

Lexical bundles in published and student disciplinary writing: Examples from history and biology

Viviana Cortes *

Department of English, Iowa State University, 203 Ross Hall, Ames, IA 50014, USA

Abstract

For more than a century, linguists have been interested in the study of frequent word combinations. The present study investigated a special type of word combination, lexical bundles, defined as a sequence of three or more words that co-occur frequently in a particular register [Longman Grammar of Spoken and Written English, Longman, London, 1999]. The goal was to improve the understanding of the function of lexical bundles in academic prose by comparing the use of such bundles by published authors in history and biology and by students at three different levels in those disciplines. The first part of this study focused on a corpus of published writing from history and biology journals. The most frequent four-word lexical bundles in that corpus (called *target bundles*) were identified and classified structurally and functionally. The second part concentrated on students' use of those target bundles. The findings of this study indicated that students rarely used these target bundles in their writing. In addition, when students did use certain bundles, their use did not correspond to the uses of bundles employed by professional authors. Finally, some pedagogical implications of the findings, as well as suggested paths for future research, are introduced and discussed.

© 2004 The American University. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Frequent fixed phrases that perform particular discourse functions are believed to play an important role in fluent linguistic production, particularly in spoken

* Tel.: +1-515-292-5862; fax: +1-515-294-6814.

E-mail address: viviana@iastate.edu (V. Cortes).

language (Pawley & Syder, 1983). The use of collocations and fixed expressions has been considered a marker of proficient language use of a particular register, including academic writing (Bamberg, 1983; McCulley, 1985). Haswell (1991, p. 236) maintained that “there can be little doubt that as writers mature they rely more and more on collocations and that the lesser use of them accounts for some characteristic behavior of apprentice writers”. These qualities stress the necessity to continue investigating frequent word combinations in the production of different discourse communities. The frequent use of lexical bundles, for example, seems to signal 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. Despite the apparent importance of the fixed phrase in English for specific purposes, significant questions remain about the degree to which fixed phrases are unique to particular registers and the extent to which expert disciplinary writing differs from that of novices with respect to the use of discipline-specific frequent word combinations. This study begins to address these issues by investigating lexical bundles in published academic prose in two disciplines, history and biology, and by comparing the findings to the lexical bundles found in the writing of university students who are studying these academic disciplines.

1.1. *The study of word combinations*

The study of frequently used word combinations has attracted the attention of researchers and instructors in the linguistic field since the beginning of the last century. The origins of the word “collocation” to refer to lexical co-occurrence, seems to go back to the nineteenth century, when the Oxford English Dictionary (Simpson & Weiner, 1873) reported that all languages use a wider variety of collocations in poetry than they do in prose. Jespersen (1917, 1924) studied collocations and fixed expressions, distinguishing between formulas and free expressions. In 1933, Palmer (1933, title page) Palmer reported that the term “collocation” was already used as a technical but rather vague term in the linguistic field. He defined collocation as “a succession of two or more words that must be learned as an integral whole and not pieced together from its component parts”. His *Interim report on English collocations* (1933) consisted of a carefully conducted classification of word combinations in English. In the 1950s, Firth (1951) also used the terms “collocation” and “collocability” to describe the habitual occurrence of a word with another word or other words.

The identification of frequent expressions in early studies of lexical co-occurrences was completed from a rather impressionistic point of view. Drawing on studies of English conversation, the seminal research conducted by Pawley and Syder (1983) argued that “fluent and idiomatic control of a language rests to a considerable extent on knowledge of a body of ‘sentence stems’ which are ‘institutionalized’ or ‘lexicalized’” (1983, p. 191). Before identifying a list of presumably frequently used conversational word combinations, these authors reported that “A few minutes’ reflection produced the following sample of clauses that are familiar to the writers as habitually spoken sequences in Australian and New Zealand

English” (1983, p. 206). The way in which the examples in this study were selected was not empirical, but this did not diminish its value in any way.

In the 1970s and 1980s, studies on the acquisition and use of different types of frequent word combinations have drawn on a number of research methods: ethnography (Fillmore, 1979; Hakuta, 1974; Peters, 1983), conversational analysis (Manes & Wolfson, 1981; Tannen, 1987) and quantitative research of the use of multi-word expressions (Altenberg, 1993; Kjellmer, 1991; Sinclair, 1991), among others. Some of these studies focused on combinations of words that are completely obscure, pure idioms, in which the meaning of the whole word combination does not resemble the meaning of the individual parts of the new expression (Chafe, 1968; Moon, 1998). Other investigations have focused on word combinations that are used as fixed expressions in oral language; particularly conversational routines or expressions which are closely related to certain speech events (Coulmas, 1981; Yorio, 1980).

Within the quantitative studies based on the analysis of language corpora, there are two different approaches to the identification of frequent word combinations that have been used in the last three decades. The first approach relies on a group of expressions selected prior to the study (Nattinger & DeCarrico, 1992), which might be considered significant because they can be recognized as familiar by native speakers of the language or because they have been considered frequent in the related literature. The second approach relies on a search tool that identifies lexical co-occurrences of different length, at different cut-off frequency points (Altenberg, 1993; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Butler, 1997).

Finally, there are many studies that have focused on the use of various types of fixed expressions in the written production of native and non-native speakers of English (Bahns, 1993; Granger, 1998; Howarth, 1998). Much of this research ties the importance of the appropriate use of frequent word combinations to the acquisition of pragmatic competence. Granger (1998, p. 145) maintained that one of the reasons why the use of recurrent word combinations has become very popular in the teaching of English as a Foreign Language, for example, is “the formulaic nature of many pragmatic rules”.

In their corpus-based analysis of lexical bundles in academic writing and conversation, Biber and Conrad (1999, p. 188) stated that “because they are so common, it might be assumed that lexical bundles are simple expressions and that they will therefore be acquired easily in the natural course of language learning”. If these recurrent word combinations were so easily acquired, then advanced language learners, such as university students, should get to master the use of these expressions in their language communities. Making an attempt to test this hypothesis and following on several studies which have compared the use of different linguistic features in the written production of published authors and university students (Conrad, 1996; Hewings & Hewings, 2002), the present research employed a quantitative, corpus-based approach to examine fixed phrases called lexical bundles in published and student academic writing in history and biology.

1.2. Lexical bundles

Lexical bundles are extended collocations, sequences of three or more words that statistically co-occur in a register. Some examples of these word combinations in academic prose are *as a result of*, *on the other hand*, *in the case of the*, *the context of the*, and *it is likely to*. Biber et al. (1999, p. 990) defined lexical bundles as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status”, simple sequences of words that commonly co-occur in natural language use. Frequency of occurrence is the defining characteristic for lexical bundles: a bundle has to occur more than 20 times in a million words to qualify, but many bundles occur over 100 times in a million words in certain registers. This frequency is very high when compared to the frequency in which some other word combinations occur, as in the case of pure idioms (e.g., *kick the bucket*), which occur 0.5 times in a million words.

Idiomaticity and fixedness, other qualities which are frequently used to describe word combinations, can also be used to describe lexical bundles (Cortes, 2002). Many lexical bundles are not idiomatic; rather, their meaning is transparent, fully retrievable from the meaning of the individual words that make up the bundle. This is the case of bundles such as *in the presence of*, *as a result of*, *I want you to*, and *what do you mean*, among many others. Lexical bundles are fixed, because the computer program that yields these word combinations at the established cut-off frequency does so for only one form of the bundle. This fixedness, however, is different from that of other word combinations, because it is determined by frequency. For example, in the corpus of academic prose in biology used for this study, only the bundle *these results suggest that* appears as a frequent recurrent word combination. The singular expression *this result suggests that* did not come up as a bundle for the established frequency cut-off point when the computer program was run.

