same field, "with less than half of the top 50 bundles in each list occurring in any other list" (p. 20). The functional classification of the bundles in his four subcorpora into three main categories (*research-oriented*, *text-oriented,* and *participant oriented* bundles [all discussed later in this paper]) equally revealed variations across fields with research-oriented bundles prevailing in the biology and engineering texts, and text-oriented bundles dominating the applied linguistics and business science corpora.

The author suggested that such variations reflected the specificities and requirements of academic writing in each of the two fields, namely, pure sciences and social sciences. He argued that while in hard sciences the *empirical* prevails over the *interpretative* with a focus on the research itself, in social sciences, the presentation of research is "more discursively elaborate" with a heavy reliance on text-oriented bundles to "provide familiar and shorthand ways of engaging with a literature, providing warrants, connecting ideas, directing readers around the text, and specifying limitations" (p. 16)

In addition to comparing the use of lexical bundles across different fields and disciplines, both Hyland (2008a) and Cortes (2004), as well as Hyland (2008b), also compared the use of such multiword expressions by expert and novice writers. While Cortes used student writings from undergraduate to graduate levels, Hyland's student subcorpora consisted of collections of Master's theses and PhD dissertations. Both studies revealed that novice writers used lexical bundles differently than expert writers.

Cortes found that students made very rare use of bundles used by expert writers; and in cases where they actually used some of these bundles, their uses differed from those of expert writers. For example, she found that the expression *at the same time* was used by history students to express addition instead of simultaneity which corresponds to its typical function in academic prose. Cortes also noted that the frequency in use of bundles at graduate level was not higher than that in lower levels, and that graduate students – who logically have had more exposure to reading materials in their field – still had difficulty in using the bundles appropriately. She concluded that mere exposure to reading materials was not sufficient for students to start using a variety of lexical bundles in the same way as expert writers. She pointed out that there is, therefore, a necessity to raise students' awareness of the frequent use of lexical bundles and the different discourse functions these expressions serve in their specific academic disciplines.

In the same vein, Hyland (2008b) found that many lexical bundles frequently used in his research articles corpus were never or only rarely used in Masters' theses and PhD dissertation corpora, sharing only six bundles in the top 15 in the PhD corpus and only five in the whole Master's list. However, Hyland also found a wider variety of bundle types in Masters' and PhD texts than in published research articles; with research articles containing 71 different types of bundles while PhD dissertations yielded 95 different types and Master's theses 149. Most of the bundles in students' texts, he noted, did not appear in research articles. He suggested that such variations could reflect a "greater reliance on formulaic expressions by less confident or proficient students in constructing their texts", but also that novice and expert writers draw from different resources to construct their texts (p. 60). Hence the need to familiarize learners with the linguistic preferences of expert writers in their new discourse communities.

These considerable variations of lexical bundles in forms, structures, and functions across registers, disciplines, and genres, as well as the differences in use by novice and expert writers, underscore the pedagogical value of studies that investigate the use of lexical bundles in specific genres of academic writing. Such studies can contribute to providing realistic models for novice writers because, as suggested by Hyland:

> Bundles occur and behave in dissimilar ways in different disciplinary environments and it is important that EAP course designers recognise this, with the most appropriate starting point for instruction being the student's specific target context. (Hyland, 2008a; p. 20)

The present study and the ones discussed in the next section may contribute to such work.

1.3 Corpus-based Studies in the Field of Medicine

There have been a handful of corpus-based studies, over the past two decades, that have looked at how discourse is constructed in MRAs. Most of those studies have focused on individual lexical items (e.g., Chen & Ge, 2007; Mungra & Canziani, 2013; Wang, Liang &Ge, 2008). A few others investigated the structural moves in different sections of MRAs (e.g., Fryer, 2012; Li & Ge, 2009; Nwogu, 1997). There have also been studies that focused on particular rhetorical functions of some grammatical features such as modals in expressing epistemic modality (Yang, Zheng, & Ge, 2015), the use of reporting verbs in medical articles (Jirapanakorn, 2012), or conditionals in medical articles (Ferguson, 2001). To my knowledge, only two studies (Gledhill, 2000; Marco, 2000) expressly addressed the use of multiword combinations in medical research articles.

Both Gledhill and Marco based their studies on Sinclair's (1991) idiom principle which suggest that each language has "a large number of semi-preconstructed phrases that constitute single choices" (as cited in Marco, 2000; p. 13). Therefore, they focused on the identification of lexicogrammatical items that commonly collocate with pre-identified words or frequently occur within preset frameworks.

Gledhill's (2000) study looked at the phraseology of Introduction sections of *Cancer* articles and investigated the discourse functions of collocations comprising the verb forms *has, have, been,* and *is;* and the prepositions *of* and *to*. On the other hand, Marco (2000) explored the collocational frameworks in medical articles focusing on intermediate words that fill in the following three frameworks: *the ... of, a ... of,* and *be ... to.*

Of relevance to the present study, was Marco's work which identified the frameworks *the ... of* and *a ... of* as constituting the most frequently used structures in medical research articles. The analysis of her corpus of 298,457 words built from 100 medical articles from *The New England Journal of Medicine* and the *British Medical Journal* revealed that together, the frames *the ... of* and *a ... of* occurred with 1,248 different collocates, all of which were nouns. It should be noted that Marco considered each frame separately, with the frame *the ... of* appearing in her corpus with much more fillers (1,150 of the 1,258 collocates). However, her findings are relevant to the present study with regard to both frameworks as each of them produced multiword expressions that followed the structural pattern *Noun Phrase + embedded of-phrase fragment* (discussed in the next chapter) which was identified as one of the most frequent structures of lexical bundles in the present study.

Marco classified the fillers of the frameworks according to the meanings they conveyed in the medical articles in her corpus. She identified six groups for the fillers of the framework *the ... of*, and two for the framework *a ...of.* Below is a brief summary of these groups.

The framework *the ... of* frequently collocated with items that express:

- *measure* or *quantification* (e.g., degree, dose, number, proportion);
- *medical procedures* or *steps of treatment* (e.g., diagnosis, administration);
- *qualities* or *properties* (e.g., efficacy, availability, ability.);
- *existence / non-existence* (e.g., presence, absence, occurrence.);
- *moment of time in a process* (e.g., beginning, development, end.); and
- *relationships between elements of research* (e.g., effect, result, characteristics, cause)

The two groups of fillers of the framework *a ... of* that somehow overlap with some of the groups described above. Items in these two groups mainly expressed:

- *quantity* (measure, dose, total, subgroup); and
- *properties that can be quantified* (sensitivity, accuracy, specificity)

It should be noted that the great majority of the fillers identified by Marco appeared in the lexical bundles of the corpus under study here. Some elements of her classification were used in the analysis of the meanings of bundles following the structural pattern *Noun Phrase + of-phrase fragment* that were identified in the study presented in the remaining sections of this paper.

CHAPTER 2. METHODOLOGY

2.1 Corpus Collection

The corpus in this study consists of medical research articles (MRAs) published in five of the leading journals in the field of medicine (*Science, The Lancet, The New England Journal of Medicine, Journal of American Medical Association, The Journal of Clinical Investigation).* The choice of these journals was guided by suggestions from an experienced and distinguished Professor, Head of the Health Department at Thies University (*field- expert advisor* hereafter). These internationally recognized journals are relevant to the students' English use situation as they are representative of the types of journals where students will later seek international publication for professional advancement. Moreover, articles from these journals are part and parcel of the reading requirements of students who are often asked to refer to such texts to complete content area assignments, though the latter are typically written in French.

In an aim to build a corpus that is representative of recent practices in the genre being studied (i.e., MRAs), the selected articles were limited to those published within the last 10 years (2006 – 2015), with 5 articles from each year. To ensure that the texts used to build the corpus, were comparable, only quantitative studies, which appear more common in the field of medicine, were considered for article selection. Three topics (diabetes, hypertension, tropical disease) were suggested by the field-expert advisor. However, even though such topics were relevant to the Senegalese context[1] and certainly justified the field expert's suggestion, for generalizability concerns, other common themes featured in the selected journals were included as well.

[1] There is a high prevalence of diabetes and hypertension in Senegal. A survey conducted by Mbaye et al. (2011) in Saint Louis – the third region of the country – revealed a diabetes prevalence of 10.4% including 7.8 known cases and 2.6 new cases. Hypertension prevalence was found at an alarming 46% known in only half of the cases.

The target size of the corpus was 1,000,000 running words based on Biber's and Conrad's (1999) suggestion that a multi-word unit needs to occur at least 20 times per million words to qualify as a "recurrent lexical bundle" (p. 184). Given that on average, the articles in the target journals comprised a little more than 4,500 running words each, 50 research articles were selected from each of the 5 journals with word counts between 3,000 and 7,000 for each article. In total, 250 articles, selected on the basis of the criteria listed below, were used to build a corpus of 1,177,611 running words (Table 1).

Table 1

The MRA Corpus (MRAC)

Journals	Number of articles (5/year)	Publication years	Mean number of words / article	Total number of words/ journal
JAMA	50	2006-2015	4,225	21,1231
JCI	50	2006-2015	5,676	28,3779
The Lancet	50	2006-2015	4,665	23,3245
NEJM	50	2006-2015	3,837	19,1852
Science	50	2006-2015	5,150	25,7504
Totals	250			1,177,611

2.1.1 Article selection criteria.

Loi (2010) identified three main variables to control for when building corpora: genre, authorship and journals under investigation. Taking into account these three points, the following guidelines were set for the selection of articles used to build the Medical Research Article Corpus (MRAC):

1. With regard to genre, only quantitative studies research articles in the field on General Medicine are considered for selection. Given the relevance of the recommended topics,

articles on diabetes, hypertension, and tropical diseases are represented at least once in each of the 5 selected journals. Other themes were randomly selected among those featured in the selected journals (see Appendix B for all article themes).

2. For the reputation of the journals under investigation, the author sought the opinion of the field-expert advisor and all articles used to build the corpus are from the following five internationally recognized peer-reviewed medical journals: *Science, The Lancet, The New England journal of Medicine (NEJM), Journal of American Medical Association (JAMA),* and *Journal of Clinical Investigation (JCI).* To ensure a representation of the most recent writing practices in the target field, only articles published in the last 10 years (2006 – 2015) were considered. Five articles were randomly selected from each year for each journal. To reach the target corpus size of at least 1,000,000 running words, only articles comprising 3,000 – 7000 words were considered for selection.
3. As regards to authorship, Wood's (2001) strict criterion was not considered for article selection. Briefly summarized, Wood suggested that articles written by native speakers better reflect appropriate language use. Therefore, to ensure that a piece of writing is representative of the way language is used by native speakers, at least its first author should be a native speaker and affiliated to an institution in a country where the language is used as first language (as cited in Wang et al., 2008; p. 446). The main point behind Wood's criterion was that non-native speakers do not have the same intuitions as native speakers about appropriate language use. An argument which, in some situations may have some proof of validity, but the issue of "nativeness" and/or "non-nativeness" was not considered in the present study for three main reasons: (1) in a context of world Englishes and English as a Lingua Franca, it might be problematic to determine who is a native speaker of English

and who is not, (2) it appears irrelevant in such context to consider a piece of writing as representative of the practices of a given discourse community on the basis of the "nativeness" or "non-nativeness" of its author(s); and (3) where writers receive their medical or research training may have a greater influence on their writing than the language they grew up using. Therefore, in this study, the articles that were deemed good enough for publication by the reviewers of the above mentioned renowned journals were considered as representative of the medical research article genre regardless of their authors' first languages.

On the basis of the above guidelines, the 250 MRAs were selected to build the database to be analyzed using the software Antconc (Anthony, 2005).

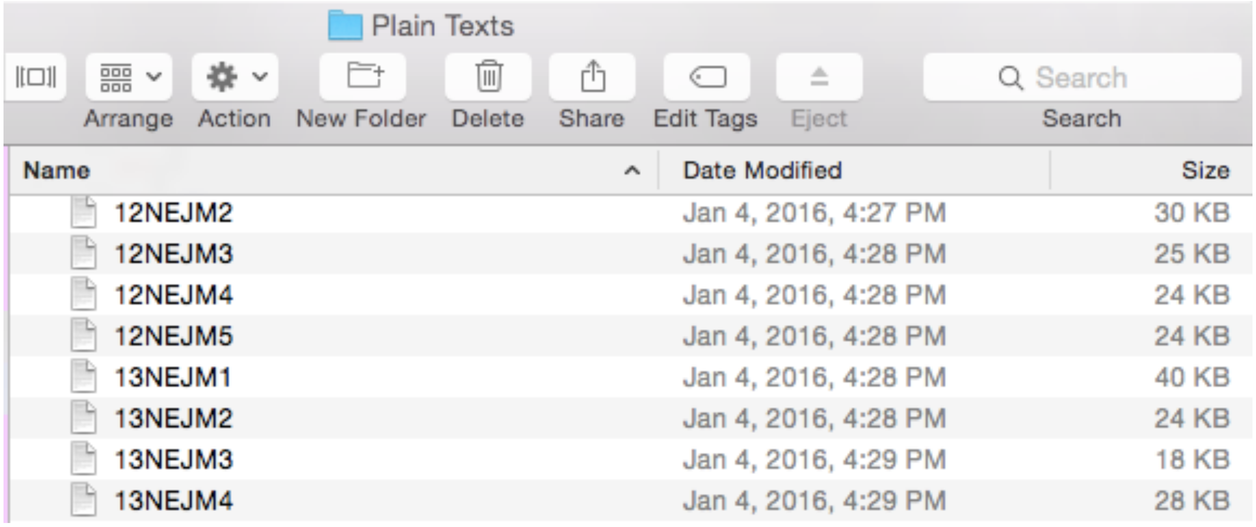
2.1.2 Building database for the software.

The 250 MRAs were downloaded from the websites of the five target journals freely accessible from the Michigan State University Library website. Each article was downloaded in its PDF version – for future reference – and in its HTML version. The HTML version of each article was copied, pasted, and saved as a Word document. In an aim to later identify at what range a given bundle occurred, each article was saved separately under a file name that included the name of the journal and the year of publication. For example, the first of the five articles published in 2015 in JAMA was saved as 15JAMA1; 15JAMA2 for the second, 15JAMA3 for the third, and so on. The same process was followed for all 250 articles from the five journals.

The next stage after downloading the 250 articles was the cleaning of the articles; that is, as suggested by Swales (1990), the removal of all reference lists, tables, appendices, footnotes, and acknowledgements; and also titles, authors' names, headers, footers, and mathematical equations (Cortes, 2004). The final word count of the MRAC was based on the cleaned articles.

