Lexical bundles in scientific English: A corpus-based study of native and non-native writing

Danica Joy Lorenzo Salazar

UNIVERSITAT DE BARCELONA

Lexical bundles in scientific English: A corpus-based study of native and non-native writing

Tesi doctoral presentada per

Danica Joy Lorenzo Salazar

com a requeriment per a l'obtenció del títol de

Doctora per la Universitat de Barcelona

Programa de Doctorat: Lingüística Aplicada
Bienni 2006-2008

Dra. Isabel Verdaguer Clavera
Directora

Departament de Filologia Anglesa i Alemanya
Facultat de Filologia
Universitat de Barcelona
2011

Abstract

The present dissertation is a corpus-based investigation of the frequency, structure and functions of lexical bundles in published scientific writing in English, whose main objective is the creation of an inventory of the most frequent and pedagogically useful lexical bundles in scientific prose, one that can be utilized in a variety of teaching applications.

In this study, three- to six-word lexical bundles were extracted from a 1.3 million-word sample from the Health Science Corpus, a collection of published articles in biology and biochemistry. This initial list was filtered and enhanced through the application of the Mutual Information (MI) statistic and of a set of exclusion criteria established to satisfy the pedagogical objectives of the study. Following the SciE-Lex investigation (Verdaguer et al., 2009) the remaining lexical bundles were grouped together using like keywords. The present study additionally used the concept of prototypical bundle, which is based on Sinclair's (2004) notion of canonical units of meaning, to tackle the semantic and structural connections between similar bundles. The structural and functional characteristics of the lexical bundles were explored through careful concordance analysis, which made it possible to categorize the bundles using modified versions of Biber et al.'s (1999) structural framework and Hyland's (2008a) functional taxonomy.

These quantitative and qualitative analyses reveal how native expert writers employ recurrent word strings in the construction of a coherent, well-structured and convincing scientific text that conforms with the conventions of the genre. They bring to light the different functions that lexical bundles perform in scientific discourse, and how these functions enable writers to address their research concerns, achieve their communication goals and elicit the desired reaction from their target audience. They also show the typical structural realizations of these bundle functions, as well as important aspects of usage that non-native writers need to be aware of to be able to incorporate these expressions in their own writing.

The study also compares the results obtained from the corpus of published scientific articles to the lexical bundles found in a smaller corpus of biomedical research articles written by native Spanish-speaking scientists, who are all non-native users of English. In accordance with the methodology proposed by Cortes (2004), the lexical bundles identified in the HSC were treated as target bundles and subsequently searched for and analyzed in the corpus of non-native writing. This comparison uncovered non-native writers' overuse of certain bundles, a tendency that results in unnecessary repetitiveness and lack of variation, as well as their restricted use of participant-oriented bundles, which points to their limited awareness of the usage and importance of this particular function.

The dissertation also discusses the pedagogical implications of its final product, a practical list of lexical bundles in scientific English for use in teaching applications, and how it addresses the six major challenges that hinder the successful introduction of lexical bundles in EAP classrooms and teaching materials, as identified by Byrd and Coxhead (2010).

To Enrico

For Breaking Bad*, and everything else you were just so right about*

Acknowledgments

First, I would like to acknowledge the predoctoral scholarship granted by the Government of Catalonia through the Agency for Management of University and Research Grants (AGAUR, grant no. 2008FI 00044), and the funding provided by the Spanish Ministry of Science and Innovation and FEDER for the SciE-Lex project (project no. HUM2007-64332/FILO), in which I was fortunate enough to take part.

I am deeply grateful to my thesis adviser, Dr. Isabel Verdaguer, who has not only provided me with constant guidance and kind encouragement throughout the course of this work, but has also believed in me enough to include me in her projects and help me advance in my academic career. None of my recent accomplishments would have been possible without her support. It would be difficult to find a kinder, more generous person, and I will always be thankful for having her as my mentor.

I would also like to thank my other colleagues in the GReLiC research group: Emilia Castaño, Elisabet Comelles, Dr. Trinidad Guzman, Dr. Joseph Hilferty, Dr. Natalia Judith Laso and Aaron Ventura. I feel blessed to have had the opportunity to work with them, and I look forward to our continued collaboration.

My thanks go to all my professors and classmates in the Applied Linguistics doctorate program of the University of Barcelona. I would especially like to thank my dear friends, Dr. Claudia Marcela Chapetón and Mireia Ortega, for all the great times we have spent together. I thank them for their friendship and for sharing some of the most important events of my life.

Thanks as well to all my other friends in Barcelona, who have made a girl from the

other side of the world feel right at home. Special thanks go to my dear friend Marisa de Prada, for everything she has done for me.

I wish to thank everyone at the Centre for English Corpus Linguistics at the Catholic University of Louvain in Belgium, especially Dr. Sylviane Granger, Dr. Magali Paquot, Dr. Sylvie De Cock, Dr. Gaëtanelle Gilquin and Dr. Fanny Meunier, for their valuable insight and the warm hospitality they always give me in my research visits to Louvain-la-Neuve.

I would also like to acknowledge Prof. Iliana Martínez of the National University of Río Cuarto (UNRC) in Argentina, for allowing me to use her corpus of Argentinian research writing.

I am also grateful to Robert Ranieri and everybody at Prime Management, whose kindness and generosity made it possible for me to attend three AAAL conferences, truly unforgettable experiences that have made a positive impact on my development as a researcher.

In the past three years, I have been lucky enough to travel to many places around the world to attend linguistics conferences, where I have met some truly remarkable people who, by generously giving me their time, have helped steer my research in the right direction. I am indebted to them all, but most of all to those who ended up being my friends: Rachel Wicaksono and Dr. Christopher Hall from York St. John University in the United Kingdom, and Dr. Shirley Dita, Dr. Ariane Borlongan and Dr. Danilo Dayag from De La Salle University in the Philippines.

My travels have also made me realize the true meaning of the Filipino diaspora. It means having first cousins, second cousins, aunts, uncles, in-laws and friends in every country you can possibly go, who are all guaranteed to welcome you and make you feel loved. My thanks go to all these first cousins, second cousins, aunts, uncles, in-laws and friends, all of whom I cannot wait to see again!

To my immediate family, who has been there from the beginning, I give all my love and gratitude. My deepest, sincerest thanks go to my mother, for being my very best friend and my inspiration. Nothing motivates me more than the idea of making her proud.

Finally, to Enrico. Thank you so much for being you, and for sharing this wonderful life with me and Lilly. I really cannot ask for anything more. This is all because of you.

Maraming, maraming salamat sa inyong lahat.

Danica Salazar

Barcelona, October 20, 2011

Table of contents

List of tables

List of figures

Chapter I

Introduction

It is undeniable that English has established itself as a language of international prestige, given its status of *lingua franca* in many important fields of contemporary life (Hoffman, 2000). Among these fields is the academe, with English now playing a leading role in the dissemination of academic knowledge all over the world. The predominance of the language in higher education and research is obvious in the sheer number of academic journals being published in English, of second-language speakers studying academic subjects in English, and of non-native academics required to carry out most, if not all, of their scholarly work in English. The growth of English as the international language of academic communication is a hotly debated issue, with one side defending the language as a valuable tool that empowers its users by breaking down linguistic barriers to knowledge, and the other viewing it as "a powerful carnivore gobbling up the other denizens of the academic linguistic grazing grounds" (Swales, 1997, p. 374). A large number of non-native scientists in many parts of the world are situated in this complex, English-dominated academic context, and many of them find their written production in this language falling short of academic expectations when measured against expert-writer models.

The difficulties faced by non-native writers in producing accurate, effective academic texts in English have prompted a multitude of studies on the elements that constitute well-written academic prose, and the ideal way to teach them to students learning English for use in academic contexts. A significant number of these investigations harness the power of computers to analyze language corpora—large collections of digitally stored, naturally occurring texts—with the aim of establishing linguistic and textual patterns and developing systematic descriptions of these patterns.

One of the most important findings revealed by corpus-based language studies is the fact that, instead of constantly making new combinations of individual words, native speakers often depend on a stock of prefabricated, semi-automatic word chunks (Sinclair, 1991). These results have led researchers to look beyond the word in language description and give importance to collocations and multi-word units of meaning (see reviews in Granger & Meunier, 2008b; Howarth, 1996a; Wray, 2000; Wray & Perkins, 2000).

Corpus-based research has also shown that these multi-word expressions that come so naturally to native speakers are a source of difficulty for non-native users of a language (De Cock, 2003; Granger, 1998; Howarth, 1998; Nesselhauf, 2005). Recurrent word combinations are usually fairly easy to understand, but they can hinder language production. Although ignored by traditional, word-based language descriptions, these lexical sequences are essential to achieving native-like competence and fluency, and are thus important aspects that have to be taken into account in language teaching and learning (Coxhead, 2008; Howarth, 1998b; O'Keeffe, McCarthy, & Carter, 2007; Wray, 2000). The use of words in the correct context and in the correct combinations is part of good writing, and it is important for a second or foreign-language writer to know the most frequent combinations used in specific registers, genres and disciplines. This is especially true in scientific writing, where authors are required to produce succinct, precise texts to be able to communicate their ideas and research results to a scientific audience. Scientific discourse is also governed by stylistic conventions established by community expectations. Gledhill (2000a, p. 204), for instance, speaks of the "phraseological accent" that pervades much of technical writing, a tendency manifested by the widespread use in scientific English of formulaic constructions unusual in general English. This, he claims, is evidence not only of the existence of a scientific discourse community, but also of the influence of community norms on scientific expression.

The phraseological trend in linguistic research has made an impact on the conception and design of reference tools aimed at helping non-native writers bridge the gap between their written academic output and that of their native counterparts. One such tool is the SciE-Lex Electronic Combinatory Dictionary[1], an electronic database of non-technical words used in biomedical English, conceived as a writing aid for members of the Spanish medical community (Verdaguer, Poch, Laso, & Giménez, 2008). The creators of SciE-Lex acknowledged the importance of precision and correctness in scientific discourse and recognized that to be able to provide Spanish scientists with the information needed for precise and correct writing, it was necessary to adopt a linguistic approach that considered both syntax and semantics. By compiling the Health Science Corpus (HSC), their own restricted-domain corpus consisting of four million words of scientific research articles in English from prestige journals of biology, biochemistry and biomedicine, and applying corpus-based research methods to this corpus, they were able to identify the words that were to be entered into the database, and to analyze the relevant features and interconnections of these words. This later enabled them to establish general patterns and develop systematic descriptions for each dictionary entry that include its word class, morphological variants and equivalent(s) in Spanish, as well as the entry's patterns of occurrence, a list of its collocates, some examples of the word in use as attested in the

[1] The HSC and SciE-Lex were created as part of the research project, "Creation of a Database of Lexical Combinations in Scientific English," coordinated by Dr. Isabel Verdaguer of the University of Barcelona and financed by the Spanish Ministry of Science and Education and FEDER (Project Number BFF2001-2988).

corpus and notes to clarify usage.

The contents of the first version of SciE-Lex were largely derived from co-occurrence analysis, a probabilistic, frequency-based approach that highlights instances of word co-selection, termed *collocation* (Manning & Schütze, 1999; Sinclair, 1991; Stubbs, 2002). The information supplied by SciE-Lex on the frequent collocates of non-technical words in scientific research writing can be considered its most unique and significant contribution as a writing tool, given the current shortage of reference materials that focus on the co-occurrence patterns of this type of vocabulary.

However, the SciE-Lex team soon determined that co-occurrence analysis only allowed them to see part of a much bigger picture, and that in order to achieve a more complete description of the conventionalized phraseology of scientific prose, it was also necessary to explore continuous sequences of repeatedly co-occurring words.

One landmark investigation of such highly frequent contiguous sequences of words is the large-scale study of *lexical bundles* published as a chapter of the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999, chap. 13). This chapter was based on the analysis of multimillion-word corpora representing conversation and academic prose. This study, which is founded exclusively on frequency criteria, compares spoken and written university registers and deals with uninterrupted lexical sequences with as many as six words. Biber, Conrad and Cortes (2003) later developed an analytical framework for the classification of lexical bundles according to their discourse functions. In a subsequent study, these authors investigated the use of lexical bundles in university classroom teaching and textbooks (Biber, Conrad, & Cortes, 2004). More recently, further improvements on the lexical bundle approach were offered by authors such as Hyland (2008a), who devised a functional taxonomy for lexical bundles better suited for written research genres, and Simpson-Vlach and Ellis (2010), who used a combination of statistical measures and teacher insights to build a pedagogically valid list of academic formulas similar to lexical bundles.

These studies became the springboard for the second stage of the SciE-Lex project, which involved supplementing the original database with three- to five-word lexical bundles, together with information on their composition, function and textual distribution (Verdaguer, Comelles, Laso, Giménez, & Salazar, 2009). The SciE-Lex team adopted Biber et al.'s (1999) definition of lexical bundles and used frequency criteria to identify them in the HSC. However, to eliminate bundles with no recognizable meaning or function but were frequent only because of the high frequency of their individual components, the mutual information (MI) statistic was also used to create the list, following Simpson-Vlach and Ellis (2010). The list was further refined through the application of a set of exclusion criteria that were necessitated by the pedagogical nature and objectives of SciE-Lex, and by the collocational data already included in the database. Concordance listings were then analyzed to structurally and functionally classify the bundles according to a structural taxonomy modeled after Biber et al. (1999) and a functional classification scheme based on Hyland (2008a). The qualitative part of the analysis and the subsequent linking of lexical-bundle information to SciE-Lex's headwords were facilitated by the grouping of like bundles using shared keywords (Verdaguer et al., 2009).

The present dissertation was carried out within the framework of the second phase of the SciE-Lex project[2]. The study was conducted based on the same principles used by the SciE-Lex team in producing the list of lexical bundles to be included in the second, expanded version of the dictionary. It is a similarly frequency-driven, corpus-based investigation of the frequency, structure and functions of lexical bundles in published scientific writing in English. However, the study extends its scope beyond SciE-Lex by establishing as its main objective the creation of an inventory of the most frequent and pedagogically useful lexical bundles in scientific prose, one that can be utilized in a variety of teaching applications.

In this study, three- to six-word lexical bundles were extracted from a 1.3 million-word sample of the HSC. This initial list was filtered and enhanced through the application of the MI score and of a set of exclusion criteria established to satisfy the pedagogical objectives of the study. As in the SciE-Lex investigation, the remaining lexical bundles were grouped together using like keywords. The present study additionally used the concept of prototypical bundle, which is based on Sinclair's notion of canonical units of meaning, to tackle the semantic and structural connections between similar bundles. The structural and functional characteristics of the lexical bundles were explored through careful concordance analysis, which made it possible to categorize the bundles using modified versions of Biber et al.'s (1999) structural framework and Hyland's (2008a) functional taxonomy.

These quantitative and qualitative analyses reveal how native expert writers employ recurrent word strings in the construction of a coherent, well-structured and convincing scientific text that conforms with the conventions of the genre. They

[2] The second stage of the SciE-Lex project was funded by the Spanish Ministry of Science and Innovation and FEDER (project no. HUM2007-64332/FILO). The present dissertation was financed by the same project and by a grant from the government of Catalonia (grant no. 2008FI 00044).

bring to light the different functions that lexical bundles perform in scientific discourse, and how these functions enable writers to address their research concerns, achieve their communication goals and elicit the desired reaction from their target audience. They also show the typical structural realizations of these bundle functions, as well as important aspects of usage that non-native writers need to be aware of to be able to incorporate these expressions in their own writing.

The study goes one step further by comparing the results obtained from the corpus of published scientific articles to the lexical bundles found in a smaller corpus of biomedical research articles written by native Spanish-speaking scientists, who are all non-native users of English. In accordance with the methodology proposed by Cortes (2004) in her comparative study of lexical bundles in published and student writing in history and biology, the lexical bundles identified in the HSC were treated as *target bundles* and subsequently searched for and analyzed in the corpus of non-native writing. This comparison with a non-native corpus underscores the differences between the native and non-native writers and pinpoints instance of overuse and underuse. This in turn serves to improve our understanding of the difficulties that non-native scientists may face in the use of lexical bundles, and how these difficulties can be addressed in the language classroom, as well as in language-learning materials and research-writing aids.

The main objectives of the study are reflected in the following research questions:

1. What are the most frequently occurring target bundles in the HSC?
2. What are the structural and functional characteristics of these target bundles? How can they be classified according to these features?

3. Do the target bundles also occur in the corpus of non-native scientific writing?
4. What are the differences between the native and non-native corpora in terms of the frequency, structure and functions of the target bundles?

This dissertation is structured in eight chapters. Following this first introductory chapter, Chapter II presents a review of the literature that informed the present investigation. This includes a brief overview of relevant corpus-based language studies and previous research on phraseology, with a special emphasis on lexical bundles. Chapter III explains the rationale behind using a corpus-based approach, details the corpora used in the study and provides justification for the methodological choices taken in the creation and analysis of the list of target bundles and the comparison of these findings with the non-native corpus. Chapter IV describes in greater depth the process of generating, refining and organizing the list of target bundles, centering on lexical bundle extraction, the application of exclusion criteria, the keyword analysis and the determination of prototypical bundles. Chapter V deals with target bundles and their frequency, structure and functions in the corpus of native expert writing, while the succeeding chapter, Chapter VI, gives an account of the frequency and structural and functional features of prototypical target bundles in non-native expert scientific writing. Chapter VII, which is devoted to the pedagogical applications of the study, summarizes the useful features of its final product, a practical list of lexical bundles in scientific English for use in teaching. It also refers to the six challenges to teaching lexical bundles identified by Byrd and Coxhead (2010) and discusses how the results of the investigation address each of these challenges. Finally, the dissertation closes with some concluding remarks and recommendations for further research.

Chapter II

Review of literature

1. Corpus-based language studies

In recent years, linguists have exploited increasingly sophisticated computer technology to compile ever-larger collections of text on which to base studies of naturally occurring language, thereby establishing the corpus-based approach as a methodology for linguistic analysis. John Sinclair, one of the pioneers of modern corpus linguistics, defines the term *corpus* as "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (2005, p. 16). Some of the basic techniques of analysis that can be done on corpora using standard, widely available text-handling tools are concordancing, word frequency counts or wordlists, keyword analysis, cluster analysis and lexico-grammatical profiles. Although frequency is a key issue in this type of investigation, corpus-based studies do not only rely on simple counts of linguistic features, but also involve qualitative interpretations of quantitative data. The goal of corpus-based research goes beyond merely reporting numerical findings; it also aims to uncover patterns of language use through the analysis of these results (Biber, Conrad, & Reppen, 1998).

Corpus-based analytical methods offer a different perspective of language, one that emphasizes language use rather than structure. They have opened up new avenues of research and have been applied to such diverse fields as lexicology, semantics, pragmatics, discourse analysis, dialectology, language variation studies, sociolinguistics, historical linguistics, translation, stylistics, psycholinguistics, cultural anthropology, social psychology and forensic linguistics. This review of literature gives a brief overview of the impact of corpora on some areas that are particularly relevant to the present investigation: lexicography, lexis and grammar, English-language teaching, and its subfield of English for Academic Purposes (EAP). It will then focus on the latest developments in the relatively new discipline of phraseology that have informed the design and execution of this study.

1.1. Lexicography

The advent of computers and electronic corpora has brought about a revolutionary change in dictionary making. Corpus linguistics has transformed lexicographical practice by providing entirely new sources of linguistic evidence and novel ways of handling, analyzing and presenting lexicographic data.

The first large-scale dictionary project to exploit the potential of large electronic corpora is the *Collins COBUILD English Language Dictionary* (Sinclair, 1987). This dictionary was created using evidence from the Collins-Birmingham University International Language Database (COBUILD), which has now grown to the vast and still expanding Bank of English. The original COBUILD corpus, collected in Birmingham in the 1980s under the direction of John Sinclair, has produced a number of dictionaries and grammars, including several editions of the influential *Collins COBUILD Dictionary* and the *Collins COBUILD Grammar Patterns* series (Francis, Hunston, & Manning, 1996, 1998).

What was pioneered by the COBUILD project is now the accepted practice in lexicography, as language corpora are now considered the standard tool for lexicographers (O'Keeffe, McCarthy, & Carter, 2007). The corpus method has replaced the laborious, time-consuming and highly subjective citation method as the principal means of collecting lexicographic data. All major publishers now rely on multi-million word corpora to compile dictionaries and related reference materials. The Cambridge International Corpus (CIC), for instance, has over one billion words as of the time of writing. There is also the widely used British National Corpus (BNC), a large, entirely annotated reference corpus compiled by a consortium of dictionary publishers. More and bigger language corpora are becoming available in many other languages apart from English.

By giving them access to vast amounts of authentic language data, language corpora have enabled lexicographers to count the occurrences of words and expressions and determine their relative frequency (Svensén, 2009). Corpora have also made it possible to examine the properties of a language in depth, bringing to the lexicographer's attention those instances of normal usage that ordinarily escape human perception.

Corpora provide clear, objective criteria for selecting headwords, analyzing material, writing definitions and ordering word senses. In corpus-driven lexicography (Williams, 2002), analysts depend on the patterns that emerge from the corpus, not on their intuition. This has resulted in more contextually relevant dictionaries, whose definitions of both lexical and grammatical words are based on evidence derived from real language in use.

The corpus revolution in lexicography also led to the development of more efficient means of storing, accessing, transferring and cross-referencing source material and the creation of new tools designed to handle large quantities of data (Williams, 2003). These tools facilitate the work of lexicographers and leave them free to devote their energies to writing more precise, meaningful dictionary entries (Rundell, 2002).

Computer technology has also made an impact on user access to dictionaries. Electronic and online dictionaries offer several advantages over the traditional paper dictionary, such as efficient integration of detailed information, multiple look-up routes (fuzzy searches, hyperlinks, etc.) and the possibility of user customization (De Schryver, 2003).

Like standard dictionaries, monolingual dictionaries written for language learners are now also largely corpus-driven (*Cambridge advanced learner's dictionary*, 2008; Hornby et al., 2010; Rundell, 2007). Corpus technology not only enhances the content of learner's dictionaries, but also offers novel means of information access and presentation that make these dictionaries more effective tools for both decoding and encoding. The role of the pedagogical dictionary as an encoding aid is strengthened by supplementary material such as the "Improve your writing skills" section of the second edition of the Macmillan English Dictionary for Advanced Learners (De Cock et al., 2007), which relied on the comparative analysis of native and non-native corpora (see Section 1.5. below) to provide detailed advice on academic-writing areas that often cause difficulties for English-language learners.

Currently in development is the Louvain EAP Dictionary (LEAD), a web-based EAP dictionary and writing resource targeted at non-native users. Apart from the rich descriptions of non-technical words used to perform key functions in academic discourse, this dictionary offers both semasiological and onomasiological access and an innovative customization system that automatically adapts content to users' disciplines and mother-tongue backgrounds (Granger & Paquot, 2009).

1.2. Lexis and grammar

Using corpora and corpus tools, lexicographers have been able to analyze patterns of language use that have helped them create more complete, insightful dictionary entries. This patterning of language that corpus linguistics has revealed is perhaps its most important contribution to lexis and grammar, two areas that had previously been considered separate but, thanks largely to corpus research, are now known to be highly interdependent.

Corpus linguistics has challenged the traditional dichotomy of vocabulary and syntax by providing powerful electronic means to uncover instances of lexis-grammar co-selection that used to elude the human observer. Many attempts have been made to explain and illustrate the interrelationship between lexis and grammar, and some of the most influential models are summarized in Römer (2009).

The way to corpus-driven lexico-grammatical research was paved by John Sinclair with two groundbreaking concepts: the idiom principle and lexical grammar. The idiom principle refers to the phraseological tendency of language, whereby words do not appear in isolation, but combine with each other to make meaning (Sinclair, 2004). This is in contrast to the open-choice principle, which assumes that words are individually chosen to fill certain slots in a sentence. According to Sinclair, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (1991, p. 110).

Massive corpus evidence for the inseparability of lexis and grammar led Sinclair to go beyond lexico-grammar and propose the notion of lexical grammar, "an attempt to build together a grammar and lexis on an equal basis" (2004, p. 164), where meaning and structure are considered as one.

Echoing Sinclair's concept of lexical grammar is Hunston and Francis' pattern grammar . Developed from an extensive study of the then 250-million word Bank of English, pattern grammar makes two basic claims about the grammar of individual words, or patterns: "firstly, that all words can be described in terms of patterns; secondly, that words which share patterns also share meanings" (Hunston et al., 1997, p. 209).

The first statement is exemplified by simple patterns such as V and Vn for the verb *to eat*, and by more complex ones such as those associated with the impersonal *it* pattern. Some words have various patterns for the same meaning; others have a particular pattern for a particular sense; while others have several meanings that can be disambiguated using the different patterns they occur in. The second claim is illustrated by a pattern like V *by* -ing, where the V slot is usually filled by verbs that fall into one of two meaning groups: "to start" or "to end" (*begin by saying*) and "to respond to or compensate for something" (*atone by fasting*) (Hunston et al., 1997).

Another radical new theory of language in the Sinclairian contextualist tradition is Hoey's notion of lexical priming (Hoey, 2005). Hoey put forth a theory that reverses the traditional roles of vocabulary and syntax: instead of constraining lexis, grammar is in fact only the output of a highly complex lexical structure. This is a view of grammar as the outcome of frequently associated words "primed" for use with each other in specific contexts and text types. Another central premise of this theory is that our knowledge of a word is conditioned by our encounters with it, as we use it and see it used in different language structures, textual positions and text types (Hoey, 2004, 2005).

A more recent approach that bridges the sense-structure divide also seeks to reconcile corpus and cognitive linguistics. Collostructional analysis poses the question, “Are there significant associations between words and grammatical structure at all levels of abstractness?” (Stefanowitsch & Gries, 2003, p. 211). This family of analytic methods measures the strength of association or repulsion between words and constructions, with the aim of identifying which words occur more or less frequently with particular constructions, thus demonstrating the close interaction between lexis and grammar.

Research at the lexis-grammar interface, made possible by the arrival of corpus linguistics, has drawn attention to the study of meaning beyond the word and brought phraseology to the forefront of language analysis. The studies highlighted in this section are in fact just some of the more influential research strands in the distributional approach to phraseology, which will be discussed in more detail later on in this chapter.

1.3. English language teaching

The corpus studies described above have shown that human intuitions about certain aspects of language, such as semantics and grammar, can very often be wrong. However, it is a fact that most of what is being taught in language classrooms and presented in language textbooks is still based on the intuitions of teachers and textbook authors, and is hardly an accurate reflection of how language is actually used. Corpus linguistics offers a solution by providing an empirical basis for checking our idea of language and bringing to light linguistic features that escape our perception (O’Keeffe et al., 2007).

Corpora can also help close the gap between language in and outside the classroom by giving textbook writers and course designers a means to incorporate more natural discourse features in English-language teaching materials. The *Touchstone* series of course books (McCarthy, McCarten, & Sandiford, 2005) is just one example of corpus-informed material for language learners. Several major publishers have multimillion-word corpora at their disposal, which they use to produce corpus-based grammars, course books, vocabulary books, exam practice books, teaching guides and other resources for English-language learning and teaching.

A more direct application of corpus techniques in language teaching is Computer-Assisted Language Learning (CALL), which includes the use of corpora in the language classroom. With this approach, learners themselves get to use a corpus through guided hands-on tasks or corpus-based materials such as concordance lines on handouts (Johns, 1991). This type of activity is known as data-driven learning (DDL), and many teaching materials based on this approach are currently available in print and online (Johns, 2002).

Another important development in corpus linguistics that has made a significant contribution to English-language teaching is the emergence of learner corpora, which are electronic collections of authentic texts produced by foreign-language learners (Granger, 2003). The learner corpus, however, has applications beyond language teaching, and is thus considered in its own section below.

1.4. English for Academic Purposes

The evidence-based approach of corpus linguistics is extremely useful in determining what is typical in certain genres, as it makes it possible for analysts to examine the most frequent words, phrases and structures in different domains. In the field of

English for Academic Purposes (EAP), this potential has been exploited by various researchers to identify distinctive linguistic features of academic discourse.

Studies of written academic prose have revealed that long words, nouns, nominalizations, derivational suffixes, linking adverbs, attributive adjectives and prepositional phrases are particularly frequent in this type of writing, while second-person pronouns, direct questions, present-tense verbs, private verbs, contractions and *that*-deletions occur rarely (Gilquin, Granger, & Paquot, 2007; Hyland, 2006). Frequency counts of academic vocabulary led to the construction of resources such as Coxhead's (2000) Academic Word List. Research has also pointed to the highly conventionalized nature of EAP-specific phraseology, which is characterized by a number of semantically and syntactically compositional word combinations (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hyland, 2008; Simpson-Vlach & Ellis, 2010). This aspect is particularly relevant to the present study and is further discussed below.

There is currently a debate in the field over the necessity for a general or subject-specific approach to EAP. In the area of academic vocabulary, for instance, corpus-based studies have shown the frequency across disciplines of subtechnical academic words that mainly perform organizational or rhetorical functions (Granger & Paquot, 2009, 2009; Luzón Marco, 2001; Thurston & Candlin, 1998). This finding is supported by corpus-driven work by Paquot (2010), which proves the existence of a range of non-technical words and phrases that is used in a variety of disciplines to fulfill academic functions such as defining, exemplifying, classifying, and reporting other scholars' work. In academic phraseology, Simpson-Vlach and Ellis (2010) were able to extract a number of academic formulas common to many domains. All these results seem to point towards a core academic vocabulary that transcends disciplinary boundaries.

This conclusion stands in contrast to variationist studies that have compared linguistic features across academic disciplines, subdisciplines, and even text sections (Biber & Finegan, 1994; Conrad, 1996; Fløttum, Dahl, & Kinn, 2006; Martínez, 2003; Ozturk, 2007). Authors such as Hyland (2000, 2008) challenge the idea of a core academic vocabulary and highlight the specific features of different disciplinary environments.

As for materials development, findings from corpus-based research have formed the basis of highly useful EAP-oriented resources such as textbooks (Huntley, 2006; McCarthy & O'Dell, 2008; Schmitt & Schmitt, 2005; Swales & Feak, 2004; Thurston & Candlin, 1997) and dictionaries (Major, 2006; Rundell, 2007).

1.5. Learner-corpus research

A fairly recent trend in corpus research is the compilation of learner corpora and analysis of learner language. This relatively new corpus type contains data from foreign or second-language learners compiled following strict design criteria that control a wide range of learner and task variables. Learner-corpus researchers employ various methods of analysis to quantify and examine large amounts of learner data in order to highlight significant patterns in interlanguage. One of these methods is contrastive interlanguage analysis (CIA) (Granger, 1996), a methodology that involves comparisons of learner language and one or more native-speaker reference corpora and comparisons of different varieties of learner language.

A pioneering collection of learner corpora that has generated a number of interesting studies is the International Corpus of Learner English (ICLE). ICLE contains over three million words of essay writing by advanced learners of English as a foreign language from a wide range of mother-tongue backgrounds, including French, German, Dutch, Spanish, Swedish, Finnish, Czech, Japanese, Chinese, Polish and Russian (Granger, 2003). The ICLE project is coordinated by the Centre for English Corpus Linguistics of the Catholic University of Louvain-la-Neuve in Belgium, but it is actually a collaborative effort among several universities in different parts of the world. The 21 ICLE subcorpora were compiled following the same design criteria and are thus directly comparable. This large-scale, international project has already proven to be of enormous value in the study of learner language.

Learner-corpus research holds enormous potential for many fields of linguistic inquiry, not least for EAP, which has long been shown to be a thorny area for native and non-native writers alike. Several linguists call for more learner corpus-based studies in EAP, noting the dominance of studies based exclusively on native corpora in this line of research (Gilquin et al., 2007). Flowerdew (2001), for instance, describes how learner-corpus data can shed light on three areas of difficulty for non-native academic writers: collocational patterning, discourse features and pragmatic appropriacy.

Studies on lexico-grammatical patterning have yielded interesting results. Altenberg and Granger (2001) used learner corpora to show that the anomalous use of restricted collocations and prefabricated expressions led to a high percentage of errors in non-native writers' production. In an earlier study, Milton (1999) examined his non-native students' use of fixed expressions in their essays and found that they depended on a small range of these expressions. To confirm this, he compared a student-essay corpus with a parallel corpus of native writing and proved that non-native students depended on a limited number of fixed phrases, which made their writing style noticeably repetitive.

As for learner-corpus studies on discourse features, a number of them have investigated the use of connectors in EAP writing (Altenberg & Tapper, 1998; Flowerdew, 1998; Granger & Tyson, 1996; Milton & Tsang, 1991). Other authors such as Aijmer (2002), Granger and Rayson (1998), and Hinkel (2002) have found many stylistic features in non-native essay writing that are more characteristic of informal speech than written academic discourse. Finally, pragmatic inappropriacy in non-native academic writing has been highlighted by various studies (Aijmer, 2002; Hyland & Milton, 1997; Neff, 2008).

Learner-corpus research has also uncovered that some language features are common to learners from several native-language backgrounds while others are only observed in certain learner groups. While the former characteristics may be attributed to developmental factors, the latter features may be presumed to result from first-language influence.

The fact that many corpus-based studies use novice writing in both learner and native control corpora makes the novice-writer effect another factor to be taken into account in learner-corpus research. Although many difficulties in academic writing appear to be specific to learners, others seem to be shared by native writers and non-native novice writers. For example, Cortes (2002a, 2002b) showed considerable differences between novice and professional writers in their use of lexical bundles typical in EAP, while Neff, Ballesteros, Dafouz, Diez, Martínez et al. (2004) demonstrated excessive reader-writer visibility in both groups of novice writers.

2. Phraseology

2.1. The scope of phraseology

Cowie defines phraseology as "the study of the structure, meaning and use of word combinations" (1994, p. 3168). This interest in how words combine with each other in the English language can be traced back to the early 20th century, when researchers such as Firth (1951), Jespersen (1917, 1924) and Palmer (1933) published theoretical works on collocations and fixed expressions. These were followed in the 1970s, 1980s and early 1990s by qualitative studies of formulaic expressions in both spoken and written language (e.g., Hakuta, 1974; Manes & Wolfson, 1981; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Peters, 1983; Tannen, 1987).

There is currently no shortage of interest or research activity in the comparatively recent field of phraseology, but its development is slowed down by the absence of general consensus on terminology, descriptive approaches and analytical procedures (Granger & Paquot, 2008; Howarth, 1996).

Phraseological units have been given different names by different researchers, among them *lexical phrases*, *formulas*, *routines*, *fixed expressions*, *prefabricated patterns* and *lexical bundles*, and there are as many approaches to their analysis as there are names for them. According to Biber et al. (2004), empirical studies on word combinations differ in terms of: 1) research goals (description of the full range vs. a small set of multi-word units); 2) criteria for identification of multi-word units (perceptual salience, frequency criteria, etc.); 3) formal characteristics of multi-word units (continuous sequences, discontinuous frames or lexico-grammatical patterns; two-word collocations vs. longer sequences); 4) number of text samples used (small vs. large corpora); and 5) presence or absence of register comparisons (written texts only, spoken texts only, both).

Although, as Biber et al. (2004) point out, a diversity in research methods and perspectives is needed to better understand a complex issue like phraseology, it is also true that such a situation “hinders communication between linguists and generally increases the impression of fuzziness in the field” (Granger & Paquot, 2008, p. 28).

Howarth (1996a) attributes the lack of consistency in the area to the way most researchers focus on only a part of the whole phraseological spectrum: idioms for some, collocations for others, and speech formulas for still others, to give Howarth’s examples. He also cites phraseology’s almost independent development in a wide range of disciplines: from descriptive linguistics, lexicography and discourse analysis to second language acquisition and pedagogy, language processing and even artificial intelligence (Howarth, 1996).

Granger and Paquot (2008), for their part, link phraseology’s variable scope to its vague boundaries with four related disciplines: semantics, morphology, syntax and discourse. They also outline two distinct approaches to the study of phraseology: the traditional approach and the distributional approach.

2.2. Two approaches to phraseology

The traditional approach

The traditional approach to the study of word combinations is strongly influenced by the Russian perspective on phraseology, where a set of linguistically identified multi-word expressions lies on a continuum of fixedness. At one end of this spectrum are pure idioms, which are the most rigid and least substitutable and are thus considered the "prototype of the phraseological unit" (Gläser, 1998, p. 126), while at the other end are free combinations.

The traditional approach draws a clear demarcation line between the realm of phraseology and those of syntax and semantics by disregarding variable combinations that are subject only to syntactic and semantic restrictions, as well as fully compositional multi-word units whose meanings are predictable from their constituent parts. This approach also places emphasis on units with identifiable discourse features, such as Cowie's (1988) routine and speech formulae and Mel'čuk's (1998) pragmatic phrasemes.

These two authors proposed two of the more important typologies within the traditional approach. Cowie's (1988, 1994) model distinguishes between composites and formulae. Composites are further subdivided into three categories that fall on a continuum from transparent to opaque: restricted collocations, figurative idioms and pure idioms. Formulae, subdivided into routine and speech formulae, are autonomous sentence-like units that fulfill certain pragmatic functions. Mel'čuk's (1998) model, with its dual categories of semantic and pragmatic phrasemes, is a similarly influential framework subscribing to the traditional view of phraseology.

The distributional approach

The large amounts of authentic language data and the multi-word extraction techniques afforded by modern corpus linguistics have enabled researchers to explore the phraseological tendency of language as never before. Corpus-based studies have not only confirmed the interaction between syntax and semantics, but have also shown the pervasiveness of patterns and formulaic sequences in language use. These studies prove that instead of constantly making new combinations of individual words, native speakers often depend on a stock of prefabricated, semi-automatic word chunks. As Sinclair (1991) observes:

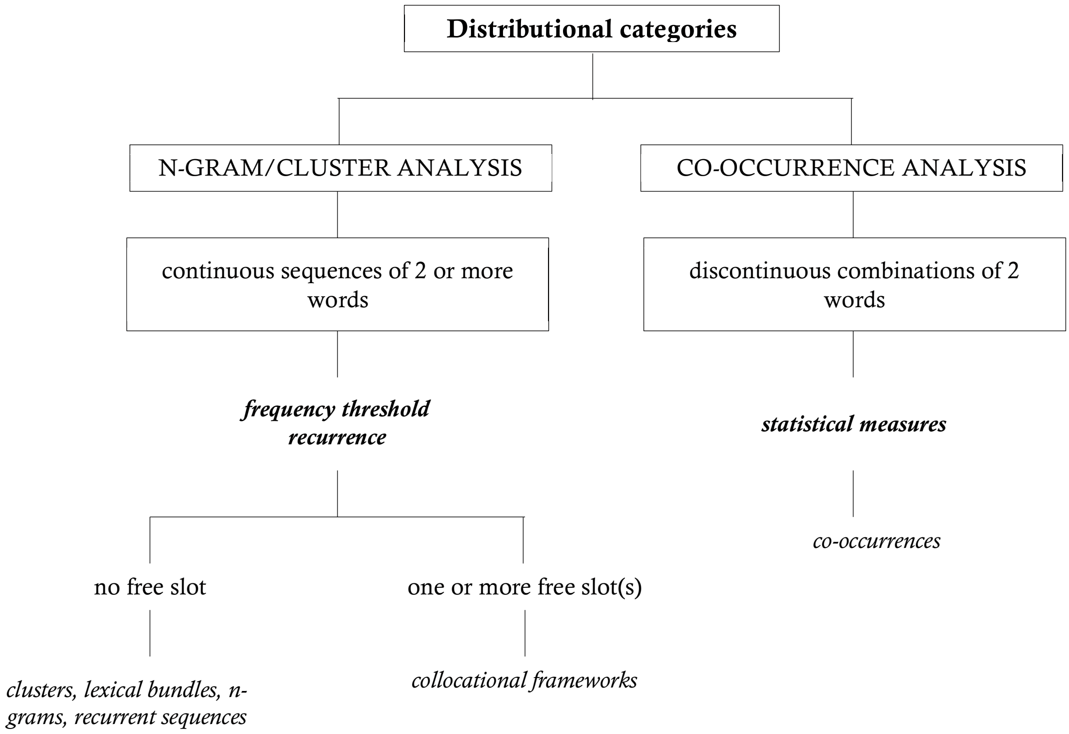
> By far the majority of text is made of the occurrence of common words in common patterns, or in slight variants of those common patterns. Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up a text. (p. 108)

These radical new findings led to the development of a new, inductive approach to phraseology, the distributional (Evert, 2004) or frequency-based (Nesselhauf, 2004) approach. Firmly rooted in Sinclair's idiom principle (see Section 1.2 above), this model considers phraseology as central instead of peripheral to language. Since it does not depend on pre-defined linguistic categories for the identification of phraseological units, this approach covers a wide range of word combinations, including those that were previously regarded as outside the bounds of phraseology, such as frames, collocational frameworks, colligations and compositional recurrent phrases (Granger & Paquot, 2008). These units were shown to be a ubiquitous feature of language, while most of the restricted units favored by the traditional approach were found to occur rarely (Biber et al., 1999).

Instead of using semantic criteria to determine what a phraseological item is, the distributional approach draws on a contextual view of meaning and explores the relationship between a word and its surrounding context, introducing such concepts as semantic preference, the "relation between a lemma or word-form and a set of semantically related words" (Stubbs, 2001, pp. 111-112) (see also Partington, 2004;

Sinclair, 1996, 1998) and semantic prosody, the "consistent aura of meaning with which a form is imbued by its collocates" (Louw, 1993, p. 157) (see also Louw, 2000). The distributional approach also embraces the lexico-grammar interface as part of phraseology, encompassing such notions as Hoey's (2005) lexical priming and Stefanowitsch and Gries' (2003) collostructional analysis (see Section 1.2 above).

Figure 1. Distributional categories (Granger & Paquot, 2008, p. 39)


Distributional categories
N-GRAM/CLUSTER ANALYSIS
CO-OCCURRENCE ANALYSIS
continuous sequences of 2 or more words
discontinuous combinations of 2 words
frequency threshold recurrence
statistical measures
co-occurrences
no free slot
one or more free slot(s)
clusters, lexical bundles, n-grams, recurrent sequences
collocational frameworks

Granger and Paquot (2008) propose a typology of the types of phraseological units obtained through the distributional method, differentiating between two main extraction procedures: co-occurrence analysis and n-gram analysis (see Figure 1).

Co-occurrence analysis focuses on the statistical associations between lexical items. Words that co-occur more frequently than expected by chance are referred to as *collocations* or *collocate* (Manning & Schütze, 1999; Sinclair, 1991; Stubbs, 2002). Other analysts use the terms *co-occurrence* and *co-occurrent* (Evert, 2004; Granger & Paquot, 2008; Schmid, 2003). Collocations reflect probabilistic events that result from repeated co-selection of words by speakers of a given language, such as the regular co-occurrence of the verb *have*, the adjectives *bad* and *recurrent* and the prepositions *about* and *in* with the noun *dream*. These strong statistical preferences are demonstrated by language corpora and are now a generally recognized aspect of vocabulary description and pedagogy (Lewis, 2000; McCarthy & O'Dell, 2005; O'Keeffe et al., 2007).