In their exhaustive study of English grammar, Biber et al. (1999) identified the most frequent “lexical bundles” in academic prose and conversation in the Longman Corpus of Written and Spoken English, along with a structural and grammatical categorization and a register comparison, which was further developed by Biber and Conrad (1999). 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. Nevertheless, lexical bundles have strong grammatical correlates, which facilitates their grouping into several basic structural types. Biber et al. (1999) found that most of the bundles in conversation were *clausal*, of the type (*pronoun*) + *Verb* + (*complement*), such as *I want you to*, *take a look at*, and *it's going to be*, among many others. In academic prose, on the other hand, 60% of the bundles are *phrasal*, parts of *noun phrases* or *prepositional phrases*, as in the case of *as a result of*, *on the basis of*, and *on the other hand*.

Lexical bundles have also been classified functionally by Cortes (2001a) and Biber, Conrad, and Cortes (2003). The functions in their taxonomy refer to the meanings and purposes of the language, functions that try to provide texture or organize the discourse according to situations or contexts (Kress, 1976). The

taxonomy was designed using the most frequent bundles identified in conversation and academic prose by Biber et al. (1999). The four core categories in this taxonomy, which are used in the current study, are: referential bundles, text organizers, stance bundles, and interactional bundles. These four categories could be tied to the linguistic functions described by Halliday (1971, in Webster, 2002). Referential bundles perform an ideational function; they help writers structure their experience and determine their way of looking at things. Expressions in this category are time, place, or text markers, such as *at the beginning of*, *the end of the*, or *at the same time*. Text organizers are word combinations used to express textual functions which are concerned with the meaning of the sentence as a message in relation to the surrounding discourse. Some of the functions performed by these expressions are contrast (e.g., *on the other hand*), inference (e.g. *as a result of*), or focus (e.g. *it is important to*), to mention only a few. The following categories, stance and interactional bundles, perform interpersonal functions. Stance bundles express attitudes that frame some other proposition; expressions such as (e.g., *I don't know why*, *are more likely to*). Finally, interactional bundles are conversational word combinations used to express politeness or to report, as in *thank you very much*, and *I said to him*, among others.

2. The study

Even though lexical bundles are frequent combinations of three or more words, the present study investigated the use of four-word lexical bundles because many four-word bundles hold three-word bundles in their structures (as in *as a result of*, which contains *as a result*). Four-word bundles were chosen over five-word bundles because the former are much more frequent (over 10 times more frequent), and they present a wider variety of structures and functions to analyze.

The following research questions helped to focus the investigation:

1. Which are the most frequent four-word combinations (lexical bundles) in published history and biology writing? How are these lexical bundles (called target bundles) classified structurally and functionally?
2. Are there any significant differences in the functions of these expressions between disciplines?
3. Are these target bundles used by students at different university levels in history and biology? If so, how frequently are these bundles used? Are the functions students try to convey with those bundles similar to those conveyed by published authors?

2.1. Corpus collection

The collection of the texts that made up the different corpora used in this study was divided into two different steps: the collection of the corpus of published writing and the collection of the corpus of student writing (see Tables 1 and 2 for a description of the corpora). The first step consisted of the collection of the corpus of published writing, which was used to identify the most frequent lexical bundles used

Table 1

Corpus of published academic writing in history and biology

Disciplines	History	Biology
Journals	Journal of American History, American Historical Review	Ecology, Evolution, Science
Years	1999–2001	1997–2000
Type of texts	Research articles	Research articles
Number of texts/writers	92/92	199/199
Mean length of texts	10,502 words	5157 words
Total number of words	966,187 words	1,026,344 words

Table 2

Corpus of student writing in history and biology

<i>Undergraduate lower division</i>			
History words		Biology words	
113 response papers	54,600	59 lab reports	72,517
29 synthesis papers	50,571	10 research papers (CC)	9436
20 research papers (CC)	31,548	21 synthesis papers (CC)	11,529
19 synthesis papers (CC)	33,500		
181 total papers (99 writers)	170,219	90 total papers (86 writers)	93,482
<i>Undergraduate upper division</i>			
History words		Biology words	
23 synthesis papers	50,789	51 research reviews	131,064
29 research papers	166,271	9 research papers (CC)	15,266
		12 synthesis papers (CC)	24,803
52 total papers (51 writers)	217,060	72 total papers (72 writers)	171,133
<i>Graduate level</i>			
History words		Biology words	
15 research papers	65,513	33 research papers	117,888
16 research papers(CC)	27,100	18 research papers (CC)	28,764
13 synthesis papers (CC)	13,217		
44 total papers (44 writers)	105,830	51 total papers (51 writers)	146,652
Total	493,109		411,267

by published authors (called *target bundles*). History and biology were selected as the target academic disciplines. These disciplines were chosen because they represent different research and methodological traditions, showing some of the diversity present among university disciplines, and because each of them considers writing to be an important skill in the development of academic competence. The review of previous studies revealed that writing is considered an important part of competence in history and biology (Conrad, 1996; Langer, 1992; Stockton, 1994, 1995; Walvoord & McCarthy, 1990; Wilkinson, 1985). Both history and biology aspire to help students become proficient writers, often providing them with writing-intensive courses

in the disciplines, which attempt to help students become members of their discourse communities.

In order to select the texts which would make up the corpus of published writing, seven professors in the Department of History and seven professors in the Department of Biology at Northern Arizona University were interviewed. The aim of these interviews was to gather information about journals which could represent the writing in those disciplines, and which professors considered model writing for their students to learn from. The survey resulted in the selection of the following journals and publications: *Journals of American History* (JAH) and *American Historical Review* (AHR) for history, and *Ecology* (EC), *Science* (SC), and *Evolution* (EV) for biology. The electronic versions of these journals were accessed, and the texts were downloaded and saved in electronic files whose names indicated the journal name and the year of publication. Both published and student corpora were completely cleaned, that is no titles, headers, footers, captions, scientific formulae, or references were included in the final word count or in the analysis conducted in the present study.

The second step consisted of the collection of a corpus of students' writing in these disciplines at three university levels: undergraduate lower division, undergraduate upper division, and graduate level. Some of the students' papers making up the corpus which Conrad (1996) used for her investigation were also used as part of the corpus for this research project, especially for those sections of students' writing which needed to be enlarged to make the different levels similar in size. Whenever Conrad's corpus is described, the letters CC (Conrad Corpus) accompany the corpus description; otherwise, the information or description comes from the corpus compiled for this study. All papers in the corpus were produced by students at Northern Arizona University.

As shown in Table 2, the corpus of student writing had about one million words in total and was made up of different types of papers. Research papers or research oriented assignments were preferred but that type of writing was not always available. In this case, other types of written assignments (such as lab reports or response papers) were collected to complete the sample.

The undergraduate lower division section of the corpus was made up of papers written by students taking freshman or sophomore classes, while the undergraduate upper division section papers were written by students taking junior and senior classes.

2.2. Computer programs and bundle identification

Two computer programs were used in the identification of lexical bundles in the different corpora of academic writing used in this study. The Lexical Bundles program (LBP) was designed for this research using Borland Delphi (1998) and was used to identify those bundles that were frequently used by published authors and which were considered target bundles in each discipline. All four-word combinations were stored in a database which also recorded the frequency with which each word combination occurred and the number of texts in which it was used. Once all the

selected texts had been processed, the program transferred all those bundles that occurred at least 20 times in a million words and in five or more texts to an output text file, which also contained the total number of words and the total number of files processed. Biber et al. (1999) considered lexical bundles all those word combinations that recurred over 10 times in a million words and repeated in five or more texts in the Longman Corpus. In the present study the most conservative cut-off point of 20 times in a million words was selected.