The latter were then saved as plain texts (.txt format) to match the requirements of the software. The plain text versions of the articles were saved under the same names as the Word versions. A folder was created for each journal and included 50 articles in plain text format (as shown in figure 1 below) to be uploaded for analysis.

Figure 1. Example of plain text files saved per year and per journal



Name	Date Modified	Size
12NEJM2	Jan 4, 2016, 4:27 PM	30 KB
12NEJM3	Jan 4, 2016, 4:28 PM	25 KB
12NEJM4	Jan 4, 2016, 4:28 PM	24 KB
12NEJM5	Jan 4, 2016, 4:28 PM	24 KB
13NEJM1	Jan 4, 2016, 4:28 PM	40 KB
13NEJM2	Jan 4, 2016, 4:28 PM	24 KB
13NEJM3	Jan 4, 2016, 4:29 PM	18 KB
13NEJM4	Jan 4, 2016, 4:29 PM	28 KB

2.2 Identification of Lexical Bundles

As mentioned in many corpus studies (e.g. Biber et al. 2004; Hyland, 2008a; Biber & Barbieri, 2007), the frequency cutoff for identifying word combinations that qualify as lexical bundles is rather arbitrary. Drawing from Biber et al.'s (1999), Hyland's (2008a), and Cortes's (2004) methodologies, the cutoff in this study was initially set at a frequency of 20 times per million words with an occurrence across at least 10% of the texts (i.e., articles). However, given that the MRAC comprised 250 texts and 10% of the texts amounted to 25 texts, any bundle identified by the software, based on the initial cutoff, would necessarily have occurred at least 25 times. This would have automatically excluded all the bundles that occurred at least 20 times per million words but in less than 25 texts. Therefore, to avoid excluding important bundles, the range was readjusted to 5% of texts in the corpus (roughly 12 texts); which was still an acceptable range if we consider that Biber et al. (1999) used a much lower range of 5 texts in a corpus of more than 10 times larger than the MRAC.

Only four-word bundles were targeted for mainly two reasons. First, as stated by Hyland (2008a), it is more interesting to study four-words bundles as they not only are more common than 5-word units, but they also "offer a clearer range of structures and functions than three-word bundles" (p. 8). In addition, many three-word bundles are embedded in four-word bundles (Cortes, 2004).

To identify the four-word bundles in the MRA Corpus, the Antconc N-Gram tool was used. The 250 articles in plain text format were uploaded, the frequency set at 20 and the range at 12. The software then processed the files and generated a list of *candidates* for four-word bundles, to use the term of Biber's and Barbieri's (2007). Indeed, as suggested by Biber et al. (1999), it is sometimes the case that the software identifies some word combinations that do not necessarily qualify as lexical bundles. For instance, they identified some lexical combinations in their corpus as *local repetitions*; that is, combinations that "are often repeated within the span of a single text [and that] reflect the immediate topical concern of the discourse" (p. 991). The conversational excerpt below, from their study, shows an example of local repetition. They provided local repetitions in *[]* and the lexical bundles in bold italics.

*C: You like Vinnie Jones, **don't you**?*
A: No.
*C: **Yes you do**.*
*A: Vinnie Jones, **he's such** a hard bloke.*
C: Yeah.
*B: **It's really** funny, that man.*
A: [I reckon he] er well, – probably [could] –
*[**I don't** reckon] Eric [Cantona could].*
B: [Don't reckon Cantona could] what?
*A: [**I don't** reckon Cantona could] beat him.* (Biber et al. 1999; p.991)

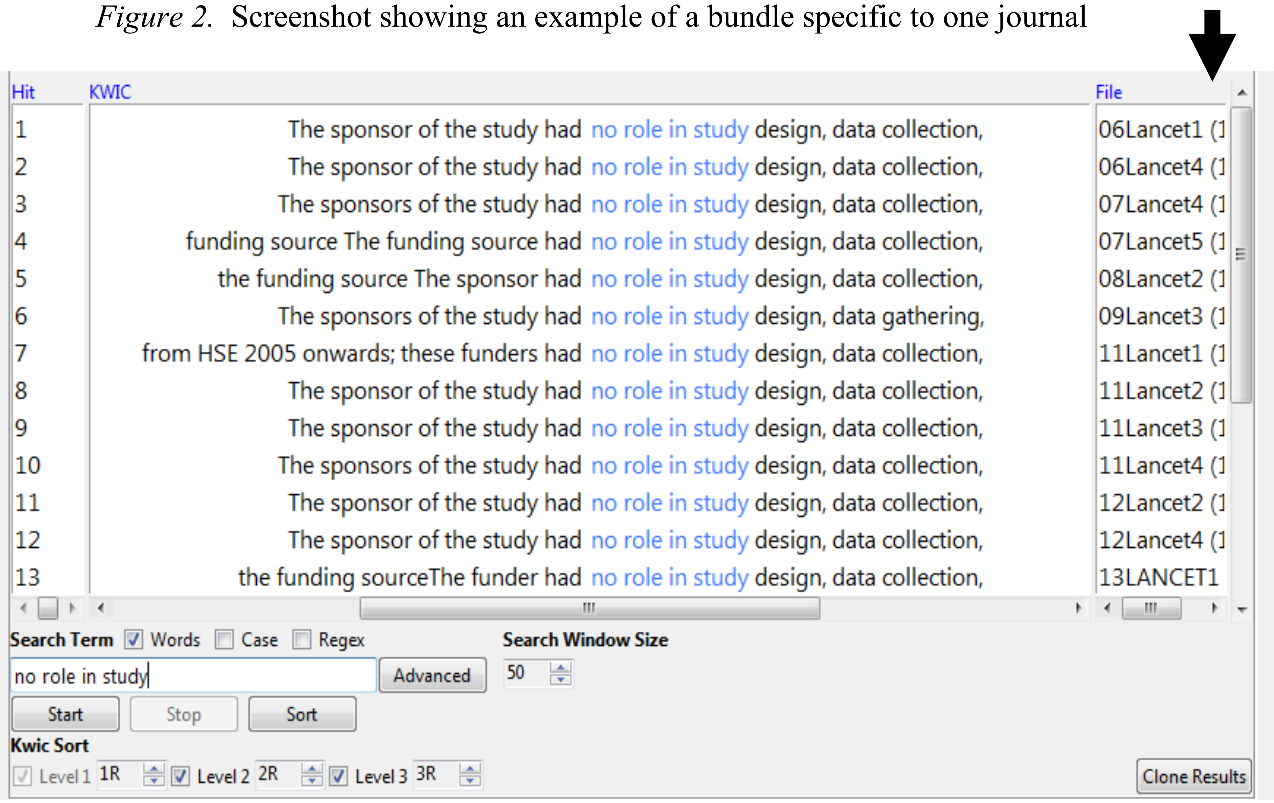
The authors argued that such combinations do not qualify as lexical bundles because they are not widely used by different authors across different texts. The initial list of candidates for four-word bundles in the the MRAC did not include local repetitions but two particular types of combinations were identified with features that distinguished them from the rest of the other identified four-word bundles. The following sections describe such types of combinations.

2.2.1 Four-word combinations specific to one journal.

Some four-word combinations, sometimes with a very high frequency, occurred only in articles from one single journal. For instance, "*in the supplementary appendix"* occurred 212 times per million words and across 29 articles, all of which were from the NEJM. In the same way, "*no role in study*" appeared 25 times in 25 different articles all from *The Lancet*. The concordance list of "*no role in study"* revealed a repetition of basically one same statement, albeit with slight variations in sentences (see figure 2). A quick look at the PDF initial articles and at the journal's submission guidelines revealed that the presence of such statements was typically due to a specific submission requirement which asked authors to clearly state the role of the sponsors in the study. As a result, in all articles published in *The Lancet,* there is a section that appears under the heading "role of the funding source" at the end of the Methods section and before the Results section. As regards *"in the supplementary appendix"* it mainly appeared in sentences that directed readers to different tables and figures available online at the journal website under the heading Supplementary Appendix. Here also, the use of the bundle appears to be a direct result of the journal's submission guidelines which require writers to include a minimum of tables and figures in the actual articles and provide any additional tables and figures in the "Supplementary Appendix" section.

Like local repetitions that are typically related to some specific issue at one point of a discourse, bundles such as "*no role in study*" and "*in the supplementary appendix"*, appeared quite constrained – as opposed to occurring in natural discourse – and reflected the writers' fulfilment of one specific requirement in one specific journal. Given that five different journals were used to build the MRAC, bundles that appeared to be specific to one particular journal were eliminated from the initial list of four-word bundle candidates.

Figure 2. Screenshot showing an example of a bundle specific to one journal



```
Hit  KWIC                                                                                      File
1                 The sponsor of the study had  no role in study  design, data collection,    06Lancet1 (1
2                 The sponsor of the study had  no role in study  design, data collection,    06Lancet4 (1
3                The sponsors of the study had  no role in study  design, data collection,    07Lancet4 (1
4         funding source The funding source had  no role in study  design, data collection,    07Lancet5 (1
5            the funding source The sponsor had  no role in study  design, data collection,    08Lancet2 (1
6                The sponsors of the study had  no role in study  design, data gathering,     09Lancet3 (1
7   from HSE 2005 onwards; these funders had  no role in study  design, data collection,    11Lancet1 (1
8                 The sponsor of the study had  no role in study  design, data collection,    11Lancet2 (1
9                 The sponsor of the study had  no role in study  design, data collection,    11Lancet3 (1
10               The sponsors of the study had  no role in study  design, data collection,    11Lancet4 (1
11                The sponsor of the study had  no role in study  design, data collection,    12Lancet2 (1
12                The sponsor of the study had  no role in study  design, data collection,    12Lancet4 (1
13          the funding sourceThe funder had  no role in study  design, data collection,    13LANCET1
```

Search Term ☑ Words □ Case □ Regex

no role in study

Search Window Size 50

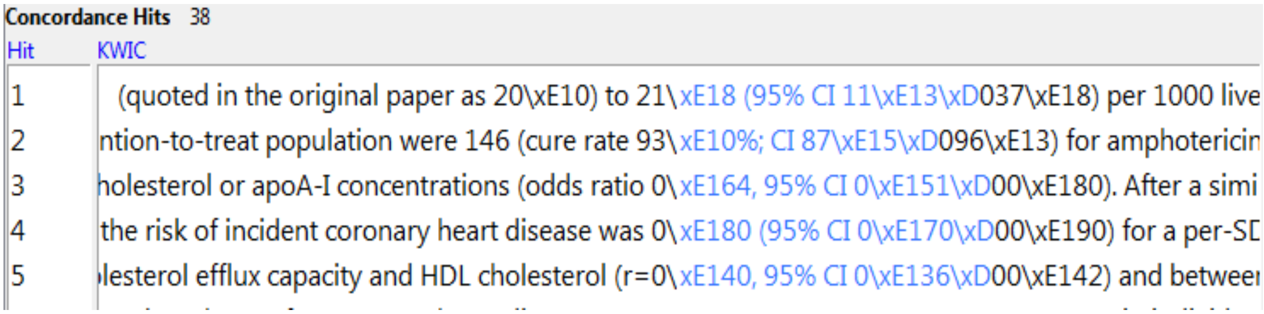
Kwic Sort

☑ Level 1 1R ☑ Level 2 2R ☑ Level 3 3R

2.2.2 Bundles with number slots.

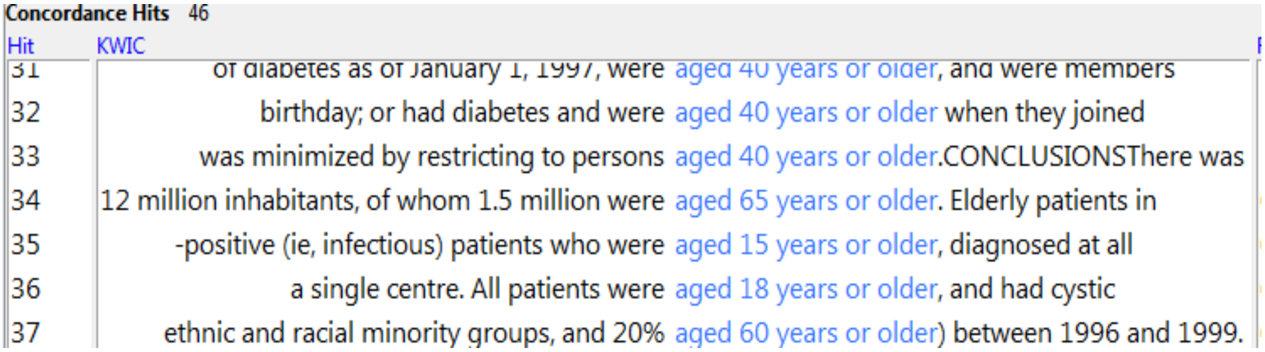
The second type of particular combination in the initial list of candidates for four-word bundles was four-word units that included slots for numbers. Indeed, such units were identified almost accidentally as this pattern emerged as a result of the functional features of the software. In the identification of N-grams or word combinations, the software does not take into account numbers and special characters (e.g. /, %, =). As a result, combinations such as those shown in Figure 3 and Figure 4 below were included in the initial list of candidates.

Figure 3. Example of a misrepresented formula (*xe ci xe xd)*



```
Concordance Hits  38
Hit   KWIC
1      (quoted in the original paper as 20\xE10) to 21\xE18 (95% CI 11\xE13\xD037\xE18) per 1000 live
2     ntion-to-treat population were 146 (cure rate 93\xE10%; CI 87\xE15\xD096\xE13) for amphotericin
3     holesterol or apoA-I concentrations (odds ratio 0\xE164, 95% CI 0\xE151\xD00\xE180). After a simi
4     the risk of incident coronary heart disease was 0\xE180 (95% CI 0\xE170\xD00\xE190) for a per-SD
5     lesterol efflux capacity and HDL cholesterol (r=0\xE140, 95% CI 0\xE136\xD00\xE142) and betweer
```

Figure 4. Example of a bundle with number slot (*aged_ years or older)*



```
Concordance Hits  46
Hit   KWIC
31              of diabetes as of January 1, 1997, were aged 40 years or older, and were members
32                      birthday; or had diabetes and were aged 40 years or older when they joined
33               was minimized by restricting to persons aged 40 years or older.CONCLUSIONSThere was
34    12 million inhabitants, of whom 1.5 million were aged 65 years or older. Elderly patients in
35              -positive (ie, infectious) patients who were aged 15 years or older, diagnosed at all
36                        a single centre. All patients were aged 18 years or older, and had cystic
37          ethnic and racial minority groups, and 20% aged 60 years or older) between 1996 and 1999.
```

Quite logically, bundles such as *xe ci xe xd,* shown in Figure 3, were eliminated from the initial list of candidates as they represented series of field-specific shorthand rather than actual lexical bundles. However, a quick look at the concordance list of *aged years or older* revealed the presence of a number between 'aged' and 'years' in all occurrences of the bundle; meaning that "aged [number] years or older" was actually a five-word bundle. Other bundles of the same type, as will be seen later, even included two slots for numbers.