N-gram analysis refers to the extraction of frequently occurring strings of two or more words variously called *n-grams* (or more specifically, *bigrams* or *trigrams*) (Stubbs, 2007a, 2007b), *clusters* (Scott, 2006), *chains* (Stubbs, 2002; Stubbs & Barth, 2003), *recurrent sequences* (De Cock, 2003), *recurrent word combinations* (Altenberg, 1998), etc. Although this type of analysis is usually associated with continuous, uninterrupted word sequences, some n-gram researchers have also studied discontinuous language patterns. Renouf and Sinclair (1991) searched a corpus for a set of these patterns, which they termed *collocational frameworks.* Collocational frameworks are composed of fixed high-frequency function words combined with free slots filled by a variety of content words (e.g., *a* + ? +*of* , *an* + ? + *of* , *be* + ? + *to*). Biber (2009) investigated similar features using a corpus-driven method that involved identifying the most common patterns in a corpus, determining the variability and fixedness of the elements within these patterns and comparing their use in speech and writing. Other recurrent multi-word sequences that allow one or more free slots are Stubbs' *phrase-frames* (2007a, 2007b) and Cheng, Greaves and Warren's *concgrams* (2006).

The concept of *lexical bundle* (Biber et al., 1999) the terminology adopted for this study, falls under the category of n-gram analysis and is explained in depth later in this chapter.

2.3. Phraseology and lexicography

Phraseology finds numerous applications in other fields of linguistic inquiry, not least in lexicography. John Sinclair's phraseological work has had particularly lasting influence on lexicography, as, in the words of Moon (2008),

> [it] challenges the viability of the traditional model of the dictionary as an ordered listing of individual words and senses, whether defined or translated. It points instead towards a radically different model, where meanings are located through and within phraseology. This has implications or dictionary design and methodology, and more broadly for the identification of the lexicon of a language and the items populating that lexicon. (p. 243)

Many of the phraseological ideas introduced by Sinclair were implemented in the *Collins COBUILD English Language Dictionary* (Sinclair, 1987), which featured a number of corpus examples showing phraseological patterns and collocates.

The new emphasis on the inseparability of meaning and context has led lexicographers to devise new ways to document lexical phenomena beyond the orthographic word, especially in pedagogical dictionaries. This resulted in innovations such as the full-sentence definition format (Hanks, 1987), the contextual glossing of headwords, and the extended descriptions of high-frequency delexicalized words (Moon, 2008).

Apart from the changes it brought to the design of dictionary entries, the importance of phraseology in lexicography is also evidenced by the publication of a large and growing number of collocation, idiom and other types of phraseological dictionaries, targeted at both native users and learners (Benson, Benson, & Ilson, 2010; *Macmillan collocations dictionary for learners of English*, 2010; McIntosh, Francis, & Poole, 2009; Moon, 1995; Parkinson & Francis, 2006; Sinclair & Moon, 1989).

2.4. Phraseology and English language teaching

Beyond pedagogical lexicography, phraseology has so far had little direct impact on English language teaching and learning, despite mounting research evidence demonstrating the pervasiveness of formulaic patterns in spoken and written language. Granger and Meunier (2008b) discuss some of the reasons for this current state of affairs. One important factor is the need to change teacher and learner attitudes towards the study of phraseology. Giving teachers and learners the motivation to look beyond the single word in the language classroom will involve making them aware of the role of common lexical sequences in the promotion of receptive mastery and productive fluency and accuracy (Coxhead, 2008; Granger & Meunier, 2008; O'Keeffe et al., 2007). Psycholinguistic research also provides evidence that automaticity achieved through the use of formulaic language facilitates comprehension and production for learners by lightening their cognitive processing load (Girard & Sionis, 2004). In addition, a few studies point towards the positive impact of phraseological competence on social integration (Adolphs & Durow, 2004) and natural interaction within cultural communities (Prodromou, 2005).

The successful introduction of phraseological units into classrooms and language-teaching materials requires more than just convincing teachers and learners of their utility. It is also necessary to allow them fast and easy access to phraseological information, which can only become widely available through the development of better statistical measures and automatic procedures for identifying multi-word units in a variety of genres and text types, as well as the creation of user-friendly modes of delivering this data (Granger & Meunier, 2008). It is this in this regard that Granger and Meunier (2008b) stress the possibilities offered by new technologies, which provides teachers with the means to simplify the presentation of information to students while still accounting for the inherent complexities of phraseology.

The same authors consider it wrong to apply the principles of input-rich, immersion-based first-language learning to second-language learning, and recommend that classroom input on phraseology be supplemented by explicit teaching using the appropriate methodologies. They also caution against the rejection of grammar teaching in favor of phraseology, and advocate principled eclecticism, wherein various approaches are combined with teacher experience and common sense in the selection of teaching items that address the realities of the teaching and learning environment and meet learners' specific needs (Granger & Meunier, 2008).

In spite of growing interest on the topic, there is as yet very little sound, research-based advice on how to teach multi-word units of meaning, and much less on their effectiveness as teaching items (Byrd & Coxhead, 2010; Coxhead, 2008). Are lexical phrases really "an ideal unit for teaching" which "prove highly motivating" and "highly memorable for learners and easy to pick up" (Porto, 1998)? Granger and Meunier call for "more empirical evidence of the actual impact of a phraseological approach to teaching and learning" (2008b, p. 249). There is also a need for more research on which types of lexical sequences are worth teaching, and which pedagogical approaches should be adopted that can lead to greater gains in phraseological competence.

2.5. Phraseology in academic writing

Recent corpus-based phraseological research has also made a significant contribution to understanding the role of frequent multi-word combinations in characterizing registers, genres and disciplines, with several studies highlighting the importance of the fixed phrase in particular discourse communities. As Hyland (2008a) notes:

> […] words which follow each other more frequently than expected by chance, [help] to shape text meanings and [contribute] to our sense of distinctiveness in a register. Thus the presence of extended collocations like *as a result of*, *it should be noted that*, and *as can be seen* help identify a text as belonging to an academic register while *with regard to, in pursuance of,* and *in accordance with* are likely to mark out a legal text. (p. 5)

Corpus investigations of academic speech and writing have provided insight on the distinctive features of formulaic language in a variety of research fields (cf. DeCarrico & Nattinger, 1988; Hewings & Hewings, 2002; Howarth, 1996b; Oakey, 2002; Paquot, 2007; Scott & Tribble, 2006; Simpson, 2004), with some placing particular focus on scientific genres (Gledhill, 1995, 2000a, 2000b; Luzón Marco, 2000; Pecorari, 2009; Verdaguer, 2003; Williams, 1998). These studies clearly establish the functional significance of highly frequent recurrent sequences of words in disciplinary discourses. As Williams maintains, "in order to understand texts, we must look at them closely to find the lexico-grammatical strategies that they adopt to assist communication within a specialized community" (2002b, p. 60).

Multi-word expressions have proven to be essential not only to lexico-grammatical competence, but also to fluency and pragmatic competence (Cortes, 2004; Granger, 1998). As early as 1983, Pawley and Syder claimed 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'' (p. 191). This becomes particularly true in specialized contexts. Hyland (2008a) argues that frequently occurring word combinations signal participation in a given community, and links appropriate use of these combinations to communicative competence in a field of study and unfamiliarity with them to inexperience and lack of expertise. This argument is supported by studies such as those by Chen and Baker (2010), Cortes (2004), Haswell (1991), Hyland (2008b) and Nesselhauf (2005), which associate infrequent and inappropriate use of formulaic sequences to novice and learner writing. These studies also stress the need to include the explicit teaching of relevant phraseology in EAP curricula.

In light of the findings produced by the research just described, phraseological units have increasingly come to be seen as essential building blocks of coherent communication in the academe. In the following section, we will turn to one type of multi-word unit that has been the subject of several groundbreaking studies in different settings, the academic context among them: the lexical bundle.

3. Lexical bundles

Lexical bundles were first defined and explored in detail by Biber, Johansson, Leech, Conrad and Finegan in a chapter of the *Longman Grammar of Spoken and Written English* (LGSWE) (1999), their exhaustive corpus-based study of English grammar.

In this chapter, Biber and colleagues define lexical bundles as “bundles of words that show a statistical tendency to co-occur” (1999, p. 989) and as “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (1999, p. 990).

Lexical bundles are identified through empirical means, as these contiguous combinations of words are automatically extracted from a given corpus using a computer program. In the case of the LGSWE, its authors identified frequently occurring lexical sequences in the conversation and academic-prose sections of the Longman Spoken and Written English Corpus (LSWE), with each section containing around five million words.

The LGSWE chapter on lexical bundles is distinctive for relying mainly on frequency criteria for the identification of multi-word units of meaning. However, frequency cut-offs are somewhat arbitrary and depends on the scope of each study: work on lexical bundles has used cutoff ranges between ten and 40 instances per million words. The minimal cut-off set by Biber et al. (1999) was at least ten times per million words, but a lower cutoff was used for less common five- and six-word lexical bundles.

Another condition that must be satisfied for a recurring lexical sequence to qualify as a lexical bundle is dispersion, meaning that it must occur in multiple texts within a register. This criterion is important in order to avoid individual speaker/writer idiosyncrasies. Biber et al.’s (1999) lexical bundles are spread across at least five different texts in each register, but the minimum dispersion can vary across studies.

Studies on lexical bundles have found that the longer the bundle, the lower is its frequency (Hyland, 2008a; Simpson-Vlach & Ellis, 2010). In both the conversation and academic-prose sections of the LSWE, there are almost ten times as many three-word lexical bundles as four-word lexical bundles, and about ten times as many four-word lexical bundles as five-word lexical bundles. Three-word bundles occur over 80,000 times per million words in conversation and over 60,000 times per million words in academic prose, while four-word bundles occur over 8,500 times per million words in conversation and over 5,000 times per million words in academic prose (Biber et al., 1999).

Lexical bundles also include fixedness among its distinguishing characteristics. But as Cortes (2004) points out, this fixedness is a result of the frequency criteria applied during the bundle extraction process and is thus different from the fixedness that characterizes other word combinations. Only the form of the bundle that meets the cut-off frequency qualifies as a bundle, regardless of its other forms. In the present study, for example, only the bundle *are expressed as* occurs frequently enough to be considered a lexical bundle, not its singular form *is expressed as*.

Lexical bundles are also different from idioms and other invariable, non-compositional phraseological items. Many lexical bundles are not idiomatic, as their meaning is derivable from the words they contain. Consider, for example, *in the presence of, studies have shown that* and *the result of,* just some of the most frequent lexical bundles found in this study, all of which are fully compositional.

With regard to their structure, lexical bundles are, in most cases, not complete structural units, but rather parts of phrases or clauses with other fragments embedded in them. Biber et al. (1999) found that only 15% of lexical bundles in conversation and 5% in academic prose represent complete structural units, and that most lexical bundles bridge two units, that is, the last word of the bundle is often the first element of the following structure.

However, Biber et al. (1999) also observe that lexical bundles have strong structural correlates that make it possible to classify them according to several basic structural types. These grammatical correlates differ considerably depending on the register: bundles in conversation are most commonly clausal, of the type pronoun + verb + complement (e.g., *I want you to*, *it's going to be*), while in academic prose, most lexical bundles are phrasal, parts of noun phrases or prepositional phrases (e.g., *as a result of, on the other hand*) (Biber et al., 1999). These authors propose a structural classification for lexical bundles based on these typical grammatical correlates. The structural categories corresponding to academic prose are summarized in Table 1.

Table 1. Structural classification of lexical bundles in academic prose (Biber et al., 1999, pp. 1015-1024)

STRUCTURE	**EXAMPLES**
Noun phrase with *of*-phrase fragment	*the end of the, the beginning of the, the base of the, the point of view of*
Noun phrase with other post-modifier fragments	*the way in which, the relationship between the, such a way as to*
Prepositional phrase with embedded *of*-phrase fragment	*about the nature of, as a function of, as a result of the, from the point of view of*
Other prepositional phrase (fragment)	*as in the case, at the same time as, in such a way as to*
Anticipatory *it* + verb phrase/adjective phrase	*it is possible to, it may be necessary to, it can be seen, it should be noted that, it is interesting to note that*
Passive verb + prepositional phrase fragment	*is shown in figure/fig., is based on the, is to be found in*
Copula *be* + noun phrase/adjective phrase	*is one of the, may be due to, is one of the most*
(Verb phrase +) *that*-clause fragment	*has been shown that, that there is a, studies have shown that*
(Verb/adjective +) *to*-clause fragment	*are likely to be, has been shown to, to be able to*
Adverbial clause fragment	*as shown in figure/fig., as we have seen*
Pronoun/noun phrase + *be* (+…)	*this is not the, there was no significant, this did not mean that, this is not to say that*
Other expressions	*as well as the, may or may not, the presence or absence*

In addition, shorter lexical bundles are usually subsumed in longer sequences. For example, the four-word bundle *it should be noted* forms part of the five-word bundle *it should be noted that*, which is in turn incorporated into the six-word bundle *it should be noted that the*.

Some attempts have also been made to classify lexical bundles according to their function. Biber, Conrad and Cortes (2003, 2004) put forward a preliminary taxonomy that reflects the meanings and purposes of lexical bundles in text and distinguishes among three primary functions: 1) stance expressions, 2) discourse organizers and 3) referential expressions (see Table 2). They provide the following definition of each category (Biber et al., 2004):

Stance bundles express attitudes or assessments of certainty that frame some other proposition. Discourse organizers reflect relationships between prior and coming discourse. Referential bundles make direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as especially important. (p. 384)

Table 2. Functional classification of lexical bundles (Biber et al., 2004, pp. 384-388)

I. Stance expressions Express attitudes or assessments of certainty that frame some other proposition	**II. Discourse organizers** Reflect relationships between prior and coming discourse	**III. Referential bundles** Make direct reference to physical or abstract entities, or to the textual context itself	**IV. Special conversational functions**
A. Epistemic stance *I don’t know if, I think it was, are more likely to, the fact that the* B. Attitudinal/modality stance B1) Desire *if you want to, I don’t want to* B2) Obligation/directive *you might want to, it is important to* B3) Intention/prediction *I’m not going to, it’s going to be* B4) Ability *to be able to, can be used to*	A. Topic introduction/focus *what do you think, if you look at* B. Topic elaboration/ clarification *I mean you know, on the other hand*	A. Identification/focus *that’s one of the, of the things that* B. Imprecision *or something like that, and stuff like that* C. Specification of attributes C1) Quantity specification *there’s a lot of, how many of you* C2) Tangible framing attributes *the size of the, in the form of* C3) Intangible framing attributes *the nature of the, in the case of* D. Time/place/text reference D1) Place reference *in the United States* D2) Time reference *at the same time, at the time of* D3) Text deixis *shown in figure N, as shown in figure* D4) Multifunctional reference *the end of the, the beginning of the*	A. Politeness *thank you very much* B. Simple inquiry *what are you doing* C. Reporting *I said to him/her*

This initial framework became widely adopted and was later extended and modified by other authors, notably by Hyland (2008a). This author investigated the frequency, forms and functions of lexical bundles in a large corpus composed of research articles, Master’s theses and doctoral dissertations from four different disciplines. He then modified Biber et al.’s (2004) classification to create categories that better represent the lexical bundle functions he found in his corpus of research writing. The resulting taxonomy assigns each bundle to one of three broad categories of research, text and participants, which are further divided into several subcategories (see Table 3).

Table 3. Functional classification of lexical bundles in academic writing (Hyland, 2008a, pp. 13-14)

Research-oriented bundles Help writers to structure their activities and experiences of the real world	**Text-oriented bundles** Concerned with the organization of the text and its meaning as a message or argument	**Participant-oriented bundles** Focused on the writer or reader of the text
Location Indicating time/place *at the beginning of, at the same time, in the present study* **Procedure bundles** *the use of the, the role of the, the purpose of the, the operation of the* **Quantification** *the magnitude of the, a wide range of, one of the most* **Description** *the structure of the, the size of the, the surface of the* **Topic** related to the field of research *in the Hong Kong, the currency board system*	**Transition signals** Establishing additive or contrastive links between elements *on the other hand, in addition to the, in contrast to the* **Resultative signals** Mark inferential or causative relations between elements *as a result of, it was found that, these results suggest that* **Structuring signals** Text-reflexive markers which organize stretches of discourse or direct the reader elsewhere in text *in the present study, in the next section, as shown in figure* **Framing signals** Situate arguments by specifying limiting conditions *in the case of, with respect to the, on the basis of, in the presence of, with the exception of*	**Stance features** Convey the writer's attitudes and evaluations *are likely to be, may be due to, it is possible that* **Engagement features** Address readers directly *it should be noted that, as can be seen*

It is clear that lexical bundles, as "a fundamentally different kind of linguistic construct from productive grammatical constructions" (Biber et al., 2004, p. 399), have made a significant impact on research in multi-word units of meaning, and has so far been used to investigate textual organization and differences between registers, text types and native- and non-native speaker output (Römer, 2009).

3.1. Lexical bundles in academic writing

The register comparisons carried out by Biber and colleagues (1999) in their pioneering study of lexical bundles have shown the extent to which recurrent language is used, not only in conversation, but also in academic prose. Lexical bundles have proven to be pervasive in academic genres, and to have certain features particular to academic texts. For instance, Biber et al. (1999) found almost no lexical bundles representing complete structural units in the academic section of their corpus. Most of the bundles they identified in academic prose span two structural units, such as a noun phrase or beginning of a prepositional phrase. Most of these bundles therefore end in a function word, such as an article or a preposition (e.g., *the end of the, as a result of*). The few structurally complete bundles are usually prepositional phrases that function as discourse markers (e.g., *for the first time*, *in the first place*). In addition, most lexical bundles in academic prose were found to consist of nominal or prepositional elements that co-occur in highly productive frames, such as *the _ of the _*. The two empty slots in the frame can be filled by many words to make several different lexical bundles (e.g. *the size of the, the structure of the, the purpose of the, the nature of the).* Biber's (2009) investigation of the patterns represented by recurrent multi-word sequences likewise uncovered a preference in academic writing for formulaic frames with variable slots, which makes this register distinctive from conversation, where continuous fixed sequences are preferred.

This and further research on lexical bundles in academic writing have provided strong evidence of the central role of fixed phrases in this type of discourse. These studies indicate that the frequent and appropriate use of lexical bundles is an important component of fluent linguistic production in academic environments, "helping to shape meanings in specific contexts and contributing to our sense of coherence in a text" (Hyland, 2008a, p. 4).

Several corpus studies in EAP have sought to identify the most important lexical bundles in the academic setting and the extent to which they differ by genre, register and discipline. Biber (2006), for instance, found a much higher density of lexical bundles in classroom teaching in comparison to conversation and textbooks. He attributed this result to classroom talk's reliance on both oral and written genres. Other studies similarly strive to describe the phraseological features that characterize particular discourse types (Biber et al., 2004; Pickering & Byrd, 2008; Stubbs & Barth, 2003).

Hyland's (2008a) cross-disciplinary study of lexical bundles in research articles, doctoral dissertations and Master's theses found variations in their frequencies and preferred uses in the diverse fields of biology, electrical engineering, applied linguistics and business studies. His findings led him to question the notion of a core academic phrasal lexicon and call for a discipline-specific approach to the teaching of lexical bundles.

Hyland's results stand in contrast to those of Simpson-Vlach and Ellis, who used an "innovative combination of quantitative and qualitative criteria, corpus statistics and linguistic analyses, psycholinguistic processing metrics, and instructor insights" (2010, p. 4) to create an empirically derived, pedagogically useful list of formulaic sequences[3] for academic speech and writing they named the Academic Formulas List

[3] Simpson-Vlach and Ellis (2010) use the terms *formula* and *formulaic sequence* instead of *lexical bundle*, but the word combinations they include in their list are similar to lexical bundles in that they are repeated contiguous lexical sequences identified using frequency criteria.

(AFL). In building the AFL, these authors were able to identify frequently recurring word combinations that cover a wide range of academic genres.

Other studies aimed to improve our understanding of lexical bundles in academic discourse by comparing the use of bundles by writers of different first languages and levels of expertise. Cortes (2004) analyzed the forms and functions of the most frequent four-word lexical bundles in published history and biology articles, which she called *target bundles*, and examined their use in texts written by students at three different levels in the same disciplines. Her findings showed that students rarely used target bundles in their writing, and those that they used were employed in a different way than in professionally written texts.

In addition to the novice-writer effect, Chen and Baker (2010) explored the influence of non-nativeness in lexical bundle use in their comparative investigation of published academic texts and L1 and L2 student writing. They discovered a small range of lexical bundles in L2 student texts in comparison to published academic texts, as well as instances of overuse and underuse of certain expressions in both L1 and L2 student writing.

Salazar (2010) investigated the use of lexical bundles in two different varieties of English through an analysis of lexical bundles with verbs retrieved from two corpora of medical research articles: one with texts from a Philippine English-language journal and another from the *British Medical Journal*. Her quantitative results showed a lower amount of verbal lexical bundles in the Philippine corpus compared to the British, while her qualitative findings uncovered certain structural and functional differences between the bundles used in the two corpora.

3.2. Lexical bundles and EAP pedagogy

Research on lexical bundles generally agrees on the pedagogical value of recurrent word combinations, and many studies endeavor not only to shed light on the theoretical status of lexical bundles, but also to discuss specific suggestions for teaching. Simpson-Vlach and Ellis' (2010) work on the AFL, for instance, was carried out with a view to facilitate the inclusion of formulas into EAP curricula.

Descriptive and comparative studies such as those of Hyland (2008a) and Cortes (2004) conclude with the pedagogical implications of their findings, where they advocate the design and implementation of consciousness-raising tasks and productive exercises that can encourage learners to notice multi-word units in their reading and introduce these units into their writing. Cortes (2006) even took her investigation directly to the classroom when she planned and taught a series of micro-lessons on lexical bundles to a group of university students in a writing-intensive history class, then conducted pre- and post-instruction analyses on the students' class assignments. The students' limited gains in lexical bundle use even after the micro-lessons led the author to suggest the need for longer and better exposure to lexical bundles in a corpus-enhanced disciplinary writing course. Neely and Cortes (2009) focused their attention to the use of a small set of lexical bundles in academic lectures, on which they based the design of a series of academic listening lesson plans.

Byrd and Coxhead (2010) built their own list of 21 four-word lexical bundles used in arts, commerce, law, and science through the analysis of a corpus of academic writing and the comparison of their results to published results of similar data. Through this investigation, they were able to identify six key challenges in taking lexical bundle data into the EAP classroom. First among these issues is how lists of lexical bundles found in research reports can be used as a basis for selecting multi-word units for teaching and learning. Another difficulty is determining the length of lexical bundle to teach, in those cases where bundles form part of longer ones. Additional challenges include the inadequate contextual information that current lists of lexical bundles provide, and the lack of face validity of these items for EAP students. Finally, the authors comment on the challenge of teaching lexical bundles in spite of the contradiction between an analytical teaching approach and the use of bundles as unanalyzed chunks, and students' limited exposure to authentic examples of lexical bundles in use, given the logistic constraints of the EAP classroom. These challenges and the possible solutions for them will be elaborated on in Chapter VII.

The literature outlined in this section leaves little doubt that frequently recurring lexical sequences are a prevalent feature of academic language, and that their mastery is crucial to fluent and idiomatic production. The research summarized here provides justification for investigating these sequences, operationally defined as lexical bundles, with a view to creating a list of bundles that can be used to guide principles and decisions for EAP pedagogy.

Chapter III

Methodology

1. Rationale for the lexical bundle approach

In the previous chapter, we have taken a detailed look at lexical bundles, which are defined as fixed and largely compositional sequences of words that are identified using frequency and dispersion measures and classifiable by their structural and functional correlates.

The aim of the present study is to create a list of pedagogically useful multi-word units in scientific writing and compare their use in native and non-native texts. The lexical bundle approach was chosen for this purpose for a variety of reasons. Primary among them is the fact that lexical-bundle identification is an objective, straightforward means of extracting multi-word units with a certain level of fixedness. Lexical bundles also offer the advantage of being empirically derived, as they are identified on the basis of frequency criteria. The pedagogical value of this approach is based on the widely held assumption that the most frequent vocabulary items are of the highest currency and usefulness and are therefore deserving of attention, especially in vocabulary teaching (Nation, 2001).

The process of retrieving lexical bundles can also bring to light word combinations that cannot be noticed by introspection or intuition alone, thus providing a new perspective on formulaic language. In the words of Conrad and Biber (2004), it shows

> […] whether there are multi-word sequences that are used with high frequency in texts, whether different registers tend to use different sets of these sequences, and if so, to what extent the bundles fulfill discourse functions and thus play an important part in the communicative repertoire of speakers and writers. (p. 58)

Lexical bundles can provide insight into the characteristic phraseology of specific contexts of language use, such as the scientific research genre with which this study is concerned. As Scott and Tribble (2006) assert, although fixed distributional phraseological units automatically exclude such features as widely spaced collocational items, they are still useful for understanding expert texts and how they are produced, and how the output of apprentice and/or non-native language users might compare to that of expert and/or native users. These authors underline the potential of these items "to enhance our appreciation (and that of learners) of what works in particular kinds of text, and what has a better chance of being accepted by experienced readers in a specific field" (Scott & Tribble, 2006, p. 132). Moreover, the fact that lexical bundles present identifiable structural characteristics and textual functions, as demonstrated by a number of exploratory studies (Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2003; Biber, Conrad, & Cortes, 2004; Hyland, 2008), makes them a good starting point for exploring phraseological differences between registers, genres, disciplines and writer groups (Römer, 2009).

2. Corpus of published scientific writing

The corpus on which this study is principally based is a two million-word sample from the Health Science Corpus (HSC). The HSC consists of close to four million words of published research writing in English in the health sciences. The corpus was collected by the University of Barcelona's SciE-Lex research team to be used for the lexico-grammatical and phraseological analyses that resulted in the SciE-Lex Electronic Combinatorial Dictionary.

The HSC is a collection of scientific research articles taken from leading journals in the fields of biology, biochemistry, biomedicine and medicine. The corpus is composed of 718 research articles published in English in the years 1998 to 1999, which are attributed to authors with English first and last names and to those affiliated to universities in native English-speaking countries. Although it cannot be definitely ascertained whether these articles were written by native speakers, prior to their publication these papers underwent a rigorous peer-review and editing process to ensure that they conformed to the standards and style of a scholarly journal. They can thus be considered representative of accepted, legitimated and institutionalized research writing in the health sciences, and ideal writing models for any scientist wishing to publish in English.

Research papers in the health sciences were chosen for this corpus primarily because of the rhetorical structure of research writing in this domain. Publications in biology, medicine and other related disciplines generally have the hour-glass macro structure described by Swales (1990), in which papers begin with an overview of the subject matter, then narrow down on a particular research question that is later answered by a specific experiment, and finally broaden out again to relate the results of the experiment to a wider field. This rhetorical structure is what is usually considered typical of scientific reports, especially of experimental research (Tarone, Dwyer, Gillette, & Icke, 1998). The results from a health-science corpus can therefore be extended to a large number of other scientific fields that lend themselves to experimentation.

The HSC articles were downloaded from the online versions of the selected journals and converted into plain text files. To ensure smooth and accurate data processing, the files were cleaned of headers, footers, diagrams, images, captions and references, as well as anomalous capitalizations, paragraph breaks and columnar layouts.

For this particular study, a sample of the HSC amounting to roughly two million words was used. To maintain a high level of structural uniformity, only those articles from journals that strictly follow the general scientific format of abstract, introduction, materials and methods, results and discussion were included in the sample. Table 4 presents a summary of the journals and articles in the corpus sample and their respective word counts.

Table 4. Corpus of published scientific writing (Health Science Corpus sample)

JOURNAL TITLE	SPECIALIZATION	NUMBER OF TEXTS	MEAN LENGTH OF TEXTS IN WORDS	TOTAL NUMBER OF WORDS
Biochemical Journal	Biochemistry, cell and molecular biology	53	5,829	308,937
EMBO Journal	Molecular biology	40	11,223	448,933
Genes and Development	Molecular biology, molecular genetics, cell biology and development	64	4,720	302,126
Genetics	Heredity, genetics, biochemistry, molecular biology	54	7,149	386,068
Journal of Cell Biology	Cell biology	26	8,391	218,184
Journal of Clinical Investigation	Biomedicine	53	7,889	418,161
TOTAL NUMBER OF WORDS IN CORPUS				**2,082,409**

3. Creating and analyzing the list of target lexical bundles

3.1. Lexical bundle identification

The first step of the analysis was to create a list of the most frequent and pedagogically useful lexical bundles in the published scientific corpus. These bundles are referred to in this study as *target bundles,* following Cortes (2004).

In accordance with Biber et al., (1999), a lexical bundle is defined in the present study as a frequently recurring sequence of words. Two-word sequences were excluded here, since they are too numerous and usually represent recurrent collocations. Included in the data set are highly frequent three-word bundles, whose pedagogical importance Simpson-Vlach and Ellis (2010) clearly showed in their own study of academic formulas. These three-word strings, together with four-word bundles and comparatively rare five- and six-word sequences were all considered for a more complete list.

Lexical bundles were identified using orthographic word units, and only word strings uninterrupted by punctuation marks were included. In addition, to qualify as a recurrent lexical bundle, lexical sequences must occur at least ten times per million words.

Another important metric used to create the list of target bundles is the mutual information (MI) score. MI is a measure of the strength of association between words, as it "compares the probability of observing *x* and *y* together (the joint probability) with the probabilities of observing *x* and *y* independently (chance). If there is a genuine association between x and y, the joint probability […] will be much larger than chance" (Church & Hanks, 1990, p. 23).

A higher MI score means a stronger association and thus a more coherent and interesting relationship between words. This additional metric was applied in order to weed out those bundles that do not have identifiable meanings or functions but occur often because of the high frequency of the words that they contain. It was also used to avoid discounting useful but less frequent phrases that tend to end up at the bottom of frequency-ordered lists (Simpson-Vlach & Ellis, 2010). As the frequency measure confirms the utility of certain lexical bundles, the MI statistic ensures greater coherence that correlates with distinctive function and meaning. Frequency and MI therefore combine to make a more reliable metric for producing a list of bundles for pedagogical applications (Simpson-Vlach & Ellis, 2010).

The computer program *Collocate* (Barlow, 2004) was used to produce a list of three-, four-, five-and six-word bundles that occur at least ten times per million words in the HSC sample, filtered by MI score. The list of 1,732 lexical bundles generated by the program was then saved and ranked by frequency and MI score.

3.2. Exclusion criteria

The quantitative and statistical measures described above provided a reliable, straightforward method for creating a manageable master list of lexical bundles. However, it was evident that not everything in this long list of recurrent word sequences was of pedagogical relevance, and that further sifting was needed in order to produce a more refined set of lexical bundles for teaching. The fact that all the word combinations in the master list meet certain frequency and coherence criteria does not necessarily mean that they will all be of equal benefit to language teachers or learners, or that they all fall within the scope of this study.

Thus, to further narrow down the frequency- and MI-based master list, certain exclusion criteria were established to eliminate those lexical bundles that could not be included due to some of their characteristics. Such an intuitive selection process can be considered "methodologically tricky and open to claims of subjectivity" as Simpson-Vlach and Ellis (2010, p. 4) point out, but this additional step was found to be necessary for the study to achieve its primary objective of creating a list of only the most pedagogically useful bundles in scientific writing.

It is important to stress at this point that applying these exclusion criteria was more of a methodological and pedagogical decision than a theoretical one. Excluding a number of word sequences from the final list does not imply that they are not, in fact, lexical bundles. They undeniably are, because they fit the operational definition of *lexical bundle* described in detail above. Their exclusion only serves to limit the scope of this study, whose aim is not to make an exhaustive list of the most frequently recurring word sequences in a particular genre, but to make a lexical-bundle list that is clear, organized, comparable, and most importantly, manageable for someone wishing to present these bundles in a classroom, a textbook or a pedagogical dictionary.

Table 5 presents these exclusion criteria along with some examples. The exclusion analysis is explained at length in the following chapter.

Table 5. Exclusion criteria

Fragments of other bundles	*on the basis, in the case, by the addition*
Bundles ending in articles	*consistent with the, results in a, indicated by an*
Topic-specific bundles	*amino acid residues, the crystal structure, decapping in vivo*
Bundles composed exclusively of function words	*have also been, but did not, there was no*
Bundles with random numbers	*at least one, of the two, for the first*
Time bundles	*for 30 min, for 1 h, 15 min at*
Temperature, volume and length bundles	*min at 30 8c, 1 ml of, in 20 mm*
Random section titles	*fig 1 a, figure 4 a, table 1 in*
Meaningless bundles	*are means s e m, presence of 0, h at room*
Web noise	*response to this, of this article, has been cited by*

3.3. Structural classification

The next stage of the analysis of the target bundles found in the published scientific corpus was to explore their structural characteristics. Biber et al., (1999) showed that lexical bundles have strong grammatical correlates and created a classification that group them into several basic structural types. A section of this framework corresponds to the most common structural patterns of lexical bundles in academic prose. This categorization, summarized in Table 6, was adopted in the present study to sort the target bundles according to their grammatical structure. Five new categories were added: other noun phrases, other adjectival phrases, verb phrases with personal pronoun *we*, other passive fragments and other verbal fragments. The bundles were assigned to different categories after they had been examined in context using the concordance program *Antconc* (Anthony, 2006).

Table 6. Structural patterns more widely used in academic prose (adapted from Biber et al., 1999, pp. 1015-1024)

Noun phrase with *of*-phrase fragment	*a variety of, the association of, the total number of*
Noun phrase with other post-modifier fragment	*no effect on, a role in, the difference in*
Other noun phrase	*lines of evidence, the present study*
Prepositional phrase + *of*	*in the presence of, as a consequence of*
Other prepositional phrase (fragment)	*in addition to, as a result, with respect to*
Passive + prepositional phrase fragment	*are shown in, was associated with*
Other passive fragment	*has been reported, similar results were obtained*
Anticipatory *it* + verb or adjectival phrase	*it is likely that, it has been proposed that*
Copula *be* + adjective phrase	*is consistent with, are representative of*
(Verb phrase or noun phrase) + *that*-clause fragment	*this suggests that, the possibility that*
(Verb or adjective) + *to*-clause fragment	*shown to be, is likely to, to account for*
Adverbial-clause fragment	*as described previously, as seen in*
Verb phrase with personal pronoun *we*	*we found that, we were unable to*
Other verbal fragment	*for review see, does not require*
Other adjectival phrase	*similar to that, not due to*
Other expression	*in order to, as well as*

3.4. Functional classification

The next step in the analysis of target bundles was to categorize them in terms of their primary discourse-pragmatic functions. Hyland's (2008a) classification scheme was found to be particularly useful for the present study, as it is adapted to the specific concerns of research-focused written genres (see Chapter II, Section 3 above). However, this framework was treated only as a starting point, as it was necessary to make some changes to the categories in order to more accurately reflect the functions performed by the lexical bundles in the HSC.

Hyland's (2008a) three broad groupings were maintained, but the subcategories were modified and added to. The research-oriented subcategories of *location*, *procedure*, *quantification* and *description* were preserved, but the topic subcategory was eliminated, given that topic-specific bundles had been previously disregarded. In its place is a new category called *grouping*, which includes bundles related to the grouping, categorization, classification and ordering of research elements.

The text-oriented subcategories underwent a number of changes. Hyland's (2008a) *contrastive* and *resultative* functions were substituted by the narrower subcategories *additive* and *comparative,* and *inferential* and *causative*, respectively. This is to show more clearly the differences between the four functions that Hyland had previously collapsed into two categories. *Structuring* and *framing* were retained, and three new subcategories were added: *citation*, for bundles used to cite sources and supporting data; *generalization*, for bundles that signal generally accepted facts or statements; and *objectives,* for bundles that introduce writer aims.

Finally, in the participant-oriented category, the only change made was the addition of the *acknowledgment* subcategory for bundles that serve to recognize people or institutions that have participated in or contributed to the study being described.

Table 7 lists the functional categories in this modified taxonomy, along with definitions and examples.

Table 7. Functional taxonomy of target bundles (adapted from Hyland, 2008a, pp. 13-14)

Research-oriented bundles Help writers to structure their activities and experiences of the real world	**Text-oriented bundles** Concerned with the organization of the text and its meaning as a message or argument	**Participant-oriented bundles** Focused on the writer or reader of the text
Location Indicate place, extremity and direction *at the site, the tip of, on the left*	**Additive** Establish additive links between elements *on the other hand, in addition to, in concert with*	**Stance** Convey the writer's attitudes and evaluations *is likely to, is necessary for, it is possible that, it is clear*
Procedure Indicate events, actions and methods *the onset of, was carried out, used to identify*	**Comparative** Compare and contrast different elements *as compared with, in contrast to, significantly different from*	**Engagement** Address readers directly *it should be noted that, see figure 1, as seen in*
Quantification Indicate measures, quantities, proportions and changes thereof *total volume of, a large number of, the ratio of, a decrease in*	**Inferential** Signal inferences and conclusions drawn from data *found to be, these results suggest that, we conclude that*	**Acknowledgment** Recognize people or institutions that have participated in or contributed to the study *a gift from, kindly provided by*
Description	**Causative**	

Indicate quality, degree and existence *the appearance of, the extent of, the presence of* **Grouping** Indicate groups, categories, parts and order *a wide range of, this type of, the sequence of, a portion of*	Mark cause and effect relations between elements *as a result of, is caused by, by virtue of* **Structuring** Text-reflexive markers that organize stretches of discourse or direct the reader elsewhere in text *as described previously, as shown in figure, in the materials and methods section* **Framing** Situate arguments by specifying limiting conditions *in the case of, with respect to, on the basis of, in the presence of, with the exception of* **Citation** Cite sources and supporting data *it has been proposed that, as reported previously, studies have shown that* **Generalization** Signal generally accepted facts or statements *little is known about, is thought to be* **Objective** Introduce the writer's aims *we asked whether, to show that, in order to*	

A concordance program was again used to analyze the target bundles in their corresponding contexts and determine the specific functions they perform. However, an initial attempt to apply the classification to this corpus revealed a significant number of lexical bundles with multiple functions. It soon became obvious that in order to provide a more accurate, detailed picture of the functions of lexical bundles in scientific texts, it was necessary to implement an alternative approach that took the multifunctionality of bundles into account. Such an approach inevitably involved analyzing all instances of every target bundle on the list in its context of use, so that the corresponding discourse functions could be assigned to it. This provides even further justification for narrowing the scope of the study and creating a more concise list of lexical bundles.

The multifunctionality of lexical bundles is covered in depth in Chapter V.

3.5. Keyword and prototype analysis

Initial qualitative analyses of the list of target bundles uncovered a number of relationships between these word combinations. One main observation is that shorter bundles are often incorporated into longer lexical bundles, which is consistent with the findings of other lexical-bundle researchers (Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2003; Biber, Conrad, & Cortes, 2004; Hyland, 2008). For instance, the three-word bundle *the presence of* is part of the four-word bundle *in the presence of,* while the three-word bundle *as described in* is a fragment of the six-word bundle *as described in materials and methods.*

A range of semantic and structural relationships was also detected between the lexical bundles. There are bundles that share the same keyword, but have singular and plural, positive and negative, active and passive and past and present forms, as well as varying subjects, adjectives, prepositions and degrees of certainty.

Table 8 summarizes these semantic links and provides examples.

Table 8. Relationships between lexical bundles

Singular and plural forms	*is found in, are found in / was present in, were present in /* *the difference in, the differences in*
Past and present forms	*appear to be, appeared to be/ is based on, was based on /* *we find that, we found that*
Positive and negative forms	*it is clear, it is not clear / was detected in, was not detected /* *is due to, not due to*
Active and passive forms	*we propose that, it has been proposed that / studies have shown that, we show that, it has been shown that*
Different prepositions/conjunctions	*in contrast to, in contrast with / as described in, as described above / to determine whether, to determine if*
Different verbs	*shown in table, summarized in table /* *shown in figure, described in figure*
Different subjects	*results indicate that, data indicate that /* *this indicates that, results indicate that*
Different adjectives	*the level of, high levels of, low levels of /* *a role in, an important role in*
Different degrees of certainty	*is due to, may be due /* *it is likely that, it seems likely that*

To address these variations and semantic relationships and facilitate the functional classification, the remaining bundles on the list were grouped by keyword, with each group headed by a *prototype* of the bundle (Sinclair, Jones, & Daley, 2004). In this study, the status of prototypical bundle is usually designated to the most frequently occurring form of a bundle.

At this stage of the analysis, frequency and MI score become of secondary importance. Careful analysis of the semantic relationships between lexical bundles was carried out in order to determine which bundles are prototypical and which are just components or variations of a prototype. After an examination of concordance lines for each lexical bundle, it was decided that lexical bundles with distinct meanings, functions and lexico-grammatical preferences were to be regarded as separate prototypical bundles, while the rest were to be considered variations of these prototypes.

As for lexical bundles that form part of other bundles, those that have the same frequency as a prototypical bundle where eliminated, while the rest were treated as variations and grouped with the corresponding prototype. For example, *absence or presence of* occurs 60 times in the corpus, exactly the same frequency as the complete bundle *in the absence or presence of,* meaning that the two bundles pertain to the same instances of the same sequence. *Absence or presence of* was therefore considered a fragment of a longer bundle and was deleted.

The following chapter contains a discussion of the results of this part of the analysis.

4. Comparison with the non-native corpus

The final phase of the study involved comparing the use of lexical bundles in published scientific writing to their use in non-native writing.

4.1. Corpus of non-native scientific writing

The non-native corpus used in this study is composed of 43 biology articles that together make a total of 120,718 words.

Finding the right non-native texts for comparison with the Health Science Corpus was a main priority at the beginning of this study. Since one of the research goals of this investigation was to identify non-native scientists' deviant uses of lexical bundles in the papers they write for publication so that these particular difficulties could be addressed, it was considered essential to control for topic, text type and author profile when choosing texts for the non-native corpus. It was decided that the non-native corpus, like the HSC sample, should include research articles in the health sciences following the abstract-introduction-materials and methods-results and discussion format, written by scientists with ample knowledge of the discipline. When these criteria are applied, it is more likely that the dissimilarities between the corpora are due to linguistic factors and not to differences in subject matter, register, genre or scientific competence.

The articles that comprise the non-native corpus were kindly provided by Prof. Iliana Martínez of the National University of Río Cuarto (UNRC) in Argentina. The articles are part of a corpus of biology manuscripts that Prof. Martínez is currently compiling. These original, uncorrected manuscripts were written in English by native Spanish-speaking researchers of the UNRC and submitted to Prof. Martínez for revision, so they could later be submitted to a journal for publication.

The articles included in this corpus have all been accepted for publication after revisions, and their authors are experienced researchers with numerous publications in reputable English-language journals. However, despite being skilled biologists capable of reading highly technical and specialized literature in their field, these authors' writing difficulties are evidenced by the many language revisions journal editors demand of their submitted work (Martínez, 2005). Given the language and knowledge profile of these non-native scientists, it can be said that any differences found between their written reports and those in the HSC can be attributed to a gap in linguistic awareness rather than a lack of scientific knowledge (Martínez, 2005).

As with the texts in the published scientific corpus, the manuscripts in the non-native corpus were processed as plain text files and cleared of all unnecessary textual and formatting elements as described above.

4.2. Analysis of non-native scientific writing

In her comparative analysis of the use of lexical bundles in published and student disciplinary writing, Cortes (2004) took a more qualitative approach, treating the lexical bundles she found in published texts as *target bundles* to be searched for in her smaller corpus of student texts. The same strategy was adopted in the present study: the target bundles found through the analysis of the HSC were identified in the corpus of article manuscripts written by native Spanish-speaking scientists, and their frequencies, structures and functions were recorded and compared to the HSC results using relative frequencies per 100,000 words. Cases of overuse and underuse were identified through the results of log-likelihood tests, calculated using the UCREL log-likelihood calculator (http://ucrel.lancs.ac.uk/llwizard.html). Examples were also studied in context to determine qualitative differences between native and non-native use of lexical bundles in scientific writing.