The second program, MonoConcPro, (Barlow, 2002) was used for two different tasks. The first task was to analyze the use of the target bundles in their corresponding contexts in order to identify the major functions these bundles performed in published academic writing. The second task consisted of the identification of the target bundles in the corpora of student writing at the three academic levels present in this study. The frequency and number of texts in which the target bundle was found was recorded and the examples in contexts produced by this concordancing program were used in the analysis of the functions students gave to these target bundles.

3. Results and discussion

As the subsections of the student corpus (divided into university levels) were much smaller than the published corpus, the identification of bundles had to follow a more qualitative approach. The most frequent four-word combinations which matched the cut-off point in this study were considered lexical bundles and became the *target bundles* which would be used for the comparison between published and student writing. A more quantitative approach, which would identify the most frequent four-word combinations in the corpus of student writing for further comparison with those bundles identified in the published corpus, would not have been reliable due to the corpus size difference previously explained (Cortes, 2001b).

3.1. Target bundles in history: structural and functional classification

Appendix (Table 7) shows a comprehensive list of the 54 bundles identified as target bundles in history writing. Target bundles in each discipline were structurally classified using the same classification used by Biber et al. (1999). Table 3 shows target bundles in history classified according to those categories. As shown in Table 3, the structural characteristics of these bundles follow those of the bundles identified in previous studies of academic prose (Biber & Conrad, 1999; Biber et al., 1999), reinforcing the concept that academic writing is structurally more phrasal in nature than conversation, and is strongly marked for post-nominal modification, as in the case of genitive expressions or other prepositional phrases which are post-nominal modifiers.

Using the taxonomy designed by Cortes (2001b), which was later revised in Biber et al. (2003), the target bundles in history were classified according to the functions

Table 3

Target bundles in history classified according to their structural correlates

1. <i>Noun Phrase with “of” phrase fragment</i>
a member of the, both sides of the, one of the most, the beginning of the, the course of the, the creation of the, the end of the, the eve of the, the history of the, the image of the, the majority of the, the nature of the, the power of the, the rest of the, the turn of the, turn of the century
2. <i>Noun phrase with post-nominal clause fragment</i>
the extent to which, the fact that the, the ways in which
3. <i>Prepositional phrase with embedded “of” phrase</i>
as a result of, as part of a, as part of the, at the beginning of the, at the end of the, at the time of, at the turn of, at the university of, by the end of, from the perspective of, in the case of, in the context of, in the course of, in the face of, in the form of, in the hands of, in the history of, in the midst of, in the name of, in the wake of, in the words of, on the basis of, on the eve of, on the part of, over the course of
4. <i>Other prepositional fragments</i>
at a time when, at the same time, for the first time, in the first place, on the one hand, on the other hand, to the extent that
5. <i>Verb (be) + complement (noun phrase)</i>
was one of the
6. <i>Other expressions</i>
as well as a, as well as the

they performed in context. Some new categories had to be created in order to classify the rest of the target bundles in this corpus because these bundles were not present in the corpora used in previous studies of lexical bundles. These categories were referential quantifying bundles, referential subject-bound bundles, and other bundles in biology.

As shown in Table 4, target bundles in history performed a referential function or were used as text organizers. Referential bundles were mainly used as time or place markers, or to introduce quantities or amounts, as in the following examples:

In 1802, *at a time when* two public buildings, the old “State House” as well as “Federal Hall”, were vacant, the ordinarily parsimonious Common Council voted to construct a civic building all their own (AHR).

Having grown used to living, working, and playing in physically separated communities on *both sides of the border*, members of these transnational social networks often feel as “at home” in one place as another (JAH).

Many in this group had initially expressed uncertainty about resettlement, fearing white hostility. *The majority of the* second group had also grown up among other Japanese Americans, and few enjoyed close interracial friendships (JAH).

Text organizers in history covered different functions, indicating comparison/contrast or inference, or, in the case of framing bundles, identifying textual

Table 4

Target bundles in history classified according to their functions in context

Categories	Sub-categories	Bundles
Referential bundles	Place markers	at the university of, both sides of the
	Quantifying bundles	a member of the, as part of a, as part of the, one of the most, the majority of the, the rest of the, was one of the
	Subject-bound bundles	in the history of, the history of the
Text organizers	Contrast/Comparison	on the one hand, on the other hand, in the first place
	Inferential	as a result of, on the basis of, in the first place
	Framing	from the perspective of, in the case of, in the context of, in the form of, on the part of, the extent to which, the fact that the, the nature of the, the ways in which, to the extent that, as well as the, as well as a
Other bundles in history		in the words of, in the name of, in the hands of, the creation of the, the power of the, the face of the, the image of the

conditions used to interpret prior or forthcoming discourse. Here are some examples of text organizers in history:

Any examination of American nationalism must, sooner or later, contend with its contradictory character. *On the one hand*, it offers a civic creed promising all Americans the same individual rights irrespective of color, religion, or sex (JAH).

On the basis of these letters and the published debates, it would seem clear that the paucity of a rhetoric of conspiracy among the patriot elites observed for the prerevolutionary period persisted through the first weeks of the Estates General and the National Assembly (AHR).

It was previously mentioned that the corpus of published academic writing in history contained several bundles that had not been identified as frequent word combinations in the related literature (Biber et al., 1999). Some of these bundles reflect an inherent connection to the social disciplines, especially those related to social events or issues (*the power of the, the creation of the, in the context of, the image of the*). Here are some examples of this type of bundles in their contexts:

Before the strike, CR was, *in the words of* Business Week, “the fair-haired boy of liberalism”, the vanguard of a rapidly expanding consumer movement that was seen as central to the movement culture of the depression decade (JAH).

In the context of the Communist International’s “Third Period”, of intensified class struggle and anticipation of imminent socialist revolution, the Communists cast the NAACP in the dual role of class-and-race traitor adopting the vocab-

ulary hitherto reserved for “reformists” and “social fascists” within the labor movement (AHR).

3.2. Target bundles in biology: Structural and functional classification

The 109 target bundles identified in the corpus of academic writing in biology belonged to various structural groups (see Appendix (Table 8) for a complete list of these bundles). As shown in Table 5, the structures of the majority of such bundles were also associated with academic prose like the target bundles in history, as in the case of noun phrases or prepositional phrases followed by embedded fragments of prepositional phrases or clauses (e.g., *the mean of the, in the presence of*). There were

Table 5

Target bundles in biology classified according to their structural correlates

1. Noun Phrase with “of” phrase fragment

a function of the, a large number of, a measure of the, a wide range of, a wide variety of, an order of magnitude, each of the three, presence or absence of, one of the most, the beginning of the, the center of the, the course of the, the depth of the, the direct effects of, the effect of the, the effects of the, the end of the, the genetic basis of the, the length of the, the magnitude of the, the mean number of the, the mean number of, the mean of the, the nature of the, the number of the, the number of individuals, the number of species, the population dynamics of, the position of the, the relative importance of, the rest of the, the results of the, the shape of the, the size of the, the slope of the, the total number of the

2. Noun phrase with post nominal clause fragment

the degree to which, the extent to which, the possibility that a

3. Other noun phrases

an important role in, an increase in the, spatial and temporal variation, the relationship between the