If numbers were taken into account by the software, there would be little or no chance of identifying such combinations even as five-word bundles. Given that the age of patients vary from one study to another, there would be no guarantee that one age group would be targeted frequently enough for potential combinations such as *aged 20 years or,* or *20 years or older* to occur at least 20 times and across more than 12 articles. Yet, it clearly appears that the combination *aged years or older* has an important function in the MRAs as it provides key information on the study population. Such four-word combinations (i.e. with slots for numbers) were excluded from the list of four-word bundles – because they actually were five-word or six-word bundles – and are presented separately in this study given their potentially important rhetorical functions.

2.2.3 Working definition.

On the basis of the definitions of lexical bundles in the literature (e.g., Hyland, 2008a; Biber & Barbieri, 2007; Cortes, 2004; Biber et al., 1999) and in view of the above mentioned considerations, for the present study, were considered as four-word lexical bundles all recurring four-word combinations that (a) occurred at least 20 times in the MRAC; (b) were used across at least 12 different texts; (c) occurred in articles from at least two of the target journals; and (d) were uninterrupted by any other character.

2.3 Structural Classification Framework

The identified four-word lexical bundles were classified following the 12 categories identified by Biber et al. (1999). The authors suggested that though most lexical bundles in academic prose do not represent complete structural units, they can be classified according to their grammatical patterns. The following 12 categories were identified by Biber et al. (1999) as the most most frequent patterns in academic prose and the bundles listed are illustrative of examples from the MRAC.

- Noun phrase with *of*-phrase fragment: *a combination of, an increased risk of, the course of the, the proportion of patients, the results of the.*
- Noun phrase with other post-modifier fragment: *an important role in, an increase in the, a*

reduction in the, loss to follow up, death from any cause.

- Prepositional phrase with embedded *of*-phrase fragment: *at a dose of, for the treatment of, in the pathogenesis of, in the proportion of, by a factor of.*
- Other prepositional phrase fragments: *in accordance with the, with an increased risk, in contrast to the, of the primary outcome, in addition to the.*
- Be + noun/adjective phrase: *is consistent with the, is one of the, were similar to those.*
- Passive verb + prepositional phrase fragment: *are shown in figure, has been associated with.*
- Anticipatory it + verb/adjective phrase: *it is possible that, it is important to.*
- (Verb phrase) + that-clause fragment: *our data suggest that, studies have shown that, these findings demonstrate that, we found that the.*
- (Verb/adjective) + *to*-clause fragment: *is likely to be, to determine whether the, we were able to, more likely to have.*
- Adverbial clause fragment: *as compared with placebo, as shown in figure.*
- Pronoun/noun phrase + be (+…): *there was no difference, this is the first.*
- Other expressions: *did not differ between, intention to treat analysis, the presence or absence.*

As can be seen, there are instances in each of the 12 categories of lexical bundles in the MRAC. However, some categories occurred at a much higher frequency than others as will be seen in the Results section.

2.4 Functional Classification Framework

The analysis of the functions of the lexical bundles in the present study is based on Hyland's (2008a) functional framework developed on the basis of Biber's (Biber, 2006; Biber et al., 2004) taxonomy. Biber's taxonomy included three main functional categories summarized below with examples for each category drawn from the list of MRAC bundles.

- **Stance expressions** described as bundles used to express "attitudes or assessments of certainty that frame some other proposition" (p. 384). Some examples include *is likely to be, it is important to, were more likely to*;
- **Discourse organizers** used to indicate relations between prior and upcoming discourse. (e.g*., as well as the, on the other hand)*; and
- **Referential bundles** that directly refer to the textual context or to physical and abstract entities. (e.g. *is one of the, on the basis of, at the same time, at the end of, at the time of, as shown in figure)*

However, Biber et al. developed their functional taxonomy on the basis of a corpus that included both spoken (conversation and classroom teaching) and written (textbooks and academic prose) registers; and though the initial objective was to identify potential functions of lexical bundles in any given register, the authors found "dramatic differences" (p. 396) in the function types across the registers in their study. Indeed, Biber and Barbieri (2007) suggested that in the case of ESP, it is necessary to consider each register "on its own terms" (p. 265) in order to provide an adequate description of how the lexical bundles are used in a given register.

In that respect, Hyland (2008a), drawing from Biber's et al.'s taxonomy, developed a framework which, in his words "collects bundles into the three broad foci of research, text, and participants, and introduces sub-categories which specifically reflect the concerns of research writing" (p.13). That framework, summarized in table 2, was found more relevant to the genre under investigation in the present study (i.e., medical research articles) and therefore, was used to analyze the functions of the identified lexical bundles. The examples provided in Table 2 are taken from the list of bundles identified in the MRAC.

Table 2

Functional classification of lexical bundles in the field of research according to Hyland (2008a)

Category	Overall Function	Subcategory	Examples from the MRAC
Research-oriented	Structure writers' experiences / activities	Location (in time/place)	*at the time of, at the end of, in the present study.*
		Procedure	*the role of the, the purpose of the, the use of the.*
		Quantification	*a wide range of, the total. number of, one of the most.*
		Description	*the structure of the, the baseline characteristics of.*
		Topic (related to field of study)	*animal care and use, food and drug administration.*
Text-oriented	Organize text and its meaning	Transition signals (contrastive/ additive links)	*on the other hand, in addition to the, in contrast to the.*
		Resultative signals (inferential/ causative relations)	*these findings suggest that, as a result of, our data indicate that.*
		Structuring signals (organize discourse/ direct readers elsewhere in text)	*in the present study, as shown in figure, are shown in table.*
		Framing signals (specify limiting conditions)	*on the basis of, in the case of, in the presence of, in the case of, with respect to the.*
Participant-oriented	Focus on writer or reader	Stance features (writer's attitudes/evaluations)	*it is possible that, were more likely to, were consistent with the, is likely to be.*
		Engagement features (address reader directly)	*it is important to, studies are needed to, would be expected to.*

CHAPTER 3. RESULTS

3.1 Frequency of Four-word Bundles in the MRAC

Overall, 204 four-word bundles were identified in the 1,177,611 word MRAC for a total of 8,053 individual cases (see Appendix A for complete list of identified four-word bundles) based on the criteria set out above. Table 3 shows the top 50 bundles ranked by frequency, as well as their range or occurrences across different articles. As can be seen, the top 10 bundles all occurred over 100 times per million words and the most frequently used lexical bundle, *on the basis of,* occurred 306 times, across 122 articles – almost half of the total number of articles – and in all five target journals.

Table 3

Top 50 four-word bundles in the MRAC

Bundles	F*	R**	Bundles	F	R
1. on the basis of	306	122	**23.** there were no significant	57	36
2. in the placebo group	235	25	**24.** for the treatment of	56	36
3. with the use of	228	57	**25.** the basis of the	56	42
4. in the control group	199	27	**26.** the intention to treat	55	31
5. in the presence of	139	50	**27.** end of the study	54	25
6. at the time of	137	83	**28.** as compared with the	52	32
7. the end of the	129	59	**29.** as well as the	51	42
8. in the absence of	117	67	**30.** the results of the	51	39
9. in the intervention group	115	13	**31.** did not differ between	50	22
10. at the end of	109	49	**32.** the use of a	50	35
11. an increased risk of	83	36	**33.** were included in the	50	37
12. of patients in the	79	27	**34.** presence or absence of	49	24
13. years of follow up	74	26	**35.** randomly assigned to		
14. were randomly assigned to	72	43	receive	49	31
15. in the number of	71	39	**36.** the proportion of patients	48	19
16. was associated with a	71	46	**37.** associated with an		
17. in the context of	69	46	increased	47	25
18. the total number of	66	41	**38.** in the present study	47	29
19. the primary end point	65	23	**39.** have been shown to	46	39
20. the use of the	65	35	**40.** with an increased risk	45	24
21. has been shown to	64	45	**41.** in accordance with the	44	39
22. did not differ significantly	63	39	**42.** between the two groups	42	24

Table 3 (Cont'd)

Bundles	F*	R**	Bundles	F	R
43. there was no significant	63	41	**47.** was defined as a	41	32
44. was approved by the	59	55	**48.** were more likely to	41	26
45. as a result of	58	46	**49.** are shown in table	39	33
46. these data suggest that	58	39	**50.** in the risk of	38	22

* *F = Frequency* ** *R = Range*

3.2 Structures of Identified Bundles

Data regarding the structures of the identified bundles are shown in tables 4 and 5. Table 4 provides a classification of the 204 bundles according to their structural correlates; and table 5 shows the proportion of each structure in the list of identified bundles. As can be seen, bundles beginning with prepositional phrases and noun phrases were far more frequently used in the medical articles in the present study. With 31.37% of all identified bundles being prepositional phrases and another 23.03% being noun phrases, these two categories account for more than half (54.40%) of the four-word bundles in the MRAC.

Table 4

Structural classification of four-word bundles in the MRAC

Noun phrase with* of*-phrase fragment

a p value of, a result of the, an increased risk of, and the number of, and the risk of, duration of follow up, efficacy and safety of, end of the study, median follow up of, or a combination of, p value of less, presence or absence of, proportion of patients with, the absence of a, the baseline characteristics of, the basis of a, the basis of the, the course of the, the design of the, the effect of the, the end of the, the number, of patients, the presence of a, the proportion of patients, the results of our, the results of the, the time of the, the total number of, the use of a, the use of the, years of age or, years of follow up.

Noun phrase with other post-modifier fragment

a median follow-up, a reduction in the, a significant increase in, a significant reduction in, an important role in, an increase in the, death from any cause, increase in the number, loss to follow up, no significant difference in, no significant differences between, one of the most, power to detect a, significant differences between the, significant increase in the.

Table 4 (Cont'd)

Prepositional phrase with embedded **of-*phrase fragment***

as a result of, as part of the, as the number of, at a dose of, at the end of, at the level of, at the time of, by a factor of, by the end of, by the presence of, for a total of, for each of the, for the presence of, for the treatment of, in the absence of, in the case of, in the context of, in the development of, in the incidence of, in the number of, in the pathogenesis of, in the presence of, in the proportion of, in the rate of, in the regulation of, in the risk of, in the setting of, in the treatment of, in view of the, of the number of, on the basis of, over the course of, to the development of, with a history of, with the exception of, with the use of.

Other prepositional phrase fragments

as well as in, at baseline and at, at the same time, by intention to treat, during the study period, for the primary outcome, in accordance with the, in addition to the, in contrast to the, in kilograms divided by, in the control group, in the current study, in the general population, in the intervention group, in the placebo group, in the present study, in the study and, in the two groups, in this study we, of participants in the, of patients in the, of the patients in, of the patients who, of the patients with, of the primary outcome, on the other hand, with an increased risk, with respect to the.

Be + *noun/adjective phrase*

are consistent with the, is consistent with the, is one of the, was similar to that, were more likely to, were similar in the, were similar to those.

Passive verb + prepositional phrase fragment

are shown in figure, are shown in table, associated with an increased, been shown to be, compared with those in, has been associated with, has been shown to, have been associated with, have been shown to, lost to follow up, to be associated with, was added to the, was approved by the, was associated with a, was associated with an, was considered statistically significant, was defined as a, was defined as the, was not associated with, was obtained from the, was performed with the, were approved by the, were excluded from the, were included in the, were randomly assigned to, would be expected to.

Anticipatory* it + *verb/adjective phrase

it is possible that, it is important to.

(Verb phrase)* + that-*clause fragment

here we show that, our data indicate that, our data suggest that, studies have shown that, these data suggest that, these findings suggest that, these results demonstrate that, these results indicate that, these results suggest that, we found that the, we have shown that.

(Verb/adjective)* + to-*clause fragment

is likely to be, more likely to be, more likely to have, randomly assigned to receive, studies are needed to, to assess the effect, to determine whether the, we were able to, we were unable to.

Adverbial clause fragment

as compared with placebo, as compared with the, as shown in figure.

Pronoun/noun phrase* + be *(+…)

there was a significant, there was no difference, there was no evidence, there was no significant, there were no differences, there were no significant, this is the first.

Other expressions

an intention to treat, and in the control, and in the placebo, animal care and use, as well as the, assess the effect of, between the two groups, data and safety monitoring, did not differ between, did not differ significantly, food and drug administration, had no effect on, informed consent was obtained, intention to treat analysis, intention to treat population, less than was considered, play a role in, primary end point was, reduce the risk of, study was approved by, than in the control, than in the placebo, the efficacy and safety, the follow up period, the intention to treat, the presence or absence, the primary end point, the primary outcome was, we did not observe.

Table 5

Distribution of bundle structures in the MRAC

Structure	Number of bundle types	% of total number of bundles
Prepositional phrase with embedded *of*-phrase fragment	36	31.37
Other prepositional phrase fragments	28	
Noun phrase with *of*-phrase fragment	32	23.03
Noun phrase with other post-modifier fragment	15	
Other expressions	29	14.22
Passive verb + prepositional phrase fragment	26	12.75
(Verb phrase) + *that*-clause fragment	11	05.40
(Verb/adjective) + *to*-clause fragment	09	04.41
Be + noun/adjective phrase	07	0 3.43
Pronoun/noun phrase + *be* (+…).	06	02.94
Adverbial clause fragment	03	01.47
Anticipatory *it* + verb/adjective phrase	02	00.98
Totals	204	100

3.3 Functions of lexical bundles in the MRAC

As mentioned in the previous chapter, the functional analysis of the MRAC lexical bundles was based on the taxonomy developed by Hyland's (2008a) which included three main categories: research-oriented bundles, text-oriented bundles, and participant-oriented bundles. Each of these categories included a number of subcategories with (a) research-oriented bundles expressing location (in time and place), procedure, quantification, description, and topic; (b) text-oriented bundles used to serve the functions of transition signals, resultative signals, structuring signals, and framing signals; and (c) participant-oriented bundles including the functional subcategories of stance features and engagement features.

To facilitate the classification of the MRAC bundles, Cortes's (2004) taxonomy was used to supplement some of the subcategories identified by Hyland (2008a). For example, the subgroup *statistical* was added to the category of quantifiers to accommodate bundles such as *no significant difference between, did not differ significantly, there was no evidence,* and so on. In the same way, the function *comparison* was added to the subcategory of resultative signals to include bundles such as *as compared with the, compared with those in, than in the placebo,* so on and so forth.