5. Concluding remarks

This methodological section elaborated on how a combination of frequency criteria and statistical measures were used to extract a pedagogically oriented list of target lexical bundles from a multimillion-word corpus of native scientific writing. It also described the structural and functional classification of the target bundles, and explained the quantitative and qualitative comparisons made between target-bundle occurrences in the native corpus and those in a smaller but similar corpus of non-native scientific articles. The following chapter will explain in greater detail the most important methodological issues addressed briefly in this chapter, issues that were involved in the creation, filtering and organization of the final list of target lexical bundles.

Chapter IV

Creating and organizing the list of target bundles

This chapter provides an in-depth discussion of the steps taken to generate, refine and organize the list of target bundles in the Health Science Corpus.

1. Extracting lexical bundles from the HSC: Frequency and MI score

Lexical bundles as originally conceived by Biber et al. (1999) are based solely on frequency criteria. The approach is grounded in the view of frequency as evidence of the typical combinations and central meanings of words in particular contexts (Hunston, 2006), and it has indeed been useful in analyzing and describing the structure and functions of fixed lexical sequences in different registers and genres (see Chapter II, Section 3 above).

However, authors such as Simpson-Vlach and Ellis (2010) have recently recognized two inherent weaknesses of a purely frequency-based method of identifying multi-word units of meaning: first, the fact that frequency of occurrence alone does not always ensure semantic or functional coherence; and second, frequency's tendency to favor lexical sequences that occur often because of their highly frequent individual components, which are usually function words. This led them to propose the Mutual Information (MI) score as an additional metric for formula identification. The MI score compares the frequency of a multi-word unit to the overall frequencies of each of its component words, thereby reflecting the likelihood that the two words occur together for a reason and not just by random chance (Church & Hanks, 1990;

Manning & Schütze, 1999; Oakes, 1998). It is a statistical measure of association that has been used by a number of word co-occurrence studies to gauge the collocational strength of word pairs. In recent years, it has also been applied to multi-word combinations in studies such as those by Ellis, Simpson-Vlach and Maynard (2008) and Simpson-Vlach and Ellis (2010). Its use for this purpose is facilitated by software such as *Collocate* (Barlow, 2004), which automatically computes MI scores for longer sequences.

After applying the MI statistic to their spoken and written academic corpora, Simpson-Vlach and Ellis (2010) found that high MI scores tend to correspond to distinctive function and meaning, as the measure highlights functional formulas such as *does that make sense* and *you know what I mean* (in their spoken corpus) and *due to the fact that* and *there are a number of* (in their written corpus), while relegating to the bottom generally non-functional phrases such as *the um the* and *okay and the* (in their spoken corpus) and *to be of*, *as to the* and *of each of* (in their written corpus) (p. 8). In the same study, these authors performed a correlation analysis of frequency and MI score with teacher insights on the formulaicity, functionality and teaching worth of a selected sample of formulas from their data. The results of the analysis suggest that, compared to raw frequency, the MI score is a better determinant of which sequences instructors judge "worthy of teaching, as a bona fide phrase or expression" (Simpson-Vlach & Ellis, 2010, p. 10).

Prior to these encouraging findings in favor of the MI score, Biber (2009) expressed some concerns regarding its use as a test of formulaic status for sequences longer than two words. One of these is that the MI score does not take into account the order of the words in the string. This may be of no consequence to the word pairs for which it was initially used, but according to Biber, it may be problematic for multi-word sequences whose formulaicity is partly determined by their fixed word order.

Biber (2009) also considered an important issue the way the MI score privileges relatively less frequent combinations of content words, while disfavoring sequences with high-frequency words, particularly grammatical elements. In his opinion, this proves the point that the MI approach and the frequency approach bring to light two different kinds of associations, which he describes in the following manner (italics mine):

> […] multi-word sequences with high MI scores tend to be technical referring expressions (usually extended noun phrases) composed of lexical/content words; these can be regarded as *multi-word collocations*. In contrast, the most frequent word sequences (lexical bundles) usually incorporate both function words and lexical words; these can be regarded as *multi-word formulaic sequences*. (p. 289)

The explorations carried out by Biber (2009) and Simpson-Vlach and Ellis (2010) clearly show the different advantages and disadvantages of using frequency and MI score for lexical bundles extraction. After taking their results into consideration, it was decided to combine both metrics in the present study in order to capture both types of associations identified by Biber.

As mentioned in the previous chapter, three-, four-, five- and six-word lexical bundles were extracted from the two million-word Health Science Corpus (HSC) using *Collocate* (Barlow, 2004). The program's full extract command was used to process the whole corpus and produce a list of n-grams with the span and the statistical filter set by the user, which in this case was the MI score, with a default minimum of 0.5. Of the 8,457 lexical bundles identified by *Collocate*, 1,737 met the previously established frequency cut-off of ten instances per million words. These candidate bundles were then ranked, first by their individual frequencies, then by their MI scores.

An initial inspection of the computer-generated list indicated that the combined metrics were able to strike a satisfactory balance between the opposing tendencies of frequency and MI score. Of the 1,737 lexical bundles on the list, only 72 or 4% are technical terms composed entirely of lexical words, meaning that this type of sequences were not unduly prioritized as Biber (2009) predicted. On the other hand, the 82 bundles consisting exclusively of function words constitute only 5% of the list, suggesting that these items, which usually have no pedagogically compelling meaning or function, were appropriately pushed to the bottom of the list as Simpson-Vlach and Ellis (2010) also found. Finally, no negative effects to the list were observed as a result of the MI score's disregard of word order.

It had to be acknowledged, however, that the automatically created list was still too long to be manageably analyzed for structure and function, much less to be meaningful to teachers or lexicographers. It was thus treated only as the basis for further refinement.

Please refer to Appendix 1 for the complete list of bundles extracted using *Collocate*, ranked by frequency and MI score.

2. Applying the criteria for exclusion

In order to narrow down the list of lexical bundles to be included in the dictionary, the SciE-Lex team devised a number of exclusion criteria, taking into account the pedagogical objectives of the dictionary and the collocational information it contains

(Verdaguer et al., 2009). The same principle was adopted in the present study, where a set of exclusion criteria was established to further refine the original list automatically generated by the *Collocate* program and limit the number of target bundles to be investigated.

It is worth repeating here that the exclusion of certain target bundles on the basis of these criteria was a methodological and pedagogical decision taken in consideration of the scope and aims of the present study. Although some categories such as random section titles (*fig 1 c, figure 4 a*) and meaningless sequences (*mg ml in, containing 0 5)* can be considered noise and be readily deleted, there were other eliminated bundles that could be interesting for studies of a different nature, but were found to contribute little to the effectiveness of the present list as a pedagogical tool. It should be noted, however, that their elimination does not take away from their status as a lexical bundle in general, since they possess the characteristics of lexical bundles as described in the literature.

The following lines describe the exclusion criteria in further detail.

Fragments of other bundles. Biber et al. (1999) observe that a number of common lexical bundles can be extended to form longer sequences, and the same observation can be made about the present list of target bundles. Here, however, lexical bundles that are incorporated into longer bundles and have a similar frequency as these bundles were excluded. Cases like these were eliminated to avoid unnecessary repetition and make the list as brief and concise as possible. Consider, for example, the three-word bundle *is likely that*, which is part of the four-word bundle *it is likely that*. Both bundles occur 66 times in the HSC, meaning that in all instances, *is likely that* occurs as a fragment of *it is likely that*. Similarly, the three-word bundle *by the* *addition* appears 85 times, only one occurrence more than the related four-word bundle *by the addition of. It is likely* and *by the addition* were therefore disregarded.

In contrast to these examples, the three-word bundle *are consistent with* was preserved, even though it clearly overlaps with the longer bundles *results are consistent with* and *these results are consistent with*. This is because *are consistent with* occurs 93 times, much more frequently than the four- and five-word bundles of which it forms part (which occur 28 and 21 times, respectively). A look at the concordance lines revealed that *are consistent with* collocates with several other nouns apart from *results*, including *data, findings, observations* and *studies.* Additionally, apart from the demonstrative *these,* these nouns can also co-occur with the possessive pronoun *our*. Other overlapping bundles such as *consistent with this, consistent with previous* and *consistent with our* were maintained, since they provide additional information about this particular group of bundles that the others do not. Closer inspection of the concordances showed that all these related bundles can be strung together in different ways, with *are consistent with* as the central, invariable fragment:

> [these, our] [results, data, findings, observations, studies] *are consistent with* [this, our, the] (previous) [data, idea, hypothesis, observations, notion, reports, results, studies, work]

All other shorter bundles that do not provide such information and are merely fragments of longer bundles were disregarded in the study.

There is also the case of bundles such as *the presence of*, which forms part of the longer bundle *in the presence of* (1), but can also function as an independent bundle (2) (3):

(1) *In the presence of* CoA and ATP, incorporation of [3H] myristic acid into mature GIPL species (iM2, iM3, iM4) occurred in the same fractions that contained highest DPMS activity (Figure 4A). [45]

(2) *The presence of* multiple forms of Upd in the untreated cells most likely reflects partially glycosylated intermediates. [37]

(3) Primers 2 and 9 amplified a 498 bp fragment of wild-type DNA and did not amplify either mutant allele due to *the presence of* a Mu transposon between the primer-binding sites. [29]

This type of subsumed bundle was also maintained on the list.

Bundles ending in articles. Lexical bundles ending in the articles *a, an* and *the* were discarded after it was found that most of them were already part of shorter bundles, as in the case of *in the presence of a* and *as described in the*. Similar to other bundle fragments, they do not provide any additional information that makes them worth including in the list of target bundles. Since they are also very numerous, amounting to 483 items or 28% of the list, it was decided that the detail that will be maintained with their inclusion is less important than the brevity and clarity that will be gained from their exclusion.

Topic-specific bundles. Lexical bundles such as *a conformational change, cells were transfected with* and *the x chromosome* are beyond the scope of the present study, given its goal to find word combinations that occur across a range of subjects and disciplines in the health sciences and similar scientific fields, not just in the specific topics of the papers that were selected for the HSC. Moreover, understanding domain-specific vocabulary requires a certain degree of scientific knowledge, and teaching them is usually the role of specialists in the field, not of language teachers (Nation, 2001).

Topic-specific bundles were labeled as such when they have one or both of the following characteristics: 1) they appear in a limited number of articles and/or only in a specific journal; and 2) their keyword is found as a headword in the second edition of the *Oxford Dictionary of Biochemistry and Molecular Biology* (Cammack, 2006). While bundles like *amino acid residues* and *the crystal structure* are clearly technical, others such as *ability to bind* and *a final concentration of* were not. However, a check of the corresponding concordance lines showed that the latter examples and other similar bundles are used largely in their terminological sense in the corpus.

The final categories of deleted lexical bundles are examples of sequences that made it to the list because of the high frequency of their component words, not because they hold particularly interesting meanings or functions. They are exactly the kind of bundles favored by frequency-based ordering that was intended to be kept to a minimum by the use of the MI score.

Bundles composed exclusively of function words. These are repetitive series of function words such as *to that of, may not be* and *we have not.*

Bundles with random numbers. These bundles are usually composed of prepositions and random cardinal and ordinal numbers like *in the two, of the first* and *at least three.*

Random section titles. These consist of the words *figure*, *fig* and *table* and a series of random numbers and letters, such as *figure 2 a, fig 1 c* and *table 1 in.*

Bundles that express time. These are made up of prepositions, cardinal numbers and the time abbreviations *min, h* and *hr*, like *for 30 min, 4 h in* and *for 1 hr.*

Bundles that express temperature, volume and length. These comprise prepositions, cardinal numbers and abbreviations of various measurement units, like *min at 30 8c*, *1 ml of* and *in 50 mm.*

Meaningless bundles. These bundles, with examples such as *1 2 and, are means s e m* and *mm tris hcl*, are completely devoid of any identifiable meaning.

Web noise. The bundles *this article has, to this article, response to this, of this article* and *has been cited by* were found to be part of website links that were originally in the downloaded corpus articles. These managed to escape the cleaning of the corpus text files and had to be manually deleted from the list.

With the application of the above exclusion criteria, the original list of 1,732 was narrowed down to a more manageable size of 769 lexical bundles. Of over a thousand bundles, these 769 items are the ones that best serve the purpose of this study. They have the most potential to yield interesting results in the subsequent qualitative analyses, results that can be incorporated into a pedagogical description of lexical bundles that both instructors and learners will find useful.

See Appendix 2 for a complete list of the excluded bundles, and Appendix 3 for the list of bundles after the application of exclusion criteria.

From a methodological point of view, the use of exclusion criteria argues in favor of bringing human intuition to bear in the selection of phraseological items for analysis. Although computer-aided extraction processes based on quantitative criteria are extremely useful for highlighting phraseological patterns that elude our intuition, there is never any assurance that all the results they provide meet the needs of the researcher, and, in the case of pedagogically motivated studies, the needs of teachers and learners. Computers can offer leads, but it is up to the analyst to decide whether they are worth pursuing. As Wray asserts, some questions "cannot be answered without the application of common sense and a clear idea of the direction of one's research: the latter automatically creates bias in the interpretation of the raw data"

(2002, p. 28).

Ad hoc intuitive decisions are nothing new to the study of multi-word units of meaning. Several phraseological studies have used human judgment as methodological support for corpus-based procedures (Altenberg & Eeg-Olofsson, 1990; Butler, 1997; De Cock, Granger, Leech, & McEnery, 1998), chiefly to determine which items to prioritize and to eliminate results that are "phraseologically uninteresting" (Altenberg & Eeg-Olofsson, 1990, p. 7). Especially in studies that aim to identify word combinations for teaching, an intuition-based selection process is necessary. Even a largely quantitative study such as Simpson-Vlach and Ellis (2010) had to depend on teacher insights to come up with a formula that can reliably predict if a lexical sequence is worth teaching. It seems clear that until our corpus tools have become sophisticated enough to recognize which word patterns are most relevant for classrooms, textbooks and pedagogical dictionaries, subjective judgment cannot be completely avoided in pedagogically motivated phraseological analyses. As O'Keeffe et al. point out, although corpus analysis has given us the means to overcome the difficulties involved in the retrieval of formulaic sequences, "the automatic retrieval of recurrent strings is only the beginning, and a good deal of inferential analysis is still necessary to see meaning in the lists spewed out by the computer" (2007, p. 79).

3. Analyzing keywords and determining prototypical bundles

Once a reasonably manageable number of target bundles had been reached, the only question that remained was how to organize the remaining bundles in a manner that would facilitate the structural and functional analysis.

Another methodological procedure adopted from the SciE-Lex analysis is the use of keywords. In the SciE-Lex study, since the lexical bundles to be included in the dictionary would later have to be linked to its headwords, the SciE-Lex team decided to group the lexical bundles by their keywords (Verdaguer et al., 2009). The term *keyword* refers here to the word that carries the meaning of the entire lexical sequence.

In a study of formulaic sequences and the way they are accessed and utilized in a multilingual context, Spöttl and McCarthy (2003) found that students presented with unfamiliar chunks taken from a corpus tended to focus on a "strong" lexical verb or noun in or near the chunks as they attempt to retrieve their meaning. Grouping the bundles by keyword takes advantage of the presence of these strong lexical words. It also uncovered certain semantic and structural relationships among the lexical bundles that were not as obvious when they were presented in a frequency-ordered inventory.

Lexical bundles with shared keywords were revealed to be variations of a set of nouns, verbs and adjectives. Bundles with noun keywords had singular and plural forms (*in this experiment, in these experiments*) and different collocating verbs (*shown in figure, described in figure*) and adjectives (*an important role, an essential role, a critical role*). Those with verbal keywords have the most variation: there are singular and plural forms (*is associated with, are associated with; has been reported, have been reported*), positive and negative forms (*is known about, is not known*), active and passive forms (*results suggest that, it has been suggested that)* past and present forms (*can be detected, could be detected; we find that*, *we have found, we found that*), as well as diverse co-occurring subjects (*results demonstrate that, we demonstrate that*), prepositions (*was used as, was used* *for*) and conjunctions (*to determine whether, to determine if*). Bundles with adjectival keywords have positive and negative forms (*it is clear, it is not clear*) past and present forms (*is dependent on, was dependent on*) and varying degrees of epistemic certainty (*is due to, may be due, it is likely that, it seems likely that*).

There is clearly a new perspective to be gained from grouping the bundles based on shared keywords. Frequency and MI score become of secondary importance as bundles with common nodes are analyzed together, shedding light on typical patterns and variations. This method of analysis also provides evidence in support of John Sinclair's idea of canonical units of meaning. In an interview conducted by Wolfgang Teubert in 2003, published in Sinclair et al. (2004), Sinclair discussed an innovative model of language where

> […] there would be, for each lexical item, one canonical form amid all the variation. The computer would be the tool that distilled this canonical form. One such form might be a phrase like *get in touch with,* where *in touch with* is invariable and *get* is the default collocate. There are all sorts of other verbs that could be substituted for *get*: *bring, be, keep, remain*, etc. […] for every distinct unit of meaning there is a full phrasal expression which is differentiated from all other full expressions of units of meaning, and which we call the canonical form. We find it conflated in the short form (e.g., *in touch*), which is perhaps all the student must remember; but the short form must always be related to the full canonical form. (p. xxiv)

Sinclair's notion of the canonical form was adopted in the present study to address the semantic and structural links that connect the target bundles. These canonical forms are here referred to as *prototypical bundles,* using the term suggested by Teubert in the interview (Sinclair et al., 2004, p. xxiv).

In order to differentiate the prototypical bundles from those that are just components or variations of a prototype, concordance lines were carefully examined for each group of related bundles. It was discovered that although some bundles are merely different forms of a single prototype, there are others that either have distinct lexico-grammatical environments, or signal differences in usage or function that are important enough to merit explicit marking. To see this distinction more clearly, consider the following examples.

As can be seen in sentences (4) to (7), the bundles *an important role, an essential role* and *a critical role* simply represent variations of the prototypical bundle *a role in* but do not change the fundamental meaning or function of the prototype:

(4) In addition to *a role in* DNA binding, Mg2+ may also assist the topoisomerase VI DNA cleavage and religation reactions. [67]

(5) An interesting possibility suggested by these data is that signals from stromal progenitor cells may have *an important role* in maintaining a population of nephrogenic mesenchyme at the tips of the branching ureter. [22]

(6) Activin has been shown to have *an essential role* in mesoderm and neural induction in Xenopus development. [27]

(7) These observations led us to hypothesize that p16 elevation plays *a critical role* in senescence cell cycle arrest and that overcoming this block is an important step in tumorigenesis in vivo, as well as immortalization in vitro. [113]

This is in contrast with a pair of related bundles: *it is clear* and *it is not clear.* It is obvious from sentences (8) and (9) that the positive bundle *it is clear* is functionally distinct from its negative form *it is not clear*, since the latter is used to lend more epistemic commitment to a statement, while the former is used as a hedge for an unproven hypothesis:

(8) *It is clear* that Sid2p's own kinase activity does not play a role in directing it to the cleavage site. [93]

(9) *It is not clear* whether these immune responses constitute the means of protection against HIV infection. [79]

Finally, a comparison of sentences (10) to (13) shows how *we were able to* co-occurs with a different set of words than the like bundles *is able to* and *was able to*. *Is able to* and *was able to* collocate with nouns pertaining to research subjects, forming sentences that describe research findings. *Were able to*, on the other hand, collocates with the pronoun *we* and the noun *colleague* to refer to researchers and what they were able to accomplish in their studies.

(10) Secondly, we *were able to* show that recombinant CKI phosphorylates immunoprecipitated mTNF at the site that is naturally phosphorylated in vivo (Figure 4). [106]

(11) This compares favourably with the traditional purification procedure in which Wetterau and colleagues *were able to* isolate between 0.5 and 3.0 mg of the heterodimer from 600 g of bovine liver. [78]

(12) Recombinant CKI *is able to* phosphorylate mTNF in vitro. [106]

(13) We found that the cdc7-1 strain RM14-3a *was able to* grow at temperatures up to 27°C, although slightly more slowly than at 23°C. [21]

On the basis of these patterns, the criterion for separating prototypical bundles was established. Lexical bundles with distinct meanings, functions and lexico-grammatical preferences were to be regarded as separate prototypical bundles, while the rest were to be considered variations of these prototypes. Thus, in the above examples, *an important role, an essential role* and *a critical role* are variations of the prototypical bundle *a role in*; *it is clear* and *it is not clear* are two separate prototypical bundles; and *were unable to* is a prototype distinct from *is able to* and *are able to*.

It was also determined that, in a group of bundles with shared characteristics, the status of prototypical bundle was to be designated to the most frequently occurring form of a variable string. For example, in a set of like bundles comprising of *is associated with* (*n* = 61), *are associated with* (*n* = 28), *was associated with* (*n* = 25) and *be associated with* (*n* = 20), *is associated with* was assigned prototypical status.

Another important observation made through the keyword and prototype analysis is that lexical bundles tend to string together in more unpredictable ways than originally described by Biber et al. (1999). For instance, in the discussion of exclusion criteria, we looked at the bundle *are consistent with*. This three-word bundle, together with other bundles featuring the adjective *consistent* as keyword, form several possible combinations:

> [these, our] [results, data, findings, observations, studies] are consistent with [this, our, the] (previous) [data, idea, hypothesis, observations, notion, reports, results, studies, work]

It was also mentioned that in all these combinations, *are consistent with* is the central, invariable fragment, similar to the canonical form proposed by Sinclair. In this example, *are consistent with* is the prototypical bundle, and all other overlapping bundles (*results are consistent with, these results are consistent with, consistent with this, consistent with previous* and *consistent with our*) are treated as variations of the prototypical form.

It should be noted that the information provided by the target bundles are complemented by additional information gleaned from concordance analyses of the lexico-grammatical environment of the prototypical bundle. This is in line with Hunston's conceptualization of a multi-word semantic unit as a "progressively lengthening sequence, where each additional item collocates with the preceding items taken together" (2009, p. 143). With *are consistent with* as a starting point, and *results are consistent with, these results are consistent with, consistent with this, consistent with previous* and *consistent with our* as further clues, it was possible to find less frequent collocates of the prototypical bundle and capture a fuller picture of what turned out to be a much longer and more variable sequence.

Table 9 presents three more examples of prototypical bundles with several possible combinations. Words in italics are collocates that are not incorporated into any particular bundle but were discovered through additional concordance analysis.

Table 9. Examples of prototypical bundles with possible combinations

Prototypical bundle	shown in table
Keyword	table (n.)
Related bundles	are shown in is shown in summarized in table are summarized in are shown in table in table 1 in table 2 in table 3
Possible combinations	[is, are] [*described*, *given*, *listed*, *presented*, shown, summarized] in table [1,2,3…]

Prototypical bundle	results suggest that
Keyword	suggest (v.)
Related bundles	these results suggest these results suggest that data suggest that taken together these these data suggest these data suggest that together these results together these data taken together these results
Possible combinations	(taken together) [these, our] [data, *experiments*, *findings*, *observations*, results] (*strongly*) suggest (that)

Prototypical bundle	is due to
Keyword	due (adj.)
Related bundles	be due to was due to may be due
Possible combinations	[is, was, [*could*, may, *might*] be] (*likely, mainly, possibly, presumably, probably*) due to

One surprising finding is that some bundles that at first did not seem to have any connection to each other were actually part of one long bundle, such as the case of the bundles *in combination with* and *alone or in,* and *to note that* and *it is important to*, which in fact combine to form the longer sequences *alone or in combination with* and *it is important to note that.*

It can be concluded that the keyword and prototype analysis constitutes an important step in the methodological process, as it uncovered relationships, patterns and tendencies among the target bundles that would otherwise had been left unexplored. A number of bundle variations that were not revealed by the quantitative criteria were discovered, facilitating the subsequent structural and functional analyses and contributing to a much richer description of target bundles for pedagogical purposes.

Please refer to Appendix 4 for the complete list of target bundles, including both prototypical and non-prototypical forms, grouped by keyword and containing information on possible variations and combinations.

4. Concluding remarks

This chapter provided justification for the methodological choices made in this study, and discussed how the use of the MI statistic, the application of exclusion criteria and the concepts of keyword and prototypical bundle helped filter and enhance the final list of target lexical bundles. The next chapter will concentrate on these target bundles in the HSC, and explore their frequency and structural and functional characteristics.

Chapter V

Target bundles: Frequency, structure and functions

This chapter focuses on the three main features of the target bundles found in the corpus of expert native scientific writing: frequency, structure and function.

1. Frequency of target bundles

After the application of the exclusion criteria, a total of 769 lexical bundles of varying lengths remained on the list of target bundles. These 769 bundles amount to a total of 37,909 individual cases, which make up 2% of the more than two million words in the HSC.

As can be expected, the list is largely composed of three-word strings, which account for 83% or 640 of the 769 target bundles. They are followed by 113 four-word bundles, which equal 15% of the total. The list is rounded out by the much rarer five-word and six-word bundles, both of which represent just 1% of all target bundles, with just eleven and five bundles respectively. Apart from the fact that the length and frequency of lexical sequences are inversely related, the predominance of three-word bundles can be explained by the pedagogically motivated decision to exclude bundles that end with articles, which significantly reduced the number of four-word target bundles.

Table 10 shows the 50 most commonly used target bundles in order of frequency. It can be seen that all but the last five bundles in the top 50 occur at least 60 times per million words. The bundle *the presence of* is the most frequent, occurring over 450 times per million words, 30% more than the second-ranked bundle, *data not shown*.

Table 10. Top 50 lexical bundles in order of frequency

Rank	Lexical bundle	Tokens	MI Score
1	the presence of	906	8.518913
2	data not shown	625	15.556469
3	in the presence of	541	13.109891
4	the absence of	481	8.218921
5	in the absence of	387	13.240078
6	as well as	307	14.240235
7	the number of	273	7.14912
8	the effect of	259	6.858231
9	as described previously	244	15.403582
10	the ability of	237	7.730166
11	as described in	227	10.177912
12	shown in figure	216	10.021748
13	been shown to	209	11.443076
14	the addition of	203	6.676684
15	is required for	194	11.402583
16	was used to	190	9.596848
17	in response to	189	9.46708
18	a number of	183	8.239267
19	results not shown	180	13.490686
20	the effects of	176	7.03375
21	the level of	168	7.466129
22	it is possible	165	14.306728
23	to determine whether	164	15.343361
24	the role of	164	6.491655
25	the fact that	158	10.366571
26	has been shown	156	14.604337
27	is consistent with	154	11.591088
28	in addition to	154	8.558108
29	the amount of	154	8.021226
30	the formation of	149	6.72299
31	in this study	148	10.799778
32	it is possible that	146	20.813609
33	at room temperature	146	18.976404
34	the activity of	145	4.660801
35	was added to	144	10.970233
36	the possibility that	143	9.830042
37	the rate of	142	6.836724
38	the basis of	139	8.326431
39	for review see	137	16.903517
40	were incubated with	136	10.896266
41	we found that	130	12.172597
42	on the basis of	129	16.29173
43	in order to	128	10.124116
44	have shown that	126	11.192163
45	the present study	124	12.172034
46	was determined by	119	11.0729
47	shown to be	119	9.70822
48	were carried out	118	17.079535

49	in the same	116	6.625662
50	as shown in	113	8.323654

Table 11 compares the top 50 HSC target bundles with the top 50 most frequent four-word bundles in Hyland's (2008a) almost 800,000-word corpus of research articles, PhD dissertations and MA/MSc theses in biology. Despite the fact that Hyland concentrated on one bundle length and used a much smaller corpus with a wider variety of text types, there are still striking similarities between his top 50 and those of the present study. The bundles in bold are among the 50 most frequent in both Hyland's corpus and the HSC, while bundles that are in bold and underlined are those that are in the top 50 in Hyland's corpus but are less frequent in the HSC. The italicized bundles are in the top 50 in both corpora, except that in the HSC, the shorter bundle without the article or preposition is the one included (e.g., the HSC's *the presence of* vs. Hyland's *the presence of a* and *the presence of the*). The same applies to italicized bundles that are also underlined, only that they do not count among the HSC top 50 but appear further down in the frequency ranking.

The consistency between Hyland's (2008a) list and the list of target bundles in the HSC is a clear indication of the validity of both studies' findings. It demonstrates that these lexical bundles are indeed characteristic of the disciplinary discourse of the life sciences, and that they are the ones that will be most useful to individuals who wish to comprehend and produce research-focused texts in this particular field.

Table 11. Comparison of HSC findings with Hyland's (2008) biology corpus results

Rank	HSC	Hyland Biology
1	*the presence of*	**in the presence of**
2	data not shown	*in the present study*
3	**in the presence of**	**on the other hand**
4	the absence of	*the end of the*
5	**in the absence of**	is one of the
6	*as well as*	*at the end of*
7	the number of	it was found that
8	*the effect of*	at the beginning of
9	as described previously	*as well as the*
10	the ability of	as a result of
11	as described in	**it is possible that**
12	*shown in figure*	**are shown in figure**
13	*been shown to*	**was found to be**
14	the addition of	*be due to the*
15	is required for	**in the case of**
16	was used to	**is shown in figure**
17	in response to	*the beginning of the*
18	a number of	*the nature of the*
19	results not shown	*the fact that the*
20	*the effects of*	*may be due to*
21	the level of	**are summarized in table**
22	*it is possible*	**has been shown to**
23	to determine whether	**an important role in**
24	the role of	**at room temperature for**
25	*the fact that*	**at the same time**
26	*has been shown*	*can be used to*
27	is consistent with	**in the absence of**
28	in addition to	**as shown in figure**
29	the amount of	*with respect to the*
30	the formation of	**used in this study**
31	*in this study*	*was added to the*
32	**it is possible that**	*a result of the*
33	*at room temperature*	*in addition to the*
34	the activity of	the quality of the
35	*was added to*	are listed in table
36	the possibility that	*is due to the*
37	the rate of	*the presence of a*
38	the basis of	*the results of the*
39	for review see	as found in the
40	were incubated with	**were found to be**
41	we found that	**a wide range of**
42	on the basis of	*the effect of the*
43	in order to	*the presence of the*
44	have shown that	to the presence of
45	*the present study*	*was used as a*
46	was determined by	*as a result the*
47	shown to be	**have been shown to**
48	were carried out	*in this study the*
49	in the same	*it is possible that the*
50	*as shown in*	the base of the

The following sections are dedicated to the structural and functional characteristics of lexical bundles in the HSC. From this point on, only the frequencies assigned to structural and functional categories of prototypical bundles will be considered. Non-prototypical forms were disregarded in the quantitative analysis due to the presence of overlapping sequences and of those belonging to more than one prototypical bundle. Since these bundles appear multiple times on the list, counting their tokens could inflate the quantitative results. Limiting the frequency analysis to prototypical bundles[4] guarded against skewed data and afforded a less detailed yet more accurate and reliable picture of the structure and functions of lexical bundles in the native scientific corpus.

The type-token distinction is another important issue when comparing different categories, as one category can be represented by a large number of different bundle types that each occurs infrequently. The reverse can also be true, where a category is assigned to a few bundle types, with each one having a large number of individual occurrences. It is for this reason that frequency counts are provided for both bundle types and tokens for each structural and functional category.

2. Structural characteristics of target bundles

Several other studies on lexical bundles agree with Biber et al.'s (1999) observation that instead of representing complete structural units, bundles tend to consist of syntactic fragments that extend across structural units (Biber et al., 2004; Byrd & Coxhead, 2010; Hyland, 2008; Simpson-Vlach & Ellis, 2010). This is especially true

[4] The prototypical bundles *the basis of, a consequence of, the context of* and *the presence of,* which can function as independent bundles but also form part of the longer bundles *on the basis of, as a consequence of, in the context of* and *in the presence of,* respectively, were excluded from the quantitative analysis for the same reasons.

of academic prose, where Biber et al. found almost no bundles representing a syntactic whole. Lexical bundles do, however, fall into several basic structural types, which these authors used to create a widely adopted structural taxonomy of lexical bundles.

When Biber et al.'s (1999) structural framework was applied to the target bundles in the HSC, it was found that their categories covered most of these bundles' structural correlates. Only five new categories were added to the original classification scheme: other noun phrases, other adjectival phrases, verb phrases with personal pronoun *we*, other passive fragments and other verbal fragments.

Table 12 presents the structural classification of prototypical target bundles with the corresponding type and token frequencies. Figures 2 and 3 show the distribution of the different structural types and tokens.

Table 12. Structural classification of target bundles

STRUCTURE	TYPES	%	TOKENS	%
Noun structures				
Noun phrase + *of*-phrase fragment	107	24%	5828	25%
Noun phrase with other post-modifier fragment	17	4%	915	4%
Other noun phrase	9	2%	408	2%
Verb structures				
Passive + prepositional-phrase fragment	84	19%	3695	16%
Other passive fragment	18	4%	1234	5%
Verb phrase with personal pronoun *we*	10	2%	513	2%
Other verbal fragment	12	3%	522	2%
Prepositional-phrase fragments				
Prepositional phrase + *of*	28	6%	2041	9%
Other prepositional phrase (fragment)	58	13%	2689	12%
Other structures				
Verb or adjective *to*-clause fragment	28	6%	1360	6%
Verb phrase or noun phrase + *that*-clause fragment	18	4%	1016	4%
Adverbial-clause fragment	15	4%	804	4%
Copula *be* + adjective phrase	17	4%	753	3%
Other adjectival phrase	8	2%	335	2%
Anticipatory *it* + verb or adjectival phrase	10	2%	439	2%
Other expression	3	1%	457	2%
TOTAL	**442**	**100%**	**23009**	**100%**

Figure 2. Distribution of structural types

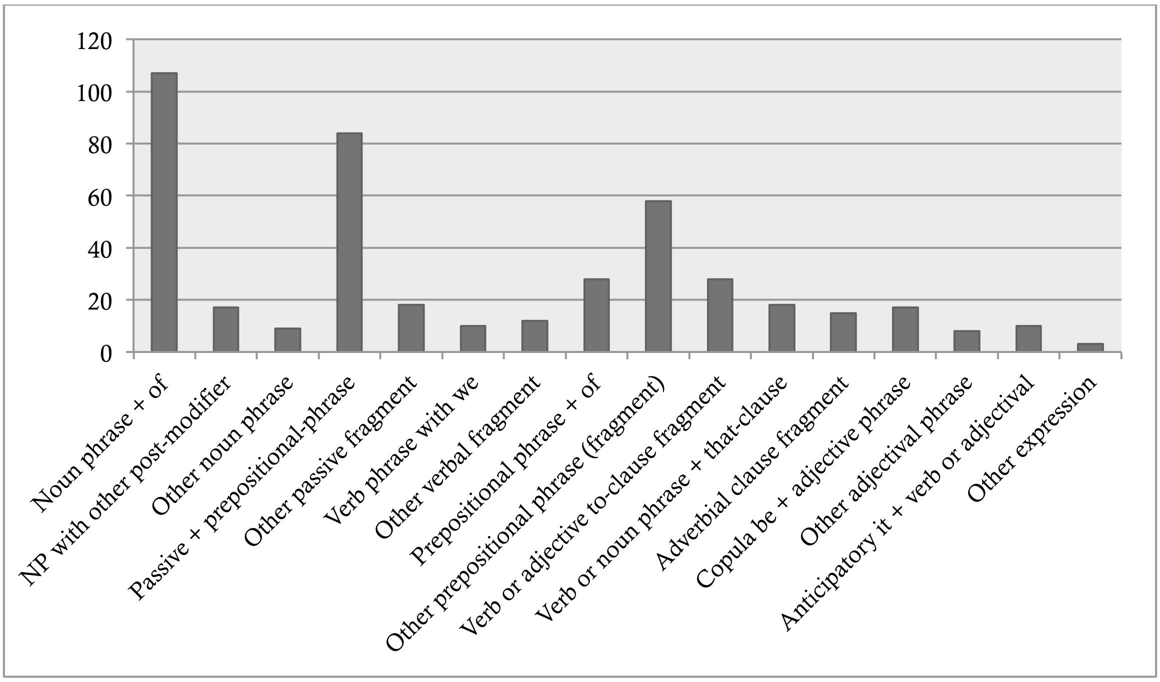
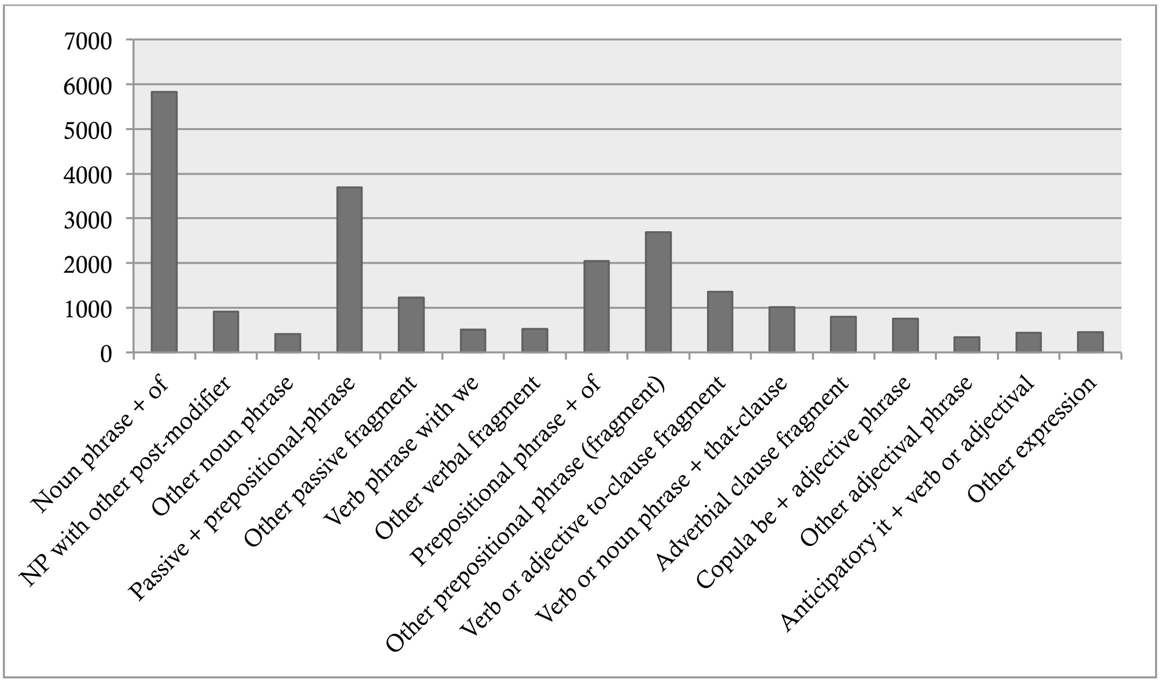


Figure 3. Distribution of structural tokens



Noun structures

Table 13 lists all target noun structures, including non-prototypical forms, by their alphabetically ordered keywords.

Table 13. Noun structures

Noun phrase + *of*-phrase fragment	the ability of, the absence of, the accumulation of, the action of, the activities of, the activity of, the addition of, the amount of, increasing amounts of, the analysis of, the appearance of, the assembly of, the association of, the average of, an average of, the basis of, the beginning of, the behavior of, the bottom of, a combination of, the combination of, a comparison of, a component of, a consequence of, the context of, the control of, the course of, high degree of, the degree of, a deletion of, a density of, the detection of, the development of, the distribution of, the effect of, the effects of, the efficiency of, the end of, the evolution of, the existence of, the extent of, a family of, the formation of, a fraction of, the fraction of, the frequency of, a function of, the function of, the generation of, the growth of, the identification of, the identity of, the importance of, the inability of, the incorporation of, the intensity of, the interaction of, the introduction of, the lack of, the length of, at the level of, high levels of, low levels of, the level of, the levels of, the localization of, the location of, a loss of, the loss of, the majority of, the mechanism of, a member of, is a member of, other members of, the method of, a mixture of, the nature of, a large number of, large number of, a number of, small number of, the number of, the total number of, total number of, the onset of, the organization of, the origin of, the pattern of, a percentage of, the percentage of, a portion of, the position of, the positions of, the presence of, the process of, the product of, the products of, the production of, the properties of, the proportion of, , the possibility of, the question of, a range of, the range of, a wide range of, the rate of, the rates of, the ratio of, the region of, this region of, the release of, the remainder of, the removal of, the rest of, the result of, the results of, a result of, the role of, the sequence of, a series of, a set of, the significance of, the site of, the size of, the stability of, the structure of, the study of, a subset of, the surface of, the time of, the timing of, the tip of, the top of, a total of, this type of, two types of, the use of, the value of, a variety of, an equal volume of, equal volume of, total volume of, the yield of
Noun phrase with other post-modifier fragment	a change in, a decrease in, a defect in, the difference in, the differences in, the difference between, no effect on, no evidence for, a gift from, an increase in, the increase in, the interaction between, the interaction with, its interaction with, a model for, model in which, a reduction in, the reduction in, the relationship between, the requirement for, a requirement for, a response to, a role in, an important role in, important role in, a role for
Other noun phrase	the ability to, their ability to, its ability to, lines of evidence, several lines of evidence, according to the manufacturer's, according to the manufacturer's instructions, the manufacturer's instructions, mechanism by which, a small number, similar results were, the results presented, the results obtained, an important role, an essential role, a critical role, previous studies have, a previous study, the present study, this work was, the indicated times, the same time, an equal volume, the present work

It can be seen from Table 12 above that the noun phrase with *of*-phrase fragment is the most common structure in the HSC, accounting for a quarter of all prototypical target bundles in the corpus. Together with noun phrases with other post-modifier fragments and other types of noun phrases, they comprise over 30% of all prototypical tokens and types. This result coincides with recent findings (Biber et al., 1999; Byrd & Coxhead, 2010; Hyland, 2008) and supports the view of academic writing as being "noun-centric" (Swales, 2008, p. v).

Noun structures feature 129 different keywords, the widest variety among all other lexical-bundle structures. They therefore carry a broad range of meanings in the scientific texts. Noun structures are commonly used to denote qualities (*a function of, the nature of, the ability to*), degree (*the degree of, the extent of*) and existence (*the presence of, the absence of*); to describe events (*the beginning of, the loss of*) and actions (*the addition of, the production of*); to indicate measurements (*an equal volume, the size of*), quantities (*the amount of, a small number*) and proportions (*a fraction of, the percentage of*); to mark location (*the region of, the site of*); and to signify groupings (*a set of, a wide range of*) and group membership (*a member of, a component of*).

It is also interesting to note that most noun structures are variations of the highly productive frame *the* __ *of* and *a* __ *of*, where the blank slot is filled by a number of words, e.g. *action, bottom, combination, development* and *evolution*.

Verb structures

Table 14 displays all target verb structures, including non-prototypical forms, by their alphabetically ordered keywords.