4. Prepositional phrase with embedded “of” phrase

as a consequence of, as a function of, at the beginning of, at the end of, at the university of, for each of the, in each of the, in the absence of, in the case of, in the context of, in the evolution of, in the form of, in the number of, in the presence of, of the distribution of, of the effects of, of the number of, on the basis of, on the evolution of, on the number of, on the order of, with the exception of, with the number of

5. Other prepositional phrases

at the same time, in addition to the, in contrast to the, in the present study, in the study area, of the relationship between, of the variance of, on the other hand, with respect to the, in the life cycle

6. *it + Vbe + adjective + (clause fragment)*

it is difficult to, it is important to, it is likely to, it is possible to

7. *(modal) Vbe + (complement noun phrase/adj. phrase)*

be the result of, is likely to be, is the number of, is a function of, is consistent with the, may be due to, are likely to be, are more likely to, there is no evidence

8. *(noun phrase/pronoun) + V + (complement)*

not appear to be, can be used to, did not differ significantly, has been shown to, have been shown to, used in this study, was found to be, was positively correlated with, studies have shown that, play an important role, these results suggest that

other types of bundles, though, that showed clausal structures, as in the case of those bundles that have the structure *it* + *Vbe* + *Adjective* + (*to clause-fragment*), such as *it is important to*, or (*noun phrase*) + (*aux*) *V* + (*complement*), as in *these results suggest that*. The structures in these bundles also reflected academic prose style: most of the phrasal bundles had phrasal or clausal post modification, and the clausal bundles had verbs in the passive or “*it*” + *Vbe* + *adj.* structures, which reflected a depersonalized mode frequently found in academic prose.

Four categories (and their corresponding sub-categories) from the same taxonomy used for the classification of target bundles in history (Biber et al., 2003; Cortes, 2002) were used for the target biology bundles: referential bundles (time markers and place markers), text organizers (contras/comparison bundles, inferential bundles, focus bundles, and framing bundles), and stance bundles (epistemic–impersonal/probable–possible). The following categories had to be created in order to functionally classify the rest of the bundles identified in this corpus: referential quantifying bundles, referential quantifying statistical bundles, referential descriptive bundles, other stance bundles, and other bundles in biology. These categories are described and illustrated in the following sections. Table 6 presents the target bundles in biology classified into these functional categories.

Referential bundles in biology writing were also used as time or place markers, as descriptive bundles, or to introduce quantities or amounts. In this last sub-category, there was a sub-group labeled *statistical*, which contained expressions used to report statistical procedures which were often found in the Results section of biology papers. These are some examples of referential bundles in biology in their contexts:

At the time of the planting, the soil was naturally moist and did not require watering (EV).

One could estimate the breeding propensity, but not the proportion of breeders *in the study area* a given time within the season (EC).

In this study, *the size of the* disturbed area did not affect the rate of recovery of the turf in either the abrasion or the removal treatments (EC).

Immigrants that were unrelated to other group (22) members *did not differ significantly* in time spent on raised guard from individuals that were related to other group members (Wilcoxon signed ranks test: $T = 13$, $n = 7$ groups, $P = 0.5$) (SC).

Text organizers sub-categories covered one more function (focus bundles) than the target bundles identified in the history corpus. Text organizers in biology were also used to imply comparison/contrast and inference, to qualify a certain element of the succeeding discourse in terms of its importance or degree of difficulty (focus bundles), or to connect preceding and forthcoming discourse under a given condition (framing bundles). The following are examples of text organizers in published biology writing:

Table 6
Target bundles in biology classified according to their functions in context

Category	Sub-category	Bundles
Referential bundles	Time markers	a the beginning of, at the end of, at the same time, at the time of, the beginning of the, the end of the, in the course of, in the life cycle
	Place markers	at the university of, the center of the, the position of the, in the study area, in the present study
	Descriptive bundles	the depth of the, the length of the, the shape of the, the size of the, the slope of the
	Quantifying bundles	a large number of, a measure of the, a wide range of, an increase in the, an order of magnitude, in the number of, of the number of, of the variance of, of the variance in, on the number of, on the order of, one of the most, that the number of, the average number of, the magnitude of the, the mean number of, the mean of the, the number of individuals, the number of species, the rest of the, the total number of, as the number of, with the number of, is the number of, for each of the, each of the three, in each of the <i>Statistical:</i> was positively correlated with, of the distribution of, not significantly different from, did not differ significantly
Text organizers	Contrast/Comparison Inferential	on the other hand, in contrast to the as a result of, as a consequence of, on the basis of, the results of the, be the result of, the effect of the, the effects of the, the direct effects of
	Focus	it is important to, it is difficult to, there is no evidence, these results suggest that, an important role in, play an important role, the relative importance of, the results of the, these results suggest that
	Framing	in the absence of, in addition to the, in the case of, in the context of, in the form of, in the presence of, as a function of, presence or absence of, the nature of the, with respect to the, the degree to which, the extent to which, a function of the, as well as the, with the exception of
Stance bundles	epistemic–impersonal/ probable–possible	may be due to, it is possible that, it is likely that, is likely to be, are likely to be, are more likely to be, the probability that a
	Other stance bundles	has been shown to, have been shown to, was found to be, not appear to be, can be used to, studies have shown that, used in this study
Other bundles in biology		for the evolution of, in the evolution of, is a function of, is consistent with the, more closely related to, not appear to be, of the effects of, of the relationship between, on the evolution of, spatial and temporal variation, the genetic basis of, the population dynamics of, the relationship between the

However, *it is important to* note that sometimes the optimization may fail due to lack of convergence in parameter estimation (EC).

In contrast to the majority of *Acacia* species, *A. thomasi* and *A. etbaica* flowered only during the dry season, at a time when very few individuals of other species were in flower (EC).

Females raised on each diet varied in their ages at first reproduction and in total reproductive output; *as a result of* these differences, individual fitness values differed within each diet (EC).

The progeny derived from the last two generations of inbreeding will express both inbreeding components, as *in the case of* high *F*-values (EV).

Stance bundles are a category of expressions which did not occur in the history corpus. Most of these bundles convey a certain degree of probability/lack of probability and they are mainly used as hedges, introducing a degree of tentativeness to what is being reported. Examples include:

If sex expression and pollen filling are genetically determined, then *it is possible that* inbreeding has an influence over such male fitness components (EV).

The SIC method *was found to be* the most appropriate information criterion for identifying the best model among six alternative models considered (EC).

Finally, the category “Other bundles in biology” contained mainly bundles which were subject-specific, including words closely related to the biological field. Here are some examples of this type of bundles:

Thus, female plumage has changed independently of male plumage throughout the history of this group, and alternative explanations *for the evolution of* female plumage need to be considered (EV).

Genome sciences, combined with bioinformatic tools, allow us to dissect *the genetic basis of* multifactorial diseases and to determine the most suitable points of attack for future medicines, thereby increasing the number of treatment options (SC).

3.3. Comparison between target bundles in published history and biology writing

A structural comparison between target bundles in history and in biology showed that lexical bundles in history only belonged to two structural groups: noun phrases and prepositional phrases. Lexical bundles in biology, on the other hand, covered a wide variety of structural groups: noun phrases, prepositional phrases, *it + Vbe + adjective* clause fragments, *Vbe + complement* clause fragments, and *noun phrase + V + complement* clause fragments. Even though clausal bundles had been related more to conversation than to academic prose in the literature, the

clausal bundles identified in the biology corpus shared structures which had also been previously identified as frequently occurring in academic prose (Biber et al., 1999). One of the sub-groups consisted of expressions holding the structure *it + Vbe + adjective + clause fragment* (e.g., *it is important to*), which are extraposed structures with a predicative adjective controlling a complement clause. The other group consists of bundles that contain verbs in the passive voice, which are fragments of extraposed structures with verb predicates, passive constructions that are followed by *that*-clauses (e.g., *has been shown that*).