The functions of the 204 four-word lexical bundles were determined by analyzing the concordance lists generated by Antconc and which provided several instances of each bundle used in context. Table 7 shows the distribution of bundle functions in the MRAC. As can be seen, Research-oriented bundles were far more frequent in the MRAC; with a total of 122 bundles representing more than half (58.65%) of all the four-word bundles identified in the MRAC. On the other hand, data in table 5 indicate a quite low frequency of participant-oriented bundles in the medical research articles used in the present study; with only 14 bundles in this category, representing less than 7% of all identified bundles in the MRAC. The classification of the 204 bundles according to their functions in the MRAC is provided in table 6.

As can be seen in table 6, some bundles appear in more than one category or subcategory. This is because the analysis of the concordances revealed that they could serve different functions depending on syntactic constructions. For example, *at the same time* mostly functioned as a research-oriented bundle indicating simultaneous procedures. In such cases, it was often followed by "*as"*; as shown in example 6 below.

(6) Mice that received 100,000 $CD4^{+}CD25^{+}$ T cells *at the same time* as receiving naïve $CD4^{+}$ T cells served as healthy controls. (08Science5)

However, when used at the beginning of a sentence, *at the same time* appeared to be a text-oriented bundle used as a transition signal to express some contrastive relationship between observations and/or findings of the study.

(7) As reported in Table 3, costs associated with medications not part of the dialysis treatment but provided at the time of the dialysis visit decreased significantly after the imposition of budget caps. *At the same time*, the data revealed a substantial increase in non-dialysis-related outpatient visits and the costs for drugs provided to dialysis-dependent patients during their non-dialysis-related outpatient visits. (15NEJM2)

Table 6

Distribution of bundle functions in the MRAC

	Research-Oriented	Text-Oriented	Participant-oriented	Totals
Number of Bundles	122	72	14	208*
%	58.65	34.62	6.73	100

* total is higher than 204 because of certain bundles appearing in more than one category.

Table 7

Classification of identified bundles according to their functions in the MRAC.

Category	Subcategory	Bundles
Research-oriented	Location (in time/ place)	at the end of, at the level of, at the time of, between the two groups, by the end of, over the course of, at baseline and at, **at the same time**, duration of follow up, during the study period, in the control group, **in the current study**, **in the present study,** in the general population, in the intervention group, in the placebo group, in the present study, in the study and, in the two groups, in this study we, end of the study, the course of the, the end of the, the time of the, and in the control, and in the placebo, the follow up period.

Table 7 (cont'd)

	Procedure	an important role in, assess the effect of, by intention to treat, by the presence of, for the presence of, for the primary outcome, for the treatment of, **in the treatment of,** informed consent was obtained, power to detect a, play a role in, randomly assigned to receive, reduce the risk of, study was approved by, the design of the, the use of a , the use of the, to assess the effect, to determine whether the, was added to the, was approved by the, was considered statistically significant, was defined as a, was defined as the, was obtained from the, was performed with the, were approved by the, were excluded from the, were included in the, were randomly assigned to, with the use of.
	Quantification	a median follow-up, median follow up of, years of follow up, a p value of, an increase in the, an increased risk of, and the number of, a reduction in the, as part of the, as the number of, at a dose of, by a factor of, for a total of, for each of the, in kilograms divided by, increase in the number, is one of the, less than was considered, of the number of, p value of less, one of the most, proportion of patients with, the number of patients, the proportion of patients, the total number of, years of age or. ***Statistical:*** did not differ between, did not differ significantly, no significant difference in, no significant, differences between, significant differences between the, significant increase in the, there was a significant, there was no difference, there was no evidence, there was no significant, there were no differences, there were no significant, a significant increase in, a significant reduction in.
	Description	and the risk of, baseline characteristics of the, or a combination of, primary end-point was, the baseline characteristics of, the basis of the, the basis of a, the primary outcome was, this is the first, with a history of, with an increased risk.
	Topic	an intention-to-treat, animal care and use, data and safety monitoring, death from any cause, efficacy and safety of, food and drug administration, intention-to-treat analysis, intention-to-treat population, loss to follow up, lost to follow up, the efficacy and safety, of the primary outcome, the intention-to-treat, the primary end-point
Text-oriented	Transition signals (contrastive/ additive links)	**at the same time**, as well as the, in addition to the, in contrast to the, on the other hand

Table 7 (cont'd)

	Resultative signals (Comparison, Inferential/ Causative Relations)	a result of the, are consistent with the, as a result of, as compared with placebo, as compared with the, associated with an increased, compared with those in, had no effect on, here we show that, is consistent with the, our data indicate that, our data suggest that, than in the control, than in the placebo, the effect of the, the results of our, the results of the, these data suggest that, these findings suggest that, these results demonstrate that, these results indicate that, these results suggest that, to be associated with, to the development of, was associated with a, was associated with an, was not associated with, was similar to that, we did not observe, we found that the, we have shown that, we were able to, we were unable to, were similar in the, were similar to those.
	Structuring signals	**in the current study, in the present study**, as shown in figure, are shown in figure, are shown in table.
	Framing signals (specify limiting conditions)	as well as in, in accordance with the, in the absence of, in the case of, in the context of, in the development of, in the incidence of, in the number of, in the pathogenesis of, in the presence of, in the proportion of, in the rate of, in the regulation of, in the risk of, in the setting of, **in the treatment of**, in view of the, of participants in the, of patients in the, of the patients in, of the patients who, of the patients with, on the basis of, presence or absence of, the absence of a, the presence of a, the presence or absence, with respect to the, with the exception of.
Participant-oriented	Stance features (writer's attitudes/ evaluations)	is likely to be, it is possible that, more likely to be, more likely to have, were more likely to.
	Engagement features	been shown to be, has been associated with, has been shown to, have been associated with, have been shown to, studies have shown that, studies are needed to, would be expected to, it is important to

3.4 Technical and Subtechnical Lexis in the MRAC Bundles

The aim in this section is not to discuss the importance of technical vocabulary, described by Chung and Nation (2003) as words "that are largely unique to a particular specialized field" (p. 111); nor subtechnical vocabulary – academic vocabulary elsewhere – which was defined in Lam (2001) as "words that have one or more 'general' English meanings and which in technical contexts take on extended meanings" (p. 1). Instead, this section simply reports a finding that might be of interest to EAP/ESP instructors and learners.

One legitimate and understandable expectation from the present study was the presence of technical lexical items in the MRAC bundles. However, as can be seen in the tables 3 and 4 above, lexical items comprised in the identified bundles are typically subtechnical. The vast majority of these subtechnical items in the MRAC bundles can be found in the Medical Academic Word List (MAWL) established by Wang et al. (2008) from a corpus of over a million words from 288 medical articles. It is often the case in ESP that there is a thin line between technical and subtechnical vocabulary, but with the help of field specialists and experienced ESP practitioners, Wang et al., were able to identify and exclude from the MAWL 27 word families that were considered by they consulted as "purely technical".

A quick search of these purely technical 27 word families revealed a high frequency of such words in the MRAC with 17 words (out of 27) occurring more than 100 times and across a wide range of texts. For instance, a search of the headword *cardia* yielded 1,171 tokens across a range of 94 articles in all five journals (see Table 8 for the frequency and range of the 27 words in the MRAC). However, in spite of their high frequency in medical articles – at least in those used in the present study – it appears that technical vocabulary items do not collocate with other words in patterns regular enough to be identified as lexical bundles. Only one lexical item (*pathogenesis*) out of the 27 technical words appeared in the four-word bundles identified in the MRAC (*in the pathogenesis of)*. The implications of such finding are discussed later in the present paper.

Table 8

Distribution of "purely technical" word (Wang et al., 2008) in the MRAC

Rank*	Headword	F**	R***	Rank	Headword	F	R
1	Cardia	1171	94	15	stent	162	7
2	Aortic	487	20	16	cerebral	156	29
3	phenotypic	465	82	17	**pathogenesis**	**112**	**45**
4	pulmonary	409	40	18	Carcinoma	95	16
5	vivo	392	14	19	cutaneous	87	7
6	Epithelial	355	36	20	hemorrhage	57	19
7	cytokine	322	41	21	exogenous	47	23
8	ischemia	300	32	22	Posterior	45	15
9	Lymphoid	290	49	23	Necrosis	38	18
10	Vitro	265	75	24	Anterior	38	17
11	Mitochondrial	205	20	25	dorsal	35	12
12	Lysis	204	61	26	Pathophysiology	34	20
13	hepatic	190	31	27	Situ	33	17
14	Ligand	177	35				

** Rank in the MRAC*
*** F = Frequency (number of tokens)*
**** R = Range*

3.5 Bundles with number slots

In total, 10 different bundles with number slots were identified in the MRAC; which amounts to 316 individual cases. Table 9 shows the identified bundles with slots indicated with an underscore sign (_). These bundles did not follow one specific structure but appeared to be all research-oriented bundles that mainly fell under two subcategories: quantification and description.

The quantifying bundles in the list of bundles with number slots were mainly used to provide information on the study population (example 1) or the study procedure (example 2). Some bundles like the ones in examples 1b and 4 required two number slots. The bundles as identified by the software age shown in italics and the interposed numbers, revealed by the concordance lists, are in bold type.

(1) a- *A total of* **398** *patients* proceeded to the randomization phase (11JAMA1)

b- Participants had to be **30** *to* **70** *years of age* and have a body-mass index (the weight in kilograms divided by the square of the height in meters) of 25 to 40. (9NEJM5)

(2) A total of 3083 deaths from all causes were observed over a mean *follow-up of* **15.8** *years*, with a maximum *follow-up of* **36** *years* among women. (14NEJM1)

On the other hand, the descriptive bundles were used exclusively to provide information on field-specific experiments (example 3) or laboratory procedures (examples 4).

(3) Montelukast at *a dose of* **4** *mg* daily was added for children (07NEJM2)

(4) Samples were further postfixed in 1% osmium tetroxide *for* **2** *hours at* **4°***C*, dehydrated with graded concentrations of alcohol, and embedded in Epon. (10JCI1)

Table 9

Bundles with number slots in the MRAC

Rank	Bundle	Freq	Range
1	children younger than_ years	60	12
2	aged _ years or older	46	20
3	overnight at _º c	36	23
4	a total of _ patients	32	23
5	follow up of _ years	29	17
6	_ to _ years of age	28	16
7	for _ hours at _ c	22	16
8	for _ minutes at _ c	22	14
9	a dose of _ mg	21	14
10	stored at _ º c	20	14

CHAPTER 4. DISCUSSION OF FINDINGS

The purpose of the present study was to identify four-word lexical bundles frequently used in medical research articles and to analyze the functions they serve in such articles. Chapter 3 introduced the four-word bundles identified in the MRAC as well as the classification of such bundles according to their functions in context. The present chapter proceeds to the discussion of the results presented in chapter 3.

4.1 Structures of the MRAC Lexical Bundles

As mentioned in the previous chapter, the list of four-word bundles identified in the MRAC included all 12 structural categories of lexical bundles identified by Biber et al. (1999) as frequently used in academic prose. Clearly, bundles in each of these categories play important roles in medical articles as attested by their frequency and range in the MRAC. However, for time and space constraints, not all categories can be discussed in the present study. Therefore, this section particularly focuses on three main categories of bundles identified in the MRAC: (1) bundles starting with noun phrases (i.e., NP + *of*-phrase fragment, and NP + other post-modifier fragment); (2) bundles starting with prepositional phrases (i.e., PP + *of*-phrase fragment, and other prepositional fragments); and (3) the category labeled *Other Bundles*, which according to Biber et al. (1999) include "lexical bundles that do not fit neatly into any of the other categories" (p.1024). While the focus on the *Other Bundles* category is mainly motivated by the predominance of topic-specific bundles in this category, bundles in the forms of prepositional phrases and noun phrases are discussed in this section first and foremost because of their high frequency (hence their importance) in the MRAC.

Indeed, as shown in table 4 in the previous chapter, four-word bundles beginning with noun phrases and prepositional phrases accounted for more than half of all identified bundles in the MRAC. These results are consistent with findings of previous studies of lexical bundle such as Biber et al. (1999), Cortes (2004), and Hyland (2008a), to name but a few. Biber et al. found that more than 60% of lexical bundles identified in academic prose were parts of prepositional or noun phrases. In the same vein, Hyland identified noun phrases with embedded *of*-fragments as the most common structure across his corpus of academic texts from four different disciplines: Electric Engineering, Biology, Business Studies, and Applied Linguistics. Similar results were found by Cortes (2004) in her analysis of lexical bundles identified in her corpora of published texts in History and Biology.

In all three studies, the authors appeared to attribute the high frequency of prepositional and noun phrases in their different corpora to the fact that "academic writing is structurally more phrasal in nature than conversation" (Cortes, 2004; p. 404). As such, academic writing involves frequent use of post-nominal modifications which include (a) genitive expressions, with the use of *of*-phrase fragments as a post-modifiers of noun phrases (e.g., *the end of the, the total number of, the basis of the, the presence of a)*, or of nouns in prepositional phrases (e.g., *as a result of, in the treatment of, in view of the, with the use of*); and (b) other post-modifier fragments other than *of*-phrases that can be used with noun phrases (e.g., *an important role in, a reduction in the, death from any cause, power to detect a),* or prepositional phrases (*in addition to the, in accordance with the, of patients in the, of patients with a).*

As already mentioned, the results of the present study are no exceptions to the findings of previous studies of the structures of lexical bundles in academic writing. The four structures comprising prepositional phrases and/or noun phrases (NP + *of*-fragment, NP + other post-modifier fragments, PP + *of*-fragment, and other PPs) were used much more frequently than the other structure in the MRAC; and the lexical bundles following these structural patterns were used to convey a wide range of meanings, some of which are presented below.

4.1.1. Noun phrases + embedded *of*-phrase fragment.

A look at concordances of lexical bundles in this category revealed that they were mainly used to convey meanings already identified by Marco (2000) in her analysis of collocates that usually fill the framework *the … of* and *a… of* (i.e., *the/a + noun [= NP] + of)* in medical research articles. Her findings were indeed consistent with the earlier claim by Biber et al. (1999) that lexical bundles in this category are mainly used to (1) describe place, size, and amount; (2) mark existence or presence; (3) identify abstract qualities; and (4) describe events or processes. Marco's classification was more specific to medical research articles and better corresponds to the meanings of noun phrases with embedded *of*-fragments in the MRAC. It included six groups – four of which correspond to the four categories identified by Biber et al. – which are discussed below with reference to bundles identified in the MRAC.

(1) Expressing quantity/measure.