Table 14. Verb structures

Passive + prepositional-phrase fragment	was added to, were added to, was analyzed by, were analysed by, were analyzed by, was assessed by, is associated with, are associated with, was associated with, be associated with, is based on, was based on, carried out at, carried out in, carried out with, were carried out at, is caused by, be caused by, were collected from, compared with control, when compared with, is composed of, was confirmed by, described in the experimental section, was performed as, were performed as, prepared as described, was performed as described, were performed as described, carried out as, performed as described previously, were prepared as described, carried out as described, are described in, described in figure, was detected by, were detected by, was detected in, be detected in, were detected in, was determined as, was determined by, were determined by, was digested with, was dissolved in, was examined by, be explained by, were exposed to, are expressed as, is found in, are found in, was found in, were fixed in, was generated by, were generated by, were grown at, were grown in, were grown to, were identified by, have been identified in, been identified in, been identified as, been implicated in, has been implicated in, were incubated for, were incubated with, is indicated by, are indicated by, are indicated in, was induced by, was introduced into, be involved in, is involved in, are involved in, to be involved in, was isolated from, were isolated from, the isolation of, is known about, little is known about, is localized to, were made by, was measured by, is mediated by, was mixed with, was observed in, has been observed, also observed in, been observed in, was obtained by, was obtained from, were

	obtained by, were obtained from, expressed as a percentage of, was performed as, was performed by, was performed in, was performed on, was performed with, were performed as, were performed in, were performed using, were performed with, was prepared by, was prepared from, were prepared as, were prepared by, were prepared from, were processed for, kindly provided by, was purchased from, were purchased from, was purified from, referred to as, were removed by, was replaced with, is required for, are required for, be required for, to be required for, was required for, also required for, that are required for, not required for, is not required, is not required for, were obtained with, were obtained in, was resuspended in, were resuspended in, were separated by, were separated on, data not shown in, are shown as, is shown in figure, are shown in figure, shown in figure, shown in figure 1, shown in fig, shown in figure 2, shown in figure 3, shown in table, are shown in, is shown in, are shown in table, were stained with, was subjected to, were subjected to, summarized in table, are summarized in, is supported by, was supported by, were tested for, tested for their ability to, were transferred to, were treated for, were treated with, was used as, was used for, was used in, was used to, were used as, were used for, were used in, were used to, were washed in, were washed twice with, were washed with
Other passive fragment	were allowed to, carried out using, was carried out, were carried out, has been demonstrated, performed as described, been described previously, have been described, has been described, can be detected, could be detected, was not detected, was determined using, activity was determined, would be expected, results are expressed, have been found, have been identified, has been implicated, have been implicated, at the indicated, of the indicated, little is known, is not known, activity was measured, was performed using, analysis was performed, experiments were performed, extracts were prepared, has been proposed, to be determined, has been reported, have been reported, to be required, similar results were obtained, results were obtained, can be seen, to that seen, data not shown, results not shown, has been shown, been shown previously, to that observed, has been suggested, medium supplemented with, be used to, been used to, can be used, has been used, used in this study, used to amplify, used to determine, used to identify, were washed three times
Verb phrase with personal pronoun *we*	we asked whether, we conclude that, we demonstrate that, we found that, we find that, we have found, we have identified, we propose that, we show that, we have shown, we have shown that, here we show that, we suggest that, we tested whether, we were unable to, we have used
Other verbal fragment	did not affect, does not affect, did not appear, does not contain, may contribute to, had no effect, had no effect on, exclude the possibility, does not require, would result in, not result in, play a role, play a role in, for review see, for reviews see, see figure 1, see figure 2, see table 1, see materials and methods, these results suggest, these data suggest, suggesting that this

Verb structures represent 25% of all prototypical target-bundle tokens and 28% of all types in the corpus. Although they feature only 80 individual keywords, fewer than those found in noun structures, verbal constructions present more structural variation.

The majority of verb structures are composed of a verb in the passive voice followed by a prepositional-phrase fragment. Passive expressions that incorporate a present-tense verb typically denote locative or logical relations between elements. They mainly serve to label data presented in tables and graphs (14) (15), or to identify the basis of an argument (16) (17).

(14) Bacterial strains used in this study *are described in* Table 3. [2]

(15) The location of the probe used for genotyping *is shown in* A. [60]

(16) The analysis *is based on* the oxidation of glucose 6-phosphate, which is formed following the phosphorylation of fructose to fructose 6-phosphate by hexokinase, and its subsequent isomerization by phosphoglucoisomerase. [72]

(17) This hypothesis *is supported by* the reduction in the percentage of BSA-gold positive phagosomes in cells that were incubated at 13°C, a temperature that is known to inhibit early-late endosome fusion. [109]

This finding is consistent with that of Hyland (2008a), who claims that:

> Identifying tabular or graphic displays of data and the bases of an assertion are typically constructed through formulaic passive constructions in the hard sciences. This both highlights the research or text feature being discussed and can help downplay the personal role of the scientist in the interpretation of data to suggest that the results would be the same whoever conducted the research. (p. 11)

However, in addition to present-tense passive constructions, there is a also a marked prevalence of passive structures with past-tense verbs, most of which are found in the Experimental, Materials and Methods or Methods section of the research articles. These past-tense passive constructions are associated with a different set of verbs than their present counterparts, as their keywords tend to be activity verbs referring to specific experimental procedures, as in the following examples:

(18) For arrest in S phase, hydroxyurea (Sigma Chemical Co.) *was added to* 0.1 M to log-phase cells in liquid YPD, pH 5.8, and incubated at 26°C until >70% of cells were large-budded. [16]

(19) The chorions *were removed by* immersion in 50% bleach in Triton-NaCl for 2 min. [85]

Here the passive is used to shift the focus from the scientist to the action itself, in order to emphasize that the generally accepted procedures are being respected, and that the outcome of such procedures will be the same regardless of the human agents carrying them out. This lends credence to Tarone et al.'s (1998) generalization that authors of scientific articles (in her particular case, of astrophysics journal papers) use the passive when they wish to indicate that they are simply following established or standard procedure.

It is also remarkable that many past-tense passive constructions have corresponding noun phrase + *of* structures, as shown in these examples:

(20) The importance of the 5'-untranslated region of the oli1 mRNA in the biogenesis of subunit 9 was first recognized by *the analysis of* a temperature-sensitive strain h45 shown to contain a single base insertion 87 nucleotides (nt) upstream of the oli1 coding region (OOI et al. 1987). [25]

(21) In this series the hyposmotic solution was made by *the removal of* 25 mM NaCl. [107]

The noun phrases seem to be just another depersonalization technique used by scientists to complement passive structures.

In their analysis of the patterns of use of active and passive constructions in medical expository texts in English, Salazar, Ventura and Verdaguer (2011) found that the empirical nature of the medical field requires authors to use passive structures to objectively describe experimental procedures, and personal constructions to express conclusions drawn from the results of these experiments. These observations are supported not only by the frequency of both present- and past-tense passive bundles in the HSC, but also by the occurrence of bundles consisting of a verb collocating with the personal pronoun *we*. If the authors of these health-science research articles use the passive to talk about scientific methodology and the logical bases of their assertions, they use the highly personal form *we* + verb to discuss their objectives (22), observations (23), achievements (24) and conclusions (25).

(22) To explore this hypothesis, *we asked whether* conditions that would obviate the need for the initial viral attachment, such as bringing CypA-deficient viruses into close contact with target cells, would rescue their infectivity. [82]

(23) In our experiments *we found that* the phosphate contents of the starches were reduced in plants where both the SSII and SSIII isoforms were reduced, and that this was dependent on the total reduction in soluble SS activity. [52]

(24) *We have identified* a J-binding protein in nuclear extracts of T.brucei bloodstream form and the related kinetoplastids C.fasciculata and L.tarentolae. [17]

(25) Thus, *we conclude that* GlcN-(2-O-octyl)PI is not a substrate for HeLa MT-I. In addition, neither this compound nor its N-acetyl derivative affected the processing of exogenous GlcN-PI to glycolipid H5. [91]

It can be seen from the above examples that lexical bundles including the personal pronoun *we* are mainly employed by scientific writers to claim ownership for their results and affirmations. In this manner, they are able to stress the novelty and importance of their work and build a “credible authorial identity” (Hyland, 2001, p. 219) as an ‘‘opinion holder’’ and an ‘‘originator’’ of new ideas (Tang & John, 1999, pp. 228–229). *We* + verb bundles allow researchers to firmly establish their position and gain recognition for their views, something that they themselves consider essential when writing a research article for publication (Hyland, 2001, pp. 222-223).

Through the structural analysis of the verbal target bundles, it was possible to establish certain usage patterns that demonstrate the importance of both active and passive, personal and impersonal expressions in scientific writing. Passive, impersonal constructions are employed in the objective discussion of experimental methods and justification of claims, so as to build a sound and universally acceptable foundation for the author’s subsequent assertions. Active, personal structures, on the other hand, are used by scientists to explain their aims, findings, accomplishments and conclusions, as a way to underscore their original contribution to the field of research. The judicious use of personal and impersonal expressions reflects the dual role of the scientist as conductor of research and claim maker, and plays an essential role in the construction of an effective research article. These choices of voice and tense constitute a subtle yet important rhetorical function of which non-native or novice writers should be made aware.

Prepositional-phrase fragments

Table 15 shows all target prepositional-phrase fragments, including non-prototypical forms, by their alphabetically ordered keywords.

Table 15. Prepositional-phrase fragments

Prepositional phrase + *of*	in the absence of, in the absence or presence of, by the addition of, by addition of, in the amount of, on the basis of, in the case of, as a consequence of, in the context of, at a density of, at the end of, with the exception of, in the formation of, as a function of, by the method of, in a number of, in the number of, of a number of, as part of, as a percentage of, in the presence of, in the presence or absence of, for the presence of, by the presence of, for the production of, in the production of, at a flow rate of, in the regulation of, as a result of, in support of, at the surface of, on the surface of, in terms of, by use of, with the use of, in the vicinity of, by virtue of
Other prepositional phrase (fragment)	for their ability to, in accordance with, in addition to, for an additional, in agreement with, in the bottom, in this case, in all cases, in each case, in some cases, in combination with, in comparison with, in concert with, under these conditions, under the same conditions, in conjunction with, as a consequence, in contrast to, in contrast with, as a control, in the control, under the control of, in the dark, with the exception, in these experiments, in this experiment, as in figure, in figure 1, in figure 2, in figure 5, in figure 3, in fig 1, in figure 7, with the following, on the other hand, on ice for, of a large, on the left, to the left, in a manner, by the method, as a model, in this model, in this paper, in the present, in the present study, in the present study we, in this process, in the region, in this region, in this report, with respect to, in response to, as a result, to the right, in the same , at the same, to the same, in the materials and methods section, in the experimental section, in a similar, at the site, in this study, in this study we, in support of this, at the surface, on the surface, in table 1, in table 2, in table 3, at room temperature, at room temperature for, at the same time, at the time, at various times, at this time, in the top, in a total, of the total, by treatment with, for up to, in the upper

In accordance with the results of Biber et al. (1999) and Hyland (2008a), most of the target lexical bundles with a prepositional phrase, especially those with embedded *of*-phrases, commonly signify abstract, logical relationships between propositional elements:

(26) *On the basis of* soft tissue morphology, Lemelin predicted that Ateles has the ability to hyperextend the tail. [30]

(27) There are several possible explanations for this discrepancy. First, the conditions or protein constructs we chose may have prevented binding or the interaction may only occur *in the context of* the complete translation machinery. [21]

(28) We have shown that DMPK mice develop late-onset skeletal myopathy *as a consequence of* abnormal excitation/contraction coupling. [5]

(29) All wild-type N. meningitidis strains tested were able to use human Hb as an iron source *with the exception of* strain 2844. [44]

(30) The BimC motor Cin8p is required to assemble and elongate the bipolar spindle, probably *by virtue of* its ability to cross-link and slide microtubules. [16]

Some prepositions are characterized by a specific meaning. Some bundles with the preposition *by* are associated with methods (*by the method of, by use of*), many with the preposition *in* denote processes (*in the formation of, in the regulation of*) and amounts (*in the amount of, in a number of*), and some bundles with the preposition *at* serve to introduce measurements (*at a density of, at a flow rate of*).

There are numerous other prepositional-phrase fragments. Several are used to refer to the study or text itself (31) or to different sections of the article (32).

(31) *In the present study*, a contribution from unlabeled hepatic lipid stores to TG synthesis may be less likely, because the subjects had been fasted for 24 hours by the end of the infusion test, which should substantially reduce hepatic lipid stores. [70]

(32) Standard methodologies were employed as outlined *in the Materials and methods section*. [53]

Others serve to identify place (*in the region, at the site*), extremity (*in the bottom, at the surface*) and orientation (*on the left, to the right*).

Many others have more figurative meanings:

(33) The anchor also acts as a co-chaperone *in concert with* D-VI, ensuring proper folding of the catalytic subunit. [42]

(34) *On the other hand*, several amino acids which form a second, non-catalytic pocket in mammalian ACs were conserved in the protozoan cyclase, i.e. were like those in ACs. [51]

(35) Thus, although there is good evidence that M protein is involved in evasion of phagocytic killing in vitro, the data remain inconclusive *with respect to* the role of M protein in bacterial virulence during in vivo infection. [4]

The frequent and varied use of prepositional-phrase fragments in the HSC clearly indicates that in scientific writing, the sense of English prepositions goes beyond the concrete adverbial meanings traditionally presented in English-language classes (Byrd & Coxhead, 2010).

Other structures

Table 16 presents all other target structures, including non-prototypical forms, by their alphabetically ordered keywords. The following findings closely match Biber et al.'s (1999) own description of these forms.

Table 16. Other structures

Verb or adjective *to*-clause fragment	are able to, be able to, is able to, was able to, were able to, to account for, to act as, to address this, appear to be, appears to be, appeared to be, not appear to, does not appear to, not appear to be, did not appear to, to associate with, to confirm that, to demonstrate that, to determine whether, to determine if, to distinguish between, to ensure that, be expected to, would be expected to, expected to be, found to be, was found to, were found to, was found to be, been found to, were found to be, to interact with, known to be, is known to, are known to, is likely to, likely to be, is likely to be, are likely to, are likely to be, to note that, is predicted to, predicted to be, been proposed to, remains to be, remains to be determined, been reported to, is required to, be required to, been shown to, has been shown to, shown to be, have been shown to, was shown to, has been shown to be, to show that, to test whether, to test this, to test this hypothesis, is thought to, thought to be, are thought to, is thought to be, are unable to, was unable to, were unable to, is unlikely to, unlikely to be
Verb phrase or noun phrase + *that*-clause fragment	the conclusion that, results demonstrate that, have demonstrated that, the fact that, by the fact that, have found that, the finding that, the hypothesis that, the idea that, this implies that, this indicates that, results indicate that, these results indicate that, data indicate that, these data indicate that, the notion that, the observation that, the possibility that, possibility is that, been proposed that, studies have shown that, have shown that, has shown that, results show that, shown previously that, this suggests that, results suggest that, these results suggest that, data suggest that, these data suggest that, have suggested that, has been suggested that
Adverbial-clause fragment	as compared with, as described previously, as described above, as previously described, as described in the experimental section, as described in materials and methods, as described by, as described in, essentially as described, as described for, as determined by, as shown in figure, were as follows, as indicated by, as judged by, as measured by, as opposed to, as reported previously, as seen in, as shown in, as shown by

Copula *be* + adjective phrase	is capable of, is consistent with, which is consistent with, are consistent with, be consistent with, is dependent on, was dependent on, is difficult to, is due to, be due to, was due to, may be due, is essential for, are essential for, is important for, be important for, is an important, is independent of, is necessary for, is also possible, are representative of, is responsible for, be responsible for, are responsible for, be the result of, is sensitive to, is similar to, are similar to, was similar to, is subject to, is sufficient to
Anticipatory *it* + verb or adjectival phrase	it appears that, it is clear, it is not clear, it is likely, it is likely that, it seems likely that, it should be noted, it should be noted that, it is important to, it is possible, it is possible that, it has been proposed that, it has been shown that, it has been shown, it was shown, it has been suggested, it is unlikely
Other adjectival phrase	alone or in, consistent with this, consistent with previous, consistent with our, significantly different from, not due to, little or no, also present in, are present in, is present in, was present in, were present in, closely related to, the same as, similar to that, similar to those, similar to that of, very similar to, only a small
Other expression	this is consistent with, results are consistent with, these results are consistent with, in order to, there are several, taken together these, together these results, together these data, taken together these results, as well as, as well as in

Verb or adjective + *to*-clause fragment

Lexical bundles of this structure can be simple *to*-clauses or *to*-clauses preceded by a predicative adjective or a verb phrase.

Bundles with verb phrases before the *to*-clause are most frequently used to refer to previous findings (36) (37) or known and accepted facts (38) (39). The verb phrase is typically in the passive voice.

(36) The figure-of-eight DNA molecules *were found to be* cleaved by EcoR124II at the same positions as the -structure when assayed for cleavage in the mixture with the other DNA species produced by Xer recombination (not shown). [46]

(37) In addition, the toxicity of the carcinogenic metal compound, cadmium chloride, was investigated, since glutathione has *been proposed to* have a direct role in its detoxification. [98]

(38) A microenvironment that is relatively deficient in FN may therefore allow monocytes to differentiate into the tissue macrophages that *are known to* orchestrate repair of the damaged myocardium (13, 55, 56). [99]

(39) However, to date, HIV-1 entry *is thought to be* mediated exclusively by gp120. [82]

Bundles featuring predicative adjectives controlling a *to*-clause express ability (40) and likelihood (41).

(40) Thus, the c-Jun S63/73A mutant *is able to* support cell proliferation at levels similar to wild-type, but is completely inactive with regard to protection of cells from UV-induced apoptosis. [111]

(41) The methylated PAI2 and PAI3 genes in the fluorescent pai1-pai4 deletion mutant *are likely to be* relics of a de novo methylation event in the parental strain WS that persist solely through efficient maintenance methyltransferase activity. [47]

Simple *to*-clauses commonly indicate methodological aims (42) (43) (44). They are usually found in sentence-initial position.

(42) *To confirm that* the ability of mFlagAx to activate TCF-dependent transcription was dependent on its ability to bind GSK-3, a leucineproline mutation was introduced into the putative hydrophobic interface of the coiled-coil domain at position 521. [89]

(43) *To determine whether* cortical-associated p34cdc2 influences cortical myosin II activity during cytokinesis, we labeled eggs in vivo with [32P] orthophosphate, prepared cortices, and mapped LC20 phosphorylation through the first cell division. [87]

(44) Initially, assuming that such tails would also block access to the DNA by RecBC enzyme, our strategy was to resect the DNA at one end with Exo III, perform Exo I protection assays, and use Southern hybridization with strand-specific oligonucleotide probes *to distinguish between* the top and the bottom strands. [13]

Verb phrase or noun phrase + *that*-clause fragment

Lexical bundles comprising a *that*-clause can have either a noun or a verb phrase in the main clause.

That-clauses introduced by the nouns *conclusion, fact, finding, hypothesis, notion, observation,* and *possibility* serve to highlight a propositional statement, especially when presenting facts or findings corroborating the claim (45) (46) (47).

(45) *The fact that* cyclases exist with C1a and C2a arranged in both ways strongly supports the hypothesis that initially membrane-anchored monomers formed a homodimeric AC. [51]

(46) Transfection of cells with upd lacking a signal sequence does not result in Hop phosphorylation (lane 3), consistent with *the notion that* Upd is required extracellularly for signaling to occur. [37]

(47) Therefore our results provide circumstantial evidence in favor of *the hypothesis that* the discrepancies in estimates are due to differences in the mutation rate per germline replication between different parts of the genome. [90]

Verb phrases followed by *that*-clause fragments are commonly used to preface inferences drawn from the author's own results (48) or from those of other studies (49).

(48) *These results suggest that* loss of silencing events during development are common, whereas shifts from a nonsilenced to a silenced state are extremely rare or do not occur. [47]

(49) Previous *studies have shown that* some hnRNPs are also extractable from nuclei at 0.5 M NaCl. [63]

Adverbial-clause fragment

Lexical bundles beginning with the subordinator *as* frequently appear in text-reflexive markers that direct the reader to different parts of the article (50) (51) and to related literature (52).

(50) Briefly, cyclin D1 immune complexes were prepared from 600 μg of whole-cell extract prepared *as described previously* and incubated with 1 μg of GST-Rb in the presence of kinase buffer (20 mM MgCl2, 50 mM Tris pH 7.5, 20 μM ATP, 10 μCi [-32P] ATP). [111]

(51) Extra-long chains are those eluting earlier than the B4 fraction, *as shown in* Figure 4. [52]

(52) Acid extracts of GAS surface M protein were prepared from 100-ml broth cultures *as described by* Lancefield. [4]

They are also employed in stating the basis of an assertion (53) (54) and making comparisons (55) (56).

(53) When the spc42-10i mutation was present, the plasmid loss rate was reduced 1,000-fold *as judged by* the absence of colony growth at high dilutions, but was unaffected by the presence of the spc110-1i or spc110-2i mutations. [1]

(54) In initial experiments we found that GFP-DPMS was catalytically active when expressed in both Escherichia coli and in L.mexicana promastigotes, *as indicated by* a 50% increase in enzyme activity over endogenous wild-type levels in the latter (unpublished data). [45]

(55) Moreover, all lyso-PC doses elicited significant pulmonary edema *as compared with* lungs from LPS-pretreated animals perfused with saline. [88]

(56) In xrn1 strains, the 5' ITS1 signal is distributed throughout the cytoplasm (Fig. 2f) *as opposed to* the mostly nucleolar localization observed in XRN1 wild-type cells (Fig. 2b). [66]

Copula *be* + adjective phrase

These lexical bundles are combinations of the copula *be* and an adjective phrase. They are used to express causative (57) and comparative (58) relationships, as well as the author's evaluative assessment of a proposition (59) (60).

(57) This *may be due to* differences in strain background or partially toxic effects of the disruption mutant used in that study. [62]

(58) During this time period, the cells undergo morphologic changes that *are similar to* those detected in senescent fibroblast cultures: they become enlarged, flat and spread out. [111]

(59) The occurrence of a single exchange near each end of the linear fragment would result in positive interference of genetic exchanges; such interference *is difficult to* measure in E. coli crosses, but is well documented in most eukaryotes. [95]

(60) We showed that CypA *is essential for* the initial attachment of HIV-1 to target cells. [82]

Anticipatory *it* + verb or adjectival phrase

Lexical bundles that introduce extraposed structures in the anticipatory *it* pattern are controlled by an adjective or a verb phrase.

The majority of bundles with the anticipatory *it* structure feature predicative adjectives followed by a *to-* or *that*-clause. They are employed by writers in the appraisal of possibility (61), likelihood (62) and importance (63).

(61) Based on its relationship with Wnt and APC, *it is possible that* ß-catenin may positively regulate cellular proliferation or inhibit apoptosis. [69]

(62) In wild type, *it is likely that* the persistent CycE observed beginning in stage 10B inhibits assembly of new prereplication complexes at most origins during this period. [11]

(63) *It is important to* emphasize that our studies only pertain to HIV-1 infection in older adults, and not to HIV-1 infection in children or to adults < 20 years old. [38]

Lexical bundles with extraposed structures comprising a verb predicate are typically passive constructions followed by a *that*-clause. Although they also communicate the writer's stance, they do so by presenting the proposition as an obvious and widely accepted fact (64) (65) (66).

(64) *It is clear* that we need more investigations into the total ferritin genes in one species. [105]

(65) Although the temperature shift is drawn as having taken place two-fifths of the way through S phase, *it should be noted that* this is arbitrary; it was not possible to determine the time in S phase at which cells were shifted. [21]

(66) *It has been shown that* the heterochromatin-binding protein HP1 interacts with the ND10 component sp100, thereby suggesting for the first time a link between ND10 and the chromatin compartment. [26]

Other adjectival phrases

These are lexical bundles formed by different adjectival fragments that do not fall into the other categories, most of which express comparative relations (e.g., *significantly different from, closely related to* and *similar to that*).

Other expressions

This category includes all other target bundles that do not fit into the previously described categories (e.g., *this is consistent with, in order to, as well as*).

3. Functions of target bundles

The results presented above confirm what previous studies have shown: that in spite of their fragmentary nature, lexical bundles follow certain structural patterns that provide insight into the nature of biomedical research articles. This section demonstrates that the same is true with regard to lexical bundles and their functions.

O'Keeffe et al. (2007) use the term *pragmatic integrity* to denote the pragmatically specialized roles that lexical chunks fulfill in discourse, a notion of functional adequacy that is independent of structural completeness. They argue that "it is in pragmatic categories rather than syntactic or semantic ones that we are likely to find the reasons why many of the strings of words are so recurrent [...] by *pragmatic categories* we mean the different ways of creating speaker meanings in context" (p. 71, italics mine). Indeed, in the previous section, to explain why certain structures are more frequent than others, it was necessary to link lexical bundles to pragmatic categories such as discourse and stance marking. This section will show that all target bundles found in the HSC fall into coherent functional categories that form part of a systematic descriptive framework.

From a pedagogical perspective, the functional analysis of lexical bundles is essential to their value as teaching items. Even though bundles are largely incomplete units that include words already familiar to most advanced-level EAP students, their functions afford them a certain degree of face validity for teachers and students. The fact that bundles can be used to do things such as introduce topics, compare and contrast elements, quote sources and draw conclusions gives instructors and learners enough incentive to teach and learn these multi-word expressions. This, in turn, makes it of utmost importance to provide an accurate yet accessible functional description of lexical bundles that can help EAP students master certain functions that are crucial to academic writing.

3.1. Multifunctionality of lexical bundles

No attempt at functionally classifying lexical bundles can be made without tackling the issue of their multifunctionality. Biber et al. (2004) acknowledge that a single lexical bundle can serve multiple functions in different contexts, such as *the beginning of the* and *at the end of*, which can be time, place or text-deictic references depending on the textual environment; or even in a single occurrence, such as *take a look at* and *let's have a look*, which can be considered directives as well as topic introducers (pp. 383-384). The solution that they propose is to examine concordances of potentially multifunctional bundles and classify them according to their most common use.

However, determining the primary function of a lexical bundle through frequency comparisons is not always straightforward. It is very difficult to determine exactly what the most frequently used function of a bundle is without analyzing all concordances and categorizing every single one of its occurrences. And in the case of bundles with overlapping functions in the same instance, this method is downright impossible. Assigning functions in this way also makes it easy to overlook uses that may be less frequent but not less pragmatically interesting. For example, Byrd and Coxhead (2010) note that of the 281 occurrences of the bundle *the end of the* found in their corpus, 17 point the reader to a specific section of the text, while the rest indicate the end of a process or an event. The much lower frequency of the text-deictic function of this particular bundle does not necessarily make it less important than the time-reference function.

Among the target bundles on the list, 153 were found to be multifunctional. Of these, 101 have multiple functions in a single occurrence, as in the case of *may contribute to* (67)*, this suggests that* (68)*, is thought to be* (69) and *is unlikely to* (70).

(67) The morphological changes suggest that enhanced motility *may contribute to* this dramatic increase in colony size, but this is speculative. [69]

(68) *This suggests that* EcoR124II promoted branch migration to the end of the region of 290 bp homology and then introduced a double-strand break at the site where the further branch migration was blocked by DNA heterology. [46]

(69) In addition to a role in transport, the plant proton pump *is thought to be* involved in signal transduction and responses to the environment. [114]

(70) As suggested by the conditioned taste aversion paradigm (Table 1), the inhibitory effect of CCK-8 in the mice *is unlikely to* be the result of an aversive stimulus (e.g., nausea). [48]

May contribute to expresses a causative relation, while *this suggests that* and *is unlikely to* indicate inferential relations, but all three bundles can also be considered stance markers. *Is thought to be* conveys a generalization as well as an inference.

A closer look at bundles with context-dependent functional variations reveals that there are different factors that influence these variations. One of these factors is the bundle's position in a sentence. Consider the following uses of the bundle *at the same time,* which both Cortes (2004) and Hyland (2008a) classify as a time marker:

(71) The nuclei became enclosed by an intact nuclear envelope *at the same time* as control nuclei (~30 min) but did not increase in size for at least 4.5 h. [32]

(72) When the steady state is achieved, the mean phenotypic value does not lie at zopt, but lags behind zopt by an amount denoted by S (i.e., S is the difference between the optimum and the mean phenotype). *At the same time*, the genetic variance (VG, S), the heritability (h2S), and the mean death rate (S) all depend on the rate of environmental change. [108]

It can be seen from (71) that, as Byrd and Coxhead (2010) point out, the meaning of *at the same time* is more about simultaneity than actual time. In this example, the bundle acts more as a descriptor of a specific condition than a time marker. In the second example (72), *at the same time* appears at the beginning of the sentence, where, instead of indicating time or simultaneity, it serves a discourse-marking function that can be likened to *in addition* or *similarly*. This demonstrates that the function of a lexical bundle can change depending on where it is placed in a sentence.

A bundle's position in the text can also have an impact on its use. For instance, most occurrences of the bundle *as indicated by* mark the inferential relationship between two elements. This is exemplified by the following extracts, one taken from the Results section of an article and another from the Conclusions section:

(73) In initial experiments we found that GFP-DPMS was catalytically active when expressed in both Escherichia coli and in L. mexicana promastigotes, *as indicated by* a 50% increase in enzyme activity over endogenous wild-type levels in the latter (unpublished data). [Results] [45]

(74) It is subject to activation by phosphorylation, *as indicated by* its sensitivity to protein phosphatase 2Ac. [Conclusions] [10]

However, some instances of the same bundle have an entirely different use when found in the captions of figures. In these parts of the text, they serve a text-reflexive function:

(75) Ca2+-binding results in a conformational change in the N-terminal helices,
as indicated by red arrows (PDB accession code 1DVI). [Figures] [42]

(76) Approximately 75 protein spots were enriched in the IGC fraction *as indicated by* the circled regions. [Figures] [63]

Another important conditioning factor is the lexical bundle's immediate co-text. The words surrounding the bundle sometimes determine its function in the sentence. The bundle *is supported by*, for example, has two distinct functions: one is to provide justification for an argument (77), and the second is to acknowledge research funding (78).

(77) To date, this model *is supported by* observations in vitro using rat and human hemoglobin and whole erythrocytes. [33]

(78) Dr. Badley *is supported by* a grant from Physicians Services Incorporated Foundation and the AIDS Program Committee of Ontario. [20]

The specific use of *is supported by* is easily recognizable from the words that follow it. It serves the first function when followed by the words *data, experiments, findings* and *observations*, and the second when followed by the words *fellowship* and *grant*.

The bundle *is consistent with* is another good example. Some of its occurrences are used to compare one element to another, as in the following extract:

(79) The value derived for the shape-dependent Mark-Houwink parameter (a f10.41) *is consistent with* the condensed or branched morphology observed in the electron microscope. [86]

But when co-occurring with the nouns *data*, *evidence, reports, results, studies* and *work*, which are sometimes modified by the adjectives *earlier*, *other*, *previous* and *published*, its function becomes that of citing previous research whose results agree with the author's findings:

(80) This conclusion *is consistent with* earlier data showing that efficient stimulation of processive DNA polymerase activity requires the simultaneous presence of all three subunits [...] (Onrust et al., 1991). [101]

(81) Specifically, the identification of 21 chromosomal segments that contribute to reduced pollen viability *is consistent with* other studies that have identified a large number of factors that affect male sterility (e.g., TRUE et al. 1996; WU et al. 1996). [77]

(82) The decline in E2F1 and E2F3 DNA-binding activities reflects post-transcriptional regulation (Fig. 2B) and, at least for E2F1 activity, *is consistent with* previous work that has demonstrated an ability of cyclin A/cdk2 to bind to the amino-terminus of the E2F1 protein, phosphorylate the associated DP1 protein specifically, and result in the inactivation of the E2F1 DNA-binding activity (Krek et al. 1994; Xu et al. 1994; Krek et al. 1995; Dynlacht et al. 1997). [50]

(83) This result *is consistent* with published reports, which state that the N-terminus of MCP-1 is involved indimerization thought to be necessary for MCP-1 signalling. [76]

The same applies to the bundle *in agreement with*, which shares the dual comparative-citation function:

(84) These results are *in agreement with* the lower total content of Ca#+-ATPase in our 'slow' preparations, quantified by densitometry of Coomassie Blue-stained gels and Western immunoblots (Table 1). [comparative] [102]

(85) These observations are *in agreement with* earlier findings by our group and Machwate et al. indicating that agents that increase cAMP production, such as PTH and prostaglandins, suppress apoptosis of osteocytes/osteoblasts and periosteal cells, respectively. [citation] [73]

(86) The more efficient processing of GlcNAc-PI compared with GlcN-PI is *in agreement with* previous reports that suggest substrate channelling between the de-N-acetylase and MT-I in the trypanosomal pathway (Smith et al., 1996, 1997b; Sharma et al., 1997). [citation] [91]

The influence of discipline can be seen in the use of the bundle *in the presence of,* classified by Cortes (2004) and Hyland (2008a) as a text-organizing framing bundle. The following example from the HSC supports this classification, where the bundle is used to specify the conditions of an experimental procedure:

(87) For a partial crosslinking of EEA1 from cytosol, 100 ll (300 lg) of HeLa cytosol was incubated *in the presence of* 5 mM bismaleimidohexane (BMH) for 1 h at 4°C. [10]

The following examples taken from scientific texts in the British National Corpus (BNC) give further evidence of the framing function of *in the presence of*:

(88) And here is a recording then of the channel activity *in the presence of* ten to the minus eight molar calcium [...] [BNC spoken: natural science lecture]

(89) Many polymerizations proceed best *in the presence of* catalysts. [BNC written: academic, technical, engineering]

(90) *In the presence of* malignancy it is known that different tissues can respond in different ways. [BNC written: academic, medicine]

(91) Results of three sets of experiments *in the presence of* calmodulin are shown with standard deviations marked by errors bars. [BNC written: non-academic, natural science]

But when the above examples are compared to extracts from non-scientific texts in the BNC, a difference in the use of this bundle becomes obvious:

(92) He had discussed with the parents *in the presence of* the plaintiff [...] [BNC spoken: courtroom]

(93) It is as if, while *in the presence of* a dead man, the poet is reverent and sad [...] [BNC written: essay, school]

(94) He did not want to bring her in to talk to him, nor did he want to interview her *in the presence of* her devoted but sharp-eyed husband. [BNC written: fiction]

Although *in the presence of* also functions in the non-scientific examples as a framing bundle, there is a clear difference between the scientific and non-scientific extracts with regard to meaning. In the first set of examples, the keyword *presence* is used to denote existence, while in the other set it signifies the attendance or appearance of a person. This observation and the high frequency of *in the presence of* in the HSC (ranking third most frequent with 541 occurrences) suggest the importance of this bundle as a formula for presenting the different elements involved in an experiment. They indicate a specialized use of *in the presence of* in scientific texts that is worth pointing out to language learners or novice writers with particular interest in the sciences.

All this highlights the importance of recognizing all attested functions of lexical bundles, regardless of their frequencies. In the present study, instead of determining a single function to be assigned to multifunctional bundles, bundles with multiple uses were assigned to multiple functions. This more comprehensive approach was made possible by the prior filtering process, which narrowed down the list of target bundles to a manageable number of individual types.

3.2. Distribution of target-bundle functions

The target bundles were classified according to a modified version of Hyland's (2008a) functional taxonomy, discussed in Chapter III, Section 3.4 above. This classification scheme made it possible not only to organize the lexical bundles based on their typical meanings and uses, but also to determine the extent to which each functional category is used in scientific writing, thereby gaining a better awareness of the particular concerns of this type of discourse.

Table 17 lists the functional categories with their respective type frequencies. Figures 4 and 6 illustrate the functional distribution of bundle types, while Figures 5 and 7 represent the functional distribution of prototypical bundle tokens.

Table 17. Functional classification of target bundles

FUNCTION	TYPES	%	TOKENS	%
Research-oriented bundles	**216**	**43%**	**10141**	**39%**
Location	22		774	
Procedure	111		5137	
Quantification	36		1906	
Description	28		1535	
Grouping	19		789	
Text-oriented bundles	**242**	**48%**	**13734**	**52%**
Additive	7		639	
Comparative	21		1113	
Inferential	67		3062	
Causative	23		1490	
Structuring	32		2402	
Framing	51		3094	
Citation	24		1166	
Generalization	4		145	
Objective	13		623	
Participant-oriented bundles	**48**	**9%**	**2348**	**9%**
Stance	36		1818	
Engagement	9		425	
Acknowledgement	3		105	
TOTAL	**506**	**100%**	**26223**	**100%**

Figure 4. Distribution of research-, text- and participant-oriented categories by type

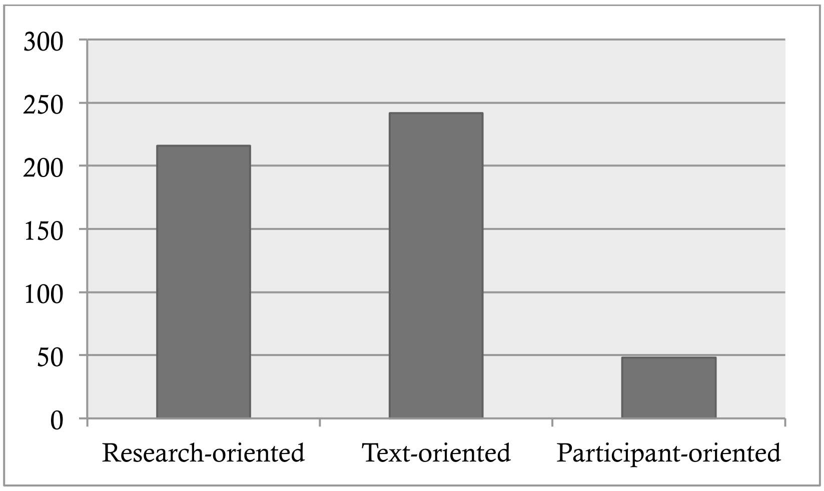
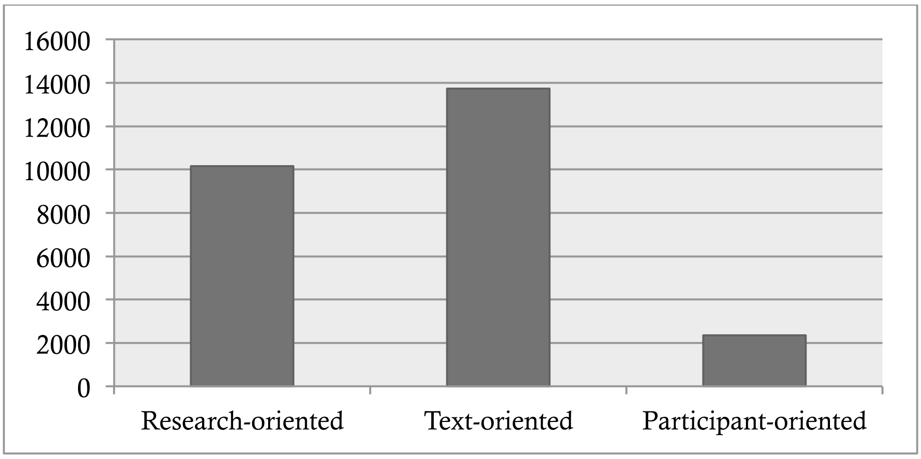


Figure 5. Distribution of research-, text- and participant-oriented functions by token



As can be observed, of the three main functional categories, text-oriented bundles are the most frequent, accounting for 48% of prototypical bundle types, with 242, and 52% of prototypical bundle tokens, with 13,734. Research-oriented bundles follow with 216 types (43%) and 10,141 tokens (39%). Participant-oriented bundles are the least frequently used, with 9% of both types (*n* = 48) and tokens (*n* = 2,348).

These numbers differ from Hyland's (2008) results, which show research-oriented bundles to be the predominant functional category in his science and technology corpora. This seeming contradiction can be explained by the decision to discard lexical bundles ending with articles from the present list of target bundles. Many of these disregarded bundles are noun phrase + *of* structures that fulfill research-oriented functions, and their elimination consequently reduced the number of this type of bundle. This notwithstanding, research-oriented bundles still occur with high enough frequency to be considered an important characteristic of scientific writing.

In fact, upon turning to the distribution of the more specific functional subcategories, it can be seen that research-oriented procedure bundles are the most common. A total of 111 types and 5,137 tokens have this function, representing 22% of all prototypical target-bundle types and 20% of all tokens. They are joined by four text-oriented functions: framing (51 types, 10%; 3,094 tokens, 12%), inferential (67 types, 13%; 3,062 tokens, 12%), structuring (32 types, 6%; 2,402 tokens, 9%) and causative (23 types, 5%; 1,490 tokens, 6%). Two research-oriented functions also place high on the frequency list: quantification, which accounts for 36 types (7%) and 1,906 tokens (7%) and description, which represents 28 types (6%) and 1,535 tokens (6%). Another frequently used category is that of participant-oriented stance bundles, with 36 types (7%) and 1,818 tokens (7%). The top eight most frequent functions account for more than 75% of all bundle types and tokens, a large part of the total.

Figure 6. Distribution of functional categories by subcategory

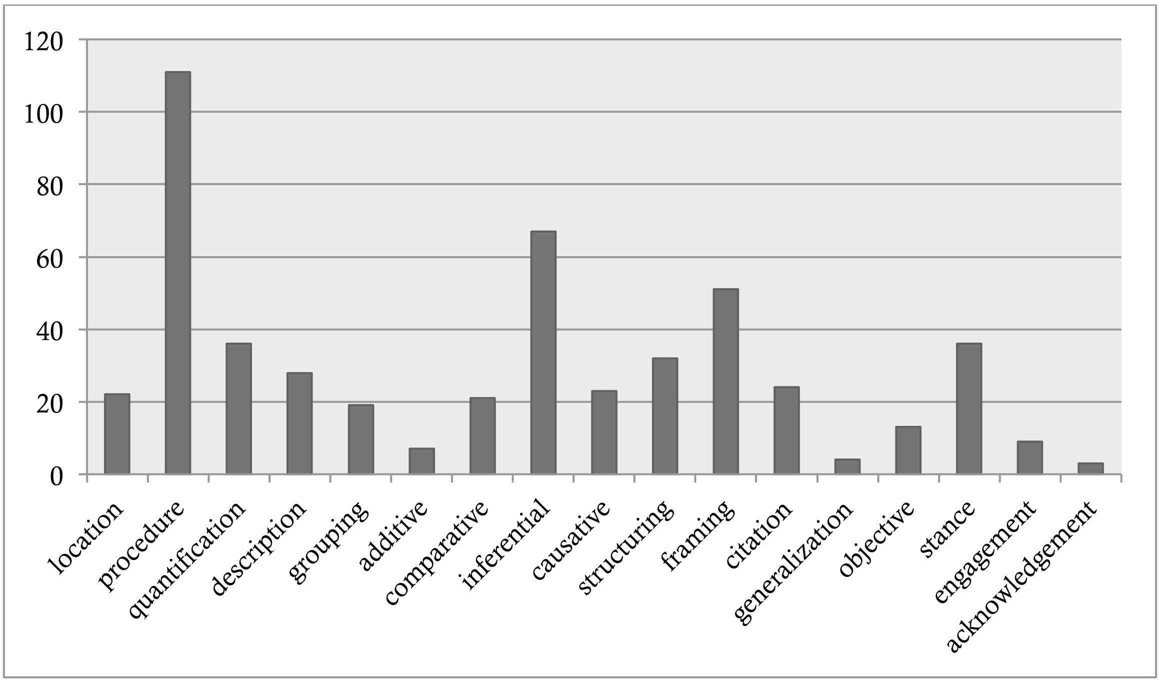
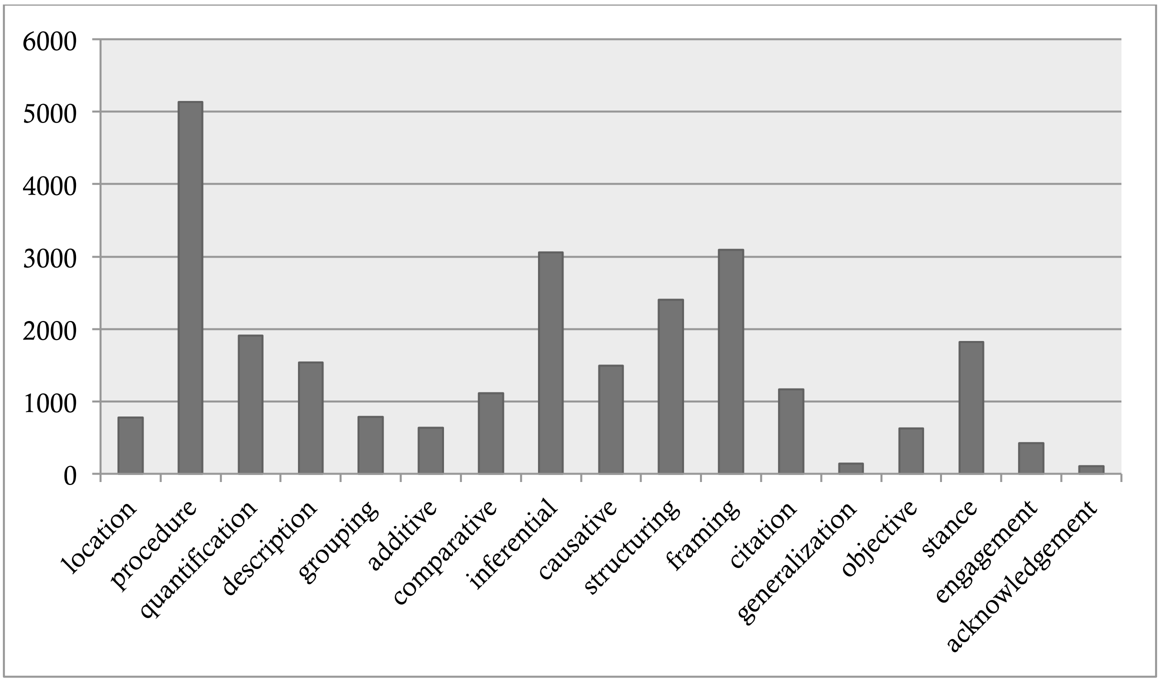


Figure 7. Distribution of functional categories by token



Research-oriented bundles

Table 18 shows all research-oriented bundles, including non-prototypical forms, by their alphabetically ordered keywords.