Functionally, different lexical bundles identified in both corpora were used to perform similar functions. In the case of referential bundles, it seems logical to find that all bundles classified as time markers in the history corpus referred to years or time periods in which historical events took place, while in the biology corpus, those bundles were used to refer to different periods of time, mainly stages in the evolutionary or developmental processes of different biological phenomena.

Another difference in the use of bundles in these disciplines is the variety of epistemic–impersonal/probable–possible stance markers used in biology. These bundles are often clausal and included expressions like *are likely to be*, *is likely to be*, *it is possible that*, and *the probability that the*. In the corpus of published biology, these expressions were used in many cases to hedge the effect of an affirmation, or to make an affirmation or argument more tentative. It is important to point out that these bundles are not frequently used by published history writers, nor are other lexical bundles with these functions. This finding may suggest that history authors make use of single vocabulary items or other expressions (perhaps shorter than lexical bundles) in order to express the same function.

The bundles identified and classified in the last sections were the target bundles that were used as a basis for comparison between the production of bundles by published authors and by university students in history and biology.

3.4. Target bundles in student writing in history

This section analyzes the use of target bundles by students in history (see Appendix (Table 7) for a complete alphabetical list of all the target bundles in history and their frequencies in the corpus of student writing in history in the three levels analyzed in this study). It is necessary to note that when dealing with bundles frequency in student writing, the number of papers in which a bundle appeared was used as a measure of overall use instead of the total frequency of bundles because recurrent use of a bundle by a single student in a paper could inflate a frequency count. Sometimes a bundle seemed frequent, because the number of occurrences was high, but a closer look showed that the bundle was used by only a few students, who repeated the use of the expression several times in the same paper. This is why the numbers in Appendix (Table 7) show the total frequency of the bundle and the number of texts in which the bundle was used.

Two considerations are important when discussing the use of lexical bundles by students at different university levels in history: frequency and bundle function.

Regarding frequency, many lexical bundles frequently used by published authors were rarely or never used by students.

Three lexical bundles identified in the target corpus were never used by students in history: *from the perspective of*, *on the eve of*, and *the eve of the*. In the case of the bundles including the word *eve*, a logical explanation for the absence of these bundles in students' writing could be the fact that this expression is related to a specific period of time that perhaps is not present in any of the topics of the student papers in this corpus. The bundle *from the perspective of*, on the other hand, seems to be closely connected to this discipline, and students may not be accustomed to using it in their papers, needing perhaps more exposure to texts that present the correct use of this bundle. Other bundles frequently used in published writing (*in the context of*, *in the course of*, *the extent to which*, *the ways in which*, *to the extent that*, to mention just a few) were used in less than five papers in the whole corpus of student writing in history.

Many other bundles were also rarely used. In fact, 15 (27%) out of the 54 bundles in the target corpus were used in two or fewer papers at all levels, and 14 (25%) were used in three to five papers at all levels. Lexical bundles like *as well as a* or *as part of a* were rarely used, but parallel constructions with *the* (*as well as the* and *as part of the*) were slightly more frequent.

Some referential bundles like *at the turn of*, *in the course of*, *in the wake of* were rarely found in student corpora. Students seemed to rely on the use of other referential bundles which may be more familiar to them such as *at the beginning of*, *the beginning of the*, *at the end of*, and *the end of the*, which were more frequent at all levels of student writing. This also occurred with some text organizers like *the extent to which* and *to the extent that*, which were used to express the notion of degree. Text organizers were also rarely used by students across all levels, as in the case of *on the one hand*, which was used only twice in the lower division and once in the upper division.

The repetitive use of bundles by an individual author in a single paper occurred at all levels, but seemed to be more frequent in upper division student writing. In short papers, the repetition of a bundle makes the discourse sound redundant, especially in the case of text organizers. Perhaps repetition is just a means that students employ to try out the use of a bundle as a form of overgeneralization (using the bundle when it is not necessary) in an attempt to reach a more developed balance in the use of bundles present in the writing of published authors in this discipline. This overgeneralization could follow the same steps of other cases of overgeneralization in the acquisition of a language, and may be a stage in moving towards a more appropriate use of the bundle in the future.

Among the most frequent bundles used by students at the three levels were expressions like: *as a result of*, *as well as the*, *at the same time*, *in the case of*, *on the other hand*, *one of the most*, *the beginning of the*, *the end of the*, *the history of the*, and *the rest of the*. These bundles served a variety of functions which lead to the second consideration of the comparison across levels.

Regarding function, students at different academic levels sometimes used lexical bundles to convey functions different from those identified in published history

writing. This is the case of bundles like *at the same time*, used for addition instead of simultaneity. The following examples showed that students tended to use this bundle metaphorically.

However, once I divulged myself into her past history, I began to wonder if I should even write this paper at all. I found information which was heartbreaking and shocking *at the same time*. (History paper, Undergraduate lower division).

He was tactful in his letter, being firm, but *at the same time polite*. (History paper, Undergraduate lower division).

These uses of this bundle deviate from the straightforward, concrete style that was more typical of academic prose. Perhaps students have been using this bundle in narrations or more creative types of writing and that is why the function they conveyed through their use was more metaphorical or creative than what was typically found in academic writing.

Intuitively, it might be easy to believe that exposure to reading materials which includes these functional uses of lexical bundles could help students in the acquisition of bundles and promote their use in writing tasks in history. The development in the use of lexical bundles toward target-like functions did not follow an expected progression in student writing in history: some bundles used to perform certain functions in lower division writing conveyed different functions when used by students in upper division or graduate level courses (e.g., *at the same time* and *both sides of the*). When looking at graduate level writing, for example, frequency in the use of the bundles was not higher than in the lower academic levels, yet students still found difficulty using those bundles in target-like ways. Exposure to reading materials alone, then, did not seem to be enough for students to use a variety of lexical bundles or to use bundles conveying functions like those found in published academic writing.

3.5. Target bundles in student writing in biology

Frequency and function were also considered when analyzing the use of target bundles at different university levels in biology. In regards to frequency, biology students followed the same pattern as history students: many target bundles, identified as frequently used in published academic prose in biology were never or rarely used by students at all levels. Appendix (Table 8) shows the target bundles in biology identified in the written production of students in biology at the three levels analyzed in this study.

The least frequent target bundles in student writing in biology were referential bundles (*the beginning of the*, *the shape of the*, *with the number of*), text organizers (*the extent to which*, *the degree to which*, *with respect to the*), and expressions that belong to the category *other bundles in biology*. A high number of these bundles are specific and perhaps the absence or the low number of occurrences of these bundles in students' written production may be due to the fact that they belong to

published texts that present specific biological issues, as in the case of bundles such as *spatial and temporal variation*, *on the evolution of*, or *an order of magnitude*. The absence of other least frequent bundles, especially those that include the word *number*, or those that are classified as *statistical bundles*, reflects that students may be reporting the results of their studies, using expressions that differ from those used in the results sections of published research. In addition, many bundles in this category which were used for less specific or discipline-bound functions, as in the case of *on the basis of*, *the extent to which*, *the degree to which*, *with respect to the*, were also rare in students' writing.