Bundles in this group are used to provide information about elements of the study that need to be quantified. In the MRAC, such elements mainly included:

- objects of the study;

(1a) However, in the *Dnahc5*^{del593} mutants, which are genetically homogeneous as compared with the genetic heterogeneity of human patients, the heart anomalies are variations of a defined set of defects, with DORV being *one of the most* common cardiac malformation. (07JCI1)

(1b) Other diabetes medications were also associated with *an increased risk of* pancreatic cancer, suggesting reverse causality because an early manifestation of this cancer is hyperglycemia. (15JAMA3)

- participants in the study; and

(1c) Once patients were deemed to be virological failures, they were counted as failures at the time of failure and at all subsequent time points when assessing *the proportion of patients* achieving virological response. (07Lancet3)

(1d) In order to increase the power for secondary end points, *the total number of* enrolled patients was increased to 18 and materials/research related observations (08NEJM1)

- materials used in the study.

(1e) The PSS indicated *the total number of* drugs in the optimised background regimen to which the patient's viral isolate showed phenotypic sensitivity. (07Lancet3)

(1f) In an independent cohort of patients, we found that *an increased number of* CD68 macrophages was correlated with a shortened progression-free survival (P=0.03). (10 NEJM3)

(2) Explaining medical/experimental procedures

(2a) The anesthetic technique was based on either total intravenous anesthesia *or a combination of* intravenous opioids and muscle relaxants in combination with volatile anesthetics (12JAMA2)

(2b) Hypokalemia can be managed with *the use of a* potassium-sparing agent or supplemental potassium chloride. (09NEJM3)

(3) Expressing qualities or properties.

(3a) The primary objective of this study was to determine the clinical *efficacy and safety of* chaperonin 10 in patients with moderate-to-severe active rheumatoid arthritis, even with concurrent treatment with standard DMARDs. (06LANCET5)

(4) Marking existence/non existence.

(4a) All recorded episodes of parasitemia were included, irrespective of *the presence or absence* of symptoms of malaria (11NEJM5)

(4b) From a therapeutic point of view, *the absence of a* single *Gpr3* allele in *APP/PS1; Gpr3*$^{+/-}$ mice was sufficient to improve cognition to a similar level as deletion of both *Gpr3* alleles. (15Science2)

(5) Indicating a moment of time in a process

(5a) Finally, M70 was demonstrated to be safe and well tolerated during *the course of the* 7-day trial. (14 Science5)

(5b) Simple effects analyses revealed no significant change in sleep onset latency for all 4 treatment groups from *the end of the* acute 6-week phase to *the end of the* extended 6-month phase. (09JAMA5)

(6) Expressing relationship among elements.

Lexical bundles in this group were used to express a variety of relationships among elements of the research mainly cause, effect, and results.

(6a) The multiorgan nature of disease in *AIRE*-deficient animals appears to be *a result of the* spectrum of self-antigens whose thymic expression relies on Aire. (09Science4).

(6b) This finding was expected from the known pharmacology of the drug and specifically from its use dependence-ie, that *the effect of the* drug in blocking sinoatrial node I^sub f^ channels is greatest when these channels are most likely to be open, that is, when heart rates are highest. (10Lancet5)

4.1.2. Noun phrases with other post-modifier fragments.

Bundles in this category were a little less frequent than those in the group discussed above. Based on their meanings in context, they can be divided into three main groups as suggested by Biber et al. (1999).

(1) Expressing the way a process or event occurred.

(7a) Using a logrank test (α=0.05 and β=0.10 for a two-tailed test), and assuming a 10% *loss to follow up*, we estimated that the necessary sample size would be 240 patients in each group. (06Lancet3)

(7b) The primary outcome measure was *death from any cause*, assessed 28 days after enrollment in the study. (09JAMA1)

(2) Indicating the relationship between elements of the study

(8a) Our studies add to and integrate these findings by demonstrating that mitochondria play *an important role in* the regulation and effector phases of CS-induced responses. (15JCI3)

(8b) The sample size also provided 90% *power to detect a* treatment difference of 2 mm Hg in the low-dose comparisons (ie, at week 4). (07Lancet2)

(3) Quantifying objects of the study, materials, or participants when necessary.

(9a) When W2mefΔEBA175 was cultured with neuraminidase-treated erythrocytes to inhibit any residual SA-dependent interactions, there was *an increase in the* difference in the mean inhibition of W2mef-WT versus W2mefΔEBA175 by samples… (08JCI2)

(9b) In the CAPRISA 004 trial, pericoital administration of 1% TFV gel was associated with *a reduction in the* risk of HIV-1 acquisition of 39% (95% CI, 6 to 60). (15NEJM4)

(9c) We then determined whether there was *an increase in the* number of IFN-γ–producing T cells with specificity for VM in *Aire*$^{o/o}$ mice. (09Science4)

An interesting observation in this section is the way the bundle *an increase in the* collocates with a variety of nouns to quantify different elements of the study. Its counterpart, which learners could easily think to be *a decrease in the* did not appear in the list of identified four-word bundles. Instead, the bundle *a reduction in the (9b),* appeared to be the preference of MRA writers over *a decrease in the*, even though the two bundles presumably convey similar meanings. Although the reasons for such preference are unclear, cases such as this one clearly point at the importance of raising novice writers' awareness of how discourses are constructed in their specific fields. As suggested by Cortes (2006), learners often use alternative word combinations to express functions similar to those conveyed by expressions used in texts written by professionals. Therefore, it is important to introduce them – native and non-native speakers of English alike – to the language practices of their academic communities in order to help them become members of such communities.

4.1.3. Prepositional phrases with embedded *of*-phrase fragment.

This category includes the most frequently used bundle in the MRAC *on the basis of.* The analysis of the concordance list of this highly frequent bundle revealed that *on the basis of* was exclusively used as a framing device in a variety of contexts including, but not limited to: describing research methodology *(10a),* describing experimental procedures *(10b)*, or reporting findings *(10c).*

> *(10a)* These levels were chosen *on the basis of* expert consultation, the guideline from the American Academy of Pediatrics,[16] and data from Johnson et al.[29] (15NEJM1)

> *(10b)* For example, SOMs were used to analyze hematopoietic differentiation *on the basis of* gene expression data, and this method was also applied to gene expression data from human lung cancers as a means of differential diagnosis. (08Science2)

> *(10c)* In addition, there was no significant interaction *on the basis of* age, the degree of existing malnutrition, the severity of illness, or the timing of the initiation of nutritional support. (14NEJM5)

The most frequent bundles in this category, however, did not start with the preposition *on* like *on the basis of* (the latter is actually the only one in this case). Indeed, the majority of bundles in the MRAC following the structure *PP + embedded of-phrase* started with the preposition in (e.g., *in the presence of, in view of the, in the setting of).* Like *on the basis of,* all of them were used in the MRAC mainly to set boundaries for arguments being presented.

(11a) Then, we built on our knowledge that the membrane protein asialoglycoprotein receptor1 (ASGR1) is a slow-maturing protein whose secretion is decreased under ER stress conditions such as *in the setting of* obesity[17]. (15Science1)

(11b) Focal adhesions (hsa04510) control cytoskeletal or adhesion dynamics and thus affect both leucocyte motility within intima and interactions between platelets and endothelium, all of which play a part *in the pathogenesis of* coronary artery disease. (12Lancet5)

In addition to two groups of bundles mentioned above, one more bundle in this category (*with the exception of)* was also used to perform similar functions as a framing signal in a variety of contexts.

Finally, other groups in this category included bundles used for (a) quantification (e.g., *at a dose of, for a total of);* (b) marking time relations (e.g., *at the time of, by the end of);* and (c) description of procedure (e.g. *with the use of, for the treatment of).*

4.1.4. Other prepositional phrases.

Bundles in this category also conveyed a variety of meanings, most of which were similar to those discussed above. The main differences lied in the prepositions at the beginning of the expressions. For example, bundles used to indicate boundaries of events or arguments being presented mainly started with the preposition *of* (*of participants in the, of patients with, of patients who*)*,* while expressions starting with *in* were mainly used to specify a location in the description of procedures and findings (*in the control group, in the study population, in the placebo group)*.

Other bundles in this category, starting with various prepositions were used for a variety of other functions, including: the identification of specific time period (12a, 12b), and the description of study procedures or methods (13)

(12a) Neither tipranavir nor darunavir were selected as part of the optimised background regimen *during the study period* reported here. (07Lancet3)

(12b) The following studies were performed for a subset of patients *at baseline and at* week 76 or at the time of early termination: (13NEJM5)

(13) Subgroup analyses *for the primary outcome* and for major cardiovascular events were evaluated with the use of tests for interaction for prespecified baseline features. (08NEJM5)

It should also be noted that this category included all transition signal bundles – discussed later in this paper – identified in the MRAC. These bundles served mainly the purposes of:

(1) expressing contrastive relationships between events, elements of the study, or observations

(14a) Adjusted *P* values using false discovery rate (FDR) analysis to account for multiple testing remained significant in the AA population (Table 2). *At the same time*, we did not observe an association for heart rate in the FC or AA cohorts. (09JCI5)

(14b) The Dallas Heart Study used a newer fluorescently labelled (BODIPY) cholesterol method to assay cholesterol efflux capacity, *in contrast to the* more established radiolabelled cholesterol assay that we used in the present study. (13Lancet1)

(2) adding supplementary information or supporting arguments to previously stated events or observations.

(15a) The proteases that control cell death and cell movement of the host cell, *as well as the* factors that mediate Ca^{2+} accumulation and prevent PS exposure, appear to be essential to parasite survival. (06Science5)

(15b) Strikingly, *in addition to the* positive indicators of oxidative stress, MAFA and FOXO1, only the nuclear content of the islet β cell–enriched NKX6.1 transcription factor was affected (Figure 4A). (13JCI2)

The last but not least category of bundles to be discussed in this section is the *Other Bundles* category to which I now turn.

4.1.5. Other bundles.

As mentioned earlier, the majority of the bundles in this category were subject-specific bundles, that is, bundles especially related to the field of research (Hyland, 2008a) or including words that are closely related to the field (Cortes, 2004). Based on their structures, these topic-specific bundles were divided into the two following groups: (1) bundles including coordinated binominal phrases; and (2) nominalizations.

(1) Bundles including coordinated binominal phrases

Biber et al. (1999) defined binominal phrases as “two words from the same grammatical category, coordinated by *and* or *or”* (p. 1030). In other words, binominal phrases consist of pairs of words (from the same part of speech) that commonly collocate in written or oral discourse. Biber et al. further explained that binominal phrases can comprise words from all four major grammatical categories, namely nouns, verbs, adjectives, and adverbs; thus giving way to the following four possible combinations: *noun and noun, verb and verb, adjective and adjective,* and *adverb and adverb.*

The binominal phrases in the MRAC subject-specific bundles consisted exclusively of *noun and noun* combinations; which is in line with the findings of Biber et al. that *noun and noun* combinations were by far the most common binominal phrases in academic prose. Consistent with Hyland’s and Cortes’s descriptions of subject-specific bundles, the noun pairs in the MRAC subject-specific bundles were closely related to the field of medicine. Below are some examples of bundles from this group, used in context. As can be seen, some of these bundles represent field-related institutions or organizations (16c).

(16a) We did a phase 3 study to assess *the efficacy and safety* of VTD versus TD as induction therapy in preparation for double autologous stem-cell transplantation in newly diagnosed multiple myeloma. (10Lancet2)

(16b) Safety data were reviewed by the *data and safety monitoring* board at the midpoint of the study (June 2013). (15NEJM1)

(16c) In 2003, the US *Food and Drug Administration* and the manufacturer agreed to this 10-year observational study to evaluate the potential risk of bladder cancer with pioglitazone use in humans. (15JAMA3).

(2) Nominalizations

Hyland and Tse (2007) described nominalization as "the way that writers in the sciences regularly transform experiences into abstractions to create new conceptual objects" (p. 244). Examples from the MRAC bundles include expressions such as *the intention-to-treat, intention-to-treat population, loss to follow-up.* As suggested by Hyland and Tse, the importance of such nominalizations lies in the fact that they enable writers to give stable names to new concepts and deal with them with no further explanations. This lends support to the claim by Wray and Perkins (2000) that lexical bundles may serve as processing shortcuts since they are stored and retrieved as chunks, thus enabling speakers and listeners to focus on processing only new information (p. 16). Indeed, each of the examples provided above represents a whole concept, which, once internalized may not require deep processing during reading or writing. However, Hyland and Tse also suggested that novice writers and readers, native and non-native speakers of English alike, might not be able to pick up such nominalizations on their own, and might need help to unravel their meanings in their specific fields. Below are some examples of these nominalizations used in context.

(17a) The *intention-to-treat analysis* compared each intervention with placebo on a modified product-limit life-table distribution with a log-rank test statistic. (09Lancet2)

(17b) An exploratory analysis of prospectively defined secondary outcomes included each separate component of *the primary end point* (i.e., death, MI, stroke, renal failure, or respiratory failure, within the first 30 days); […] (12JAMA2)

(17c) We defined *loss-to-follow-up* as confirmed information that a participant had moved beyond the possibility of visiting, usually to India. (08Lancet5)

This section provided an overview of the structures of the most frequent and possibly, most relevant four-word bundles identified in the MRAC, as well as the range of meanings they cover in the medical research articles used in this study. The next section discusses the three main functional categories served by the MRAC bundles and the relationships between these functional categories and the structural patterns of the bundles in these categories.

4.2. Functions of the MRAC Lexical Bundles

As mentioned in the previous chapter and as can be seen in in Table 5 and Table 6, the functional analysis of the 204 identified lexical bundles revealed a clear predominance of research-oriented bundles in the MRAC. Such an observation is consistent with the findings in Hyland (2008a) which revealed that almost half of all 4-word bundles identified in his science and technology corpora consisted of research-oriented bundles. Hyland attributed the predominance of research-oriented bundles in his science and technology texts to the very nature and requirements of writing in hard sciences – which, of course, include medicine. He argued that in scientific ideology, the *empirical* prevails over the *interpretative* with a focus on on empirical demonstrations and experimental results (i.e. on the research itself: practices, methods, materials, procedures, and results), rather than on individual interpretations. Therefore, research-oriented bundles which mainly contribute to the description of the research objects, contexts, and procedures “function to impart a greater real-world, laboratory-focused sense” to the scientific texts (p. 14). The next section looks at what MRAC lexical bundles were used to serve such function, that is, lay the emphasis on how the research was conducted.

4.2.1 Research-oriented bundles in the MRAC.

As can be seen in Table 6, research-oriented bundles served a wide range of functions going from indicating location in time and place to quantifying research materials and participants to describing procedures and/or research objects, to name but a few. Most of these functions have already been discussed in the previous sections in examples where they were realized with bundles including prepositional phrases and noun phrases. Hyland (2008a) found strong relationships between some functional and structural categories, with a great predominance of *NP + of-phrase fragment* in the research-oriented bundles, a high frequency of prepositional phrases in text-oriented functions, and *anticipatory it* dominating participant-oriented bundles.