Table 18. Research-oriented bundles

Procedure	the accumulation of, the action of, the activities of, the activity of, was added to, were added to, the addition of, by the addition of, by addition of, were allowed to, the analysis of, was analyzed by, were analysed by, were analyzed by, the assembly of, was assessed by, the beginning of, carried out at, carried out in, carried out using, carried out with, was carried out, were carried out, were carried out at, a change in, were collected from, compared with control, a comparison of, was confirmed by, the control of, as a control, in the control, a deletion of, was detected by, were detected by, the detection of, was determined as, was determined by, was determined using, were determined by, activity was determined, the development of, was digested with, was dissolved in, the evolution of, was examined by, were exposed to, were fixed in, the formation of, in the formation of, was generated by, were generated by, the generation of, were grown at, were grown in, were grown to, the growth of, on ice for, the identification of, were identified by, the incorporation of, were incubated for, were incubated with, was induced by, to interact with, the interaction between, the interaction of, the interaction with, its interaction with, was introduced into, the introduction of, was isolated from, were isolated from, the isolation of, a loss of, the loss of, were made by, according to the manufacturer's, according to the manufacturer's instructions, the manufacturer's instructions, activity was measured, as measured by, was measured by, mechanism by which, the mechanism of, is mediated by, the method of, by the method, by the method of, was mixed with, was obtained by, was obtained from, were obtained by, were obtained from, the onset of, the organization of, the origin of, the pattern of, was performed by, was performed in, was performed on, was performed using, was performed with, were performed in, were performed using, were performed

	with, analysis was performed, experiments were performed, was prepared by, was prepared from, were prepared as, were prepared by, were prepared from, extracts were prepared, were processed for, the process of, the production of, for the production of, in the production of, was purchased from, were purchased from, was purified from, in the regulation of, the release of, the removal of, were removed by, was replaced with, was resuspended in, were resuspended in, were separated by, were separated on, were stained with, the study of, was subjected to, were subjected to, medium supplemented with, we tested whether, were tested for, tested for their ability to, were transferred to, were treated for, were treated with, by treatment with, by use of, with the use of, the use of, be used to, been used to, can be used, has been used, was used as, was used for, was used in, was used to, were used as, were used for, were used in, were used to, we have used, used in this study, used to amplify, used to determine, used to identify, were washed in, were washed three times, were washed twice with, were washed with
Quantification	for an additional, in the amount of, the amount of, the average of, an average of, a decrease in, a density of, at a density of, the efficiency of, a fraction of, the fraction of, the frequency of, an increase in, the increase in, increasing amounts of, of a large, the length of, little or no, the majority of, a large number of, large number of, a number of, a small number, small number of, in a number of, in the number of, of a number of, the number of, the total number of, total number of, a percentage of, as a percentage of, the percentage of, the proportion of, the rate of, the rates of, at a flow rate of, the ratio of, a reduction in, the reduction in, the size of, only a small, at room temperature, at room temperature for, the time of, a total of, in a total, of the total, for up to, the value of, an equal volume of, an equal volume, equal volume of, total volume of
Description	the ability of, the ability to, their ability to, for their ability to, its ability to, are able to, be able to, is able to, was able to, the absence of, to act as, the appearance of, the behavior of, is capable of, does not contain, a defect in, high degree of, the degree of, the existence of, the extent of, a function of, the function of, the identity of, the importance of, the inability of, the intensity of, the lack of, at the level of, high levels of, low levels of, the level of, the levels of, the nature of, the presence of, also present in, are present in, is present in, was present in, were present in, the properties of, the significance of, the stability of, the structure of, the timing of, are unable to, was unable to, were unable to
Location	in the bottom, the bottom of, in the dark, at the end of, the end of, on the left, to the left, the localization of, is localized to, the location of, the position of, the positions of, the region of, this region of, in the region, in this region, to the right, the site of, at the site, the surface of, at the surface, at the surface of, on the surface, on the surface of, the tip of, the top of, in the top, in the upper, in the vicinity of
Grouping	a combination of, the combination of, a component of, the distribution of, a family of, a member of, is a member of, other members of, a mixture of, as part of, a portion of, a range of, the range of, a wide range of, the remainder of, the rest of, the sequence of, a series of, a set of, a subset of, this type of, two types of, a variety of

As mentioned previously, bundles depicting experimental procedures and scientific phenomena make up most of the research-oriented target bundles found in the HSC. Procedure bundles are mostly past-tense passive structures that describe research activities and experimental techniques:

(95) RAPD markers *were generated by* polymerase chain reactions with 10-nucleotide DNA primers of arbitrary sequence and separated on agarose-Synergel gels (Diversified Biotech). [6]

(96) To identify proteins that might be involved in second and third chromosome telomeric gene silencing, a survey *was performed using* Drosophila stocks with mutations in known chromosomal proteins, exclusive of Su(var)s. [18]

(97) After a 15 min fixation at room temperature, cells *were washed twice with* 3.5 ml of PBS, then resuspended in 3.5 ml of 5% goat serum in PBS. [68]

There are also several noun phrase + *of* constructions that refer to specific events (98), actions (99) and methods (100).

(98) This possibility was investigated by measuring *the accumulation of* Cd2+ by control and TaPCS1 expressing cells grown at Cd2+ concentrations that do not significantly affect the growth of even the control cells. [15]

(99) The fractional contribution of gluconeogenesis to endogenous glucose production was determined from *the incorporation of* [2-13C1] glycerol into plasma glucose, using mass isotopomer distribution analysis to calculate the isotopic enrichment of the triose-phosphate precursor pool. [70]

(100) Here, we show that *the use of* adherent cells as targets for attachment assays is necessary to demonstrate the crucial role of CypA in HIV-1 attachment under ‘standard' washing procedures. [82]

Many of the noun phrase + *of* structures that denote concrete actions (102) (104) have passive-verb counterparts (101) (103).

(101) Heparin (5 U/mL), wild-type or mutant recombinant VIIa (10 nM), or factor Xa (10 nM) *was added to* the cell suspension before plating, as indicated. [28]

(102) After *the addition of* water (8 ml) to each tube, they were boiled for 5 min, then cooled to room temperature and the absorbance was read at 546 nm. [65]

(103) The tryptic cleavage sites *were identified by* NH2-terminal microsequencing. [81]

(104) Here we describe *the identification of* J-binding proteins from T.brucei and the related kinetoplastid parasites Crithidia fasciculata and Leishmania tarentolae that specifically bind double-stranded DNA containing J. [17]

As commented in Section 2 of this chapter, the large number of passives and noun phrases depicting research procedures, many of which are found in the Experimental, Materials and Methods and Methods sections of the biomedical research articles, suggest the great importance of this function in scientific writing. It shows the scientists' preoccupation for carefully relaying the various steps involved in research and experimentation. But the scientists' consistent use of depersonalized constructions such as passive verbs and noun phrases is also a sign of their efforts to document their research activity in the most objective way possible.

The rest of the research-oriented bundles (quantification, description, location and grouping) are typically realized by a wide variety of prepositional and noun phrases. Although they appear in smaller numbers, they still contribute to the accurate summation of the research process by identifying location (105) and orientation (106), specifying amounts (107), measurements (108) and proportions (109), and describing research objects, models, equipment and materials (110).

(105) Since core X is structurally similar to the E and I sites of HML and HMR, we might expect to see high levels of silencing *in this region*. [74]

(106) When TGF is depicted as a curve in the plane defined by VLP and SNGFR, this TGF adaptation is represented by a shift in the TGF curve that is upward and *to the right* (Figure 6). [96]

(107) There was marked variation *in the amount of* PrP detected; one case showed only minor diffuse PrP immunoreactivity limited to a small focal area of the thalamus, while another with an incubation period of only 1 day more had extensive PrP accumulation. [55]

(108) The dialysate was applied and reapplied four times at room temperature to a 4 ml prepared TALON metal-affinity column *at a flow rate of* 1 ml/min. [78]

(109) *The majority of* the ovules (42%) contained embryo sacs where the primary endosperm nucleus had divided once or twice, and in some the zygote had initiated the embryonic mitotic divisions. [43]

(110) *The properties of* this joint molecule suggested that it could be composed of a linear dsDNA molecule that was invaded by homologous ssDNA; the resultant joint molecule would resemble the letter K and, hence, is referred to as a Kappa intermediate. [36]

The widespread use of research-oriented bundles is a reflection of the fundamental concern of scientific research articles: that of giving an objective, unbiased and precise account of experimental procedures, so that the subsequent data interpretation can be established as verifiable, reproducible and grounded in empirical reality. This is in line with Hyland's (2008a) argument that

> [The] significantly greater use of research-oriented bundles in the hard knowledge fields also expresses something of a scientific ideology which emphasizes the empirical over the interpretive, minimizing the presence of researchers and contributing to the ''strong'' claims of the sciences. Highlighting research rather than its presentation places greater burden on research practices and the methods, procedures and equipment used, and this allows scientists to emphasize demonstrable generalizations

> rather than interpreting individuals. New knowledge, then, is accepted on the basis of empirical demonstration and experimental results designed to test hypotheses related to gaps in knowledge. (p. 15)

Text-oriented bundles

Table 19 presents all text-oriented bundles, including non-prototypical forms, by their alphabetically ordered keywords.

Table 19. Text-oriented bundles

Additive	in addition to, in combination with, alone or in*, in concert with, in conjunction with, on the other hand, at the same time, as well as, as well as in
Comparative	in agreement with, as compared with, when compared with, in comparison with, is consistent with, consistent with this, this is consistent with, consistent with previous, consistent with our, which is consistent with, are consistent with, results are consistent with, these results are consistent with, be consistent with, in contrast to, in contrast with, the difference in, the differences in, the difference between, significantly different from, on the other hand, to that observed**, as opposed to, similar results were obtained, similar results were, results were obtained, were obtained with, were obtained in, the same as, in the same, at the same, to the same, to that seen**, similar to that, similar to those, similar to that of, is similar to, are similar to, very similar to, was similar to, in a similar
Inferential	were able to, to account for, it appears that, appear to be, appears to be, appeared to be, not appear to, does not appear to, not appear to be, did not appear, did not appear to, is associated with, are associated with, was associated with, be associated with, to associate with, the association of, we conclude that, the conclusion that, results demonstrate that, we demonstrate that, have demonstrated that, has been demonstrated, was detected in, be detected in, can be detected, could be detected, were detected in, was not detected, as determined by, lines of evidence, several lines of evidence, no evidence for, exclude the possibility, the possibility of, be expected to, would be expected, would be expected to, expected to be, be explained by, found to be, was found to, were found to, was found to be, been found to, were found to be, was found in, we found that, we find that, we have found, have found that, have been found, is found in, are found in, the finding that, the hypothesis that, we have identified, have been identified in, have been identified, been identified in, been identified as, been implicated in, has been implicated, has been implicated in, have been implicated, this implies that, this indicates that, results indicate that, these results indicate that, data indicate that, these data indicate that, as indicated by, be involved in, is involved in, are involved in, to be involved in, as judged by, is likely to, likely to be, is likely to be, it is likely, are likely to, are likely to be, it is likely that, it seems likely that, the observation that, was observed in, has been observed, also observed in, been observed in, the possibility that, possibility is that, it is possible, it is possible that, is also possible, is predicted to, predicted to be, we propose that, closely related to, the relationship between, are representative of, the results presented, the results obtained, can be seen, as seen in, there are several, we show that, we have shown, we have shown that, here we show that, as shown by, been shown to, has been shown to, has been shown, shown to be, have been shown to, was shown to, has been shown to be, been shown previously, it has been shown that, it has been shown, it was shown, shown previously that, this suggests that, results suggest that, these results suggest, these results suggest that, data suggest that, taken together these***, these data suggest, these data suggest that, together these results, together these data, taken together these results, we suggest that, suggesting that this, is supported by, was supported by, in support of, in support of this, is thought to,

	thought to be, are thought to, is thought to be, we were unable to, is unlikely to, unlikely to be, it is unlikely
Causative	did not affect, does not affect, is caused by, be caused by, as a consequence, as a consequence of, a consequence of, may contribute to, is due to, be due to, was due to, may be due, not due to, no effect on, had no effect, had no effect on, the effect of, the effects of, be explained by, be involved in, is involved in, are involved in, to be involved in, the product of, the products of, in response to, a response to, is responsible for, be responsible for, are responsible for, as a result, the result of, be the result of, the results of, a result of, as a result of, would result in, not result in, the role of, a role in, play a role, play a role in, an important role, an important role in, important role in, an essential role, a critical role, a role for, by virtue of, the yield of
Structuring	as described previously, as described above, as previously described, described in the experimental section, as described in the experimental section, as described in materials and methods, was performed as described, were performed as described, essentially as described, carried out as, performed as described previously, were prepared as described, carried out as described, are described in, as described for, in these experiments, in this experiment, are expressed as, results are expressed, as shown in figure, is shown in figure, are shown in figure, shown in figure, are shown in, is shown in, shown in figure 1, shown in fig, shown in figure 2, described in figure, shown in figure 3, as in figure, in figure 1, in figure 2, in figure 5, in figure 3, in fig 1, in figure 7, were as follows, with the following, is indicated by, are indicated by, are indicated in, at the indicated, the indicated times, of the indicated, as indicated by, in this paper, expressed as a percentage of, in the present, in the present study, in the present study we, referred to as, in this report, in the materials and methods section, in the experimental section, for review see, for reviews see, see figure 1, see figure 2, see materials and methods, see table 1, data not shown, results not shown, data not shown in, as shown in, are shown as, in this study, in this study we, the present study, shown in table, are shown in, is shown in, summarized in table, are summarized in, are shown in table, in table 1, in table 2, in table 3, the present work
Framing	in the absence of, in the absence or presence of, in accordance with, is based on, was based on, the basis of, on the basis of, in the case of, in this case, in all cases, in each case, in some cases, in combination with, alone or in*, is composed of, in concert with, under these conditions, under the same conditions, in conjunction with, the context of, in the context of, under the control of, the course of, is dependent on, was dependent on, with the exception of, with the exception, the fact that, by the fact that, as a function of, the idea that, is independent of, in a manner, a model for, model in which, as a model, in this model, the notion that, in the presence of, in the presence or absence of, for the presence of, by the presence of, in this process, the question of, is required for, are required for, be required for, to be required, to be required for, was required for, also required for, that are required for, does not require, not required for, is not required, is not required for, is required to, be required to, the requirement for, a requirement for, with respect to, is sensitive to, there are several****, is subject to, is sufficient to, in terms of, at the same time, the same time, at the time, at various times, at this time
Citation	in accordance with, in agreement with, is consistent with, consistent with this, this is consistent with, consistent with previous, consistent with our, which is consistent with, are consistent with, results are consistent with, these results are consistent with, be consistent with, has been demonstrated, as described by, as described in, performed as described, was performed as, were performed as, prepared as described, was performed as described, were performed as described, essentially as described, carried out as, performed as described previously, were prepared as described, carried out as described, been described previously, have been described, has been described, are described in, as described for, have been found, found to be, have been identified in, have been identified, been identified in, been identified as, been implicated in, has been implicated, has been implicated in, have been implicated, it has been proposed that, has been proposed, been proposed that, been proposed to, has been reported, have been reported, been reported to, as reported previously, studies have shown that, have shown that, has shown that, previous studies have, a previous study, results show that, been shown to, has been shown to, has been shown, shown to be, have been shown to, was shown to, has been shown to be, it has been shown, been shown previously, it was shown, it has been shown that, shown previously that, have suggested that, it has been suggested, has been suggested, has been

	suggested that
Generalization	is found in, are found in, is known about, little is known about, little is known, known to be, is known to, are known to, is not known, is thought to, thought to be, are thought to, is thought to be
Objective	to account for, to address this, we asked whether, to confirm that, to demonstrate that, to determine whether, to determine if, to distinguish between, to ensure that, in order to, remains to be, remains to be determined, to be determined, to show that, to test whether, to test this, to test this hypothesis

LEGEND
* alone or in – combines with *in combination with* to form the additive and framing bundle *alone or in combination with*
** to that observed, to that seen – combine with adjectives such as *similar* to form comparative bundles such as *similar to that observed* or *similar to that seen*
*** taken together these – combines with nouns such as *data* and *results* and verbs such as *suggest* and *show* to form inferential bundles such as *taken together these results suggest*
**** there are several – combines with nouns such as aspects, mechanisms, explanations and reasons to form various framing bundles

Text-oriented functions are associated with nearly half of target-bundle types and tokens, making them the most widely represented of the three main functional categories. Hyland (2008a) considers text-oriented bundles as particularly characteristic of the more interpretative and less empiricist soft-knowledge fields such as applied linguistics and business studies, but the present findings demonstrate that they also play a central role in the discursive practice of scientific genres.

The results of this study agree with Hyland (2008a) in that there is a large concentration of resultative markers in biology writing, a category divided here into two separate categories: inferential and causative. Inferential bundles are heavily used by scientists to convey their interpretations of relevant data and to highlight the conclusions that both reader and writer can draw from the study (111) (112), while causative markers are employed to highlight cause-and-effect relationships (113).

(111) A proposed further stage in the duplication process *was found in* some cells where the duplication plaque was partly inserted into the nuclear membrane so that it appeared to be in direct contact with the nucleoplasm. [1]

(112) Indeed, addition of excess recombinant Scythe on its own never triggered apoptosis, *suggesting that this* excess Scythe could not adopt an activated C312-like pro-apoptotic conformation in the absence of Reaper. [97]

(113) After 4 h there is an increase in these long glycosaminoglycan chains in all three experimental conditions (Figure 3B), suggesting that during this portion of the chase period the primary loss of cell-associated heparan sulphate *is due to* shedding of cell-surface molecules. [100]

Another widespread group of bundles is that of framing signals, the most frequent function in the text-oriented category. These bundles are essential to the effective elaboration of arguments, as they enable science writers to establish connections (114), set conditions (115) and define limitations (116). Framing functions are usually performed by bundles with prepositional-phrase structures.

(114) We thus conclude that the mcm genes are indeed regulated *as a function of* cell growth and that they are also subject to control by E2F, coincident with the control of many other genes encoding DNA replication activities. [50]

(115) As protease protection assays provided evidence for the membrane topology of H,K-ATPase flu tags only *in the context of* Sf9 microsomes, we sought confirmation of this topology at the cellular level by immunocytochemical labelling of intact and permeabilized Sf9 cells. [92]

(116) The linker lacks secondary structure, *with the exception of* three residues (516-518) that form a short anti-parallel -sheet with three residues (636-638) from D-IV. [42]

Structuring bundles, on the other hand, work to facilitate comprehension by providing text-reflexive explanations (117) (118) and guiding readers through the text (119). These bundles usually take the form of adverbial-clause fragments and passive structures combined with prepositions.

(117) Carbohydrate and lipid oxidation *are expressed as* grams per min. [14]

(118) The location of peptide sequences used to raise antisera EL-1 and CT-1 *are indicated in* bold type and with asterisks. [8]

(119) Details of the individual incubations *are described in* the legends to the Figures and Tables. [14]

Several structuring signals refer the reader to ancillary data, such as tables and figures, which give numerical or graphical support to the case being put forward:

(120) The data *summarized in Table* 1 suggest that INK4a -ARF+/ mice are as susceptible to RCAS-EGFR*-induced gliomas as are INK4a -ARF/ mice. [41]

(121) *As shown in Figure* 7A, the phx3 line allowed significantly more growth of the normally avirulent Pst DC3000 (avrRpm1) than Ws-0. [64]

Another key text-oriented function is that of citation. Scientists rely on citation bundles to link their findings and interpretations to prior research, simultaneously providing evidential justification to their claims and situating their own work within the wider research context:

(122) *As reported previously* (Miller and Rose, 1998), when the rare wild-type cells with anaphase in the mother were examined, the cytoplasmic microtubules nearly always extended into the bud (90%, Table V). [62]

(123) Previous *studies have shown that* stimulation with IL-12+IL-2 augments both T and NK cell IFN- production. [34]

(124) *It has been proposed that* p34cdc2 acts as the timer for cytokinesis by regulating myosin II activity (Satterwhite and Pollard 1992). [87]

Citations are frequently realized by adverbial-clause fragments, as well as by a variety of passive structures, including anticipatory-*it* and *that*-clause constructions controlled by passive verbs.

Four other text-oriented categories (comparative, additive, objective, generalization) appear in smaller quantities than the other five. They nevertheless perform the important functions of comparing and contrasting elements (125), specifying research objectives (126), prefacing statements of general knowledge (127) and providing additive links between components (128).

(125) Based on comparisons of the reported confidence intervals, the maternal estimate for genome length of loblolly pine is *significantly different from* the maternal estimates reported by ECHT and NELSON 1997. [comparative] [83]

(126) *To demonstrate that* P-gp-N280C was expressed and could be labelled by BM, the same assay was carried out in the presence of saponin, a gentle membrane permeabilizing agent. [objective] [7]

(127) Since IGCs *are known to* contain pre-mRNA splicing factors, we were interested next in determining their presence in the purified IGC fraction as a means of further assessing its purity. [generalization] [63]

(128) *In addition to* influencing the process of endocytosis, the ent1ts alleles also affect the localization of the actin cytoskeleton, in particular at cytokinesis. [additive] [110]

It is clear from these results that scientists depend heavily on text-oriented bundles to lend coherence to their writing, using them to connect, clarify and contextualize their ideas. Through the use of these bundles, they are able to communicate their own interpretations of their data while alluding to related literature and visual and mathematical evidence that warrant their claims. Text-oriented bundles also allow them to ease their readers' processing of the article by creating logically structured arguments and providing well-placed textual signposts. All these functions combine to form the foundation of effective scientific argumentation.

Participant-oriented bundles

Table 20 displays all participant-oriented bundles by their alphabetically ordered keywords.

Table 20. Participant-oriented bundles

Stance	it appears that, appear to be, appears to be, appeared to be, not appear to, does not appear to, not appear to be, did not appear, did not appear to, is associated with, are associated with, was associated with, be associated with, to associate with, the association of, be caused by, it is clear, it is not clear, we conclude that, may contribute to, we demonstrate that, have demonstrated that, be detected in, can be detected, could be detected, be due to, may be due, not due to, are essential for, be expected to, would be expected, would be expected to, expected to be, we found that, we find that, we have found, have found that, we have identified, be important for, is likely to, likely to be, is likely to be, it is likely, are likely to, are likely to be, it is likely that, it seems likely that, it should be noted, it should be noted that, to note that, it is important to, the possibility that, possibility is that, it is possible, it is possible that, is also possible, we propose that, be required for, be required to, be responsible for, be the result of, would result in, an important role, an important role in, important role in, an essential role, a critical role, we show that, we have shown, we have shown that, here we show that, this suggests that, results suggest that, these results suggest, these results suggest that, data suggest that, taken together these *, these data suggest, these data suggest that, together these results, together these data, taken together these results, we suggest that, suggesting that this, we were unable to, is unlikely to, unlikely to be, it is unlikely
Engagement	is difficult to, is essential for, exclude the possibility, the possibility that **, the possibility of **, is important for, is an important, is necessary for, it should be noted, it should be noted that, to note that, it is important to, for review see, for reviews see, see figure 1, see figure 2, see table 1, see materials and methods, can be seen, as seen in
Acknowledgment	a gift from, kindly provided by, is supported by, was supported by, this work was***

LEGEND
* taken together these – combines with nouns such as *data* and *results* and verbs such as *suggest* and *show* to form stance bundles such as *taken together these results suggest*
** the possibility that, the possibility of – combine with the verb *exclude* to form the engagement bundles *exclude the possibility that* and *exclude the possibility of*
*** *this work was* – combines with *was supported by* to form the acknowledgment bundle *this work was supported by*

This last main functional category corresponds to the dialogic interaction between the participants in the text: the writer and the reader. By expressing epistemic, evaluative and directive meanings, participant-oriented bundles help writers convey their attitudes towards their assertions and establish the appropriate relationship with their reader (Hyland, 2005).

Cortes (2004), in her functional analysis of lexical bundles in published research writing in history and biology, noted that stance markers such as *are likely to be, is likely to be, it is possible that* and *the probability that the* figure much more prominently in biology than in history. This large-scale use of stance bundles was also found in the HSC, where a large proportion of participant-oriented bundles are comprised of sequences that function to express stance.

Stance markers are linguistic devices that carry meanings such as certainty (129), possibility (130), probability (131) and necessity (132), and as such, they are effective means for writers to communicate their own assessments of certain propositions and their degree of confidence in these claims.

(129) In both cases, *it is clear that* only DNA from the 5'-labeled top strand is utilized by RecA protein in the formation of joint molecules. [13]

(130) *It is possible that* there is only a small region near the IES where alternate use of a TA can occur. [57]

(131) However, by comparison to the bacterial system, *it seems likely that* a plant homologue of the bacterial TatC protein is also involved. [103]

(132) Dbp5p accumulated in the nuclei of several strains with mutations affecting proteins involved in nuclear transport, including components of the Ran/Gsp1p system (Gsp1p, Rna1p and Prp20p), which *are essential for* nuclear import and nuclear export. [39]

It should be noted, however, that most stance expressions are realized by impersonal structures such as adjective phrases and anticipatory *it* constructions:

(133) *It is unlikely*, however, that such late signaling is important for tooth development. [27]

(134) They are *likely to be* involved directly in catalysis. [12]

(135) We hypothesized that cell-surface sialylated Lewis x might *be important for* infection by HGE. [35]

These depersonalization strategies indicate the scientific writers' efforts to soften the expression of their attitudes and opinions by means of indirect forms. This indirectness is a way for writers to protect the face of their addressees and avoid demeaning, limiting or coercing them (O'Keeffe et al., 2007). This rhetorical choice is also important for objectivity, as it "reduces the writer's role as agent and interpreter and allows research to be presented as independent of any particular scientist" (Hyland, 2008a, p. 19).

It is also interesting to observe the link between stance bundles and text-oriented inferential bundles. Several bundles simultaneously perform these two functions. In some cases, particularly those bundles that incorporate the first-person plural pronoun *we,* the inferential meaning of the bundles makes for a direct expression of stance where writers claim full responsibility for their assertions:

(136) In this paper *we have identified* a second CRE within the G6Pase promoter which is involved in the induction of G6Pase gene transcription by both cAMP and glucocorticoids. [84]

(137) *Here we show that* DivIVA is targeted to division sites late in their assembly, after some MinCD-sensitive step requiring FtsZ and other division proteins has been passed. [56]

As can be seen from the above examples, this type of stance expression is used to introduce findings and conclusions, as a way for authors to emphasize their own contributions to their field of study.

More frequently, however, writers take a more indirect approach, voicing their interpretations through impersonal constructions:

(138) *This suggests that* the conformational states of the two dimers in a tetramer are independent of each other and that these conformational states are not static in nature. [61]

They also often take a more conciliatory stance, downplaying their confidence in their contentions:

(139) Triggering of this postulated checkpoint *would result in* a general disabling of the spermatids that derive from the error-containing spermatocytes. [59]

(140) Based on previous work, an infusion rate of 4 μg/h for purified porcine RLX or rhRLX *would be expected to* produce plasma levels of 20-40 ng/ml. [19]

(141) Molecular cloning of the p62, Arp11, p27, and p25 subunits reveals a number of features that *may contribute to* interactions with membranous and other cargoes, including a RING-finger like domain within p62 (a Neurospora Ropy-2 homologue) and the alkaline isoelectric points (pIs) of p62, Arp11, and p25. [24]

This constitutes a pragmatic concept called *hedging*, another important aspect of face-protection and politeness (O'Keeffe et al., 2007). Hedges serve to mitigate the illocutionary force of the accompanying statement (Holmes, 1984) by conveying a certain degree of uncertainty or caution. Hedging thus enables writers to show modesty and deference towards their readers, as well as protect themselves from challenge and rebuttal.

The other, much less frequently occurring participant-oriented category is even more strongly implicated in the engagement of the *reader-in-the-text*, which Thompson and Thetela (1995) define as the reader construed by the writer, one who gives the necessary responses and is actively involved in the construction of discourse. Engagement markers seek to involve readers in the developing argument by addressing them directly, requesting them to focus on certain points and to see things in a particular way, and thereby persuading them to adopt the writer's position, or at least consider it valid. Scientists routinely utilize modals of obligation and evaluative adjectives of necessity and importance to perform engagement functions. Note once again how these bundles, while taking the form of directives, are softened by an indirect approach:

(142) *It should be noted*, however, that our assay would not distinguish between transcriptional switching and reciprocal recombination events in which the active and inactive expression site exchange ends upstream of the markers. [58]

(143) In these disorders identification of epitopes recognized by CD4 T cells *is important for* understanding mechanisms of disease development (molecular mimicry, for example), for enhancing diagnosis and prediction, and also for the future development of peptide-based therapies and vaccines. [71]

(144) *It is important to note that* Vmw110 did not cause a complete disintegration of centromeres since they clearly could be stained with autoimmune and anti-CENP-B sera (Figure 2), even in the absence of detectable CENP-C (data not shown). [26]

Hyland (2008a) makes similar observations regarding his hard-science corpora and stresses the formulaic nature of engagement markers and how they contribute to the precision that characterizes scientific writing:

> The relatively substantial presence of these items in the hard science corpora reflects the fact that these disciplines place considerable emphasis on precision, particularly to ensure the accurate understanding of procedures and results. The more linear and problem-oriented approach to knowledge construction found in the sciences allows arguments to be formulated in highly standardized, almost shorthand, ways which presuppose a degree of theoretical knowledge and routine practices not possible in the soft fields. As a result, directives offer writers an economical and precise form of expression which cuts more immediately to the heart of technical arguments. (p. 19)

The final category of participant-oriented bundles is acknowledgement. Lexical bundles with this classification are used to thank individuals or entities for financial assistance (145) or the provision of experimental materials (146) (147) (Pecorari, 2009).

(145) This work *was supported by* grants from the National Health and Medical Council of Australia and the Flinders Medical Centre Foundation. [72]

(146) AM-3K (a marker associated with monocytes/macrophages) was *a gift from* K. Takahashi (Kumamoto University, Kumamoto, Japan). [99]

(147) Cycle sequencing of RT-PCR products was performed on gel-isolated DNA using the CircumVent Thermal Cycle Sequencing kit (New England Biolabs, Beverly, MA) or a test cycle sequencing kit *kindly provided by* Stratagen. [40]

The use of stance bundles for evaluation, depersonalization and hedging, and of engagement bundles for reader involvement and persuasion, constitute important rhetorical strategies that language learners should master to be able to write effective academic prose. However, as Byrd and Coxhead (2010) remark, learning to use these stance and engagement markers poses not only linguistic, but also cultural challenges to non-native writers. Several studies have shown non-native writers' difficulties with expressing their judgments and the expected degrees of qualification and certainty in their academic writing (Aijmer, 2002; Hyland & Milton, 1997; Neff & Bunce, 2006; Salazar, 2008; Salazar & Verdaguer, 2009). These and other studies link these difficulties to typological mismatches between the native and foreign language and cross-linguistic variation in accepted degrees of directness and conviction (Bloch & Chi, 1995; Bloor & Bloor, 1991; Mauranen, 1993). It is obvious, therefore, that explicit teaching of the linguistic and cultural dimensions of stance and engagement is needed for non-native writers to learn how to construct an appropriate authorial voice.

4. Concluding remarks

The results of the structural and functional analysis of the target bundles are significant for two reasons. First, they offer insights into the distinctive character of scientific writing by revealing the main concerns of science writers and the ways in which they construct their arguments and pursue their agenda. Second, as commented previously, the classification of lexical bundles into structural and functional groups give them face validity for teaching, proving their value as teaching items and showing certain aspects of their use that should be brought to the attention of non-native and/or novice writers.

As evidenced by the varying frequencies and patterns of use of the different functional categories, research-oriented bundles contribute to the precise description of research objects and procedures; text-oriented bundles serve to organize, link and contextualize textual elements to express the author's interpretations of research outcomes; and participant-oriented bundles establish a positive writer-reader relationship by manipulating the reader's overall opinion of the text's validity and the writer's competence. The skilled scientific writer judiciously uses all three main functions to produce an article whose convincing, well-structured arguments are based on relevant literature and sound data derived from accurately described scientific methods, written in an engaging, non-face-threatening manner that is accessible to its audience.

And yet it is also true that just knowing what one is expected to do in a scientific article is not enough. Non-native and/or novice writers' often limited linguistic resources usually hinder their ability to perform the functions expected in their academic production, as much as they may be aware of these expectations. However, the fact that many of these functions are routinely realized through lexical bundles, and that these bundles are strongly connected to specific structural patterns (e.g., noun phrase + *of* for research-oriented functions, prepositional-phrase fragments for text-oriented functions, anticipatory-*it* structures for participant-oriented functions) can facilitate the teaching and learning of essential scientific-writing strategies, thereby enhancing non-native and non-expert writers' repertoire and giving them a wider range of options.

Chapter VI

Target bundles in non-native expert scientific writing

In this chapter, the frequency and usage patterns of target bundles in the non-native expert corpus (hereinafter NNS) are analyzed and compared to the results obtained from the native corpus, in an effort to distinguish features specific to non-native production.

At this stage of the investigation, only prototypical bundles[5] are considered in both the quantitative and qualitative analyses in order to avoid the skewing that may be caused by the presence of repeated bundle fragments and embedded sequences, and also to limit the number of bundles for comparison to a more manageable amount.

Raw counts are used in describing the frequency patterns of target bundles within the non-native corpus. However, when making comparisons between this corpus and the much larger native corpus, relative frequencies per 100,000 words are also indicated, along with results of log-likelihood tests, computed using the UCREL log-likelihood calculator (http://ucrel.lancs.ac.uk/llwizard.html).

1. Frequency of target bundles in the non-native corpus

Of the 442 prototypical target bundles, 312 were identified in the NNS corpus. However, a closer look at their individual frequencies reveals that 92 out of these 312 items occur only once. These 92 bundles, combined with the 130 others not found in

[5] The prototypical bundles *the basis of, a consequence of, the context of* and *the presence of,* which can function as independent bundles but also form part of the longer bundles *on the basis of, as a consequence of, in the context of* and *in the presence of,* respectively, were excluded from this part of the investigation for the same reasons.

the NNS corpus, make up over 70% of the list of prototypical target bundles, indicating that the majority of these bundles appear once or not at all in the non-native texts.

The non-native articles also show a more restricted range of target-bundle use, with its top 100 most frequent items constituting almost 75% of all tokens, while twice that number of items is needed to reach the same proportion in the native texts. Thus, both frequency and range point to a narrower use of target bundles in the NNS corpus.

Table 21 presents the 20 most commonly used target bundles in the NNS corpus in order of frequency. *In the presence of* is the most frequently occurring bundle, with 67 instances. It is followed by *in order to* and *the number of*, which place second and third with 54 and 53 tokens respectively. It can be observed that only six out of the 442 prototypical target bundles occur 30 or more times in the non-native texts.

Table 21. Top 20 prototypical target bundles in the NNS corpus in order of frequency

RANK	LEXICAL BUNDLE	TOKENS
1	in the presence of	67
2	in order to	54
3	the number of	53
4	as well as	44
5	the effect of	37
6	on the other hand	30
7	were carried out	29
8	with respect to	24
9	in this study	24
10	was used to	24
11	in the absence of	22
12	in agreement with	22
13	were able to	21
14	the fact that	21
15	data not shown	20
16	the present study	20
17	an increase in	19
18	in response to	19
19	in the present	18
20	carried out with	17

Table 22 compares the 20 most common prototypical target bundles in the HSC and in the NNS corpus. Shown in bold are the nine pairs of bundles that are common to the two top 20 lists, among them the top six most frequent in the HSC: *data not shown, in the presence of, in the absence of, as well as, the number of* and *the effect of.* The bundles *was used to, in response to* and *the fact that* also rank among the most frequently occurring target sequences in both corpora. The rest of the bundles represent the differences between HSC and NNS frequency patterns. Many of the most common prototypical target bundles in one corpus failed to make it into the top 20 of the other and vice versa, suggesting instances of overuse and underuse in the non-native texts with respect to the native data.

Table 22. The 20 most common prototypical target bundles in HSC and NNS

Rank	HSC	Abs	Rel	NNS	Abs	Rel
1	**data not shown**	625	30.01	**in the presence of**	67	55.50
2	**in the presence of**	541	25.98	in order to	54	44.73
3	**in the absence of**	387	18.58	**the number of**	53	43.90
4	**as well as**	307	14.74	**as well as**	44	36.45
5	**the number of**	273	13.11	**the effect of**	37	30.65
6	**the effect of**	259	12.44	on the other hand	30	24.85
7	as described previously	244	11.72	were carried out	29	24.02
8	the ability of	237	11.38	with respect to	24	19.88
9	been shown to	209	10.04	in this study	24	19.88
10	is required for	194	9.32	**was used to**	24	19.88
11	**was used to**	190	9.12	**in the absence of**	22	18.22
12	**in response to**	189	9.08	in agreement with	22	18.22
13	the level of	168	8.07	were able to	21	17.40
14	it is possible	165	7.92	**the fact that**	21	17.40
15	the role of	164	7.88	**data not shown**	20	16.57
16	to determine whether	164	7.88	the present study	20	16.57
17	**the fact that**	158	7.59	an increase in	19	15.74
18	in addition to	154	7.40	**in response to**	19	15.74
19	is consistent with	154	7.40	in the present	18	14.91
20	the formation of	149	7.16	carried out with	17	14.08

The bundles with the highest statistically significant differences in frequency are displayed in Table 23. It can be seen that four of the most frequently occurring bundles in the HSC, *the effect of, the number of, in the presence of* and *as well as*, are used in greater amounts in the NNS corpus, even showing statistically significant levels of overuse.

Table 23. Examples of prototypical target bundles overused and underused in the NNS corpus

Overused Statistically significant overuse (at p < 0.01)	were able to, in agreement with, carried out with, were carried out, the effect of, the fact that, was found in, on the other hand, an increase in, by the method, the number of, in order to, in the presence of, with respect to, at the same, in this study, in the present, the present study, was used to, as well as
Underused Statistically significant underuse (at p < 0.01)	the ability of, in the absence or presence of, data not shown, as described previously, to determine whether, the function of, the localization of, the observation that, the possibility that, is required for, the requirement for, for review see, a role for, by use of

Further examination of the overused bundles indicates the non-native writers' excessive reliance on a handful of highly frequent bundles, to the detriment of less common bundles with similar meanings. For example, the non-native authors depend heavily on *in agreement with* to relate their findings to similar results in the literature (148), *carried out with* to describe experimental materials and equipment (149), *on the other hand* to introduce a statement that contrasts with the one immediately preceding (150) and *in order to* to preface a research objective (151).

(148) Moreover, our finding is *in agreement with* the results previously demonstrated by Docampo et. al. [4], who reported Ca2+ release after addition of NH4+ and nigericin to the Fura 2-loaded T. cruzi epimastigotes. NNS033

(149) Ground state absorption measurements were *carried out with* a Hewlett Packard 8452^{a} diode array spectrophotometer. NNS006

(150) Isoforms with Rf 0.23 and 0.51 were mainly detected in the seed coat extracts, being the isoform 0.51 specific of this tissue. *On the other hand*,

isoform with Rf 0.07 (lane E-E) seems to be only detected in the embryo plus endosperm extracts (Fig. 2 A). NNS003

(151) *In order to* identify the sequence responsible of acid phosphatase activity, an insertional mutagenesis approach was employed by using the transposon Tn5::751. NNS018

These bundles, although also frequently recurring in the HSC, are not used quite as often by the native writers, who tend to employ other bundles that perform similar functions (italicized and underlined in the examples):

(152) Our finding of an additional cytoplasmic pool of KIAA0017 *is consistent with* a recent report by Watanabe et al. (1999) showing that the UV-DDB protein also interacts with the cytoplasmic domain of the Alzheimer's amyloid precursor protein (APP) in co-immunoprecipitation experiments. [63]

(153) Chromatography was *carried out using* MonoS, MonoQ and heparin-agarose columns on an FPLC system, whereas gel filtration *was performed using* a Superdex-75 column and a SMART system (Pharmacia Biotech, Uppsala, Sweden). [49]

(154) Since Dcp2p was required for both deadenylation-independent and -dependent decapping, and no 5’ to 3’ decay products were observed in dcp2 strains (Figure 3), we conclude that Dcp2p, like the DCP1 decapping enzyme (Beelman et al., 1996), is required for all mRNA decapping in vivo and is therefore likely to be a critical component of the mRNA decay machinery. This is *in contrast to* other proteins such as Mrt1p, Mrt3p and Spb8p, which affect the efficiency of decapping, but are not absolutely required for decapping (Hatfield et al., 1996; Boeck et al., 1998). [23]

(155) *To determine whether* C#-Cer stimulates PGP synthase directly, mitochondrial fractions from control H9c2 cells were prepared and PGP synthase activity was assayed in these fractions in the presence of 0±1000 lM C#-Cer. [112]

Notice especially that in (152), the writer chose the bundle *was performed using* so as not to repeat *carried out using*, which already appears in the previous clause. In (155), the author avoided *in order to* altogether and simply used the infinitive form of the verb in sentence-initial position. Alternative bundles such as those exemplified above seem to complement the use of their more frequently occurring counterparts, helping native writers achieve more variety of expression. However, these alternative phrases were found to be very rarely used in the non-native texts.