Students did repeat bundles a number of times in a single paper, which made the writing wordy or redundant, especially in short papers. Bundle repetition occurred more often in higher levels. Similar to history, perhaps this repetition is used by students as a form of overgeneralization in order to try out the use of these bundles in different parts of their texts. The more frequent use of lexical bundles in upper division and graduate biology writing reflects certain evolution across levels in biology writing. Students at higher academic levels, especially in graduate writing, used text organizers and some stance markers much more often than lower level students in this discipline. This finding may reflect the use of more sophisticated organizational patterns in upper levels, accomplished by means of the use of lexical bundles. It would be interesting to isolate and investigate the use of the text organizers that become more frequent in higher level students' writing, to identify patterns of bundles acquisition, which could result from exposure to the use of these bundles, or from the need for more elaborated writing in graduate writing assignments. The frequent use of these bundles by graduate students may also reflect that the type of writing tasks that students complete in graduate courses are more similar to professional writing. In addition, the use of bundles in higher academic levels moved, in general, in the direction of the functions that bundles perform in published writing in biology. Perhaps the more advanced students are reading more literature in the field and processing it more thoroughly because they need to use it in their own writing. In general, however, the use of target bundles at all levels still showed low frequencies in bundles that would be expected to occur more often in academic writing.

3.6. Comparison between published and student writing in history

The use of target bundles by students in history courses at different academic levels was very infrequent. The majority of the lexical bundles used by published authors in history were rarely or never used by students at all levels in this discipline. More than half of the bundles identified as target bundles (29 out of the 54 total) were used in five or fewer papers at all levels in history students' writing. This reveals the fact that students do not often use expressions such as bundles which function as text organizers, for example. It would be interesting to focus on the language forms students use instead of these expressions to convey these functions.

The target bundles more frequently used at all levels by students in history are time markers, although familiar expressions which are not exclusive to academic

writing show higher frequencies in this corpus. Expressions such as *the beginning of the*, *the end of the*, *at the end of* are more frequently used than expressions like *at the turn of*, or *in the course of*, which are lexical bundles frequently found in published history writing. The low frequency in use of these last expressions may result from lack of exposure to readings which show these expressions and the functions they convey in the writing of this discipline.

A special consideration needs to be made about the repetition of lexical bundles by a student on a paper. The repetitive use of the same bundle in a paper makes the discourse sound redundant. Particularly in history, when a lexical bundle seemed to be used frequently, it had been used by the same student over and over again in the same paper.

There seems to be no pattern of development in the use of lexical bundles at different university levels in history. Bundles that were used by undergraduate lower division students were not used by higher division or graduate students. It must be considered that the papers which made up the corpus of students' writing were collected from different students, which only allowed the comparison of the use of target bundles across levels. The only reliable way to identify patterns of development in the use of lexical bundles by students at different levels would be to conduct a longitudinal study of the same students investigating the evolution in the production of target bundles in their writing.

The task type may have also influenced the findings of the present research, because the papers used in the corpus collection for the lower levels of the history corpus are not research papers. The papers from graduate level are research papers and should, hypothetically, present more target bundles. The graduate portion of the corpus, however, does not present a considerable increase in the use of target bundles.

3.7. Comparison between published and student writing in biology

Many target bundles identified as frequently used in published academic writing in biology were never or rarely used by students in this discipline at all levels. The absence of these bundles in student writing might be due to the fact that some of these bundles are very specific and are closely related to special biological issues which might not have been present in student writing. This is the case of some bundles in the category *other bundles in biology* (such as *on the evolution of*, and *an order of magnitude*). For students to use these bundles, their writing should deal with evolutionary topics or reporting magnitudes, uses which were not present in the student writing corpus.

The absence of quantifying bundles, particularly those that hold the word *number* and statistical bundles, may reflect the fact that students do not use target bundles which are frequent in published academic writing in biology when reporting the results of their studies. In published writing, these bundles are often used in the results section of research papers. Many of the papers used for the students' corpus were not research papers. Those which were research papers did have result sections, but students did not use expressions like the target quantifying bundles in published papers.

Table 7

Target bundles in published and student writing in history

Target bundles	Published history		Student writing		
	Frequency	# of texts	Lower div. Freq.–# of texts	Upper div. Freq.–# of texts	Graduate Freq.–# of texts
a member of the	26	19	4–4	4–2	4–4
as a result of	34	25	9–7	13–8	12–7
as part of a	23	18	1–1	1–1	1–1
as part of the	24	20	4–4	2–2	2–2
as well as a	20	15	1–1	1–1	1–1
as well as the	73	35	11–9	18–14	9–8
at a time when	31	25	2–2	5–4	0
at the beginning of	25	17	6–3	7–7	4–4
at the end of	60	28	6–6	11–11	5–3
at the same time	144	61	14–12	20–14	7–7
at the time of	27	16	8–8	4–3	11–9
at the turn of	29	13	2–2	2–2	1–1
at the university of	20	14	2–2	3–3	15–10
both sides of the	20	13	0	4–3	2–2
by the end of	34	21	6–4	14–10	2–2
for the first time	43	28	8–7	12–10	1–1
from the perspective of	22	17	0	0	0
in the case of	48	29	5–4	20–14	7–6
in the context of	47	28	4–4	5–4	2–2
in the course of	26	20	1–1	2–1	2–2
in the face of	48	33	2–2	10–6	1–1
in the first place	22	13	5–5	2–2	0
in the form of	30	23	7–7	13–11	7–6
in the hands of	22	13	0	3–3	5–4
in the history of	32	24	6–6	5–5	2–1
in the midst of	34	25	2–2	6–5	5–3
in the name of	39	23	5–2	5–3	2–2
in the wake of	30	13	2–2	0	0
in the words of	20	13	0	2–2	1–1
on the basis of	37	19	5–2	6–5	1–1
on the eve of	21	14	0	0	0
on the one hand	50	28	2–2	0	1–1
on the other hand	80	40	23–19	16–14	8–7
on the part of	23	18	1–1	12–6	2–2
one of the most	53	35	20–19	23–15	11–8
over the course of	20	14	0	2–2	1–1
the beginning of the	44	28	14–13	19–15	7–6
the course of the	26	18	0	3–2	0
the creation of a	26	18	1–1	4–4	3–2
the end of the	107	47	23–16	27–19	12–8
the eve of the	20	12	0	0	0
the extent to which	22	15	0	2–2	5–3
the fact that the	28	24	8–6	11–10	1–1
the history of the	51	30	18–11	6–5	6–5
the image of the	20	13	3–2	1–1	0
the majority of the	23	17	4–4	12–10	3–3

Table 7 (continued)

Target bundles	Published history		Student writing		
	Frequency	# of texts	Lower div.	Upper div.	Graduate
			Freq.–# of texts	Freq.–# of texts	Freq.–# of texts
			Target bundles	Frequency	# of texts
the nature of the	26	22	1–1	2–2	5–4
the power of the	25	19	5–3	4–4	4–1
the rest of the	54	31	10–9	25–19	12–7
the turn of the	47	24	2–2	4–4	11–9
the ways in which	30	14	1–1	2–2	0
to the extent that	22	15	1–1	2–2	1–1
turn of the century	35	17	1–1	3–3	11–9
was one of the	20	18	8–8	10–8	5–4

Repetitive use of bundles and the use of academically oriented functions for target bundles are also relevant issues in students' writing in biology. The issue of repetitiveness is also present in students' writing at all levels. This issue seems to be more stressed in higher levels of writing in this discipline. This finding resembles the case of repetitiveness in the use of fixed expressions produced by non-native speakers of English in written essays reported by Granger (1998), in which "students 'cling on' to certain fixed phrases and expressions which they feel confident in using" (p. 156). Some bundles present the functional mismatch with target bundles functions, which was also reflected in history writing. In this case, the mismatch is more often in lower division writing and tends to disappear in higher division and graduate writing.