Such findings were only partially borne out by the results of the present study. Clearly, there were some structural patterns that appeared to be particularly related to some functional categories in the MRAC, as will be seen later. But both bundles starting with noun phrases and those starting with prepositional phrases were widely used in the research-oriented bundle category and across all its subcategories; and there were only two bundles in the whole MRAC with *anticipatory it (it is important to, it is possible that)*. Even though the latter appeared in the list of participant-oriented bundles, they were far from being the predominant structure in this functional category. On the other hand, prepositional phrases did dominate text-oriented bundles in the MRAC.

Such variations in findings are not surprising as other studies (e.g. Biber et al., 2004; Cortes, 2004; Hyland, 2008a) have shown that writers in various fields show different preferences in their choices of linguistic resources. This reminds us of Biber et al.'s (2006) caution that each field is different and should, therefore, be considered separately. The subsections below present the MRAC bundles as they were used in each of the subdivisions of the research-oriented bundles, the predominant functional category in the MRAC.

4.2.1.1 *Location in time and place.*

Bundles in this group consisted exclusively of expressions starting with prepositional phrases or noun phrases, with or without *of*-phrase fragments. This lends support to Biber et al.'s (2004) claim that most referential bundles (which include time and place reference) consist of expressions including verb or prepositional phrase fragments. It should be also noted that bundles following the structures *NP + of-phrase fragment* and *PP + of-phrase fragment* in this subcategory served exclusively the purpose of locating elements of the study in time (as shown in examples *1a* and *1b* below); thus supporting Hyland's (2008a) earlier mentioned assumption that some structural and functional categories are closely related.

(1a) Duration of follow-up was classified as short term (4 weeks, range 0-12 weeks), intermediate (26 weeks, 13-26 weeks), and long term (1 year, ≥52 weeks). (10Lancet3)

(1b) However, as salt intake is reduced, people appear to prefer food with less salt,[15] a phenomenon that is probably related to the accommodation of taste receptors *over the course of* weeks to months. (10NEJM5)

However, expressions used as time indicators were not the exclusivity of the two structures mentioned above; other noun phrases and prepositional phrases were also used to serve the same function.

> *(1c)* There was a significant interaction between follow-up time and age, which indicates nonproportionality of the effect of age over *the follow-up period*, such that the risk of death due to age decreased over time. (06JAMA4)

> *(1d)* Mean corrected plasma calcium levels did not differ between the vitamin D and placebo groups at any time *during the study period* (mean, 9.2 [SD, 0.4] mg/dL for both treatment groups and at all time points). (13JAMA2)

It should be noted that while there were equally prepositional phrases and noun phrases used to serve the function of time reference, only prepositional phrases were used to indicate location/ place. Note the recurrence of the preposition *in* at the beginning of the bundles.

> *(2a) In the placebo group*, the mean pressure gradient was 22.5±8.5 mm Hg at baseline and increased to 34.4±14.9 mm Hg at the end of the study. (08NEJM4)

> *(2b) In the current study*, the individuals who died during the intersurvey period had reported higher numbers of lifetime sexual partners at baseline than those who survived to be reinterviewed […] (06Science4)

4.2.1.2 *Procedure.*

Bundles in this category were used to describe study methodologies and experimental procedures. As suggested by Hyland (2008) these bundles play an important role in scientific texts as they help the writer in the process of providing empirical demonstrations and experimental results that will warrant acceptance of the new knowledge being presented. In the MRAC, such functions were realized mainly by bundles following the structural pattern: passive + prepositional phrase fragment.

(3a) FITC-labeled SAH-RSVF$_b$ peptide *was added to the* cells first (1 μM), immediately followed by mixing of stained virus (150 μl) with growth medium (750 μl) and addition to the cell culture. (14JCI1)

(3b) Persons receiving a diagnosis of bladder cancer before cohort entry or within 6 months of joining KPNC *were excluded from the* bladder cancer cohort, and persons with a diagnosis of any cancer before cohort entry *were excluded from the* 10-cancer cohort. (15JAMA3)

Various other structures were also used to serve the same functions, including prepositional phrases and noun phrases (with or without of-phrase fragments), Verb+ *to*-clause fragment, and other expressions featured in the *Other Bundles* structural category.

(4a) The goal of the WHI Coronary-Artery Calcium Study (WHI-CACS) was *to determine whether the* coronary-artery calcium burden differed according to randomized-group assignment among women aged 50 to 59 years after a mean of 7.4 years. (07NEJM3)

(4b) However, the patient and the follow-up team were unaware of the randomized group assignment, and a standardized follow-up protocol was implemented to *reduce the risk of* bias. (15NEJM5)

4.2.1.3 *Quantification.*

Here also, with the exception of bundles in the *Statistical* subgroup, expressions in this subcategory consisted typically of prepositional phrases and noun phrases, with or without embedded *of*-phrase fragments.

(5a) The trend toward *a reduction in the* rate of reinfarction with routine early PCI was not significant. (09NEJM1)

(5b) Indeed, treatment of the lymphoid cell line CEM-c1 with sirolimus conferred dexamethasone sensitivity to this otherwise resistant cell line, reducing the median inhibitory concentration (IC_{50}) *by a factor of* more than 50 (Fig. 6B). (06Science1)

(5c) The total number of dead cells was counted and normalized with reference to *the total number of* cells. (06JCI1)

Another category of bundles identified in the MRAC and which served the same functions as the one shown in the examples above, was the bundles with number slot. As mentioned earlier, these bundles were used essentially to describe elements of the study population or procedures that need to be quantified.

(*6a)* We enrolled women **18** *to* **45** *years of age* who were neither pregnant nor breast-feeding and who reported recent vaginal intercourse, were using effective contraception, and had normal renal, hematologic, and hepatic function. (15NEJM4)

(6b) Antigens in the peptide plates were mixed with 1:1 with PBMCs in the ELISpot plate to a final concentration of 2 or 5 µg/ml and incubated *for* **20** *hours at* **37°***C*, 5% CO_2. (13JCI4)

Bundles in the *Statistical* subgroup consisted mainly of expressions following the structures NP+ other post-modifier fragment *(*e.g. *no significant difference in, a significant increase in),* and/or Pronoun/noun Phrase + *be* (+…) (e.g. *there was no evidence, there was a significant).* A few of these bundles were from the structural category *Other Bundles* as they do not really correspond to any of the 12 structural categories identified by Biber et al. (1999) (e.g. *did not differ between*). As suggested by Cortes (2004), bundles in the *Statistical* subgroup contribute to the descriptions of statistical analyses, results, and observations; as can be seen in the examples below.

(7a) First, group differences in demographic variables and symptoms of eating disorders were assessed with one-way analysis of variance (ANOVA) followed by planned comparison tests for *significant differences between the* subgroups. (13NEJM1)

(7b) Moreover, *there was a significant* increase in p38 MAPK phosphorylation in nuclear extracts derived from *mkp-1*$^{-/-}$ mice compared with *mkp-1*$^{+/+}$ mice fed the HFD (*P* = 0.05; Figure 7A). (09JCI1)

(7c) Clinical events were lower than expected and *did not differ significantly* between groups. (08JAMA5)

4.2.1.4 *Description.*

Bundles in this group were used to provide specific characteristics of the elements of the study. They were limited in number (only 11 out of 122 research-oriented bundles in the MRAC) and typically followed the structure NP + *of*-fragment phrase or PP + *of*-fragment phrase.

(8a) Neonates with and those without colonization at 1 month by S. pneumoniae, M. catarrhalis, H. influenzae*, or a combination of* these organisms did not differ with respect to *the baseline characteristics of* sex, gestational age at birth, maternal smoking during the third trimester, maternal use of antibiotics during the third trimester, exclusive breast- feeding for at least 4 weeks, lung function (forced expiratory volume in 0.5 second), and bronchial responsiveness (provocative dose estimated from the transcutaneous oxygen tension). (07NEJM2)

(8b) We labelled cases and controls *with a history of* febrile seizure if they were discharged alive with their first febrile seizures within 0-2 years before the case died. (08Lancet1)

4.2.1.5 *Topic.*

Finally, members of this last group of research-oriented bundles corresponded to the field-specific bundles discussed earlier under the heading *Other Bundles.* As already mentioned, In the MRAC, they consisted typically of expression that included coordinated binominals and/or nominalization and were used to express field-related concepts.

(9a) PFS was defined as any one of the following criteria for progression—increased pain or analgesia or both, soft tissue disease as per RECIST (version 1.0), or the development of new lesions on the technetium-99m bone scan, confirmed with

further lesions on another bone scan not less than 6 weeks later or *death from any cause*. (07Lancet4)

(9b) In July 2011, the trial's independent *data and safety monitoring* board recommended discontinuation of the placebo group and public report of the results due to demonstration of *the efficacy and safety* of PrEP for HIV prevention in the study population. (14 JAMA5)

Bundles discussed above represent linguistic resources used by professional writers – i.e., authors of medical articles used in the present study – to present content and "real world activities and experiences" (Hyland, 2008). However, as suggested by Hyland and Tse (2004), the presentation of content in academic writing is closely related to the use of other linguistic resources that serve the functions of organizing real world claims, ideas, and experiences into "convincing and coherent texts" (p.167). In the MRAC, similar functions were realized by text-oriented bundles, to which I now turn.

4.2.2. Text-oriented bundles in the MRAC.

Bundles in this functional category were less frequent than those discussed above. Nevertheless, they still represented a little more than 34% of the MRAC bundles; which may be an indication that they are indeed important components of the medical research article. As suggested by Hyland and Tse (2004) they play an important role in the process of making content intelligible by assisting readers in connecting, interpreting, and evaluating the information being presented. As can be seen in Table 6, bundles in the MRAC mostly consisted of prepositional phrase structures and served as text signals in four functional subcategories (transition, resultative, structuring, and framing) with resultative signals being much more prominent than the three other types of text signals.

This predominance of resultative signals in the MRAC text-oriented bundles is consistent the results reported by Hyland (2008a) who found "considerable use of resultative markers" in his corpus of Biology texts (p. 16). He explained the high frequency of resultative signals in hard science texts, in general, by the fact that such bundles have a key role in the rhetorical presentation of research as they "signal the main conclusions to be drawn from the study and highlight the inferences the writer wants readers to draw from the discussion" (p 17). Bundles used for such functions in the MRAC, as well as those used as framing, structuring, and transition signals, are discussed below.

4.2.2.1 *Resultative signals.*

As already mentioned, bundles in this subcategory have important rhetorical functions. In the MRAC, they followed various structural patterns and were used mainly in the presentation of the study findings and/or the implications of such findings. They also contributed to establishing links between elements of the study or with previous studies. The rhetorical functions of bundles used to establish such links include the following:

- expressing cause and effect/result;

(10a) Twelve of those deaths (5%) occurred in the hospital *as a result of* a combination of burn injury and anoxic brain injury (n = 8) or cardiac arrest and anoxic brain injury (n = 4). (06JAMA4)

(10b) Changes in the aortic valve are *associated with an increased* risk of death from cardiovascular causes and myocardial infarction, even in the absence of hemodynamic obstruction and signs of coronary disease. (10NEJM4)

- making comparisons; and

(11a) Finally, *RIP140* mRNA and RIP140 protein levels were decreased in human colon cancers *compared with those in* normal mucosal tissue, and low levels of *RIP140* expression in adenocarcinomas from patients correlated with poor prognosis. (14JCI4)

(11b) In addition, the rate of new skin cancers at 2 years in the calcineurin-inhibitor group *was similar to that* in previous studies. (12NEJM5)

- making inferences.

(12a) These findings suggest that modified *Foxp3* mRNA may have both preventive and therapeutic applications, with implications for additional translational studies aiming beyond allergic asthma. (13JCI5)

(12b) Our data indicate that loss of sensitivity to TGF-β could be an important component of the function of *TEL-AML1*, the most frequent fusion gene in pediatric cancer. (09JCI3)

It should be noted that most of the MRAC bundles used to make inferences typically followed the structural pattern VP + *that*-clause fragment as can be seen in examples (12a) and (12b) above. The next most frequently used text-oriented bundles in the MRAC were those used as framings signal, discussed below.

4.2.2.2 *Framing signals.*

The MRAC bundles in this subcategory were typically prepositional phrases, more often with than without embedded *of*-phrase fragments. Their functions were similar to those identified in Hyland (2008), namely: specifying boundaries of an arguments being presented (13a, 13b); focusing readers on a particular instance (14a, 14b); and emphasizing/ validating an argument

(15a, 15b).

(13a) We further demonstrate that, *in the setting of Klf15* deficiency in mice, inhibition of this pathway can normalize these abnormalities in vivo. (10Sscience2)

(13b) Next, we investigated whether TEL-AML1 expression, despite its negative impact on cell proliferation rate, could, nevertheless, provide some advantage *in the context of* growth inhibitors or apoptotic stimuli. (09JCI1)

(14a) In the case of type 1 diabetes (T1D), current dogma holds that Th1 cells cause pathology, and deviating the Th1 response has been a cornerstone of immunotherapeutic efforts. (15JCI4)

(14b) In the presence of informatively missing observations, as in this study, the worst-score analysis provides an unbiased test against a restricted alternative. (08JAMA5)

(15a) The protocol was designed and completed *in accordance with the* general ethical principles outlined in the Declaration of Helsinki, 2000, and International Conference on Harmonization guidelines for good clinical practice. (11Lancet5)

(15b) In view of the capped sample size, we believe it is noteworthy that we recorded a benefit of prasugrel on ischaemic events. (09Lancet5)

4.2.2.3 *Transition signals.*

There were fewer MRAC bundles in this group than in the subcategories discussed above. The great majority of bundles used as transition signals were fairly idiomatic expressions, and also had relatively straightforward meanings (*in addition to the, in contrast to the).* They mostly consisted of prepositional phrases and were used to either add a new compatible or supporting argument (16a) or contrast two events or arguments. Among bundles used to establish contrastive links between events and arguments, Biber et al. (1999) identified two expressions *(at the same time* and *on the other hand)* with specific uses depending on how the two events being contrasted relate to each other. They suggested that while *at the same time* was used to contrast two events or arguments that are compatible or both considered true, *on the other hand* was used to contrast mutually exclusive events or arguments. Such observations held truth in the MRAC as can be seen in examples (16b) and (16c).