Several studies have reported that non-native writers make less frequent use of phraseological items in comparison to native speakers, with the exception of a few high-frequency expressions that they tend to overuse (Cortes, 2004; Granger, 1998; Howarth, 1996a; Kaszubski, 2000; Nesselhauf, 2005). Kaszubski (2000), for instance, attributes his findings to learners' tendency to go for the safest lexical options, labeled *lexical teddy bears* by Hasselgren (1994). What can be observed in the NNS corpus is a disproportionate use of a limited set of *phraseological teddy bears*, using Granger and Meunier's (2008b) paraphrase of Hasselgren's term, combined with the underuse of other possible alternatives for them. These patterns of overuse and underuse can contribute to a certain degree of repetitiveness and lack of stylistic variety in non-native writing.

2. Structural characteristics of target bundles in the non-native corpus

Table 24 provides a summary of the structural features of the prototypical target bundles identified in the non-native texts and their corresponding frequencies, while Figures 8 and 9 show the distribution of the different structural types and tokens.

Table 24. Frequency of structural categories of prototypical target bundles in the NNS corpus

STRUCTURE	TYPES	%	TOKENS		%
			ABS	REL	
Noun structures					
Noun phrase + *of*-phrase fragment	87	28%	518	429.10	29%
Noun phrase with other post-modifier fragment	12	4%	51	42.25	3%
Other noun phrase	8	2%	49	40.59	2%
Verb structures					
Passive + prepositional-phrase fragment	61	20%	299	247.68	17%
Other passive fragment	13	4%	95	78.70	5%
Verb phrase with personal pronoun *we*	7	2%	10	8.28	1%
Other verbal fragment	4	1%	5	4.14	1%
Prepositional-phrase fragments					
Prepositional phrase + *of*	21	7%	180	149.11	10%
Other prepositional phrase (fragment)	36	12%	300	248.51	17%
Other structures					
Verb or adjective *to*-clause fragment	14	4%	50	41.42	3%
Verb phrase or noun phrase + *that*-clause fragment	12	4%	50	41.42	3%
Adverbial-clause fragment	10	3%	28	23.19	1%
Copula *be* + adjective phrase	11	4%	26	21.54	1%
Other adjectival phrase	5	2%	15	12.43	1%
Anticipatory *it* + verb or adjectival phrase	8	2%	15	12.43	1%
Other expression	3	1%	99	82.01	5%
TOTAL	**312**	**100%**	**1790**	**1482.79**	**100%**

Figure 8. Structural categories of prototypical target bundles in the NNS corpus: Distribution by type

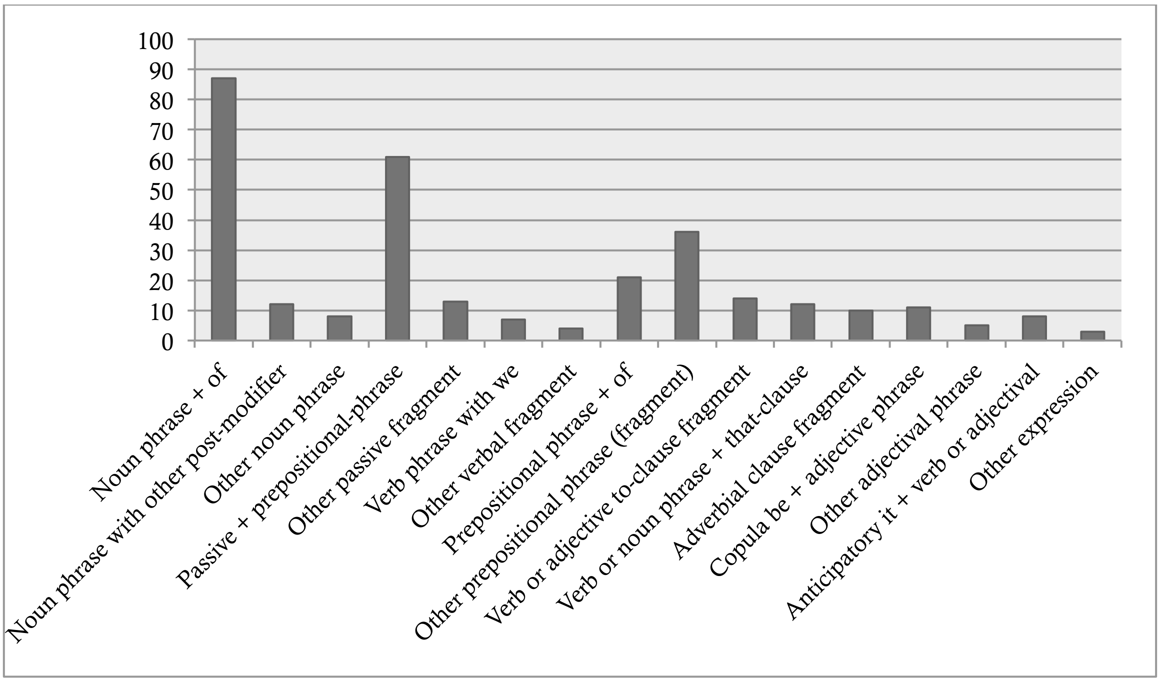


Figure 9. Structural categories of prototypical target bundles in the NNS corpus: Distribution by token

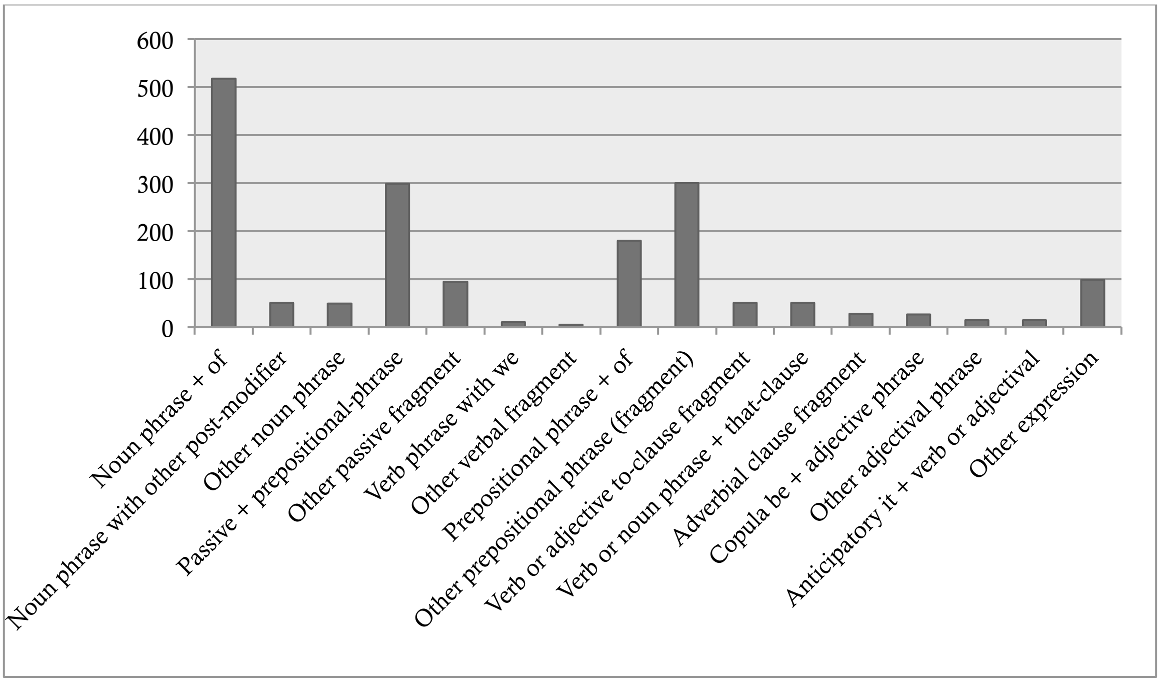


Table 25 compares the native and non-native corpora in terms of the absolute and relative frequencies of the different structural categories and displays the corresponding log-likelihood scores.

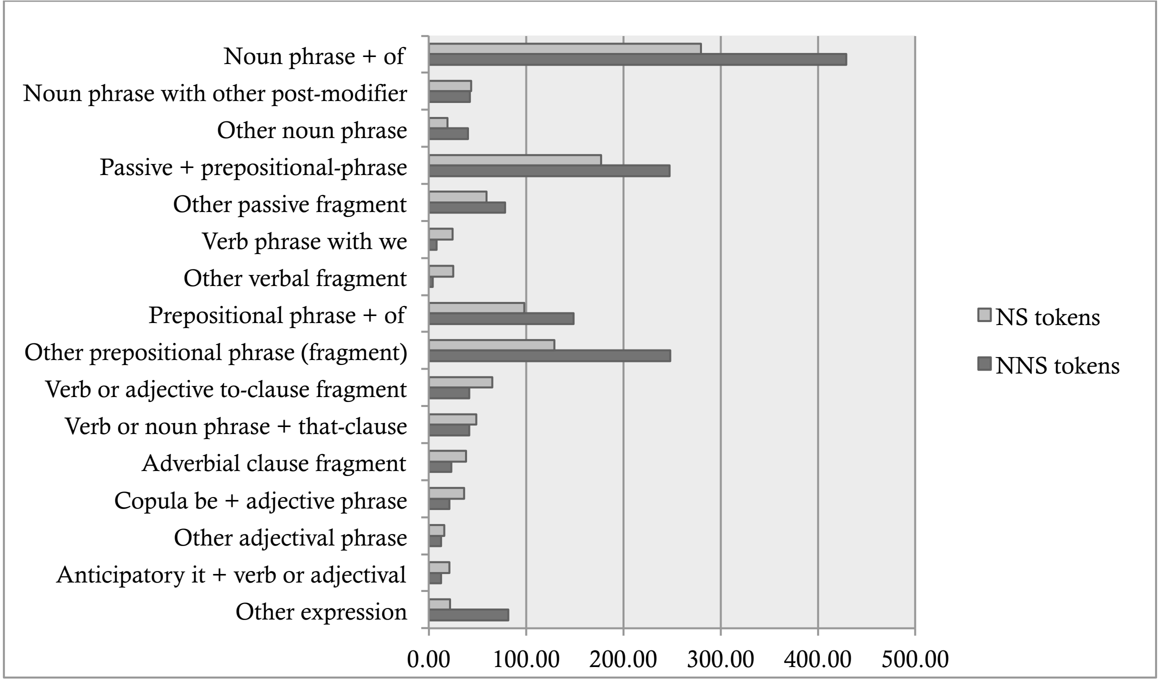
Figure 10 illustrates the relative frequencies of prototypical bundle tokens for each structural category in both corpora.

Table 25. Frequency of structural categories of prototypical target bundles in HSC and NNS

Structure	HSC		NNS		LogL
	Abs	Rel	Abs	Rel	
Noun structures					
Noun phrase + *of*-phrase fragment	5828	279.87	518	429.10	77.24 (++)
Noun phrase with other post-modifier fragment	915	43.94	51	42.25	0.08
Other noun phrase	408	19.59	49	40.59	19.22 (++)
Verb structures					
Passive + prepositional-phrase fragment	3695	177.44	299	247.68	28.03 (++)
Other passive fragment	1234	59.26	95	78.70	6.55 (+)
Verb phrase with personal pronoun *we*	513	24.63	10	8.28	16.95 (--)
Other verbal fragment	522	25.07	5	4.14	31.34 (--)
Prepositional-phrase fragments					
Prepositional phrase + *of*	2041	98.01	180	149.11	25.94 (++)
Other prepositional phrase (fragment)	2689	129.13	300	248.51	97.39 (++)
Other structures					
Verb or adjective *to*-clause fragment	1360	65.31	50	41.42	11.56 (--)
Verb phrase or noun phrase + *that*-clause fragment	1016	48.79	50	41.42	1.34
Adverbial-clause fragment	804	38.61	28	23.19	8.27 (--)
Copula *be* + adjective phrase	753	36.16	26	21.54	7.97 (--)
Other adjectival phrase	335	16.09	15	12.43	1.04
Anticipatory *it* + verb or adjectival phrase	439	21.08	15	12.43	4.80 (-)
Other expression	457	21.95	99	82.01	105.63 (++)
Total	**23009**	**1104.92**	**1790**	**1482.79**	**132.25 (++)**

LEGEND
(--) Statistically significant underuse in NNS (at p < 0.01, critical value 6.63) (-) Statistically significant underuse in NNS (at p < 0.05, critical value 3.84) (++) Statistically significant overuse in NNS (at p < 0.01, critical value 6.63) (+) Statistically significant overuse in NNS (at p < 0.05, critical value 3.84)

Figure 10. Distribution of structural categories of prototypical target bundles in HSC and NNS



As illustrated by the figures above, the prototypical target bundles in the NNS corpus follow the same structural distribution as the prototypical target bundles in the HSC, with noun-phrase + *of* structures, passive-verb fragments and parts of prepositional phrases surpassing all other structural correlates in frequency. One remarkable difference between the two corpora is the apparent overuse in the NNS corpus of these highly frequent structures, coupled with the underuse of adjectival phrases, anticipatory-*it* structures and *to*-clause, *that*-clause and adverbial-clause fragments, constructions that are of relatively low frequency in the HSC. This finding provides further evidence of non-natives' overuse of commonly used lexical sequences and underuse of comparatively less frequent strings, which in this particular case limits the structural diversity of target bundles in the non-native articles.

Table 26 lists all prototypical target bundles found in the non-native texts by their alphabetically ordered keywords.

Table 26. Prototypical target bundles found in the NNS corpus, grouped by structure

NOUN STRUCTURES	
Noun phrase + *of*-phrase fragment	the ability of, the accumulation of, the action of, the activity of, the analysis of, the appearance of, the assembly of, the beginning of, the behavior of, a combination of, the control of, the course of, the degree of, the detection of, the development of, the distribution of, the effect of, the efficiency of, the evolution of, the existence of, the extent of, the formation of, a fraction of, the fraction of, the frequency of, the generation of, the growth of, the identification of, the importance of, the inability of, the incorporation of, the intensity of, the interaction of, the lack of, the level of, the location of, the loss of, the majority of, the mechanism of, a member of, the method of, a mixture of, the nature of, a large number of, the number of, total number of, the pattern of, a percentage of, the percentage of, the position of, the process of, the product of, the production of, the properties of, the proportion of, a range of, the range of, the rate of, the ratio of, the region of, the release of, the remainder of, the removal of, the rest of, the result of, the role of, the sequence of, a series of, a set of, the significance of, the site of, the size of, the stability of, the structure of, the study of, a subset of, the time of, the timing of, the tip of, the top of, a total of, this type of, two types of, the use of, the value of, a variety of, the yield of
Noun phrase with other post-modifier fragment	a change in, a decrease in, the difference in, the difference between, no effect on, a gift from, an increase in, model in which, a reduction in, the relationship between, a response to, a role in
Other noun phrase	the ability to, lines of evidence, mechanism by which, a small number, the results presented, the results obtained, the present study, the present work
VERB STRUCTURES	
Passive + prepositional-phrase fragment	was added to, were analyzed by, is associated with, is based on, carried out at, carried out in, carried out with, is caused by, were collected from, was confirmed by, was detected by, was detected in, was determined as, was determined by, was digested with, was dissolved in, was examined by, be explained by, were exposed to, are expressed as, is shown in figure, is found in, was found in, were fixed in, were generated by, were grown at, were grown in, been implicated in, were incubated for, were incubated with, was induced by, be involved in, were isolated from, is known about, were made by, was measured by, was mixed with, was observed in, was obtained by, were obtained from, was performed by, were performed in, were purchased from, referred to as, was replaced with, is required for, were separated by, were separated on, are shown as, were stained with, were subjected to, is supported by, shown in table, were tested for, were transferred to, were treated with, was used to, was used as, was used for, were used in, were washed with
Other passive fragment	were allowed to, were carried out, has been demonstrated, was not detected, at the indicated, is not known, activity was measured, was performed using, analysis was performed, has been reported, similar results were obtained, can be seen, data not shown
Verb phrase with personal pronoun *we*	we conclude that, we found that, we have identified, we propose that, we show that, we suggest that, we were unable to
Other verbal fragment	may contribute to, not result in, see materials and methods, suggesting that this
PREPOSITIONAL-PHRASE FRAGMENTS	
Prepositional phrase + *of*	in the absence of, by the addition of, in the amount of, on the basis of, in the case of, as a consequence of, in the context of, at the end of, with the exception of, as a function of, in a number of, as part of, in the presence of, for the presence of, by the presence of, at a flow rate of, in the regulation of, as a result of, in support of, on the surface of, in terms of
Other prepositional phrase (fragment)	in accordance with, in addition to, in agreement with, in this case, in all cases, in some cases, in comparison with, under these conditions, under the same conditions, in contrast to, in the control, in the dark, in these experiments, with the following, on the other hand, by the method, in this paper, in the present, in the region, with respect to, in response to, in the same , at the same, to the same, in a similar, at the site, in this study, at the surface, at room temperature, at the same time, at the time, at various times, at this time, of the total, for up to, in the upper
OTHER STRUCTURES	
Verb or adjective *to*-	is able to, were able to, to account for, appear to be, to determine whether, to

clause fragment	ensure that, expected to be, found to be, is likely to, to note that, been proposed to, been shown to, to show that, were unable to
Verb phrase or noun phrase + *that*-clause fragment	the fact that, the hypothesis that, the idea that, this implies that, this indicates that, results indicate that, the possibility that, studies have shown that, results show that, this suggests that, results suggest that, have suggested that
Adverbial-clause fragment	as compared with, as described previously, as described by, as determined by, as shown in figure, were as follows, as indicated by, as judged by, as seen in, as shown in
Copula *be* + adjective phrase	is consistent with, are consistent with, is dependent on, is difficult to, is due to, is essential for, is an important, is independent of, is responsible for, is sensitive to, is subject to
Anticipatory *it* + verb or adjectival phrase	it appears that, it is not clear, it is likely that, it should be noted, it is possible, it has been proposed that, it has been shown that, it has been suggested
Other adjectival phrase	significantly different from, is present in, closely related to, the same as, similar to that
Other expression	in order to, there are several, as well as

Noun structures

Just like the HSC, the NNS corpus is dominated by noun structures, which account for 35% of all target-bundle tokens and types, and particularly by noun phrases featuring an *of*-phrase fragment. This structure comprises almost 30% of all target-bundle tokens and types, a larger percentage than that represented by the same structure in the native texts.

Noun constructions in the non-native articles include 92 different keywords, which are used to convey a wide variety of meanings similar to those found in the native papers:

(156) The lost of trichothecens production does not affect *the ability of* an isolate to infect wheat or maize, but it does affect the infection progression.n-decane. [quality] NNS032

(157) When Puerto Madryn was excluded from the analysis, *the degree of* differentiation was similar (Fst = 0.08; p<0.01) but the Mantel test was not significant. [degree] NNS035

(158) In some cases, *the existence of* an intramolecular hydrogen bond in the ground state, prevent photodegradation (upon direct photoirradiantion)

through a very fast intramolecular proton transfer (or H-atom transfer) between the OH group and the carbonyl group in the excited state, producing an excited phototautomer that is rapidly deactivated by thermal relaxation1-4, as shown in scheme I for the typical phenolic-type stabilizer compound methyl salicylate. [existence] NNS002

(159) Under these experimental conditions, any change in radioactivity should be interpreted as *a change in* mass for all phospholipids, since these lipids has not attained isotopic equilibrium. [event] NNS024

(160) Both interaction sites for phosphorylcholine (napp value of 1.8) were also detected by measuring *the production of* p-nitrophenol in the presence of saturating concentrations of p-NPP and variable concentrations of phosphorylcholine (Fig. 6). [action] NNS018

(161) *The rate of* oxygen consumption was greatly reduced in the comparative aerobic irradiations of: (a) A solution of Rf, but in the absence of F, (b) <The mixture Rf + CHN or FNN + 14 (g/ml SOD> (rate greatly reduced). [measurement] NNS016

(162) *The number of* replicates was appropriate since the species is apomictic and does not present intracultivar genetic variation. [quantity] NNS004

(163) On the other hand, Herman et al. (1994) demonstrated that only 27% of bacteria isolated in nearly N-free medium had the ability to fix nitrogen and *the majority of* the strains were efficient scavengers of nitrogen rather than nitrogen fixers. [proportion] NNS040

(164) A large shallow pond 1500 m away was *the site of* amphibian reproductive activity during the rainy season. [location] NNS015

(165) These sites are organized by *a set of* selected biological, physical and chemical characteristics (Reynoldson et al. 1997c.), and are used to compare with an impacted site to be assessed. [grouping] NNS023

Verb structures

Also consistent with the HSC is the frequency of verb structures in the NNS corpus, which account for 27% of all prototypical target-bundle tokens and 23% of all types. Most of these verb structures are passive expressions with a verb in the present or past tense, usually followed by a prepositional-phrase fragment. Those that include a present-verb are used to refer to tabular and graphical data (166), or to provide causal (167) or logical (168) justification for an argument.

(166) The ClustalX (33) multiple alignment for the six homologue proteins *is shown in Figure* 3. NNS018

(167) The hypersensitive phenotype *is caused by* the mutation of one gene; thus, the fact that mutants are genetically identical to the wild type with the exception of one mutated gene facilitates the study of key processes without the problem that suppose to study two cultivars with high genetic variability. NNS029

(168) This conclusion *is supported by* the fact that when Biodac plus Trichoderma species was incorporated the effect was reverted. NNS025

However, similar to the native texts, the majority of passive structures in the non-native texts incorporate a past-tense activity verb that describe scientific processes and procedures:

(169) To compare the independent variables (agricultural practices) with dependent variables, data for fungal populations *were analyzed by* ANOVA, followed by Duncan Multiple Range Test. NNS042

(170) PCR product *was digested with* endonuclease AluI (Fig 2A). NNS040

(171) Field works *were performed in* a same hour band from 10:00 PM to 13:00 PM. NNS039

(172) It is interesting to note that the cv. HF clearly modified its JA endogenous content when plants *were treated with* NaCL. NNS021

The analysis of verb structures in the HSC showed that the highly frequent use of passive bundles is complemented by a relatively lower occurrence of bundles that combine a verb with the personal pronoun *we*. This finding is indicative of the native authors' strategic use of active and passive structures, as well as personal and impersonal forms, in the construction of a convincing argument. Passive bundles are employed in the discussion of research methods and logical reasoning, so as to depersonalize these statements and make them sound more objective and universal. The *we* + verb combination, on the other hand, is a personal structure used by professional writers to directly associate themselves to their objectives, observations, achievements and conclusions, as a means of establishing their authority as researchers and promoting themselves as original, significant contributors to their discipline.

This balanced use of two contrasting forms was not observed in the NNS corpus, where there are considerably less *we* + verb bundles, in terms of both type and token. Of the ten types of *we* + verb bundles identified in the native texts, one has three occurrences (*we show that*), another has two occurrences (*we found that*), five appear only once (*we conclude that, we have identified, we propose that, we suggest that, we were unable to*) and three do not occur at all in the non-native texts (*we tested whether, we asked whether, we demonstrate that*). A statistically significant underuse of the personal pronoun *we* together with any other verb was also found throughout the whole corpus (see Table 27).

Table 27. We + verb constructions in the native and non-native corpora

	HSC		NNS		LL
	Abs	Rel	Abs	Rel	
We + verb	7669	368.28	160	132.54	232.03 (--)

LEGEND (--) Statistically significant underuse in NNS (at p < 0.01, critical value 6.63)

Of the 160 instances of *we* with an accompanying verb in the whole NNS corpus, 128 are used by the non-native authors to talk about what they have done and observed:

(173) *We were able to* isolate and identified 134 F. graminearum strains, 52, 56 and 26 from San Antonio de Areco, Alberti and Marcos Juarez respectively. NNS032

(174) On this basis, and trying to understand the role of the peptide bond in systems structurally more complex, *we carried out* in this paper a comparative study on the photodynamic action in the series tyrosine; tyrosil-tyrosine and tyrosyl-tyrosyl-tyrosine, including the methyl esters of tyr and the tripeptide (in the following all tyrosine derivatives, and the non-substituted AA will be generically named as TyrD). NNS006

(175) *We found* mainly three peroxidase isoenzymes with pI 4.8;6.3 and 9.6 (fig 1 and 2) while in crude extracts from field grown roots *we detect* mainly acidic isoperoxidases (fig.3). NNS014

The verb *find*, which is widely used by native writers with *we* and a *that*-clause to form the bundle *we find that* (176), is rarely used in this way by their non-native counterparts, who tend to employ *we find* with noun phrases (177).

(176) Here, we report that Drosophila CBP loss-of-function mutants show specific defects which mimic those seen in mutants that lack the extracellular signal Dpp or its effector Mad. Furthermore, *we find that* CBP

loss severely compromises the ability of Dpp target enhancers to respond to endogenous or exogenous Dpp. [104]

(177) *We found* a sigmoidal kinetic behaviour and a high IIA affinity in cationic isoform, that are in agreement with the main role of IAA oxidation atributed to cationic isoperoxidases by Gaspar (1986). NNS001

The verb *demonstrate* is also used differently by the native and non-native authors: the former use *demonstrate* in the present tense with *we* and a *that*-clause, forming the bundle *we demonstrate that*, to underline an important result obtained from the study being reported (178), while the latter use *demonstrate* in the past tense to refer to findings discussed in previous studies (179).

(178) *We demonstrate that* monocyte-derived CD14+ macrophages, but not monocyte derived CD83+ dendritic cells, endogenously express FasL, and that HIV infection mediated upregulation of FasL protein expression is independent of posttranslational mechanisms. [20]

(179) *We demonstrated* in previous works (29, 35), the photochemical production of the species Rf(- in several systems. It is known that under aerobic conditions Rf(-, in a subsequent step, produces the radical anion O2(-, with a reported rate constant value of 1.4x108 M-1s-1 for process 13. NNS031

Of the few *we* + verb constructions used by the non-native writers to preface their conclusions, some constitute modal and lexical-verb combinations unattested in the HSC:

(180) As a conclusion *we can say* that: the Eos or RB photosensitized oxidation of small peptides of tyr mainly occurs though a <( )>-mediated process, with the participation of an intermediate complex with polar character. NNS006

(181) Nevertheless, in a first approach *we can suggest* two main reaction pathways: they are the aerobic process and a group of other interactions that could operate in the absence of dissolved oxygen. NNS013

(182) Therefore, *we could infer* that the osmotic adjustment achieved at -0.4 and -0.8 MPa NaCl in germinating seeds could be attributed to sodium ions rather than chloride ions. NNS010

(183) Because of the experimental conditions in our study in C. venustus, where only the age or cohort differenced the individuals, *we supposed that* the physiological conditions associated with age would not be necessary and sufficient factors that caused the cohort different mortality in the field between the end of a breeding period and the beginning of the next one. NNS041

The evidence described above points to a difference between native and non-native texts, not only in frequency, but also in the usage patterns of *we* + verb constructions. It appears that the non-native authors are comfortable with using highly personal structures for self-citation, to allude to research they have previously carried out, but the same cannot be said for the use of the *we* + verb form for signaling ownership of the results and conclusions being presented.

This reluctance on the part of non-native scientific writers to assume direct responsibility for their claims using personal pronouns can be linked to the traditional view of academic prose as being distant, objective and highly impersonal, a view that is drilled into non-native and novice academic writers' minds by a number of writing manuals and style guides (Harwood, 2005a). Only recently has it been recognized that academic writing need not be totally author evacuated, that a certain degree of writer visibility is required for several important functions in research-oriented texts (Harwood, 2005a, 2005b; Hyland, 2001). One such function is foregrounding the significance and uniqueness of the work, and attributing it to the author in a self-promotional fashion. However, some studies have indicated that unlike published scientific writers such as those examined by Hyland (2001), non-native novice writers do not often use personal pronouns for highlighting original ideas (Tang & John, 1999) and non-native students at advanced levels are hesitant to use features such as first-person pronouns in academic writing, as they consider their use to be exclusive to more established scholars (Chang & Swales, 1999).

The underuse of *we* + verb sequences found in the present study's NNS corpus seem to be in contrast to the findings of some corpus-based studies of author visibility in learner academic writing, which demonstrate an overuse of personal forms in learner writing compared to native writing (Gilquin & Paquot, 2007; McCrostie, 2008; Petch-Tyson, 1998; Salazar, 2008). These studies also argue that the overabundance of personal features contributes to the speech-like quality of learner academic written production.

Far from being in contradiction, these results actually complement each other. They prove that when used thoughtfully and effectively, as professional scientific writers do, personal features such as *we* + verb bundles create a positive impression of an author who has "a confident and expert mind in full control of the material, making judgments and passing comment on issues of concern to the discipline" (Hyland, 2000, p. 123). When used excessively and in the wrong contexts, as non-native and/or novice writers tend to do, the same features can become manifestations of inexperienced writers' unfamiliarity with genre and register conventions (McCrostie, 2008). The focus, therefore, should not be on which forms to completely avoid—passive or active, impersonal or personal—but rather on stylistic and pragmatic appropriacy. Language learners and novice writers alike should learn which structures should be used with which words and in which contexts, and it is in this aspect that the phraseological approach is of particular benefit.

It must also be borne in mind that the non-native writers examined in this study are published authors writing articles for scientific journals, and have an entirely different profile from the student writers studied by the authors mentioned above. Their ideas about their audience and what a formal expository text should be like may have made these scientists adhere more strictly to the traditional, author-evacuated view of academic writing. Hence, it is important for them to be reminded that for certain purposes, the use of personal pronouns in scientific publications is recommendable, if not required.

Prepositional-phrase fragments

The prepositional phrases found in the native scripts are principally used for making abstract or logical connections between propositions. They are also useful for expressing such concepts as methods, processes, measurements, place, extremity and orientation. Fifty-seven out of the 86 types of prototypical target bundles that take the form of prepositional-phrase fragments were found in the NNS corpus. They carry similar meanings to those identified in the HSC, including those that are mainly figurative:

(184) These characteristics have been rationalized *on the basis of* a mechanism involving an intermediate compex possessing a partial charge-transfer character (scheme 1). NNS006

(185) *In the case of* roots and stems, samples were taken at two different levels to analyze the structural differences between young and adult segments of these organs. NNS022

(186) This would give a probable modification in a smaller liberation of TES in the mature life *as a result of* the gonad functionality. NNS044

(187) For this latter substrate, the second Km and Vmax are fifty and twenty fold higher *with respect to* the first Km and Vmax values. NNS018

(188) *On the other hand*, the mutants tss2 and tos showed a sligth higher JA-content in relation to the cv. Moneymaker (Fig. 1B and 1C). NNS029

(189) Moreover, our finding is *in agreement with* the results previously demonstrated by Docampo et. al. [4], who reported Ca2+ release after addition of NH4+ and nigericin to the Fura 2-loaded T. cruzi epimastigotes. NNS033

(190) Barley grains (Hordeum vulgare, cv. Carla INTA) were deembryonated, surface sterilized, and allowed to imbibe in sterile water for 4 d in the dark *at room temperature*. NNS024

(191) From each dilution, 0.1 ml of inoculum was spread in triplicate *on the surface of* different solid media. NNS042

(192) *At the end of* the egg period (ranged from 7 to 10 days) the final number of preyed egg masses was recorded. NNS034

Several other prepositional-phrase fragments serve as textual signposts that point back to the study or article itself:

(193) The fact that Trichoderma spp. used *in this study* protected peanut root in the adult plants, as was observed in the ASI decrease and increase in healthy plant, suggest a long-term protection of the subterranean portions of plants. NNS025

(194) *In this paper*, we show that tpx1 and tpx2 are induced in tomato hairy roots by elicitation with chitosan and non-autoclaved FOL conidia suspension, which is a molecular evidence of their implication in the lignification stimulated in response to plant-pathogen interaction. NNS030

(195) *In the present* work we have studied the photodynamic activity of 5-(4-trimethylammoniumphenyl)-10,15,20-tris(2,4,6-trimethoxy phenyl)porphyrin iodide (CP, Figure 1) on a human carcinoma cell line. NNS037

The token counts show a statistically significant overuse of prepositional-phrase fragments in the non-native texts with respect to the native texts. However, upon examination of individual bundle tokens, it was revealed that the higher number of prepositional bundles in the NNS corpus is largely due to the overuse of a few specific bundles, the most overused of which are listed in Table 28.

Table 28. Most overused prepositional-phrase fragments in the NNS corpus

	HSC		**NNS**		**LL**
	ABS	**REL**	**ABS**	**REL**	
in the presence of	541	25.98	67	55.50	28.27 (++)
on the other hand	51	2.45	30	24.85	73.22 (++)
in agreement with	35	1.68	22	18.22	55.70 (++)
with respect to	72	3.46	24	19.88	39.55 (++)
in this study	148	7.11	24	19.88	17.06 (++)
at the same	61	2.93	16	13.25	21.11 (++)
by the method	38	1.82	14	11.60	25.02 (++)
in the present	112	5.38	18	14.91	12.61 (++)

LEGEND (++) Statistically significant overuse in NNS (at p < 0.01, critical value 6.63)

The overuse of some of these bundles can again be attributed to the abovementioned lexical teddy-bear tendency and the non-native writers' underuse of alternative expressions in the target language. For instance, the bundles *on the other hand, in agreement with* and *by the method* can all be replaced by similar words and phrases that are attested in the native texts but are hardly or never used by the non-native authors. The contrastive meaning of *on the other hand* (196) (197) can also be expressed by the bundle *in contrast to* (198), while *in agreement with* (199) (200) can be alternated with the bundle *is consistent with* (201), and *by the method* (202) (203) with the words *according to* (204) and *following* (205).

(196) In dry seeds of mutant tss1, lower level of JA compared to the wild type was observed (Fig. 1A). *On the other hand*, the mutants tss2 and tos showed a sligth higher JA-content in relation to the cv. Moneymaker (Fig. 1B and 1C). NNS029

(197) A typical normal cell contains 25% (17-40%) protein by weight [77]. Cancer cells, *on the other hand*, contain as much as 100% more protein than normal cells. [75]

(198) In both X-only and X-Y' ends, the levels of silencing decrease both proximally and distally to the X-ACS (see Figure 1A-C). This is *in contrast to* the models of repression in which the repressive chromatin is propagated continuously from the telomere. [74]

(199) This is *in agreement with* previous observations of an increase of GA4 and GA7 and a decrease of GA3 when G. fujikuroi is grown in low oxygen concentrations (Jonhson and Coolbaugh 1990). NNS011

(200) This is *in agreement with* the study of Akiyama et al. (7), who observed that after a very long period (27 days) of fat infusion, normal rats still display a greater insulin response to glucose. [54]

(201) This *is consistent with* a model proposed by Theologis and colleagues in which transcription of auxin-regulated genes is normally repressed by the action of short-lived repressor proteins (Ballas et al. 1995; Abel and Theologis 1996). [80]

(202) The rates of evolution (either loss or generation) of primary anime reactivity in the tyrD (initial concentrations 2 per 10 M) upon eos-sensitized photooxidation were determined *by the method* described by Straight and Spikes. NNS006

(203) All media were supplemented with tryptophan (20 µg/ml). B. subtilis strains were transformed *by the method* of Anagnostopoulos and Spizizen (1961), as modified by Jenkinson (1983), or as described by Kunst and Rapoport (1995), except that 20 min after addition of DNA the transformed cultures were supplemented with 0.66% casamino acid solution. [56]

(204) Protein concentrations were determined *according to* the method of Bradford using a Bio-Rad Protein Assay Kit. [3]

(205) Holoenzyme-promoter complexes were formed at 37°C in binding buffer (40 mM Tris-HCl pH 7.5, 10 mM MgCl2, 100 mM KCl, 1 mM DTT, 100 µg/ml BSA), *following* the method of Roe et al. (1984). [9]

Other structures

With the exception of verb or noun phrases with *that*-clause fragments and other adjectival phrases, where there are no statistically significant differences between the HSC and the NNS corpus, and other expressions, which actually show evidence of overuse, all other structural categories are used less frequently in the non-native texts in comparison to the native texts.

Despite the statistically significant differences in frequency, these underused categories follow the same patterns of meaning in the NNS corpus as they do in the HSC. Simple *to*-clauses are typically employed in the expression of procedural objectives (206), while those with a verb controlling the *to*-clause introduce previous results (207), and those with predicative adjectives preceding the *to*-clause convey ability (208) and likelihood (209).

(206) *To determine whether* positively charged compounds in general might affect PA-kinase in T. cruzi and to further understand the mechanism for

activation of enzyme by NaF and no effect of Mn2+ ions in presence and absence of phosphatidic acid, we investigated the effect of polyamines on PA kinase. NNS007

(207) Genetic control of vegetative compatibility was *found to be* conditioned by numerous loci in those species where it has been investigated. NNS032

(208) In addition, we have also showed that Cch *is able to* modify phosphatidylinositol metabolism [10] and to increase InsP3 levels as a consequence of PtdIns-PLC activation in this parasite [11]. NNS033

(209) The relationship between the heterogeneity of the marginal zone and discharge (or river stage), is a functional characteristic of any river-floodplain system that *is likely to* exert a major influence on biodiversity patterns. NNS028

Adverbial-clause fragments are used to refer to different sections of the article (210) and cite relevant studies (211), as well as to make comparisons (212) and provide justification for claims (213). However, all but four of the 15 bundle types of this form occur once, twice or not at all in the NNS corpus.

(210) *As shown in Figure* 2, the photoirradiation of the mixture Rf (0.027 mM)-Iso (0.33 mM) in water, produces spectral changes that can be attributed to transformations in both components of the mixture. NNS008

(211) The specific radioactivity in these compounds was estimated in a similar manner *as described by* Domenech et al. (1996). NNS011

(212) Deuterated water was chosen as a solvent for TRPD experiments due to the convenience of prolonging the lifetime of <( )> (*as compared with* its lifetime in <( )>, given the relatively long time response<( )>. NNS013

(213) The second possibility, *as judged by* the small band that could be seen in the Western blot, is that the presence of the signal peptide that may remain also on the N-terminus could be responsible of the changes in some quantitative kinetic properties. NNS018

Bundles with the copula *be* combined with an adjective phrase serve to connect elements causatively (214) and comparatively (215), and to indicate authorial evaluations (216). Similar to adverbial-clause fragments, the majority of copula *be* + adjective phrase types are unattested or used only once in the non-native articles.

(214) The use of NaCl as the sole salinizing agent in salinity studies *is due to* the fact that generally it is the main component of the soluble salts mixture present in saline soils. NNS010

(215) This fact *is consistent with* the report of Pérez-Alfocea et al. (1993) who considered that Pera is tolerant to NaCl by its ability to accumulate ions. NNS021

(216) This *is an important* pathway in living organisms, since constitutes a source for the recovery of Rf from the semireduced species (40). NNS008

The anticipatory-*it* pattern is usually followed by a predicative adjective and is used to communicate the writer's appraisal of possibility (217) and probability (218). When it is followed by a verb predicate, commonly a passive construction preceding a *that*-clause, it conveys the writer's opinion as an evident and acknowledged fact (219).

(217) By this means *it is possible* to fluorometrically monitor the evolution of primary amino groups reactivity in a given substrare, during the course of a ( ) mediated photooxidation. NNS006

(218) These authors proposed that *it is possible* that plants exposed to K+ salts, in contrast to Na+ treatments, were not able to transport K+ into the vacuoles, causing a specific ion toxicity in the cytoplasm that inhibited both growth and glycinebetaine production. NNS010

(219) *It should be noted* that P. strombulifera roots from tolerant plants showed precocious suberization and/or lignification of the endodermal cells in the young segment. NNS022

It is interesting to note that except for *it is possible,* which occurs relatively frequently in the NNS corpus, all anticipatory-*it* bundle types are used once or not at all by the non-native writers. This seems to indicate their overreliance on *it is possible* as a marker of possibility and likelihood and lack of awareness of alternative options.

As mentioned previously, bundles with a verb or noun phrase followed by a *that*-clause fragment and those taking the form of other adjectival phrases not included in the other categories show no statistically significant frequency differences between the two corpora. These types of bundle structures have meanings similar to those identified in the native texts.

Lexical bundles with a main clause followed by a *that*-clause have either a noun or a verb phrase, with the former usually serving to emphasize an accompanying statement for the purpose of justification, and the latter functioning as references to corroborating results and studies. The bundles in the category of other adjectival phrases, on the other hand, mostly express comparative relations:

(220) *The fact that* apoptosis takes place preferential using a low dose of light could be favored in this case for the reason that CP is localized in mitochondria. NNS037

(221) The *results show that* cultivars Morpa, Don Pablo and Robusta 4047 were sensitive to the increase of environmental quality, with a specific adaptability to favourable environments because they presented a response value higher than the general mean. NNS004

(222) Other *studies have shown that* A. parasiticus account for only 10-30% of the section Flavi in peanut seed (Hill et al. 1983; Blackenship et al., 1984; Horn et al., 1995), suggesting that it is less aggressive species. NNS026

(223) These genotypes, with regression deviation *significantly different* from zero (P<0.01), are unstable in their responses. NNS004

(224) This value was *similar to that* obtained by molecular filtration through a Sephacryl S-200 HR column. NNS009

All three of these structural categories are rarely used by non-native writers, except for three overused types. The bundles *the fact that, the idea that* and *similar to that* are all used more frequently in the NNS corpus than in the HSC. Like some of the overused bundles previously described, the disproportionate use of these bundles may also be linked to the non-native authors' overdependence on familiar formulas. There is a marked absence in the non-native texts of alternative expressions to these bundles, such as *the notion that* for *the fact that* and *the idea that*, and *analogous to* and *resembling* for *similar to that*:

(225) The simultaneous loss of petD mRNA processing and translation in crp1 mutants is consistent with *the notion that* the processing event increases the efficiency with which petD mRNA is translated. [27]

(226) To determine whether the K.lactis 2 tail was interacting with the a1 homeodomain in a manner *analogous to* that of the S.cerevisiae 2 tail, we changed one of the hydrophobic residues in the K.lactis tail, isoleucine 218, to serine. [94]

(227) fam-1 mutant larvae are often lumpy in appearance and frequently develop notched heads *resembling* those seen in vab-3 or ina-1 mutants. [31]

The final structural category, other expressions, consists of three prototypical target bundles that do not fall into any of the other categories: *in order to, as well as* and *there are several.* The apparent overuse of this category is due largely to the overuse of the first two bundles. The non-native scientists seem to prefer *as well as* (228) as an addition device, over other possible candidates such as *in addition to*. And more than native speakers, they tend to use the bundle *in order to* (229) for prefacing their objectives, instead of just using simple *to*-infinitives.

(228) The high repeatability, which refers to the constancy across repeated measurements obtained for dry matter, leaf, length and crown diameter could be explained by apomitic reproduction of this species *as well as* for the absence of interaction genotypes x cuts in these characters. NNS005

(229) Competitive irradiations of nitrogen-saturated solutions of Rf in the absesnce and in the presence of <( )> showed that tis rate is dramatically siminished in the presence of Q (fig.4), and the same effect was observed in aie equilibrated solutions, although much longer irradiation times were necessary *in order to* obtain measurable absorption changes. NNS013

3. Functions of target bundles in the non-native corpus

Table 29 contains the functional classification of the prototypical target bundles found in the NNS corpus and their corresponding frequencies. Figures 11 and 12 graphically represent how types and tokens are distributed by functional category.