Perhaps the most important finding resulting from the analysis and comparison of the use of target bundles at different university levels in biology is the evolution that can be seen in the use of certain bundles, especially text organizers, across levels. Even though the papers that made up the different levels in the corpus did not belong to the same students, similar to the history corpus, a development can be traced, especially in the use of text organizers. This finding reflects the more frequent use of organizational patterns in higher level students, mainly in graduate students' writing. This evolution may reflect the need for the use of these organizational patterns in more research-oriented writing tasks, as those completed by graduate students in this study.

4. Pedagogical applications and further research

The findings of this study demonstrate that even though students might have frequently encountered these expressions in their academic reading, simple exposure to the frequent use of lexical bundles in published academic writing does not result in the acquisition of these expressions by university students. Not only was the frequency of the target bundles used by students extremely low, but also students' use of

Table 8
Target bundles in published and student writing in biology

Target bundles	Published biology		Student writing		
	Frequency	# of texts	Lower div. Freq.-# of texts	Upper div. Freq.-# of texts	Graduate Freq.-# of texts
a function of the	36	31	0	0	2–2
a large number of	38	29	3–3	1–1	1–1
a measure of the	20	19	0	0	1–1
a wide range of	49	37	3–3	6–6	3–3
a wide variety of	22	19	5–4	3–3	3–3
an important role in	31	22	3–3	6–6	1–1
an increase in the	21	19	3–3	10–7	2–2
an order of magnitude	24	18	0	0	0
are likely to be	41	29	0	0	2–2
are more likely to	20	15	0	6–6	6–4
as a consequence of	38	25	1–1	0	0
as a function of	88	35	0	1–1	4–2
as a result of	54	37	4–4	9–8	24–12
as the number of	21	14	0	0	3–3
as well as the	47	37	6–4	22–19	20–13
at the beginning of	28	22	2–2	1–1	0
at the end of	62	36	5–4	5–5	7–6
at the same time	23	17	1–1	8–6	2–2
at the time of	37	24	3–3	5–3	4–4
at the university of	25	12	1–1	7–4	0
be the result of	21	14	0	2–1	2–2
can be used to	51	38	5–5	9–7	0
did not differ significantly	35	21	0	0	6–4
each of the three	31	18	7–4	0	22–11
for each of the	56	30	4–3	0	14–10
for the evolution of	36	18	0	0	0
has been shown to	28	21	1–1	12–9	5–4
have been shown to	21	18	1–1	11–8	11–11
in addition to the	30	26	3–3	3–3	7–6
in contrast to the	33	30	0	0	1–1
in each of the	25	19	11–6	1–1	14–6
in the absence of	111	56	3–3	3–3	2–2
in the case of	62	38	5–4	10–6	7–6
in the context of	39	26	1–1	0	0
in the evolution of	37	26	2–2	0	1–1
in the form of	30	22	0	9–7	6–5
in the life cycle	27	6	0	0	0
in the number of	41	30	4–3	3–3	6–5
in the presence of	48	25	3–3	5–5	12–7
in the present study	35	21	0	0	1–1
in the study area	47	12	0	1–1	5–3
is a function of	32	18	0	1–1	2–2
is consistent with the	41	29	1–1	0	1–1
is likely to be	39	27	0	0	2–1
is the number of	44	25	0	0	1–1
it is difficult to	40	32	5–4	5–5	3–3

Table 8 (continued)

Target bundles	Published biology		Student writing		
	Frequency	# of texts	Lower div. Freq.–# of texts	Upper div. Freq.–# of texts	Graduate Freq.–# of texts
it is important to	25	20	5–5	15–12	28–12
it is likely that	24	21	1–1	2–2	6–5
it is possible that	50	34	6–6	2–2	22–14
may be due to	21	17	3–3	5–4	13–10
more closely related to	29	15	0	0	0
not appear to be	21	21	1–1	1–1	2–2
not significantly different from	22	15	0	0	2–2
of the distribution of	29	16	0	0	1–1
of the effects of	30	20	1–1	9–7	3–2
of the number of	48	34	0	0	1–1
of the relationship between	24	15	0	0	3–3
of the variance in	25	12	0	0	0
of the variation in	29	16	0	0	0
on the basis of	71	42	0	1–1	0
on the evolution of	23	17	0	0	0
on the number of	25	17	2–2	1–1	4–1
on the order of	25	12	0	0	1–1
on the other hand	65	44	1–1	9–8	9–9
one of the most	25	24	9–8	22–13	11–8
play an important role	22	17	1–1	1–1	1–1
presence or absence of	20	17	4–3	4–4	7–7
similar to that of	27	19	2–2	1–1	2–2
spatial and temporal variation	20	6	0	0	0
studies have shown that	25	21	6–6	12–11	7–7
that the number of	21	16	0	0	0
the average number of	30	19	0	0	5–4
the beginning of the	26	22	1–1	0	0
the center of the	20	13	3–2	2–2	6–6
the course of the	23	15	0	4–3	1–1
the degree to which	33	23	0	1–1	0
the depth of the	36	6	1–1	0	4–4
the direct effects of	20	5	0	0	3–2
the effect of the	22	15	1–1	0	1–1
the effects of the	23	17	4–4	1–1	11–5
the end of the	68	38	4–4	4–4	5–5
the extent to which	40	30	0	1–1	0
the genetic basis of	21	12	0	0	0
the length of the	26	19	4–3	0	4–2
the magnitude of the	20	16	1–1	0	1–1
the mean number of	37	23	0	1–1	9–4
the mean of the	34	10	1–1	0	0
the nature of the	21	18	1–1	0	1–1
the number of individuals	22	14	3–1	0	2–2
the number of species	25	12	2–1	0	2–1
the population dynamics of	21	10	0	0	2–2
the position of the	23	14	1–1	0	1–1
the probability that a	20	13	0	0	0

Table 8 (continued)

Target bundles	Published biology		Student writing		
	Frequency	# of texts	Lower div. Freq.–# of texts	Upper div. Freq.–# of texts	Graduate Freq.–# of texts
the relationship between the	21	13	0	1–1	7–7
the relative importance of	54	31	0	0	7–3
the rest of the	31	27	4–4	6–4	4–4
the results of the	31	23	7–6	6–6	4–3
the shape of the	27	18	1–1	1–1	1–1
the size of the	40	24	10–5	3–3	1–1
the slope of the	26	14	0	1–1	0
the total number of	91	47	2–2	0	23–15
there is no evidence	21	16	0	1–1	0
these results suggest that	22	20	1–1	0	10–6
used in this study	26	20	0	0	4–4
was found to be	20	15	4–4	5–4	4–3
was positively correlated with	23	11	0	0	1–1
with respect to the	29	19	3–2	0	2–2
with the exception of	30	22	0	0	5–4
with the number of	26	20	0	0	0

bundles did not always convey those functions typically associated with published academic writing in history and biology. The role of unconscious learning did not help students master the use of these expressions. A good way to help students become more familiar with lexical bundles could be to have them “notice” the frequent use of these expressions (Schmidt, 1990), becoming aware of the different contexts and discourse functions they perform in academic disciplines. Conzett (2000) stated that noticing “can refer to both accidental awareness and also to the results of deliberate focusing of attention” (p. 161). When dealing with the teaching of collocations, she suggests helping students understand the types of frequent fixed expressions they come across in their readings and become explicitly aware of their uses. This type of activity could be transferred to the teaching of lexical bundles.