(16a) In addition to the expression of these microglial markers, the transplanted CD44^{hi} cells took on a morphology and localization indistinguishable from endogenous microglia (Figure 5, B, G, J, and K), which are known to associate with vessels *as well as to* localize within intervascular areas. (06JCI5)

(16b) At 1 month, the mean creatinine level increased by 0.06 mg per deciliter (5 µmol perliter) in the telmisartan group and 0.09 mg per deciliter (8 µmol per liter) in the placebo group. *At the same time*, the mean potassium level increased by 0.22 mmol per liter in the telmisartan group and 0.13 mmol per liter in the placebo group. (08NEJM5)

(16c) Consuming moderate amounts of alcohol has been consistently associated with reduced risks of coronary heart disease, stroke, and congestive heart failure. *On the other hand*, acutely ingesting excessive amounts of alcohol (“binge drinking”) has been associated with increased risks of myocardial infarction, stroke, and atrial fibrillation. (08JAMA4)

4.2.2.4 *Structuring signals.*

This group included the lowest number of the text-oriented bundles. It included only six bundles that mainly followed three structural patterns: prepositional phrase fragments, passive + prepositional phrase fragment, and adverbial clause fragment. The low number of bundles in this category may be accounted for by the relatively short length of medical articles in this study, with an average of 4,710 running words per article. Hyland (2008a) found a much higher frequency of structuring signals in his collection of doctoral theses than in his corpus of published research articles; suggesting that the longer a piece of writing, the more need there is to keep orienting readers through the text and to indicate how a new argument or event relates to other events or arguments previously presented.

In the MRAC, structuring signals were mainly used to direct readers to additional materials; which according to Hyland (2008a) reflects a high dependence on *graphical and numerical information* and the necessity to refer to such information when presenting events or arguments, as can be seen in the examples below.

(17a) The baseline characteristics of the patients *are shown in Table* 1; domain scores reflecting the quality of life *are shown in Table* 4 of the Supplementary Appendix (available with the full text of this article at www.nejm.org). (06NEJM3)

(17b) As shown in Figure 5D (left panel), the level of β-catenin–driven transcription activity, measured as the TOP/FOP luciferase reporter ratio, was significantly lower in RIP140-overexpressing HCT116 cells (HCT-RIP) compared with that in control HCT116 cells (HCT-GFP). (14JCI4)

Finally, the last section of this discussion of functional categories in the MRAC regards participant-oriented bundles, presented below.

4.2.3. Participant-oriented bundles in the MRAC.

This category included fewer MRAC bundles than the two other functional categories discussed above. It is possible that such low frequency of participant-oriented bundles in the MRAC be a result of the very nature of the research article in hard sciences, mentioned earlier. The emphasis on the *empirical* over the *interpretational* in hard sciences may justify the lower need to use such bundles whose main function, according to Hyland (2008a) consists in providing a structure for interpreting a following event or argument.

Nevertheless, participant-oriented bundles play a non-negligible role in the research article as, according to Hyland (2005), they contribute to establishing interaction between writers and readers by helping writers connect with readers, make evaluations of their materials, acknowledge alternative views, and express their positions vis-à-vis the arguments and events being presented. He underscored the importance of such functions by contending that academic writing goes beyond mere presentation of real world activities and experiences, and includes "using language to acknowledge, construct and negotiate social relations" (p.174). He further claimed that successful research articles are those that anticipate the interpretation of their target audiences and respond to previously existing knowledge and arguments. Therefore, there is a necessity for writers to presents their findings and interpretations in convincing ways by drawing from preferred languages resources in their discourse communities to "express their positions, represent themselves, and engage their audiences" (p. 176).

According to Hyland (2008a), participant-oriented bundles in the research article convey two main meanings: (1) stance that he described as “ways writers explicitly intrude into the discourse to convey epistemic and affective judgments, evaluations and degrees of commitment to what they say”; and (2) engagement, which refers to “the ways writers intervene to actively address readers as participants in the unfolding discourse” (p. 18). In the MRAC, there was a predominance of engagement bundles over stance bundles; which is consistent with Hyland’s findings that engagement bundles highly prevailed in his science corpora. He suggested that such high presence of these expressions indicated the strong focus that hard science disciplines place on precision, and particularly on ensuring “accurate understanding of procedures and results” (p.19). The two functional features of participant-oriented bundles (engagement and stance), as expressed in the MRAC, are discussed below.

4.2.3.1 *Engagement.*

The MRAC engagement bundles were mainly used to refer to previously existing knowledge; which according to Hyland (2005) is a way of leading readers to recognize accepted disciplinary knowledge and understanding. Bundles in this subcategory typically followed the structural pattern passive + prepositional phrase fragment, as in the examples below.

(18a) Hypertension in children *has been shown to* correlate with family history of hypertension, low birth weight, and excess weight. (07JAMA2)

(18b) Although genetic variants in the region of chromosome 22 *have been associated with* various kidney diseases[14-19,37] in black patients, studies involving patients with both diabetes and kidney disease have been inconsistent.

One bundle in the MRAC, *it is important to,* was used direct readers’ attention to specific events, arguments, as shown in the examples below.

(19a) As clinical genetic testing for breast-cancer risk increasingly includes other genes in addition to *BRCA1* and *BRCA2*, *it is important to* have robust risk estimates for women who carry loss-of-function mutations in genes such as *PALB2*. (14NEJM3)

(19b) Although this inhibition assay is thought to correlate with efficacy against *P. vivax* (*20*), *it is important to* stress that this is an in vitro model around which there is some debate as to how well it will predict field performance of the vaccine. (11Science2)

4.2.3.2 *Stance.*

Stance bundles in the MRAC were used to convey meanings and serve functions similar to those identified by Cortes (2004) in her Biology corpus. Indeed, all the bundles in this category served as hedges and primarily suggested a certain level of probability; which according to Cortes conveys “a degree of tentativeness” (p. 410) to the arguments or results being presented. It should be noted that there was a limited number of such bundles in the MRAC (only five) and they typically followed the structure (Verb/Adjective) + *to*-clause fragment. Below are examples of some of those bundles used in context.

(20a) Evidence from individual-level trials of salt and blood pressure and from cross-population studies indicates that this result *is likely to be* driven partly by high salt consumption in these regions. (10Lancet1)

(20b Differences between observational studies and RCTs, notably our inclusion of frailer patients who were *more likely to have* risk factors for pneumonia hospitalization that overshadowed the risks of these medications, may explain this discrepancy. (14JAMA4)

In sum, the functional analysis of the MRAC lexical bundles revealed that research-oriented bundles constitute the dominant functional category in medical articles. As already mentioned, such high prevalence of research-oriented bundles in the MRAC is consistent with previous studies of lexical bundles in hard science research articles, and is illustrative of scientific academic writing ideology. Such ideology which emphasizes the empirical over the interpretative, very likely, also accounts for the low frequency of participant-oriented bundles in the MRAC. Text-oriented bundles were relatively frequent in the MRAC and were found to play an important role in communicating real world experiences in coherent and convincing ways. Both research-oriented and text-oriented functions were realized primarily by noun phrases and prepositional phrases. The implications of all these finding are discussed in the next section of the present study.

CHAPTER 5. CONCLUSION

5.1. Summary of Findings

The purpose of the present study was to investigate the structures and functions of the most frequently used four-word lexical bundles in MRAs. The findings are consistent with what has been said in previous studies in many regards. The identified bundles included a large number of expressions starting with prepositional phrases and noun phrases accounting for over 50% of all the MRAC bundles. This lends support to earlier claims that academic prose is more phrasal than conversation and therefore, requires more frequent use of post-nominal modifications. (Biber et al., 1999; Cortes, 2004). Other identified bundles deserving some attention included field-related bundles mostly used to express abstract field-specific concepts, and bundles with number slots which perform important functions in the description of elements of the study and of experimental procedures.

The functional analysis of the identified MRAC bundles revealed a high prevalence of research-oriented bundles in medical research articles. A finding that reflects the requirements of academic writing in hard sciences where, according Hyland (2008a), the focus is more on the empirical than the interpretative. As suggested by Hyland, new knowledge in the field of hard sciences "is accepted on the basis of empirical demonstrations and experimental results" (p. 15). With 58.65% of the four-word combinations identified in the MRAC being research-oriented bundles, it appears that lexical bundles play an important role in the validation of new knowledge presented in medical articles. They do so by highly contributing to the descriptions of study objects, methodologies, and procedures; and helping writer present sound and conclusive results.

A high majority of these functions are accomplished by bundles beginning with prepositional phrases or noun phrases; which very likely, justifies the high frequency of structures with prepositional phrases and noun phrases in the MRAC. Such findings may be of interest to ESP practitioners in the field of medicine, and more importantly to novice writers in their quests for socialization in the discourse community they are seeking to become members.

5.2. Pedagogical Implications

Studies (e.g., Cortes, 2004) have shown that explicit instruction on lexical bundles may not lead to acquisition and appropriate use of target expressions by novice writers in the short term. However, it can serve the purpose of raising students' awareness of the presence of such bundles in academic discourse; which in the long term, may lead to the production of more professional-like texts. Given the high variation of bundles across disciplines, lists like the one produced in the present study may be a good starting point by providing examples of target bundles for explicit instruction. In fact, such lists contribute to helping EAP/ESP instructors make informed instructional decisions. As suggested by Römer (2011), while ESP teachers should give priority to teaching words and expressions that will help learners deal with different genres of texts in their subject areas, they are generally not experts in the discourse used in those specific genres; hence the importance of specialized lists to inform instruction.

Instruction on lexical bundles may go beyond just raising students' awareness. Neely and Cortes (2009) suggested a model of lesson plan with scaffolding activities going from raising learners' awareness of the presence and frequency of lexical bundles in their field-specific texts; to explicit introduction of the structure and functions of such bundles; and to familiarizing students with the forms and functions of lexical bundles using concordance lines that show the target bundles used in a variety of contexts. By proposing such model of instruction, Neely and Cortes underscored the importance of corpus-based studies in specific fields, such as the present one. They demonstrated that in combination with already existing public corpora and concordance tools, the findings of such studies can be used as a basis for designing EAP/ESP instruction materials.

An earlier study by Jones and Haywood (2004) highlighted the effectiveness of scaffolding activities such as those proposed by Neely and Cortes. The authors reported increased student awareness of the use multiword expressions, and even some modest gains in learners' accurate and appropriate use of some of those expressions after repeated exposure and instruction through activities going from mere input flooding and noticing activities, to more challenging production tasks such as problem solving essays and gapped writing activities (as cited in Hyland, 2012; p. 166).

It should be noted, however, that the effectiveness of the activities mentioned above is greatly dependent on the relevance of the target items to the course and learners' goals (Neely & Cortes, 2009). As suggested by Hyland (2008a), corpus-informed lists of lexical bundles can serve as important tools for the design of relevant instruction materials, provided that they are drawn from genres relevant to students' reading and writing needs. In this regard, the findings of the present study may provide valuable insights to medical English instructors. In addition to the list of potential candidates for instruction on lexical bundles in the field of medicine, the structural and functional classifications of the MRAC lexical bundles may serve as the basis for production tasks designed to foster the retrieval and use of specific types of bundles to perform specific rhetorical functions.

Some researchers and ESP practitioners have suggested training learners to become more independent learners by investigating how language resources are used in their specific fields (Nesi, 2013). This would probably save valuable class time as it is often the case that with the limited time available for EAP/ESP instruction and the pervasiveness of lexical bundles in academic discourse, there may be only so much that instructors can do in introducing lexical bundles and providing students with the necessary repeated exposure to target expressions. However, as Conrad (1999) pointed out, the amount of information in corpus-based studies would be too much for students to deal with all at once. Carefully planned training is therefore needed to help learners focus on specific areas, one (or a few) at a time.

In their study that investigated, among other points, how successful students are in retrieving information from a corpus, Kennedy and Miceli (2001) anticipated that helping learners develop "corpus research" skills would not only help students make the most of corpora but would also help them develop other areas of their language learning and improve their capabilities with other reference tools. They proposed a gradual and guided training built around four main points:

1. formulate the question;
2. devise a search strategy;
3. observe the examples and select relevant ones; and
4. draw conclusions.

However, as mentioned above, care should be taken not to overwhelm students with the amount of information they can find in a corpus. Mini-research projects following Kennedy an Miceli's four steps could be assigned to students to investigate the functions of one to three preselected bundles. Students can then come up with their own conclusions and share their findings in class.

Finally, one more point to add to the implications of the present study relates to the "accidental" finding that lexical bundles in the MRAC primarily included subtechnical vocabulary items. This somehow calls our attention to Baker's (1988) early claim that in ESP contexts, teachers should give priority to subtechnical lexical items in vocabulary teaching. She argued that the real source of learners' difficulties in comprehending scientific text is subtechnical vocabulary which includes items used to serve specific rhetorical functions. Such claim appears to find support in the findings of the present study, as well as in Marco's (2000) study that revealed that most fillers of the frameworks *the … of* and *a … of* consisted of subtechnical items and that the meaning of a subtechnical item largely impacted the rhetorical functions of the formed multiword expressions.

It is then no coincidence that bundles used to express quantity, for example, comprised words such as *number, amount,* or *dose,* and stance bundles used to express a certain degree of tentativeness (e.g. *it is possible that, more likely to have*) include words such as *possible* and *likely,* which in themselves, convey a certain degree of uncertainty. Such words were labelled *rhetorical items* by Baker (1988) who suggested that because (1) subtechnical vocabulary items may take different meanings from one discipline to another; and (2) such items "may be used in set patterns in certain specialized genres to perform specific rhetorical functions" (p. 103), *rhetorical items* should be identified and taught not in isolation, but in context with their typical collocates. In that respect, corpus-studies such as the present one can be seen as a way of identifying typical collocates of *rhetorical items* in specialized texts.

5.3. Limitations and Further Directions

Like many other studies, the present work has its limitations, one of which is the small size of the corpus. As Hyland (2012) pointed out, small corpora tend to generate more bundles than larger ones because they generally require a lower frequency cutoff. However, despite the small size of the MRAC, the cutoff point in the present study was set at an acceptable frequency of 20 times per million words and a range of 5% of the texts (12 texts). To further avoid idiosyncratic uses and other types of *local repetitions*, care was taken to ensure that each identified bundles appeared in articles from at least two of the five journals used in the collection of the MRAC.

It should also be mentioned that the present study focused only on lexical bundles in texts produced by expert writers, and therefore, it provides only one side of the picture. The other side of the picture would be the types of bundles (if any) used by novice writers in the field of medicine, and how such bundles compare to those used by expert writers, in terms of frequency, structures, and functions. As Hyland (2008b) pointed out, evidence from learner corpora contribute to a better description of writing in a specific discipline; and can also be of invaluable support in the "selecting, sequencing, and structuring of teaching content" (p. 60).