Table 29. Frequency of functional categories of prototypical target bundles in the NNS corpus

FUNCTION	TYPES	%	TOKENS		%
			ABS	REL	
Research-oriented bundles	**157**	**44%**	**845**	**699.98**	**43%**
Location	13		34	28.16	
Procedure	80		454	376.08	
Quantification	27		212	175.62	
Description	21		79	65.44	
Grouping	16		66	54.67	
Text-oriented bundles	**165**	**47%**	**1035**	**857.37**	**53%**
Additive	4		86	71.24	
Comparative	17		136	112.66	
Inferential	49		171	141.65	
Causative	18		104	86.15	
Structuring	20		141	116.80	
Framing	32		254	210.41	
Citation	17		75	62.13	
Generalization	3		8	6.63	
Objective	5		60	49.70	
Participant-oriented bundles	**32**	**9%**	**63**	**52.19**	**3%**
stance	25		48	39.76	
engagement	5		13	10.77	
acknowledgement	2		2	1.66	
TOTAL	**354**	**100%**	**1943**	**1609.54**	**100%**

Figure 11. Functional categories of prototypical target bundles in the NNS corpus: Distribution by type

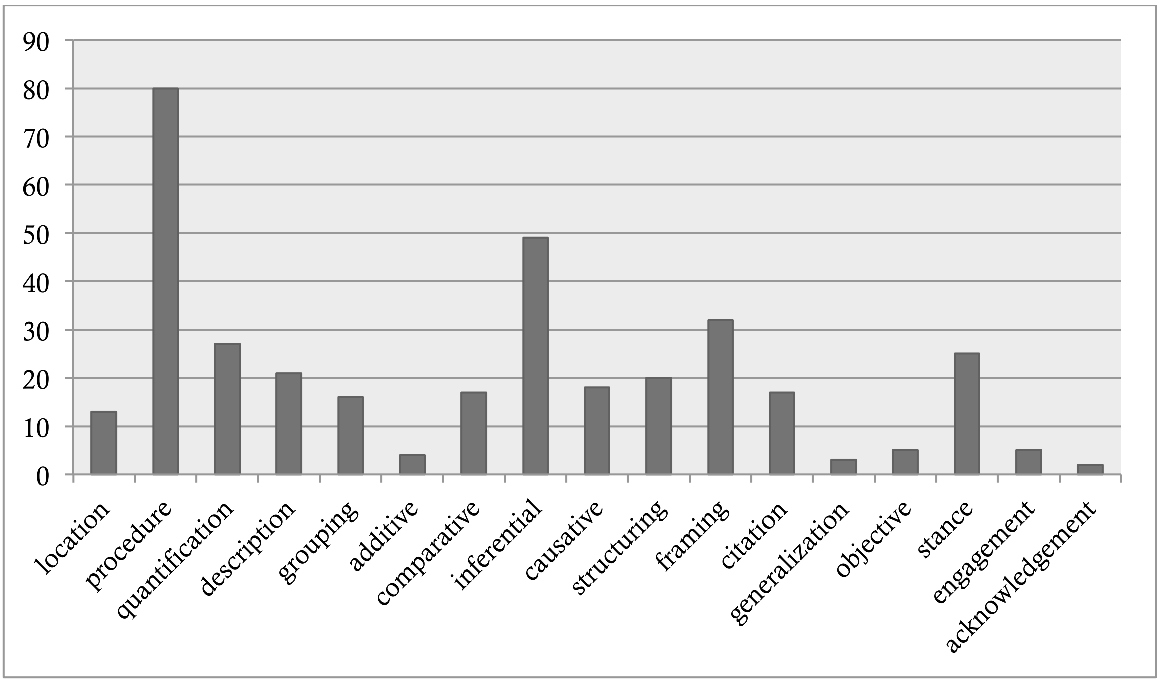
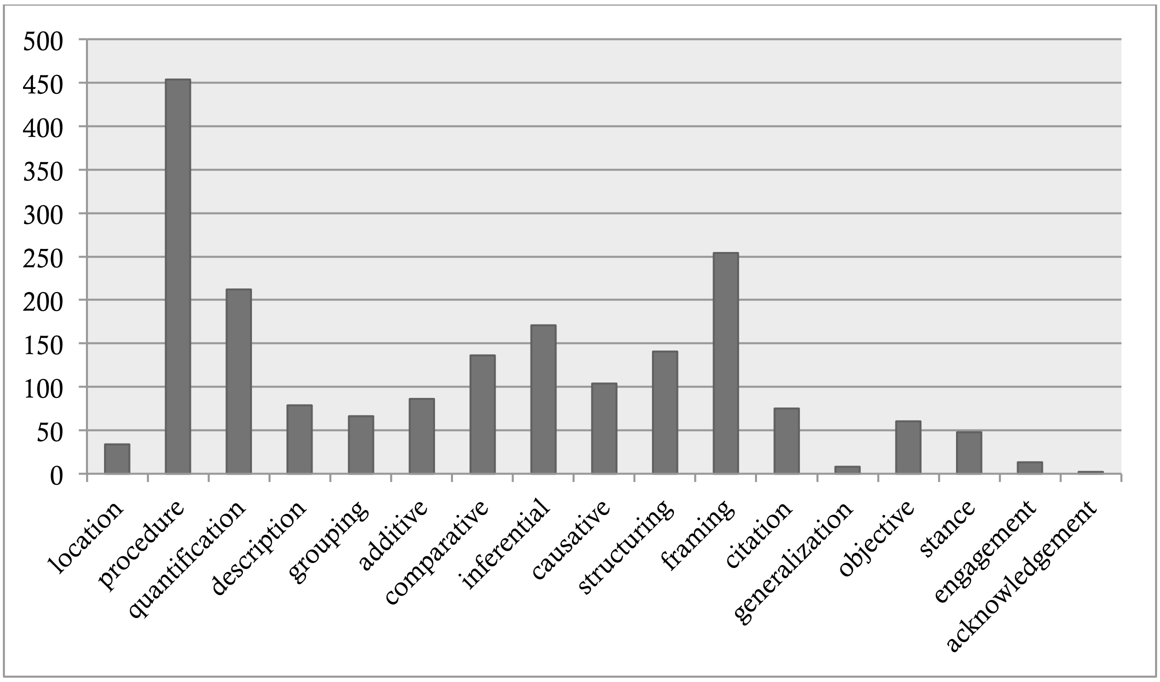


Figure 12. Functional categories of prototypical target bundles in the NNS corpus: Distribution by token



The above table and figures show that the distribution of functional categories in the non-native texts is generally consistent with their distribution in the native texts. The most widely used of the three main functional categories are text-oriented bundles, which make up 47% of prototypical bundle types, with 165, and 53% of prototypical bundle tokens, with 1,035. In close second place are research-oriented bundles with 157 types (44%) and 845 tokens (43%). Participant-oriented bundles rank a distant third, with 9% of types (*n* = 32) and 3% of tokens (*n* = 63).

With respect to the functional subcategories, the top five most common functions in the HSC are shared by the NNS corpus, with research-oriented procedure bundles ranking first in frequency in both corpora. In the NNS corpus, this category accounts for 80 types and 454 tokens. Placing second in frequency in the two corpora are text-oriented framing bundles, with 32 types and 254 tokens. In the non-native texts, framing bundles are followed by research-oriented quantification bundles with 27 types and 212 tokens. The NNS corpus' top list of five most frequent functions is rounded out by two text-oriented categories: inferential (49 types, 171 tokens) and structuring (20 types, 141 tokens).

Text-oriented causative bundles (18 types, 104 tokens), the top eight of the bundle functions in the HSC, places one spot higher in the HSC, at number seven. Participant-oriented stance bundles and research-oriented description bundles, which both made it to the top eight in the frequency rankings in the native texts, make way to text-oriented comparative (17 types, 136 tokens) and additive bundles (4 types, 86 tokens) in the non-native texts. The top eight most frequent functions in the NNS corpus account for 80% of all bundle types and tokens, an even larger chunk of the total than the top eight of the HSC.

Table 30 displays the absolute and relative frequencies of the various functional categories in the native and non-native corpora with the corresponding log-likelihood scores. Figure 13 shows the relative token frequencies of each category in each of the two corpora.

Table 30. Frequency of functional categories of prototypical target bundles in HSC and NNS

FUNCTION	HSC		NNS		LOGL
	ABS.	REL.	ABS.	REL.	
Research-oriented bundles	**10141**	**486.98**	**845**	**699.98**	**92.80 (++)**
Location	774	37.17	34	28.16	2.73
Procedure	5137	246.69	454	376.08	66.04 (++)
Quantification	1906	91.53	212	175.62	68.26 (++)
Description	1535	73.71	79	65.44	1.10
Grouping	789	37.89	66	54.67	7.40 (++)
Text-oriented bundles	**13734**	**659.52**	**1035**	**857.37**	**61.49 (++)**
Additive	639	30.69	86	71.24	43.49 (++)
Comparative	1113	53.45	136	112.66	55.61 (++)
Inferential	3062	147.04	171	141.65	0.23
Causative	1490	71.55	104	86.15	3.18
Structuring	2402	115.35	141	116.80	0.02
Framing	3094	148.58	254	210.41	25.78 (++)
Citation	1166	55.99	75	62.13	0.74
Generalization	145	6.96	8	6.63	0.02
Objective	623	29.92	60	49.70	12.29 (++)
Participant-oriented bundles	**2348**	**112.76**	**63**	**52.19**	**46.99 (--)**
Stance	1818	87.30	48	39.76	37.55 (--)
Engagement	425	20.41	13	10.77	6.35 (-)
Acknowledgement	105	5.04	2	1.66	3.57
TOTAL	**26223**	**1259.26**	**1943**	**1609.54**	**101.60 (++)**

LEGEND
(--) Statistically significant underuse in NNS (at p < 0.01, critical value 6.63) (-) Statistically significant underuse in NNS (at p < 0.05, critical value 3.84) (++) Statistically significant overuse in NNS (at p < 0.01, critical value 6.63)

Figure 13. Distribution of functional categories of prototypical target bundles in HSC and NNS

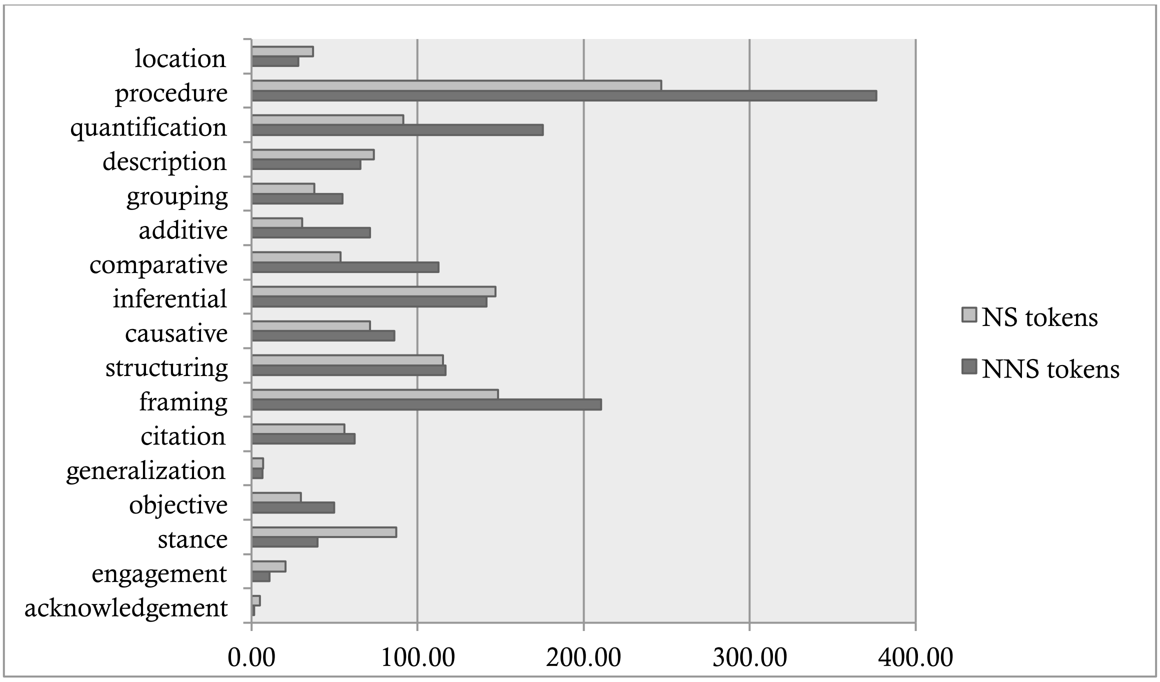


Table 31 includes all prototypical target bundles found in the non-native texts by their alphabetically ordered keywords and groups them by function.

Table 31. Prototypical target bundles found in the NNS corpus, grouped by function

RESEARCH-ORIENTED BUNDLES	
Location	in the dark, at the end of, the location of, the position of, the region of, in the region, the site of, at the site, at the surface, on the surface of, the tip of, the top of, in the upper
Procedure	the accumulation of, the action of, the activity of, was added to, by the addition of, were allowed to, the analysis of, were analyzed by, the assembly of, the beginning of, carried out at, carried out in, carried out with, were carried out, a change in, were collected from, was confirmed by, the control of, in the control, was detected by, the detection of, was determined as, was determined by, the development of, was digested with, was dissolved in, the evolution of, was examined by, were exposed to, were fixed in, the formation of, were generated by, the generation of, were grown at, were grown in, the growth of, the identification of, the incorporation of, were incubated for, were incubated with, was induced by, the interaction of, were isolated from, the loss of, were made by, activity was measured, was measured by, mechanism by which, the mechanism of, the method of, by the method, was mixed with, was obtained by, were obtained from, the pattern of, was performed by, were performed in, was performed using, analysis was performed, the process of, the production of, were purchased from, in the regulation of, the release of, the removal of, was replaced with, were separated by, were separated on, were stained with, the study of, were subjected to, were tested for, were transferred to, were treated with, the use of, was used to, was used as, was used for, were used in, were washed with
Quantification	in the amount of, a decrease in, the efficiency of, a fraction of, the fraction of, the frequency of, an increase in, the majority of, a large number of, a small number, in a number of, the number of, total number of, a percentage of, the percentage of, the proportion of, the rate of, at a flow rate of, the ratio of, a reduction in, the size of, at room temperature, the time of, a total of, of the total, for up to, the value of
Description	the ability of, the ability to, is able to, the appearance of, the behavior of, the degree of, the existence of, the extent of, the importance of, the inability of, the intensity of, the lack of, the level of, the nature of, is present in, the properties of, the significance of, the stability of, the structure of, the timing of, were unable to
Grouping	a combination of, the distribution of, a member of, a mixture of, as part of, a range of, the range of, the remainder of, the rest of, the sequence of, a series of, a set of, a subset of, this type of, two types of, a variety of
TEXT-ORIENTED BUNDLES	
Additive	in addition to, as well as, on the other hand, at the same time
Comparative	in agreement with, as compared with, in comparison with, is consistent with, are consistent with, in contrast to, the difference in, the difference between, significantly different from, on the other hand, similar results were obtained, the same as, in the same, at the same, to the same, similar to that, in a similar
Inferential	were able to, to account for, it appears that, appear to be, is associated with, we conclude that, has been demonstrated, was detected in, was not detected, as determined by, lines of evidence, expected to be, be explained by, is found in, found to be, was found in, we found that, the hypothesis that, we have identified, been implicated in, this implies that, this indicates that, results indicate that, as indicated by, be involved in, as judged by, is likely to, it is likely that, was observed in, the possibility that, it is possible, we propose that, closely related to, the relationship between, the results presented, the results obtained, can be seen, as seen in, there are several, been shown to, it has been shown that, we show that, this suggests that, results suggest that, we suggest that, suggesting that this, is supported by, in support of, we were unable to
Causative	is caused by, as a consequence of, may contribute to, is due to, no effect on, the effect of, be explained by, be involved in, the product of, in response to, a response to, is responsible for, the result of, as a result of, not result in, the role of, a role in, the yield of

Structuring	as described previously, in these experiments, are expressed as, as shown in figure, is shown in figure, were as follows, with the following, as indicated by, at the indicated, in this paper, in the present, referred to as, see materials and methods, data not shown, as shown in, are shown as, in this study, the present study, shown in table, the present work
Framing	in the absence of, in accordance with, is based on, on the basis of, in the case of, in this case, in all cases, in some cases, under these conditions, under the same conditions, in the context of, the course of, is dependent on, with the exception of, the fact that, as a function of, the idea that, is independent of, model in which, in the presence of, for the presence of, by the presence of, is required for, with respect to, is sensitive to, is subject to, there are several, in terms of, at the same time, at the time, at various times, at this time
Citation	in accordance with, in agreement with, is consistent with, are consistent with, found to be, has been demonstrated, as described by, been implicated in, it has been proposed that, been proposed to, has been reported, studies have shown that, results show that, been shown to, it has been shown that, have suggested that, it has been suggested
Generalization	is found in, is known about, is not known
Objective	to account for, to determine whether, to ensure that, in order to, to show that
	PARTICIPANT-ORIENTED BUNDLES
Stance	it appears that, appear to be, is associated with, it is not clear, we conclude that, may contribute to, is difficult to, is essential for, expected to be, we found that, we have identified, is an important, is likely to, it is likely that, it should be noted, to note that, the possibility that, it is possible, we propose that, we show that, this suggests that, results suggest that, we suggest that, suggesting that this, we were unable to
Engagement	it should be noted, to note that, as seen in, can be seen, see materials and methods
Acknowledgement	a gift from, is supported by

Research-oriented bundles

As commented previously, in both the HSC and NNS corpora procedure bundles are the most frequent of all research-oriented bundles, and the most frequent bundle function overall. Procedure bundles denote events, actions and methods and are thus useful for describing research processes and activities. They typically take the form of past-tense passive structures (230), as well as noun (231) and prepositional (232) phrases.

(230) Seeds of P. strombulifera *were collected from* an area in the Southwest in the Province of San Luis, Argentina. NNS010

(231) *The analysis of* our kinetic data in table 1 indicates both a dramatic increase in the rates constants Kt and Kr in the presence of alkali, and a remarkable solvent polarity effect on Kt. NNS002

(232) This was achieved by treating the cells or membrane fraction with exogenous phospholipase D or *by the addition of* exogenous DG. NNS007

One interesting finding with regard to procedure bundles is that, although there is a statistically significant overuse of tokens with this function in the non-native texts in comparison to the native texts, a considerable number of types, 31 out of 111, are missing. A possible reason for this is the topic-specificity of this particular bundle function, and of research-oriented bundles in general. Lexical bundles such as *the isolation of, was purified from, medium supplemented with, were washed in* and *on ice for* may refer to certain experimental techniques that were utilized by the native scientists but not by their non-native counterparts, because of the differences in topic and aims between the two sets of researchers. The reverse applies to overused procedure bundles in the NNS corpus, such as *the generation of, the growth of* and *were collected from*, which may have been used more often in the non-native articles because of the given subject matter.

This explanation, however, does not apply to all overused and underused procedure bundles. The bundles *by use of* and *with the use of*, for instance, seem to be applicable to a variety of situations but are notably absent in the NNS corpus. In this case, the underuse can be explained by the non-native writers' heavy dependence on the more familiar bundles, *carried out with* (233) and *was used to* (234), which, as shown by the following examples, can be employed in a very similar way as *by use of* (235) and *with the use of* (236).

(233) Ground-state absorption measurements were *carried out with* a Hewlett Packard 8452A diode array spectrophotometer. NNS031

(234) The wavelength of 290 nm *was used to* detect tyrD. NNS006

(235) For a given BrdU focus, incorporation was defined as the sum of all pixel intensities *by use of* IP lab spectrum software (Scanalytics). [11]

(236) First, *with the use of* transgenic mouse lines expressing tv-a in specific cell types, combinations of genes can be tested by the use of easily constructed or previously existing viral vectors. [41]

With these examples, the non-native writers once again show excessive reliance on a few known bundles and unawareness of alternative options.

The relatively less common research-oriented bundles—location (237), quantification (238), description (239) and grouping (240)—serve to describe research objects and contexts, and are usually constructed as noun and prepositional phrases. Two of these four functional subcategories, quantification and grouping, also show statistically significant overuse in the non-native texts.

(237) During the study period almost no surface activity was seen *at the site,* whereas capture rates in the pitfall traps were high. NNS015

(238) They were able to catalyse *a large number of* biochemical reactions "in vitro", but it is not yet clear which are their natural substrates "in vivo". NNS001

(239) The syncytia, as they were little developed, did not modify *the structure of* the central cylinder (Fig. 2 A). NNS012

(240) The present work focuses on *a subset of* hybrids obtained in Temple (USA) and two varieties adapted to the semi-arid regions to assess the extent of phenotypic variation for yield and other evaluated agronomic traits. NNS005

The high concentration of procedure bundles in the NNS corpus is proof that the non-native writers know the importance of reporting research practices with objectivity and precision and are capable of using many of the formulas that enable them do so, although they may be lacking a certain degree of variety.

Text-oriented bundles

Of the three main functional categories, text-oriented functions are associated with the largest number of lexical bundles in both the native and non-native texts. The top three most frequent functional categories are also the same across the three corpora: framing signals (241), which are usually realized by prepositional-phrase structures and are used for linking ideas and identifying conditions; inferential bundles (242), which help introduce or underscore results, interpretations and conclusions; and structuring bundles (243), which usually take the form of adverbial-clause fragments and passive structures combined with prepositions, and serve as text-reflexive guides for readers.

(241) *In the presence of* oxygen this process could compete with the generation of reactive oxygen species such as <O2(1(g) (process (7))>. NNS016

(242) The major PC Pase activity *was found in* the fractions obtained with 70, 80, and 90% saturation. NNS009

(243) Temperature, pH, and conductivity values that were recorded *are shown in Table* 2. NNS028

The remaining text-oriented functions were found with comparatively less frequency in the non-native texts than three most common ones: comparative (244), causative (245), additive (246), citation (247), objective (248) and generalization (249).

(244) This is *in agreement with* the higher values for the rate constants in alkaline media, accounting for the enhancement of the electron releasing ability of ionized hydroxy groups. NNS002

(245) Steady-state levels of JA and related compounds were higher in the salt-tolerant cv. Pera than in cv Hellfrucht Frühstamm (HF) and JA levels in both cultivars changed *in response to* salt-stress during the vegetative development. NNS029

(246) It is known that abiotic *as well as* biotic factors affect D. saccharalis egg survival (citas), however, exactly how these factors interact is not fully understood. NNS034

(247) Aspergillus parasiticus strains were grown at 30ºC for 7 days in 4-ml vials containing 1 ml of liquid medium (three replicates per isolate) *as described by* Horn and Dorner (1999). NNS026

(248) *In order to* obtain a reliable result and considering that the number of individuals was almost 3 times greater in one cohort with regard to the other one, 3 different comparisons were made considering the same number of animals of C2 and C3. NNS041

(249) Although considerable research has been carried out on invertebrate size spectra in freshwaters (Poff et al., 1993; Kamenir et al., 1998; Mercier et al., 1999; Feldman, 2001; Havlicek & Carpenter, 2001; Cózar et al., 2003); much less *is known about* comparisons among size spectrum of benthos, drift and marginal fauna in a river. NNS028

Despite the more or less analogous distribution of text-oriented functions in the HSC and NNS corpora, there are still some differences that are worth taking note of. Sixteen types of framing bundles out of 51, 13 inferential bundle types out of 67 and 12 structuring bundle types out of 32 are unattested in the NNS corpus. In addition, there are less inferential bundles in the native texts than in the non-native texts, although the difference is not statistically significant.

There is, however, a statistically significant overuse in the NNS corpus of additive, comparative, framing and objective bundles, the last two because of a markedly excessive use of certain bundles, namely, *as well as* (additive), *in agreement with* (comparative), *on the other hand* (additive and comparative), *in the case of, the fact that, in the presence of, with respect to* (framing) and *in order to* (objective).

Here, as with the research-oriented texts, the non-native writers demonstrate their ability to employ the basic formulas they need to perform the functions that text-oriented bundles are intended to fulfill: that is, to construct a coherent, logically constructed and easily readable text. However, there is once again the need to widen their phraseological repertoire and control their tendency to overly rely on familiar expressions.

Participant-oriented bundles

It was shown in the previous chapter that the native authors regularly employ participant-oriented bundles to shape effective reader-writer interaction, using stance bundles for such crucial rhetorical strategies as evaluation, depersonalization and hedging, and engagement bundles for convincing readers and eliciting their involvement. It was also mentioned that the expression of epistemic, evaluative and directive meanings through stance and engagement markers presents a number of linguistic and cultural challenges to non-native writers (Aijmer, 2002; Hyland & Milton, 1997; Neff & Bunce, 2006; Salazar, 2008; Salazar & Verdaguer, 2009). This observation seems to be borne out by the present study's findings with respect to participant-oriented bundles, as it is in this final functional category that the most striking differences between the HSC and NNS corpora were found.

In the HSC, participant-oriented bundles occur less frequently than the two other functional categories, representing only 9% of all bundle tokens, but this infrequency is even more pronounced in the NNS corpus, where participant-oriented functions are associated with only 3% of tokens. Additionally, there is a statistically significant underuse of stance bundles, engagement bundles and participant-oriented bundles as a whole, in the NNS corpus as compared to the HSC.

With regard to stance markers, only 25 out of 36 types were attested in the non-native texts. Of the 25 types identified, 17 appear only once, and only three have more than three occurrences: *results suggest that* (4 occurrences), *is an important* (5 occurrences) and *it is possible* (8 occurrences):

(250) Considering that tpx1 has pI 9.6 and tpx2 is even more cationic, these *results suggest that* both peroxidase isoforms have been elicited and are then responsible for the increase in peroxidase activity in the ionically bound fraction. NNS030

(251) Peanut *is an important* crop in Argentina, during the 2002/03 the production reached xxxxx ton. NNS025

(252) As a consequence, *it is possible* to deduce that Na+ transport may be involved in Ca2+ release from acidic compartments in the parasite. NNS033

There is a noticeably limited use in the non-native texts of hedging devices and depersonalized stance expressions, as realized by adjective phrases and anticipatory-*it* constructions. This is in addition to the rare occurrence of personalized stance markers incorporating the first-person plural pronoun *we,* a tendency discussed at length in the preceding section.

As for engagement markers, only five out of the nine target prototypical types with this function were found in the NNS corpus, and all except for the bundle *can be seen* (8 occurrences) appear only once or twice:

> (253) As *can be seen*, the values of the photooxidation quantum efficiencies for the reactive OHAN are in the range 0.07-0.33, being the highest values those of the isomer 1OHAN. NNS002

These results provide sufficient evidence to state that the non-native writers under analysis do not employ participant-oriented prototypical target bundles as regularly and diversely as their native counterparts. However, given the methodology of the present study and the small size of the NNS corpus, these findings are not enough to ascertain whether the non-native authors use other forms apart from the target bundles to perform participant-oriented functions, or they simply have less control of stance and engagement devices as some word-based studies indicate (Aijmer, 2002; Hyland & Milton, 1997; Kennedy & Thorp, 2007; Salazar, 2008). The findings of Chen and Baker's (2010) investigation of lexical bundles in published academic texts and L1 and L2 student academic writing seem to point in the latter direction. These authors searched for the most frequent bundles in all three corpora and discovered a much wider range of epistemic bundles in the published texts and L1 student essays than in the L2 student scripts. Both native groups demonstrate the ability to use a variety of lexical bundles to qualify their propositions, including constructions such as copula *be* + *likely to* and anticipatory-*it* + adjective fragments, as well as bundles with modal verbs, hedging verbs and hedging nouns (Chen & Baker, 2010, pp. 41-42)—structures that have also been found in the present study's HSC. The L2 student writers, in contrast, only produced four bundles that can be considered hedging expressions.

These findings emphasize the need for the explicit teaching of participant-oriented functions in academic writing, as their use proves to be a complicated task for non-native students and professional authors alike. Non-native, and even novice native writers, can benefit from teacher and material-guided reflection on how their linguistic choices can help set the correct tone for their writing and build rapport with their expected audience.

4. Concluding remarks

The analysis of the frequency and structural and functional features of prototypical target bundles in the corpus of non-native scientific writing revealed few remarkable differences between this and the native corpus as far as the use of lexical bundles is concerned. Cortes (2004), who compared expository writing in history and biology by published authors and students, found a large gap between the two writer groups she examined. Chen and Baker (2010), who dealt with native expert writing, native student writing and non-native student writing, similarly uncovered few shared features across their three groups, especially between the native and non-native writers. In comparison to these previous investigations, lexical-bundle usage in the two sets of scientific texts analyzed in the present study bear closer resemblance to each other.

The fact that this result was obtained from a comparison of equivalent text types, written by two groups of expert scientists differentiated only by their nativeness, lends support to Cortes' (2004) and Chen and Baker's (2010) claim of a developmental trend in the use of lexical bundles. Cortes observed that "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" (2004, p. 414). Since this study is concerned with non-native professional scientists who have as much experience in and knowledge of their discipline as the native scientists to whom they are being compared, it is reasonable to suppose that they have been exposed to the kind of research literature that may have familiarized them with the formulas of the genre. And since both writer groups composed exactly the same type of text, a research article to be submitted for publication, it is highly likely that the non-native writers are aiming for the same goals as their native equivalents, at least much more so than students writing research reports for class being compared to published authors.

There are, however, two important differences found between the native and non-native texts that deserve to be underscored here. First is the lesser degree of variety in non-native writing when it comes to the use of lexical bundles, brought about by the non-native writers' overuse of certain bundles. This a manifestation of the lexical teddy bear phenomenon commonly associated with learner writing, a tendency to "cling on", to use Granger's (1998) terms, "to certain fixed phrases and expressions which [learners] feel confident in using" (1998, p. 156). This habit leads to unnecessary repetitiveness and deprives non-native texts of the phraseological richness characteristic of well-written academic prose.

Second, as much as the non-native writers may be aware of the importance of research- and text-oriented bundles, and as capable as they prove to be of handling these functional categories, their limited use of participant-oriented bundles show their difficulties with this particular function. This is hardly surprising considering that this function constitutes a more subtle aspect of academic writing, one that is grounded in the established but seldom explicitly acknowledged norms of research publication. The expression of writer stance, the delicate engagement and persuasion of the reader, the proper manipulation of hedging devices and personal and impersonal forms—all these are strategies that scientists must master if they are to be successful in disseminating their work to the larger scientific community. Much of this success depends on the creation of a "competent scholarly identity" (Hyland, 2001, p. 223), and although research- and text-oriented bundles play an essential role in this process, participant-oriented bundles are key ingredients that most published scientific writers know when and how to add to achieve the desired rhetorical effect.

This chapter cannot be concluded without echoing the caveats issued by Cortes (2004) regarding limited corpus size and the target-bundle methodology adopted in this study. This method of analysis shows whether the non-native writers use the same bundle structures and functions as the native writers, but it provides no means to determine whether they are using other forms to perform the same bundle functions, or if they even wish to perform these functions at all. To determine the degree to which the target-bundle methodology reflects the actual use of lexical bundles in the NNS corpus, an independent search of three- to six-word lexical bundles that occur at least ten times in the corpus was carried out. The results of this search, after the application of the same exclusion criteria used in the extraction of target bundles from the HSC, are summarized in Table 32 below.

The findings are encouraging. It can be seen from the table that, apart from a handful of bundles, which are highlighted in bold, all of the most frequent lexical bundles in the NNS corpus are also target bundles. Some of the few bundles in bold, such as *the quenching of* and *rate constant for*, are procedure bundles whose absence on the list of target bundles can be attributed to differences in the subject matter of the HSC and

NNS corpora. Others, such as *in relation to, were found in* and *in this work* are frequent in the non-native texts and also appear in the native texts, but were not identified as target bundles because they did not meet the higher frequency cut-off applied to the larger corpus. It can also be observed that the frequency ranking of the inventory below is consistent with the frequency ordering of target bundles, with *the presence of* and *in the presence of* similarly heading the list.

Table 32. Most frequent lexical bundles in the NNS corpus

RANK	LEXICAL BUNDLE	TOKENS
1	the presence of	128
2	in the presence of	71
3	in order to	53
4	the number of	48
5	as well as	44
6	the absence of	35
7	the effect of	33
8	on the other hand	30
9	**in relation to**	26
10	in the absence of	25
11	were carried out	25
12	was used to	24
13	are shown in	23
14	with respect to	23
15	was carried out	22
16	were able to	22
17	in agreement with	20
18	in response to	19
19	were determined by	18
20	carried out with	17
21	data not shown	17
22	in this study	17
23	**the case of**	17
24	the levels of	17
25	the present study	17
26	**were found in**	17
27	an increase in	16
28	**in this work**	16
29	the fact that	16
30	**the quenching of**	16
31	**were incubated at**	16
32	shown in Fig	15
33	the effects of	15
34	at the same	14
35	the basis of	14
36	**did not show**	13
37	**it is known that**	13

38	the addition of	13
39	the amount of	13
40	the generation of	13
41	was found in	13
42	were obtained from	13
43	a mixture of	12
44	be due to	12
45	**could be observed**	12
46	has been reported	12
47	in the present	12
48	in the same	12
49	**it was observed**	12
50	**rate constant for**	12
51	the end of	12
52	**were carried out with**	12
53	were used for	12
54	**and in the presence of**	11
55	carried out in	11
56	on the basis of	11
57	similar to those	11
58	the beginning of	11
59	the evaluation of	11
60	the production of	11
61	the rate of	11
62	was observed in	11
63	**was observed that**	11
64	**were observed in**	11
65	a variety of	10
66	**an increase of**	10
67	**by means of**	10
68	of the total	10
69	shown in Table	10
70	**the determination of**	10
71	**the first order**	10
72	the formation of	10
73	the increase in	10
74	**the most important**	10
75	the use of	10
76	was used as	10

Despite these promising results, there are issues that remain that cannot be accounted for by the target-bundle methodology. One such issue is the presence of what Thewissen (2008) terms “near hits”, or close approximations of grammatically and pragmatically acceptable multi-word units that non-native writers are sometimes able to produce. Taking these near hits into consideration can lead to a better understanding of the phraseological profile of the non-native texts. However, for this to be sufficiently addressed, there is a need for a corpus of uncorrected research articles written by non-native speakers similar to the one used here, but of a comparable size to the multimillion-word native research-article corpora already in existence.

Chapter VII

Pedagogical applications of the study

In a study published in 2010, authors Byrd and Coxhead identify six major challenges that hinder the successful introduction of lexical bundles in EAP classrooms and teaching materials. This chapter will touch on each of these issues and discuss the solutions offered by the results of the present study. This discussion will not only highlight the useful features of the study's final product, a practical list of lexical bundles in scientific English for use in pedagogical applications, but also underscore its methodological contributions to research on lexical bundles.

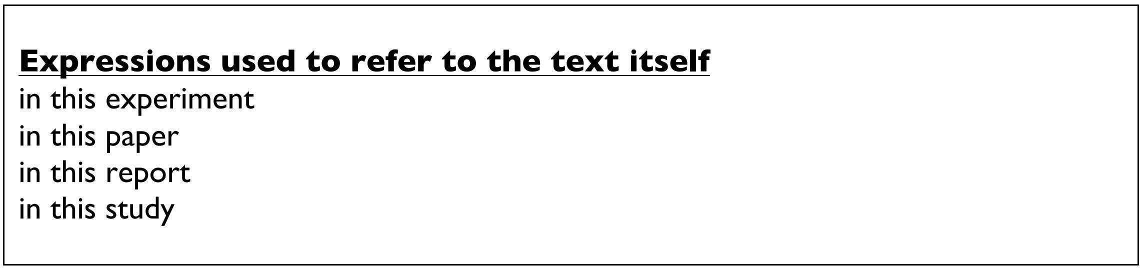
1. Working with word lists of bundles published in research reports

Byrd and Coxhead (2010) agree with Jones and Haywood (2004) on the utility of lists of lexical bundles as a basis for materials design and curriculum development, on the condition that teachers and learners are given sufficient information about how the list has been developed. From this perspective, the list provided by the present study is an ideal instrument for the selection of lexical bundles for teaching, as all the essential information relative to its creation is readily available: the type of texts from which the list was generated, its representativeness of the language required by learners, the principles of selection that were followed, etc.

The list of bundles can be sorted by frequency, structure and function, the kind of quantitative and qualitative information that can assist teachers and materials designers in deciding which multi-word units are most suited to their particular needs. In addition, the fact that the lexical bundles on the list can also be grouped by keyword and by prototypical bundle makes it more than just an inventory of discrete, frequency-ordered phraseological items. Semantic and functional relationships between like bundles are acknowledged and made explicit, and contextual examples and usage notes are provided where necessary. All this additional information simplifies the application of principles such as frequency, range, teachability, learnability and usefulness to decision-making and instruction.

Sorting lexical bundles will also make it easier for practitioners using this list to determine the level of pedagogical treatment that lexical bundles require. Some bundles can be presented in class materials, textbooks or learner dictionaries as simple lists of expressions unified by a single function (Figure 14), or they may demand a more extensive description for students to better understand the different aspects of their use (Figure 15).

Figure 14. Example of lexical bundles presented as a list



Expressions used to refer to the text itself
in this experiment
in this paper
in this report
in this study

Figure 15. Examples of full description of lexical bundles

as described in

This expression is used to refer to a process already described in detail somewhere else.

The mitochondrial fraction was prepared as described in the Experimental section.

You can use different variations of this expression depending on your purpose.

To refer to a description within the text you are writing, use the preposition *in,* then state in which section this explanation can be found.

as described in figure 1
as described in the experimental section
as described in Materials and Methods
as previously described in the experimental section

You can also use the adverbs *previously* and *above* to refer to any point in the text prior to the sentence you are writing.

as described above
as described previously
as previously described

To refer to a process described by other authors, use the preposition *by.*

as described by Smith et al. (2010).
as previously described by Smith et al. (2010).

Carry out, perform and *prepare* are just some of the verbs frequently used with this expression.

The assays were performed as described in Figure 1.
The experiment was carried out as described in Materials and Methods.
Western blots were prepared essentially as described in Smith et al. (2010).

demonstrate

The verb *demonstrate* is used in different expressions to introduce inferences drawn from a study's findings. It is frequently used with nouns such as *data, experiments, findings* and *results.*

These data demonstrate that the presence of these cells exacerbates respiratory impairment.
The above experiments demonstrate that a basal expression *of this protein is.*
Taken together, these results demonstrate that this substance plays an important role in starch breakdown.

To emphasize that the statement is your very own interpretation of your data, use first-person pronouns.

Our findings demonstrate that methylation is not required for expression.
In this report, we demonstrate that these mutants are defective at both the permissive and restrictive temperature.

Demonstrate is also useful for referring to related literature.

Recent kinetic studies demonstrate that this type of binding is a dynamic process.
It has been demonstrated that this element has potent effects.

Show is another verb that functions in a similar manner as *demonstrate.*

These results show that food transfer involves various behaviors.
In this paper, we show that the simple view does not account for this phenomenon.
It has been shown that cells can return to mitotic growth.

2. The length of lexical bundle to teach when shorter bundles are reported inside longer ones

The present study is one of the few investigations on lexical bundles that are not restricted to a given sequence length. Many researchers (Biber et al., 2003, 2004; Cortes, 2004; Hyland, 2008) focused exclusively on four-word bundles, which appear in numbers more manageable for analysis and also incorporate shorter bundles in their structures. In this study, bundles from three to six words in length are considered, and although this research decision contributed to a more complete picture of lexical bundles, it also gave rise to the problem of overlap, and the question of which unit should take priority in teaching and learning when shorter strings are embedded in longer ones.

The present investigation addressed this problem by establishing certain criteria regarding which fragments of longer bundles should be maintained and which should be excluded, following a procedure adopted from the SciE-Lex project (Verdaguer et al., 2009). In cases where shorter bundles were held within longer bundles that occurred with similar frequency, the shorter bundles were eliminated from the list. Where there was overlap, but there were considerable frequency differences between the overlapping bundles, and each fragment either could function as an independent bundle or provided additional information about the longer string, the overlapping bundles were preserved. And since the lexical bundles can be ordered by keyword, with each set of like bundles headed by a prototypical form, overlapping bundles can be grouped together and considered as a unit. For all instances of overlap, subsumption, and/or repetition, there is a column that details how related bundles are connected to each other, and how they combine to form different variations of what is basically the same canonical sequence.

Aside from adopting these criteria, it was also decided to disregard lexical bundles ending in the articles *a, an* and *the,* most of which formed part of shorter bundles and did not supply further phraseological information that could justify their inclusion.

These steps were taken to minimize excessive repetitiveness within the final list, without sacrificing any of the variational detail given by overlapping bundles. For those situations where users would like to retrieve any of the bundles affected by the exclusion criteria, they are also given access to the list of these deleted bundles (see Appendix 2).

3. Lack of information on use in context of bundles in published lists

Byrd and Coxhead (2010) also consider it essential that instructors and students be given more detailed information about the use of lexical bundles in context. However, in most published research reports on lexical bundles, there is limited room for information beyond frequency and statistical counts and a few examples of significant usage patterns within the text. Moreover, the readers of these reports usually have no access to data beyond those included by the author, since, as Byrd and Coxhead (2010) point out, much published research is based on privately held corpora.

In this regard, the present list is different from many other lists of lexical bundles. As stated previously, it is more than just an inventory of frequently occurring lexical sequences, as it offers information beyond frequency counts, MI scores and structural and functional classifications. Several possible variations of prototypical bundles are presented, and additional concordance analysis was carried out to uncover other variants beyond those shown by the lexical bundles themselves. All attested functions of multifunctional bundles are considered, and context-specific information on these multiple functions is given. Authentic examples of lexical bundles in their real contexts of use are also supplied where applicable. Usage notes are available for those bundles that require further clarification, especially in cases of variation, multifunctionality and difficulty for non-native speakers.

This level of detail is provided in order to ensure greater support for teachers and materials designers, not only for easier selection of lexical bundles for pedagogical uses, but also for more effective presentation of this type of multi-word units in classrooms and teaching materials.

4. Lack of face validity for some EAP students

Another important factor that can impede the introduction of lexical bundles into EAP courses is their apparent lack of face validity for students. Teachers wishing to work with lexical bundles in the classroom may encounter some resistance from students, who may initially find it strange to look at language phraseologically (Hill, Lewis, & Lewis, 2000), or may be unwilling to learn entire word strings when learning single words is complicated enough (Coxhead, 2008), or may not see what makes studying lexical bundles worth the effort.

The issue of face validity is addressed in this study by ensuring that the lexical bundles that make it to the final list are the most beneficial for its target users: non-native scientists aiming to write scientific reports in English, as well as language practitioners who teach courses and design writing tools and language-learning materials for this particular audience. Frequency criteria were used to identify those lexical bundles that occur most frequently in published scientific articles, and statistical criteria were used to select only those words that combine for a reason and not only by chance. Exclusion criteria were applied to eliminate as much noise as possible and preserve only those lexical bundles that have, if not structural integrity, pragmatic integrity: the specialized discourse functions performed by lexical sequences that give even grammatically incomplete strings a degree of pragmatic adequacy and pedagogical validity (O'Keeffe et al., 2007). The lexical bundles on the list are classified according to these functions, and are thus linked to such concrete textual actions as introducing topics, comparing and contrasting, citing sources and stating conclusions, which many a non-native student or professional scientist or even native apprentice writer has struggled with.