The functions that lexical bundles perform in academic writing make these expressions an important topic that needs to be exploited in the teaching of this register. For many years, writing instructors have been teaching different types of connectors and discourse markers, for example, to help their students produce more coherent writing. Many times, these discourse markers were simple words or two-word collocations (Fraser, 1999; Schourup, 1999). As lexical bundles are very frequent in published academic prose, it is necessary to encourage students to use these expressions. Instructors could introduce some of these lexical bundles (for example, those that have text organizing functions) together with some of those discourse markers or connectors, going beyond single words or two-word combinations to longer phrases, phrases fragments, or clause fragments. Native and non-native speakers of English would benefit from increased awareness of the functions performed by lexical bundles when used by published authors. As a result, students

should have a better understanding of academic readings and be able to use these bundles in a more appropriate way in academic writing. Non-native speakers of English, students who are attending English for Academic Purposes courses, for example, might have often encountered many of these word combinations. These students might possess the level of English necessary to understand the meanings of these expressions, as most of the words that make up most bundles are words which may be introduced in the early stages of second language teaching. The problem arises when students face the functions that these expressions perform in academic writing, which sometimes are not as transparent as initially expected. Both native and non-native speakers may feel more confident using other words or expressions to convey these functions. One reason students may avoid using lexical bundles might be that students do not dare to risk the chance of making mistakes by using these expressions, which are unfamiliar or may convey different functions when used in academic prose. That is one of the reasons why students should become more familiar with these expressions and be encouraged to use them with their academic-oriented functions.

A basic issue must also be taken into consideration when dealing with the implications of this study: Are students using other forms to convey the functions performed by target bundles in these disciplines? If so, what are those forms? Perhaps students are not using bundles because they are simply not trying to convey those functions. Students might also be using other bundles which were not identified as target bundles, but which could not be investigated in this study because of corpus size limitations. This issue needs further consideration.

It is possible that students were never explicitly instructed to use these target bundles when writing for their academic disciplines, and that is the reason why the use of these expressions is so rare in the student corpora. It is necessary to conduct further studies which should focus on instructor's perceptions of the use of lexical bundles by students. If papers which contain these bundles (used in the right contexts and with their academic-oriented functions) are perceived as examples of better academic writing in the disciplines, then the systematic teaching of lexical bundles should be more extensively analyzed and developed.

5. Conclusion

The main objective of this study was to identify the lexical bundles used by professional authors in history and biology and to compare the use of those bundles between published authors and university students, trying to contribute to the description of writing in these disciplines. The findings of this study revealed that the use of target bundles by students in biology and history courses at different university levels very far from resembles the use of these bundles by published authors in these disciplines. Further studies should concentrate on different ways to bridge the gap between published writing and student writing in academic disciplines, not only in the use of lexical bundles, but in the use of a wide variety of linguistic features in order to complete a better description of academic writing.

Appendix A. Target bundles in published and student writing in history and biology

See Tables 7 and 8.

References

- Altenberg, B. (1993). Recurrent word combinations in spoken English. In J. D'Arcy (Ed.), *Proceedings of the fifth nordic association for english studies conference* (pp. 17–27). Reykjavik: University of Iceland.
- Bahns, J. (1993). Lexical collocations: A contrastive view. *English Language Teaching Journal*, 47(1), 101–114.
- Bamberg, B. (1983). What makes a text coherent? *College Composition and Communication*, 34(4), 417–429.
- Barlow, M. (2002). *MonoConcPro (version 2.0): Computer software*. Houston, TX: Athelstan.
- Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honor of Stig Johansson* (pp. 181–189). Amsterdam: Rodopi.
- Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (Eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech*. Frankfurt/Main: Peter Lang.
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
- Borland Delphi Professional (version 4.0): Computer software, Imprise Corporation, 1998.
- Butler, C. (1997). Repeated word combinations in spoken and written text: Some implications for functional grammar. In C. Butler, J. Connolly, R. Gatward, & M. Wismans (Eds.), *A fund of Ideas: Recent development in functional grammar* (pp. 60–77). Amsterdam: Institute for Functional Research into Language and Language Use.
- Chafe, W. (1968). Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language*, 4, 109–127.
- Conrad, S. (1996). Academic discourse in two disciplines: Professional writing and student development in biology and history. Unpublished Doctoral Dissertation, Northern Arizona University.
- Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching collocation* (pp. 70–87). Hove: Language Teaching Publications.
- Cortes, V. (2001a). Lexical bundles in context: A new taxonomy. Unpublished Manuscript, Northern Arizona University.
- Cortes, V. (2001b). The issue of corpus size in the study of lexical bundles. Paper presented at the third north american symposium on corpus linguistics and language teaching, Boston.
- Cortes, V. (2002). Lexical bundles in published and student academic writing in history and biology. Unpublished Doctoral Dissertation, Northern Arizona University.
- Coulmas, F. (1981). *Conversational routine*. New York: Mouton Publishers.
- Fillmore, L. W. (1979). Individual differences in second language. In C. Fillmore, D. Kempler, & W. Wang (Eds.), *Individual differences in language ability and behavior* (pp. 203–228). New York: Academic Press.
- Firth, J. R. (1951). Modes of meaning. *Essays and Studies (The English Association)*, 118–149.
- Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 14, 383–395.
- Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145–160). Oxford: Oxford University Press.
- Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, 24(2), 287–297.
- Haswell, R. (1991). *Gaining ground in college writing: Tales of development and interpretation*. Dallas: Southern Methodist University Press.
- Hewings, M., & Hewings, A. (2002). “It is interesting to note that...”: A comparative study of anticipatory ‘it’ in student and published writing. *English for Specific Purposes*, 21, 367–383.
- Howarth, P. (1998). The phraseology of learner's academic writing. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and application* (pp. 161–1186). Oxford: Oxford University Press.

- Jespersen, O. (1917). *Negation in English and other languages*. Copenhagen: A.F. Host.
- Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen & Unwin.
- Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111–127). London: Longman.
- Kress, G. R. (1976). *Halliday: System and functions in language – Selected papers*. London: Oxford University Press.
- Langer, J. A. (1992). Speaking and knowing: Conceptions of understanding in academic disciplines. In A. Herrington & C. Moran (Eds.), *Writing, teaching, and learning in the disciplines* (pp. 69–85). New York: Modern Language.
- Manes, J., & Wolfson, N. (1981). The compliment formula. In F. Coulmas (Ed.), *Conversational routine: Explorations in standardized communication situations and prepatterned speech* (pp. 115–132). New York: Mouton Publishers.
- McCulley, G. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English*, 19(3), 269–282.
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford: Clarendon Press.
- Nattinger, J., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
- Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.
- Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory native like selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–230). London: Longman.
- Peters, A. (1983). *The units of language acquisition*. Cambridge, MA: Cambridge University Press.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.
- Schourup, L. (1999). Discourse markers. *Lingua*, 107, 227–265.
- Simpson, J. & Weiner, E. S. C. (Eds.). (1973). *Oxford English dictionary*. Oxford: Clarendon Press.
- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Stockton, S. (1994). Students and professional writing in biology: Disciplinary work and apprentice storytellers. *Language and Learning across the Disciplines*, 1, 79–103.
- Stockton, S. (1995). Writing in history: Narrating the subject of time. *Written Communication*, 12, 47–73.
- Tannen, D. (1987). Repetition in conversation as spontaneous formulacity. *Text*, 7(3), 215–243.
- Walvoord, B., McCarthy, L. P. (1990). *Thinking and writing in college: A naturalistic study of students in four disciplines*. Urbana, IL: National Council of Teachers of English.
- Webster, J. (Ed.). (2002). *Linguistic studies of text and discourse*. London: Continuum.
- Wilkinson, L. (1985). A freshman writing course in parallel with a science course. *College Composition and Communication*, 36, 160–165.
- Yorio, C. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly*, 14(4), 433.