Therefore, a logical next step could be an investigation of the use of lexical bundles in medical students' writings and a comparison of such bundles with those identified in the MRAC. In the meantime, the present study has provided valuable information on the use of lexical bundles in medical research articles, and it is the author's hope that it will contribute to helping medical English teachers make informed instructional decisions in EFL contexts where the medical article constitute one of the rare pieces of writing that learners will actually have to produce in English.

In EFL contexts like Senegal that inspired the present study, corpus-based studies are not common practice – if they are conducted at all. Therefore, beyond the findings presented in this paper, the methodology used in the the development of the MRAC, the identification of the bundles, and the structural and functional categorizations of identified bundles, can hopefully inspire more research that will help address learners field-specific reading and writing needs in other fields and disciplines.

APPENDICES

Appendix A: List of the 204 MRAC 4-word Bundles by Frequency

Bundles	Freq
on the basis of	306
in the placebo group	235
with the use of	228
in the control group	199
in the presence of	139
at the time of	137
the end of the	129
in the absence of	117
in the intervention group	115
at the end of	109
an increased risk of	83
of patients in the	79
years of follow up	74
were randomly assigned to	72
in the number of	71
was associated with a	71
in the context of	69
the total number of	66
the primary end point	65
the use of the	65
has been shown to	64
did not differ significantly	63
there was no significant	63
was approved by the	59
as a result of	58
these data suggest that	58
there were no significant	57
for the treatment of	56
the basis of the	56
the intention to treat	55
end of the study	54
as compared with the	52
as well as the	51
the results of the	51
did not differ between	50
the use of a	50
were included in the	50
presence or absence of	49
randomly assigned to receive	49
the proportion of patients	48
associated with an increased	47
in the present study	47
have been shown to	46
with an increased risk	45
in accordance with the	44
between the two groups	42
was defined as a	41
were more likely to	41
are shown in table	39
in the risk of	38
in the two groups	38
it is possible that	38
no significant differences between	38
of the patients in	38
the effect of the	38
the presence or absence	38
we found that the	38
with the exception of	38
been shown to be	37
lost to follow up	37
no significant difference in	37
was not associated with	37
a significant increase in	36
in the general population	36
in the incidence of	36
the course of the	36
the primary outcome was	36
to determine whether the	36
in this study we	35
than in the placebo	35
a reduction in the	34
and in the placebo	34
a p value of	33
in addition to the	33
loss to follow up	33
median follow up of	33
a median follow up	32
is consistent with the	32
than in the control	32
these results indicate that	32
were approved by the	32
with respect to the	32
during the study period	31
for the primary outcome	31
in the setting of	31
informed consent was obtained	31
studies have shown that	31
these findings suggest that	31
were similar to those	31
data and safety monitoring	30
intention to treat analysis	30
on the other hand	30
over the course of	30
the number of patients	30
these results suggest that	30
as well as in	29
for each of the	29
intention to treat population	29
were excluded from the	29
with a history of	29
assess the effect of	28
at the same time	28
by intention to treat	28
in the development of	28
in the pathogenesis of	28
significant differences between the	28
to the development of	28
was defined as the	28
and the risk of	27
death from any cause	27
efficacy and safety of	27
in contrast to the	27
in the study and	27
in view of the	27
one of the most	27
the time of the	27
to be associated with	27
was similar to that	27
a significant reduction in	26
by the end of	26
food and drug	26
administration	26
in the treatment of	26
primary end point was	26
years of age or	26
the follow up period	26
these results	26
demonstrate that	26
was performed with the	26
we did not observe	26
we were able to	26
baseline characteristics of the	25
had no effect on	25
in the current study	25
is one of the	25
more likely to have	25
of the patients with	25

Bundles	Freq
proportion of patients with	25
was associated with an	25
were similar in the	25
animal care and use	24
are consistent with the	24
in the rate of	24
increase in the number	24
less than was considered	24
of the primary outcome	24
power to detect a	24
study was approved by	24
we were unable to	24
are shown in figure	23
as part of the	23
at a dose of	23
at the level of	23
by a factor of	23
in kilograms divided by	23
more likely to be	23
of the number of	23
our data indicate that	23
the baseline characteristics of	23
there was no difference	23
an increase in the	22
in the regulation of	22
is likely to be	22
of participants in the	22
the basis of a	22
the efficacy and safety	22
the results of our	22
to assess the effect	22
a result of the	21
an intention to treat	21
and the number of	21
as the number of	21
by the presence of	21
duration of follow up	21
for a total of	21
for the presence of	21
has been associated with	21
or a combination of	21
our data suggest that	21
p value of less	21
play a role in	21
reduce the risk of	21
significant increase in the	21
studies are needed to	21
there was a significant	21
there were no differences	21
we have shown that	21
an important role in	20
as compared with placebo	20
at baseline and at	20
compared with those in	20
have been associated with	20
here we show that	20
in the case of	20
in the proportion of	20
it is important to	20
of the patients who	20
the absence of a	20
the design of the	20
the presence of a	20
there was no evidence	20
this is the first	20
was added to the	20
was considered	20
statistically significant	20
was obtained from the	20
would be expected to	20

Appendix B: Article Themes by Journal

	NEJM	***JAMA***	***JCI***	***Science***	***The Lancet***
2015	1. Neonatal health 2. Hypertension 3. Coronary Diseases 4. HIV 5. Diabetes	1. Skin Cancer 2. Hypertension 3. Diabetes 4. Global Health 5. Asthma	1. Hypertension 2. Tuberculosis 3. COPD 4. Type 1 Diabetes 5. Hearing Loss	1. Type 2 Diabetes 2. Alzheimer's Disease 3. Cancer 4. Immunology 5. Hypertension	1. Obesity 2. Cardiovascular diseases 3. Public Health 4. Liver Disease 5. Malaria
2014	1. Type 2 Diabetes 2. Tuberculosis 3. Breast Cancer 4. Asthma 5. Nutrition	1. Smoking 2. Cardiovascular Diseases 3. Diabetes 4. COPD 5. HIV	1. Virology 2. Yellow Fever 3. Hematology 4. Cancer 5. Vascular Biology	1. antibiotic resistance 2. Cancer 3. Autoimmunity 4. Cardiology 5. Liver disease	1. Tuberculosis 2. HIV 3. Pulmonary sarcoidosis 4. Diabetes 5. Global Health
2013	1. Type 2 Diabetes 2. Tuberculosis 3. Breast Cancer 4. Asthma 5. Nutrition	1. Breast Cancer 2. Malaria 3. Neuro-development 4. Cholesterol Estimation 5. Renal Dysfunction	1. Heart failure 2. Type 2 Diabetes 3. Malaria 4. AIDA/HIV 5. Asthma	1. Pulmonary Hypertension 2. Tuberculosis 3. HIV 4. Ebola 5. Type1 Diabetes	1. Coronary disease 2. Child nutrition 3. Genetic variants 4. Prostate Cancer 5. Acute Asthma
2012	1. Obesity 2. Immunology 3. Hypertension 4. Smoking & Pregnancy 5. Skin Cancer	1. Migraine 2. Cardiovascular Diseases 3. Type 2 Diabetes 4. Respiratory Infections 5. Obesity	1. Nutrition 2. Atherosclerosis 3. Hepatology 4. Pancreatic Cancer 5. Genetics	1. Breast Cancer 2. T1 Diabetes 3. Human Immunology 4. Stem Cells 5. Hepatitis C	1. Age-specific mortality 2. Years Lived with Disability 3. Diabetes 4. Hypertension 5. Genetics
2011	1. HIV-Infection 2. Primary Care Practice 3. Postnatal Depression 4. Asthma 5. Blood-Stage Malaria Vaccine	1. Cholesterol &Stroke 2. Urinary Tract Symptoms 3. Lung cancer 4. HIV 5. Heart failure	1. Developmental disorder 2. Metabolism 3. Hypertension 4. Immunology 5. Infectious Diseases	1. Viral Pathogenesis 2. Malaria 3. Cardiovascular Disease 4. Kidney disease 5. Urinary tract infection	1. Hypertension 2. Maternal and Child mortality 3. Cholera 4. Neurological Deficits 5. Multidrug Treatment
2010	1. Tuberculosis 2. Influenza 3. Cancer 4. Variations in Diagnostic Practices 5. Cardiovascular Diseases	1. Caring for Critically Ill patients 2. Pneumonia 3. Eating Disorder 4. Cerebrovascular Accidents 5. Arthritis	1. Cerebrovascular Dysfunction 2. Allergies 3. Dermatology 4. Gastroenterology 5. Arthritis	1. Diabetes-Associated Atherosclerosis 2. Cardiovascular Disease 3. DNA 4. Neurology 5. Parkinson's Disease	1. Global health 2. Stem Cell Transplantation 3. Tendinopathy 4. Prostate Cancer 5. Heart Failure

2009	1. Myocardial Infarction 2. Prostate Cancer 3. Hypertension 4. Pancreatic Diseases 5. Weight Loss Diets	1. Respiratory Infections 2. Malnutrition 3. Liver Diseases 4. Hypertension 5. Chronic Insomnia	1. Obesity 2. Urinary Tract Anomalies 3. Leukemia 4. Neonatal Brain Injury 5. Hypertension	1. Genetics 2. Obesity 3. HIV 4. Lung disease 5. biochemistry	1. H1N1 Virus 2. Diabetes Prevention 3. Neonatal mortality 4. T2 Diabetes 5. STEM Infarction
2008	1. Obsessive Compulsive Disorder 2. Type 2 Diabetes 3. Smoking and Asthma 4. Hyperlipidemia 5. Recurrent Stroke	1. Phobic disorders 2. Type 2 Diabetes 3. Addictive Behaviors 4. Atrial Fibrillation 5. Blood Pressure	1. Cerebral Malaria 2. Infectious diseases 3. Virology (measles) 4. Pulmonology 5. Genetics	1. DNA 2. Kidney development 3. Hypotension 4. Immunology 5. Type1 Diabetes	1. Febrile Seizures 2. Asthma 3. Cystic Fibrosis 4. Breast Cancer 5. Child Nutrition
2007	1. Alzheimer’s Disease 2. Childhood Asthma 3. Hormone Therapy 4. Vaccine-induced Immunity 5. Breast cancer	1. acute sinusitis 2. Hypertension 3. Colon Cancer 4. Underweight, Overweight, and Obesity 5. Migraine Disorders	1. Heart Defects 2. Ophthalmology 3. Immunology 4. Cancer 5. Type 2 Diabetes	1. Dengue & Yellow Fever 2. immunology 3. Hypertension 4. Cancer 5. Biochemistry	1. TIA & Minor Stroke 2. Hypertension 3. HIV 4. Malaria 5. DNA Testing
2006	1. Tuberculosis 2. Leukemia 3. Cystic Fibrosis 4. HIV 5. Diabetes	1. Prostate Diseases 2. Diabetes 3. Sickle Cell Disease 4. Brain injuries 5. Dehydrating Diarrheal Disease	1. AIDS/HIV 2. Neurodegenerative Diseases 3. Microbiology 4. Dermatology 5. Stem Cells	1. Genetics 2. Neurology 3. Type 2 Diabetes 4. HIV/AIDS 5. Malaria	1. Cardiovascular Diseases 2. Global Health 3. Renal Failure 4. Tuberculosis 5. Arthritis

REFERENCES

REFERENCES

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language*.

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

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

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

Chen, Q., & Ge, G. C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*, 502–514.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a foreign language*, *15*(2), 103-116.

Conrad, S. (1999). The importance of corpus-based research for language teachers. *System*, *27*(1), 1–18.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, *17*(4), 391–406.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Ferguson, G. (2001). If you pop over there: A corpus-based study of conditionals in medical discourse. *English for Specific Purposes*, *20*(1), 61–82.

Fryer, D. L. (2012). Analysis of the generic discourse features of the English-language medical research article: A systemic-functional approach. *Functions of Language*, *19*(1), 5–37.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*, 115–135.

Grabowski, L. (2015). Keywords and lexical bundles within English pharmaceutical discourse: A corpus-driven description. *English for Specific Purposes*, *38*, 23–33.

Hyland, K. (2005). Stance and engagement: a model of interaction in academic discourse. *Discourse Studies*, *7*, 173–192.

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

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, *18*(1), 41–62.

Hyland, K. (2012). Bundles in Academic Discourse. *Annual Review of Applied Linguistics*, *32*, 150–169.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: A reappraisal. *Applied Linguistics*, *25*(2), 156–177.

Jirapanakorn, N. (2012). How Doctors Report: A Corpus-based Contrastive Analysis of Reporting Verbs in Research Article Introductions Published in International and Thai Medical Journals, 39–46.

Kennedy, C., & Miceli, T. (2001, September). AN EVALUATION OF INTERMEDIATE STUDENTS' APPROACHES TO CORPUS INVESTIGATION. *Language, Learning & Technology*, *5*(3), 77. Retrieved from http://ezproxy.msu.edu.proxy1.cl.msu.edu/login?url=http://go.galegroup.com.proxy1.cl.msu.edu/ps/i.do?id=GALE%7CA78573585&sid=summon&v=2.1&u=msu_main&it=r&p=AONE&sw=w&asid=1c178f04a7c4566a7cea7f6d001d6ad6

Lam, J. K. M. (2001). A study of semi-technical vocabulary in computer science texts, with special reference to ESP teaching and lexicography. *In G. James (Ed.) Research Reports*.

Li, L.-J., & Ge, G.-C. (2009). Genre analysis: Structural and linguistic evolution of the English-medium medical research article (1985–2004). *English for Specific Purposes*, *28*(2), 93–104.

Loi, C. K. (2010). Research article introductions in Chinese and English: A comparative genre-based study. *Journal of English for Academic Purposes*, *9*(4), 267–279.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, *19*, 63–86.

Mungra, P., & Canziani, T. (2013). Lexicographic studies in medicine: Academic Word List for clinical case histories. *Iberica*, *25*(2013), 39–62.

Neely, E., & Cortes, V. (2009). A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value*, *1*(1), 17–38.

Nesi, H. (2013). ESP and corpus studies. *The Handbook of English for Specific Purposes*, 407–426.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, *11*(3), 283–304.

Nwogu, K. N. (1997). The medical research paper: Structure and functions. *English for specific purposes*, *16*(2), 119-138.

Römer, U. (2011). Corpus Research Applications in Second Language Teaching. *Annual Review of Applied Linguistics*, *31*, 205–225.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Wang, J., Liang, S. L., & Ge, G. C. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*, *27*, 442–458.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: an integrated model. *Language & Communication*, *20*(1), 1–28.

Yang, A., Zheng, S. Yuan, & Ge, G. Chun. (2015). Epistemic modality in English-medium medical research articles: A systemic functional perspective. *English for Specific Purposes*, *38*, 1–10.