The fact that all lexical bundles on the list have specific discourse functions makes it evident that phraseological competence strongly influences writing competence. Given that written texts are the main form of assessment in most universities (Jones & Haywood, 2004), and the success of academic careers continues to be measured by the number of research publications, writing proficiency remains crucial to a scientist's development. Convincing students of the value of lexical bundles thus becomes a matter of making them aware that, as Wray suggests, the functions of formulaic sequences serve "the promotion of the [user's] interests" (2002, p. 95), whether it may be to get good grades on a paper, to graduate successfully from a degree program, or to write a research article that can be accepted for publication in a journal.

5. Contradiction between analytical approach in teaching and use as unanalyzed chunks

The advent of computers has given researchers a level of linguistic observation that before was impossible. The most subliminal lexical patterns, which in the past have been ignored in favor of the most opaque, psychologically salient idiomatic units, can now be detected and analyzed using large corpora and increasingly refined corpus tools.

The most natural next step seems to be to transmit this knowledge about previously unnoticed recurrent lexical sequences to learners, in order to improve their understanding of how their target language works. However, Wray (2000) questions this practice and points out the inherent contradiction between the non-analytical nature of native-speaker use of formulaic language and teaching these same sequences through conscious analysis in textbooks or in the classroom.

Spöttl and McCarthy (2003) and O'Keeffe et al. (2007) acknowledge the validity of Wray's argument, but counter it by claiming that at least some degree of conscious linguistic analysis is required during the learning process, and that the language classroom is exactly the place where this kind of reflection can and should be encouraged. This is so that the learner can gradually acquire a repertoire of phraseological items, and as this repertoire grows, it becomes easier for the learner to use multi-word units in a more natural, native-like manner. Just as with grammatical structures or single words, acquisition can be achieved through repeated exposure, something that the present list of lexical bundles intends to facilitate and promote.

6. Having students read enough text to encounter the lexical bundles frequently enough for learning

The final, and perhaps most daunting, challenge involved in the teaching of lexical bundles is ensuring that students are given the level of exposure to lexical bundles required for efficient learning. Given the incremental nature of vocabulary acquisition (Schmitt, 2000), learning the appropriate use of lexical bundles can be achieved only after a number of exposures (O'Keeffe et al., 2007). Byrd and Coxhead stress the need for a proper understanding of learners' objectives, echoing Nation's (2009) advice to "focus on learning and teaching lexical items today that will be useful for learners tomorrow" (Byrd & Coxhead, 2010, p. 56). Researchers also agree on the importance of providing students with plenty of opportunities to encounter academic vocabulary in their chosen disciplines, such as through extensive reading activities (Byrd & Coxhead, 2010; Cortes, 2004; Coxhead, 2008; O'Keeffe et al., 2007). From this perspective, a discipline- and genre-specific approach like the one adopted in the present study can be seen as an important contribution, as the list it generated contains lexical bundles that learners in the health sciences, whether they be undergraduates or professional scientists, are most likely to come across when reading academic prose in their specific subject areas.

Multiple focused encounters with the use of lexical bundles in context should also be supported by awareness-raising activities (Byrd & Coxhead, 2010; O'Keeffe et al., 2007). Useful lexical sequences are not always the most salient, especially for learners, and teachers can draw attention to them in class materials through such means as underlining and color highlighting (Jones & Haywood, 2004; O'Keeffe et al., 2007). Students can also be instructed to keep track of lexical bundles they have learned by recording them in vocabulary notebooks (Byrd & Coxhead, 2010; Nation, 2001; O'Keeffe et al., 2007; Schmitt, 2000), class vocabulary boxes (Coxhead, 2004) or a space on the class whiteboard (Byrd & Coxhead, 2010). Such measures provide opportunities for reviewing and feedback and increase the likelihood of remembering and successful retrieval (Webb, 2007).

Encouraging learners to use lexical bundles in their own writing is also crucial to building phraseological knowledge, although several investigations have shown that this is far from being an easy task (Cortes, 2006; Coxhead, 2008; Jones & Haywood, 2004). Factors such as faulty memorization techniques, the aversion to risk-taking and committing mistakes and the tendency to rely on familiar phrases, make it difficult for learners to employ lexical bundles in their own written production (Cortes, 2004). To help students overcome these barriers and practice using lexical bundles in their output, Coxhead (2008) recommends introducing activities such as paraphrasing, summary writing and quotation practice.

The present list of target bundles can also promote the use of lexical bundles in student writing through its application as a writing aid. Since the lexical bundles are classified according to their functions, it is possible for users to access the list based on what they wish to convey in the text they are composing. The list can also be used as a basis for selecting phraseological content for more sophisticated reference tools, with the SciE-Lex Electronic Combinatory Dictionary being a notable example (Verdaguer et al., 2009).

A few studies have proposed specific teaching activities that teachers can use to teach lexical bundles to their students (Cortes, 2006; Jones & Haywood, 2004; Neely & Cortes, 2009). These exercises involve doing comprehension tasks, identifying lexical bundles and/or their functions in a source text, comparing the use of bundles in different text samples or text types, filling gaps in a text extract with the appropriate bundles, rewriting whole paragraphs using a given set of bundles and writing entire essays. Neely and Cortes (2009) even suggest the use of concordancing activities designed for lexical-bundle instruction. There is as yet very limited information on the long-term effectiveness of these teaching techniques, and so far only a few examples of these exercises with a restricted number of lexical bundles have made it to published research reports. However, the list of target bundles can facilitate the selection of lexical bundles for use with these activities, for EAP teachers who wish to use these exercises in their classrooms or materials designers who wish to include them in their textbooks and learning aids.

7. Concluding remarks

Using Byrd and Coxhead's (2010) six challenges as a framework, this chapter summarized the contributions the present investigation has made to the study of lexical bundles for pedagogical purposes. It also explained how the list of bundles the study produced can be used to effectively incorporate these multi-word units of meaning into EAP classrooms and teaching materials, an important step towards closing the gap between the language skills taught to and learned by EAP students and those they need to become successful academic writers in English.

Chapter VIII

Conclusions and recommendations

The present dissertation is a corpus-based investigation of the frequency, structure and functions of lexical bundles in English scientific writing, whose main objective was to create a list of lexical bundles of practical application to EAP pedagogy. The study, which was conducted within the framework of the SciE-Lex dictionary project, was guided by the same basic principles that the SciE-Lex team followed in the creation of a list of lexical bundles to be incorporated into in the second, expanded version of the dictionary (Verdaguer et al., 2009).

At the beginning of the study, four research questions were established in order to achieve the goal of the investigation. This concluding chapter addresses each of these questions as a summary of the dissertation's major findings and contributions to phraseology research.

Answering the first research question entailed the identification of the most frequently occurring lexical bundles in a 1.3 million-word sample of the HSC, here termed target bundles, after Cortes (2004). Creating the original list, which was carried out by a computer using frequency criteria, was only the first step in this process. The automatically generated list was also refined and enriched through the application of the MI statistic and a set of exclusion criteria defined by the pedagogical aims of the study. This highlights the importance of using statistical measures to complement frequency criteria in the identification of lexical bundles, in order to avoid generating an unnecessarily large number of items of undifferentiated value. It also confirms the necessity of using ad hoc intuitive decisions as methodological support for corpus-based procedures, especially in the case of pedagogically motivated investigations such as this study and the SciE-Lex project.

The filtering process was followed by the equally important step of organizing the lexical bundles in such a way that the semantic and structural links between similar bundles were addressed. This was made possible by grouping similar bundles together using shared keywords, following the SciE-Lex investigation (Verdaguer et al., 2009), and by using the concept of prototypical bundle, which is based on Sinclair's (2004) idea of canonical units of meaning, to head each group of like bundles.

The second research question involved the exploration of the structural and functional features of the lexical bundles through concordance analysis, and their categorization using modified versions of Biber et al.'s (1999) structural and Hyland's (2008a) functional taxonomies.

The results of this structural and functional analysis show how lexical bundles contribute to the distinctive nature of scientific writing, and how they help scientists pursue their agenda as academic writers. The frequencies and patterns of use of the different functional categories demonstrate that authors of scientific papers use research-oriented bundles to describe research objects and procedures with clarity and precision, text-oriented bundles to organize and connect their ideas and put them in the correct context, and participant-oriented bundles to establish a positive, engaging dynamic with their intended readers. The judicious use of these three main functions results in a coherent, well-structured and audience-accessible scientific article whose convincing arguments are grounded in relevant literature, sound methodological principles and reliable data.

The classification of lexical bundles into structural and functional groups is also significant in that it lends them face validity for teaching and shows their value as pedagogical items. The fact that many of the functions writers are expected to perform in academic writing are routinely realized through lexical bundles following specific structural patterns—e.g., noun phrase + *of* for research-oriented functions, prepositional-phrase fragments for text-oriented functions, anticipatory-*it* structures for participant-oriented functions—can facilitate the teaching and learning of these fundamental writing strategies.

The last two research questions are with regard to the existence of target bundles in the non-native corpus of scientific research writing, and the differences between the native and non-native corpora in terms of the frequency, structure and functions of these target bundles.

The study uncovered two significant differences between the native and non-native texts. First is the non-native writers' overuse of certain bundles, a tendency that results in unnecessary repetitiveness and lack of variation. Second is the non-native writers' restricted use of participant-oriented bundles, which points to their limited awareness of the usage and importance of this particular function. This is an issue that needs to be addressed, since participant-oriented bundles mainly serve to convey writer stance, to engage and persuade the reader, to hedge, boost and qualify propositions, and to distance oneself or claim ownership of statements, all of which are functions central to successful argumentation. It seems that non-native scientists can benefit from exposure to a wider range of formulaic sequences that can help enrich their variety of expression, and from being taught how to use participant-oriented bundles to produce a more rhetorically effective scientific article.

Apart from these two noteworthy findings, the present analysis found few differences between the native and non-native corpus in the use of lexical bundles, a result that contrasts with similar comparative studies (Chen & Baker, 2010; Cortes, 2004). This outcome seems to support the notion of a developmental trend in the use of lexical bundles, given that this study involved the comparison of equivalent text types, written by two groups of scientists that, despite being differentiated by nativeness, share the same goal of writing a scientific paper for publication, and the same degree of expertise in their fields. The expert status of the non-native authors examined here lends credence to the supposition that these scientists have had sufficient exposure to the use of lexical sequences in scientific writing to be able to incorporate these formulas into their own written production. The study's results also emphasize the need to control for topic, text type and author profile when choosing non-native texts to compare with a native corpus, so that dissimilarities between the corpora can be more readily attributed to linguistic factors and not to external features such as subject matter, register, genre or scientific competence.

By endeavoring to answer the four research questions, this dissertation has not only contributed to a better understanding of how lexical bundles are employed by native and non-native science writers, it has also produced a practical list of lexical bundles that can aid teachers, materials designers and other EAP practitioners in the introduction of these multi-word units into classrooms and teaching and learning tools. The preceding chapter discussed how this list helps overcome some of the hurdles to the successful teaching of lexical bundles identified by Byrd and Coxhead (2010). The list resolves these issues by supplying detailed information on how the list was developed, enabling users to order the list by different criteria such as frequency, structure, function and keyword, addressing the semantic and functional relationships between similar bundles, providing contextual examples and usage notes where necessary, and giving face validity to lexical bundles by linking them to specific functions. More than just being a discrete, frequency-ordered inventory of phraseological items, this study's list of lexical bundles in scientific writing is a valuable resource that can facilitate the selection of multi-word units for a variety of teaching applications.

The present dissertation builds upon the most current, innovative phraseological studies to make its own methodological contribution to the study of lexical bundles. However, it is not without its limitations. One such limitation is the restricted size of the non-native corpus used in the study, which necessitated the use of Cortes' (2004) target-bundle methodology. This procedure was able to indicate whether the non-native writers used the same bundle structures and functions as their native counterparts, but it could in no way ascertain whether they were using other forms to perform the same bundle functions, whether the target bundles they were using were indeed used with the same function as in the native texts, or whether the non-native writers were able to come up with "near hits" (Thewissen, 2008). Although an independently generated list of three- to six-word lexical bundles that occur at least ten times in the non-native corpus showed encouraging similarity to the list of target bundles, the fact remains that several questions can be sufficiently answered only by the separate extraction of lexical bundles from a non-native corpus of comparable size to the native corpus. This and similar studies could also have a lot to gain from having more than one rater for the application of exclusion criteria and assignment of lexical bundle functions, and using inter-rater reliability measures to ensure the consistency of rater judgments (V. Cortes, personal communication, March 18, 2010).

The study of lexical bundles and of phraseology in general is a relatively young and rapidly developing field with no shortage of avenues for new research. It is important to acknowledge that lexical bundles are just one piece in a large phraseological puzzle, and one essential task for those interested in this type of multi-word unit is to find out where lexical bundles fit in the bigger picture together with the many other types of lexical patterning, so as to determine how to give students and non-native academics the best possible access to the full range of formulaic language they need to communicate efficiently in academic settings (Byrd & Coxhead, 2010).

It is also necessary to take phraseology research to the classroom itself, so that the teaching approaches being proposed in pedagogy-oriented investigations can be evaluated and improved. It is only in this way that teachers and learners can fully benefit from all the groundbreaking advances in the study of multi-word units of meaning.

More research is also required to settle the debate over whether a core academic phrasal lexicon exists, as Simpson-Vlach and Ellis' (2010) results indicate, or if academic formulas are strictly discipline-specific, as Hyland's (2008a) findings suggest. As this study uses a domain-restricted corpus, the question of whether or not lexical bundles transcend disciplinary boundaries was not a problem it was designed to resolve, although it is certainly one that deserves further inquiry.

Biber et al. (2004) recognize that the complex issues surrounding the use of multi-word units in discourse can only be fully comprehended through a multiplicity of approaches and perspectives. It is hoped that this study, which has explored lexical bundles from a pedagogical perspective, represents a significant contribution towards reaching a complete understanding of the crucial role played by lexical bundles in written academic communication.

References

Adolphs, S., & Durow, V. (2004). Sociocultural integration and the development of formulaic sentences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 107-126). Amsterdam: John Benjamins.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: John Benjamins.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 101-122). Oxford: Oxford University Press.

Altenberg, B., & Eeg-Olofsson, M. (1990). Phraseology in spoken English. In J. Aarts & W. Meijs (Eds.), *Theory and practice in corpus linguistics* (pp. 1-26). Amsterdam: Rodopi.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of make in native and non-native student writing. *Applied Linguistics*, *22*, 173-195.

Altenberg, B., & Tapper, M. (1998). The use of adverbial connectors in advanced Swedish learners' written English. In S. Granger (Ed.), *Learner English on computer* (pp. 80-93). London: Addison-Wesley and Longman.

Anthony, L. (2006). Developing a freeware, multiplatform corpus analysis toolkit for the technical writing classroom. *IEEE Transactions on Professional Communication*, *49*(3), 275-286.

Barlow, M. (2004). *Collocate*.

Benson, M., Benson, E., & Ilson, R. (2010). *The BBI dictionary of English word combinations* (3rd ed.). Amsterdam: John Benjamins.

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D., & Finegan, E. (1994). Intra-textual variation within medical research articles. In N. Oostdiijk & P. De Haan (Eds.), *Corpus-based research into language* (pp. 201-221). 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* (pp. 71-93). Frankfurt/Main: Peter Lang.

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., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

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

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Bloch, J., & Chi, L. (1995). A comparison of the use of citations in Chinese and English academic discourse. In D. Belcher & G. Braine (Eds.), *Academic* *writing in a second language: Essays on research and pedagogy* (p. X-274). Norwood, NJ: Ablex.

Bloor, M., & Bloor, T. (1991). Cultural expectations and socio-pragmatic failure in academic writing. In P. Adams, B. Heaton, & P. Howarth (Eds.), *Socio-cultural issues in English for academic purposes* (pp. 1-12). Basingstoke: Modern English Publications and the British Council.

Butler, C. S. (1997). Repeated word combinations in spoken and written text: Some implications for Functional Grammar. In C. S. Butler, J. H. Connolly, R. A. Gatward, & R. M. Vismans (Eds.), *A fund of ideas: Recent developments in Functional Grammar* (pp. 60-77). Amsterdam: IFOTT University of Amsterdam.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, *5*, 31-64.

Cambridge advanced learner's dictionary. (2008). (3rd ed.). Cambridge: Cambridge University Press.

Cammack, R. (Ed.). (2006). *Oxford dictionary of biochemistry and molecular biology* (Rev. ed.). Oxford-New York: Oxford University Press.

Chang, Y.-Y., & Swales, J. M. (1999). Informal elements in English academic writing: Threats or opportunities for advanced non-native speakers? In K. Hyland & C. Candlin (Eds.), *Writing: Texts, processes and practices* (pp. 145–167). London-New York: Longman.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, *14*(2), 13 February 2011.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411-433.

Church, K. W., & Hanks, P. (1990). Word association, norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22-29.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, *20*, 56-71.

Conrad, S. (1996). Academic discourse in two disciplines: professional writing and student development in biology and history. (Unpublished doctoral dissertation). Northern Arizona University, United States.

Cortes, V. (2002a). Lexical bundles in freshman composition. In R. Reppen, S. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam: John Benjamins.

Cortes, V. (2002b). Lexical bundles in academic writing in history and biology. (Unpublished doctoral dissertation). Northern Arizona University, United States.

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*, 391-406.

Cowie, A. P. (1988). Stable and creative aspects of vocabulary use. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 126-137). London: Longman.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 3168-3171). Oxford: Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213-238.

Coxhead, A. (2004). Using a class vocabulary box: How, why, when, where, and who. *RELC Guidelines*, *26*(2), 19-23.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 149-161). Amsterdam: John Benjamins.

De Cock, S. (2003). Recurrent sequences of words in native speaker and advanced learner spoken and written English. (Unpublished doctoral dissertation). Université catholique de Louvain-la-Neuve, Belgium.

De Cock, S., Gilquin, G., Granger, S., Lefer, M.-A., Paquot, M., & Ricketts, S. (2007). Improve your writing skills. In M. Rundell (Ed.), *Macmillan English dictionary for advanced learners* (2nd ed., p. IW1-IW50). Oxford: Macmillan Education.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. *Learner English on computer* (pp. 67–79). London-New York: Addison Wesley Longman.

De Schryver, G. M. (2003). Lexicographers' dreams in the electronic dictionary age. *International Journal of Lexicography*, *16*(2).

DeCarrico, J., & Nattinger, J. R. (1988). Lexical phrases for the comprehension of academic lectures. *English for Specific Purposes*, *7*(2), 91-102.

Ellis, N. C., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly*, *42*(3), 375-396.

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations.*

Firth, J. R. (1951). Modes of meaning. *Papers in linguistics*, *1934-1951* (pp. 118–149). London: Oxford University Press.

Fløttum, K., Dahl, T., & Kinn, T. (2006). *Academic voices across languages and disciplines*. Amsterdam: John Benjamins.

Flowerdew, L. (1998). Integrating expert and interlanguage computer corpora findings on causality: Discoveries for teachers and students. *English for Specific Purposes*, *17*, 329-345.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy & R. Roseberry (Eds.), *Small corpus studies and ELT* (pp. 363-379). Amsterdam: John Benjamins.

Francis, G., Hunston, S., & Manning, E. (1996). *Collins COBUILD Grammar Patterns 1: Verbs.* London: Harper Collins.

Francis, G., Hunston, S., & Manning, E. (1998). *Collins COBUILD Grammar Patterns 2: Nouns and Adjectives.* London: Harper Collins.

Gilquin, G., & Paquot, M. (2007). Spoken features in learner academic writing: Identification, explanation and solution. *Proceedings of the Fourth Corpus Linguistics Conference* (pp. 1-12). Birmingham, United Kingdom.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, *6*, 319-335.

Girard, M., & Sionis, C. (2004). The functions of formulaic speech in the L2 class. *Pragmatics*, *14*(1), 31–53.

Gläser, R. (1998). The stylistic potential of phraseological units in the light of genre analysis. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 125-143). Oxford: Oxford University Press.

Gledhill, C. (1995). Collocation and genre analysis: The discourse function of collocation in cancer research abstracts and articles. *Zeitschrift für Anglistik und Amerikanistik*, (1), 1-26.

Gledhill, C. (2000a). *Collocations in science writing*. Tubingen: Gunter Narr.

Gledhill, C. (2000b). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115-135.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in contrast: Text-based cross-linguistic studies* (pp. 37-51). Lund: Lund University Press.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly*, *37*(3), 538-546.

Granger, S., & Meunier, F. (Eds.). (2008a). *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Granger, S., & Paquot, M. (2009a). Customising a general EAP dictionary to meet learner needs. In S. Granger & M. Paquot (Eds.), *eLexicography in the 21st century: New challenges, new applications: Proceedings of eLex 2009* (pp. 77-86). Louvain la Neuve: Presses universitaires de Louvain Cahiers du CENTAL.

Granger, S., & Paquot, M. (2009b). Lexical verbs in academic discourse: a corpus-driven study of expert and learner use. In M. Charles, D. Pecorari, & S.

Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 193-214). London: Continuum.

Granger, S., & Paquot, M. (2009c). In search of a general academic vocabulary: A corpus-driven study. *International Conference on Options and Practices of L.S.A.P Practitioners.*

Granger, S., & Rayson, P. (1998). Automatic lexical profiling of learner texts. In S. Granger (Ed.), *Learner English on computer* (pp. 119-131). London: Addison-Wesley and Longman.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*, 9-29.

Granger, S., & Meunier, F. (2008b). Phraseology in language learning and teaching: Where to from here? In S. Granger & F. Meunier (Eds.), *Phraseology in language learning and teaching* (pp. 247-252). Amsterdam: John Benjamins.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 28-49). Amsterdam: John Benjamins.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, *24*(2), 287-297.

Hanks, P. (1987). Definitions and explanations. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project in lexical computing* (pp. 116-136). London: Collins.

Harwood, N. (2005a). “We do not seem to have a theory… the theory I present here attempts to fill this gap”: Inclusive and exclusive pronouns in academic writing. *Applied Linguistics*, *26*(3), 343-375.

Harwood, N. (2005b). “Nowhere has anyone attempted...In this article I aim to do just that”: A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics*, *37*(8), 1207-1231.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237–258.

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*(4), 367-383.

Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching collocations* (pp. 88–117). Hove: Language Teaching Publications.

Hinkel, E. (2002). *Second language writers' text*. Mahwah, NJ: Erlbaum.

Hoey, M. (2004). Lexical priming and the properties of text. In A. Partington, J. Morley, & L. Haarman (Eds.), *Corpora and discourse* (pp. 385-412). Frankfurt: Peter Lang.

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Hoffman, C. (2000). The spread of English and the growth of multilingualism with English in Europe. In J. Cenoz & U. Jessner (Eds.), *English in Europe: The acquisition of a third language* (pp. 1-21). Clevedon: Multilingual Matters.

Holmes, J. (1984). Modifying illocutionary force. *Journal of Pragmatics*, *8*, 345-365.

Hornby, A. S., Turnbull, J., Lea, D., Parkinson, D., Phillips, P., Francis, B., Webb, S., et al. (2010). *Oxford advanced learner's dictionary* (8th ed.). Oxford: Oxford University Press.

Howarth, P. (1996a). *Phraseology in English academic writing : Some implications for language learning and dictionary making* (Vol. 75). Tübingen: M. Niemeyer.

Howarth, P. (1996b). How conventional is academic writing? In M. Hewings & T. Dudley-Evans (Eds.), *Evaluation and course design in EAP* (pp. 192-204). Prentice Hall Macmillan and British Council.

Howarth, P. (1998a). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161-186). Oxford: Oxford University Press.

Howarth, P. (1998b). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24-44.

Hunston, S. (2006). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2009). The usefulness of corpus-based descriptions of English for learners: The case of relative frequency. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 141–154). Amsterdam: John Benjamins.

Hunston, S., & Francis, G. (2000). *Pattern grammar : a corpus-driven approach to the lexical grammar of English* (Vol. 4). Amsterdam: John Benjamins.

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal*, *51*(3), 208-216.

Huntley, H. (2006). *Essential academic vocabulary: Mastering the complete academic word list*. Boston: Houghton Mifflin Company.

Hyland, K. (2000). *Disciplinary discourses*. Harlow: Pearson Education.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes*, *20*(3), 207-226.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. New York: Routledge.

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).

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing*, *6*(2), 183-205.

Jespersen, O. (1917). *Negation in English and other languages*. Copenhagen: A.F. Host.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen & Unwin.

Johns, T. (1991). From print out to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria*, *10*, 14-34.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus linguistics* (pp. 107-117). Amsterdam: Rodopi.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 269-291). Amsterdam: John Benjamins.

Kaszubski, P. (2000). Selected aspects of lexicon, phraseology and style in the writing of Polish advanced learners of English: A contrastive, corpus-based approach. (Unpublished doctoral dissertation). Adam Mickiewicz University, Poznań, Poland.

Kennedy, C., & Thorp, D. (2007). A corpus investigation of linguistic responses to an IELTS Academic Writing task. In L. Taylor & P. Falvey (Eds.), *IELTS collected paper: Research in speaking and writing assessment* (pp. 316–378). Cambridge: Cambridge University Press.

Lewis, M. (2000). *Teaching collocation*. Hove: LTP.

Louw, B. (1993). Irony in the text or insincerity in the writer? In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 157-176). Amsterdam: John Benjamins.

Louw, B. (2000). Contextual prosodic theory: Bringing semantic prosodies to life. In C. Heffer, H. Sauntson, & G. Fox (Eds.), *Words in context: A tribute to John Sinclair on his retirement* (pp. 48-94). Birmingham: University of Birmingham.

Luzón Marco, M. J. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes*, *19*(1), 63-86.

Luzón Marco, M. J. (2001). Procedural vocabulary: Lexical signalling of conceptual relations in discourse. *Applied Linguistics*, *20*, 1-21.

Macmillan collocations dictionary for learners of English. (2010). Oxford: Macmillan Education.

Major, M. (2006). *Longman exams dictionary*. Harlow: Longman.

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.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Martínez, I. (2003). Aspects of theme in the method and discussion sections of biology journal articles in English. *Journal of English for Academic Purposes*, *2*, 103-123.

Martínez, I. (2005). Native and non-native writers' use of first person pronouns in the different sections of biology research articles in English. *Journal of Second Language Writing*, *14*, 174-190.

Mauranen, A. (1993). Contrastive ESP rhetoric: Metatext in Finnish-English economics texts. *English for Specific Purposes*, *12*(1), 3-22.

McCarthy, M., & O'Dell, F. (2005). *English collocations in use*. Cambridge: Cambridge University Press.

McCarthy, M., & O'Dell, F. (2008). *Academic vocabulary in use*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone: Student's book 1*. Cambridge: Cambridge University Press.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal*, *32*, 97–114.

McIntosh, C., Francis, B., & Poole, R. (2009). *Oxford collocations dictionary for students of English*. Oxford: Oxford University Press.

Mel'čuk, I. (1998). Collocations and lexical functions. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 23-53). Oxford: Oxford University Press.

Milton, J. (1999). Lexical thickets and electronic gateways: Making text accessible by novice writers. In C. Candlin & K. Hyland (Eds.), *Writing: Texts, processes and practices* (pp. 221-243). London: Longman.

Milton, J., & Tsang, E. (1991). A corpus-based study of logical connectors in EFL students' writing: Directions for future research. In R. Pemberton & E. Tsang (Eds.), *Studies in lexis* (pp. 215-246). Hong Kong: The Hong Kong University of Science and Technology.

Moon, R. (1995). *Collins COBUILD dictionary of idioms*. London: Harper Collins.

Moon, R. (2008). Sinclair, phraseology and lexicography. *International Journal of Lexicography*, *21*(3), 243-254.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2009). *Teaching ESL/EFL reading and writing*. New York: Routledge.

Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

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

Neff, J. (2008). Contrasting English-Spanish interpersonal discourse phrases. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 85-100). Amsterdam: John Benjamins.

Neff, J. A., Ballesteros, F., Dafouz, E., Diez, F., Martínez, R., & Prieto, R. (2004). The expression of writer stance in native and non-native argumentative texts. In R. Facchinetti & F. Palmer (Eds.), *English modality in perspective* (pp. 141-161). Frankfurt: Peter Lang.

Neff, J., & Bunce, C. (2006). Pragmatic word order errors and discourse-grammar interdependence. In C. Mourón & T. Moralejo (Eds.), *Actas de la IV conferencia de lingüística contrastiva* (pp. 697-705). Santiago de Compostela: Universidad de Santiago.

Nesselhauf, N. (2004). What are collocations? In D. Allerton, N. Nesselhauf, & P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (pp. 1-21). Basel: Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

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

Oakes, M. P. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oakey, D. (2002). A corpus-based study of the formal and functional variation of a lexical phrase in different academic disciplines. In R. Reppen, S. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 111-129). Amsterdam: John Benjamins.

Ozturk, I. (2007). The textual organization of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes*, *26*, 25-38.

Palmer, H. E. (1933). *Second interim report on English collocations*. Tokyo: Kaitakusha.

Paquot, M. (2007). *EAP vocabulary in EFL learner writing: from extraction to analysis: A phraseology-oriented approach*. (Unpublished doctoral dissertation). Université catholique de Louvain, Belgium.

Paquot, M. (2010). Academic vocabulary in learner writing: From extraction to analysis. London: Continuum.

Parkinson, D., & Francis, B. (2006). *Oxford idioms dictionary for learners of English*. Oxford: Oxford University Press.

Partington, A. (2004). "Utterly content in each other's company": Semantic prosody and semantic preference. *International Journal of Corpus Linguistics*, *9*(1), 131-156.

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.

Pecorari, D. (2009). Formulaic language in biology: A topic-specific investigation. In M. Charles, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 91-106). London: Continuum.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107-118). London-New York: Longman.

Peters, A. (1983). *The units of language acquisition*. Cambridge, MA: Cambridge University Press.

Pickering, L., & Byrd, P. (2008). Investigating connections between spoken and written academic English: Lexical bundles in the AWL and in MICASE. In D. Belcher & A. Hirvela (Eds.), *Oral/Literate Connection: Perspectives on L2 speaking, writing and other media interactions* (pp. 110-132). Ann Arbor: University of Michigan Press.

Porto, M. (1998). Lexical phrases and language teaching. *Forum*, *36*(3). Retrieved from http://exchanges.state.gov/forum/vols/vol36/no3/p22.htm

Prodromou, L. (2005). *"You see, it's sort of tricky for the L2-user": The puzzle of idiomaticity in English as a lingua franca*. (Doctoral dissertation). University of Nottingham, Nottingham, United Kingdom. Retrieved from http://etheses.nottingham.ac.uk/1180/1/423643.pdf.

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 128-143). London: Longman.

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 140-162.

Rundell, M. (2002). Good old-fashioned lexicography: Human judgement and the limits of automation. In M. H. Corréard (Ed.), *Lexicography and natural language processing: A festschrift in honour of B.T.S. Atkins* (pp. 138-155). EURALEX.

Rundell, M. (2007). *Macmillan English dictionary for advanced learners* (2nd ed.). Oxford: Macmillan Education.

Salazar, D. (2008). *Modality in student argumentative writing: A corpus-based comparative study of American, Filipino and Spanish novice writers*. (Unpublished thesis). University of Barcelona, Spain.

Salazar, D. (2010). Lexical bundles in Philippine and British scientific English. *Philippine Journal of Linguistics*, *41*, 94-109.

Salazar, D., & Verdaguer, I. (2009). Polysemous verbs and modality in native and non-native argumentative writing: A corpus-based study. *International Journal of English Studies*, special issue.

Salazar, D., Ventura, A., & Verdaguer, I. (2011). A cross-disciplinary analysis of personal and impersonal features in Spanish and English scientific writing. Presented at the Annual Conference of the American Association for Applied Linguistics, Chicago, United States.

Schmid, H. J. (2003). Collocation: Hard to pin down, but bloody useful. *ZAA*, *51*(3), 235-258.

Schmitt, D., & Schmitt, N. (2005). *Focus on vocabulary: Mastering the academic word list*. New York: Pearson Education.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Scott, M., & Tribble, C. (2006). *Textual patterns: Keywords and corpus analysis in language education*. Amsterdam: John Benjamins.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487-512.

Simpson, R. (2004). Stylistic features of academic speech: The role of formulaic expressions. In U. Connor & T. A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 37-64). Amsterdam: John Benjamins.

Sinclair, J. (1987). *Collins COBUILD English language dictionary* (1st ed.). London: Collins.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1996). The search for units of meaning. *Textus*, *IX*, 75-106.

Sinclair, J. (1998). The lexical item. In E. Weigand (Ed.), *Contrastive lexical semantics* (pp. 1-24). Amsterdam: John Benjamins.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. London/New York: Routledge.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books.

Sinclair, J., & Moon, R. (1989). *Collins Cobuild dictionary of phrasal verbs*. London: Harper Collins.

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The OSTI report*. London: Continuum.

Spöttl, C., & McCarthy, M. (2003). Formulaic utterances in the multi-lingual context. In J. Cenoz, U. Jessner, & B. Hufeisen (Eds.), *The multilingual lexicon* (pp. 133–151). Dordrecht: Kluwer.

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209-243.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, *7*(2), 215-244.

Stubbs, M. (2007a). Quantitative data on multi-word sequences in English: The case of the word “world.” In M. Hoey, M. Mahlberg, M. Stubbs, & W. Teubert (Eds.), *Text, discourse and corpora: Theory and analysis* (pp. 163-189). London: Continuum.

Stubbs, M. (2007b). An example of frequent English phraseology: Distribution, structures and functions. In R. Facchinetti (Ed.), *Corpus linguistics 25 years on* (pp. 89-105). Amsterdam: Rodopi.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, *10*(1), 61-104.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge: Cambridge University Press.

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

Swales, J. M. (1997). English as a Tyrannosaurus rex. *World Englishes*, *16*, 373-382.

Swales, J. M. (2008). Foreword. In D. Belcher & A. Hirvela (Eds.), *The oral-literate connection: Perspectives on L2 speaking, writing, and other media interactions* (p. v-viii). Ann Arbor: University of Michigan Press.

Swales, J. M., & Feak, C. (2004). *Academic writing for graduate students: Essential tasks and skills*. Ann Arbor: University of Michigan Press.

Tang, R., & John, S. (1999). The "I" in identity: Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes*, *18*, S23-S39.

Tannen, D. (1987). Repetition in conversation as spontaneous formulaicity. *Text*, *7*(3), 215-243.

Tarone, E., Dwyer, S., Gillette, S., & Icke, V. (1998). On the use of the passive and active voice in astrophysics journal papers: With extensions to other languages and other fields. *English for Specific Purposes*, *17*(1), 113-132.

Thewissen, J. (2008). The phraseological errors of French-, German-, and Spanish speaking EFL learners: Evidence from an error-tagged learner corpus. *Proceedings from the 8th Teaching and Language Corpora Conference* (pp. 300-306). Presented at the TaLC8, Lisbon, Portugal: Associação de Estudos e de Investigação Científica do ISLA-Lisboa.

Thompson, G., & Thetela, P. (1995). The sound of one hand clapping: The management of interaction in written discourse. *Text*, *15*(5), 103-127.

Thurston, J., & Candlin, C. (1997). *Exploring academic English: A workbook for student essay writing*. Sydney: NCELTR Publications.

Thurston, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of Academic English. *English for Specific Purposes*, *17*, 267-280.

Verdaguer, I. (2003). Collocations in scientific language: A contrastive study. *Studies in contrastive linguistics: Proceedings of the Third International Contrastive Linguistics Conference* (pp. 633-639).

Verdaguer, I., Comelles, E., Laso, N. J., Giménez, E., & Salazar, D. (2009). SciE-Lex: An electronic lexical database for the Spanish medical community. *eLexicography in the 21st century: New challenges, new applications: Proceedings of eLex 2009* (pp. 325-334). Louvain la Neuve: Presses universitaires de Louvain Cahiers du CENTAL.

Verdaguer, I., Poch, A., Laso, N. J., & Giménez, E. (2008). Scie-Lex: A lexical database of collocations in scientific English for Spanish scientists. Presented at the 13th Euralex International Conference, Barcelona, Spain.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46-65.

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151-171.

Williams, G. (2002a). Corpus-driven lexicography and the specialised dictionary: Headword extraction for the Parasitic Plant Research Dictionary. In A. Braasch (Ed.), *Proceedings of the 10th EURALEX International Conference*. Denmark: Sprogteknologi.

Williams, G. (2002b). In search of representativity in specialised corpora: Categorisation through collocation. *International Journal of Corpus Linguistics*, *7*(1), 43-64.

Williams, G. (2003). From meaning to words and back: Corpus linguistics and specialised lexicography. *ASp*, *39-40*, 10 February 2011.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics*, *21*(4), 463–489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

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

Articles from the Health Science Corpus used in examples

[1] Adams, I. R., & Kilmartin, J. V. (1999). Localization of core spindle pole body (SPB) components during SPB duplication in Saccharomyces cerevisiae. *The Journal of Cell Biology*, *145*(4), 809-823.

[2] Ades, S. E., Connolly, L. E., Alba, B. M., & Gross, C. A. (1999). The Escherichia coli sigma(E)-dependent extracytoplasmic stress response is controlled by the regulated proteolysis of an anti-sigma factor. *Genes & Development*, *13*(18), 2449-2461.

[3] Albert, S., Will, E., & Gallwitz, D. (1999). Identification of the catalytic domains and their functionally critical arginine residues of two yeast GTPase-activating proteins specific for Ypt/Rab transport GTPases. *The EMBO Journal*, *18*(19), 5216-5225. doi:10.1093/emboj/18.19.5216

[4] Ashbaugh, C. D., Warren, H. B., Carey, V. J., & Wessels, M. R. (1998). Molecular analysis of the role of the group A streptococcal cysteine protease, hyaluronic acid capsule, and M protein in a murine model of human invasive soft-tissue infection. *The Journal of Clinical Investigation*, *102*(3), 550-560. doi:10.1172/JCI3065

[5] Berul, C. I., Maguire, C. T., Aronovitz, M. J., Greenwood, J., Miller, C., Gehrmann, J., Housman, D., et al. (1999). DMPK dosage alterations result in atrioventricular conduction abnormalities in a mouse myotonic dystrophy model. *The Journal of Clinical Investigation*, *103*(4), R1-7. doi:10.1172/JCI5346

[6] Beye, M., Hunt, G. J., Page, R. E., Fondrk, M. K., Grohmann, L., & Moritz, R. F. (1999). Unusually high recombination rate detected in the sex locus region of the honey bee (Apis mellifera). *Genetics*, *153*(4), 1701-1708.

[7] Blott, E. J., Higgins, C. F., & Linton, K. J. (1999). Cysteine-scanning mutagenesis provides no evidence for the extracellular accessibility of the nucleotide-binding domains of the multidrug resistance transporter P-glycoprotein. *The EMBO Journal*, *18*(23), 6800-6808. doi:10.1093/emboj/18.23.6800

[8] Bobanovic, L. K., Laine, M., Petersen, C. C., Bennett, D. L., Berridge, M. J., Lipp, P., Ripley, S. J., et al. (1999). Molecular cloning and immunolocalization of a novel vertebrate trp homologue from Xenopus. *The Biochemical Journal*, *340 (Pt 3)*, 593-599.

[9] Bowers, C. W., & A J Dombroski. (1999). A mutation in region 1.1 of sigma70 affects promoter DNA binding by Escherichia coli RNA polymerase holoenzyme. *The EMBO Journal*, *18*(3), 709-716. doi:10.1093/emboj/18.3.709

[10] Callaghan, J., Simonsen, A., Gaullier, J. M., Toh, B. H., & Stenmark, H. (1999). The endosome fusion regulator early-endosomal autoantigen 1 (EEA1) is a dimer. *The Biochemical Journal*, *338 (Pt 2)*, 539-543.

[11] Calvi, B. R., Lilly, M. A., & Spradling, A. C. (1998). Cell cycle control of chorion gene amplification. *Genes & Development*, *12*(5), 734-744.

[12] Christ, F., Schoettler, S., Wende, W., Steuer, S., Pingoud, A., & Pingoud, V. (1999). The monomeric homing endonuclease PI-SceI has two catalytic centres for cleavage of the two strands of its DNA substrate. *The EMBO Journal*, *18*(24), 6908-6916. doi:10.1093/emboj/18.24.6908

[13] Churchill, J. J., Anderson, D. G., & Kowalczykowski, S. C. (1999). The RecBC enzyme loads RecA protein onto ssDNA asymmetrically and independently of chi, resulting in constitutive recombination activation. *Genes & Development*, *13*(7), 901-911.

[14] Clay, M. A., Cehic, D. A., Pyle, D. H., Rye, K. A., & Barter, P. J. (1999). Formation of apolipoprotein-specific high-density lipoprotein particles from lipid-free apolipoproteins A-I and A-II. *The Biochemical Journal*, *337 (Pt 3)*, 445-451.

[15] Clemens, S., Kim, E. J., Neumann, D., & Schroeder, J. I. (1999). Tolerance to toxic metals by a gene family of phytochelatin synthases from plants and yeast. *The EMBO Journal*, *18*(12), 3325-3333. doi:10.1093/emboj/18.12.3325

[16] Cottingham, F. R., Gheber, L., Miller, D. L., & Hoyt, M. A. (1999). Novel roles for saccharomyces cerevisiae mitotic spindle motors. *The Journal of Cell Biology*, *147*(2), 335-350.

[17] Cross, M., Kieft, R., Sabatini, R., Wilm, M., de Kort, M., van der Marel, G. A., van Boom, J. H., et al. (1999). The modified base J is the target for a novel DNA-binding protein in kinetoplastid protozoans. *The EMBO Journal*, *18*(22), 6573-6581. doi:10.1093/emboj/18.22.6573

[18] Cryderman, D. E., Morris, E. J., Biessmann, H., Elgin, S. C., & Wallrath, L. L. (1999). Silencing at Drosophila telomeres: nuclear organization and

chromatin structure play critical roles. *The EMBO Journal*, *18*(13), 3724-3735. doi:10.1093/emboj/18.13.3724

[19] Danielson, L. A., Sherwood, O. D., & Conrad, K. P. (1999). Relaxin is a potent renal vasodilator in conscious rats. *The Journal of Clinical Investigation*, *103*(4), 525-533. doi:10.1172/JCI5630

[20] Dockrell, D. H., Badley, A. D., Villacian, J. S., Heppelmann, C. J., Algeciras, A., Ziesmer, S., Yagita, H., et al. (1998). The expression of Fas Ligand by macrophages and its upregulation by human immunodeficiency virus infection. *The Journal of Clinical Investigation*, *101*(11), 2394-2405. doi:10.1172/JCI1171

[21] Donaldson, A. D., Fangman, W. L., & Brewer, B. J. (1998). Cdc7 is required throughout the yeast S phase to activate replication origins. *Genes & Development*, *12*(4), 491-501.

[22] Dudley, A. T., Godin, R. E., & Robertson, E. J. (1999). Interaction between FGF and BMP signaling pathways regulates development of metanephric mesenchyme. *Genes & Development*, *13*(12), 1601-1613.

[23] Dunckley, T., & Parker, R. (1999). The DCP2 protein is required for mRNA decapping in Saccharomyces cerevisiae and contains a functional MutT motif. *The EMBO Journal*, *18*(19), 5411-5422. doi:10.1093/emboj/18.19.5411

[24] Eckley, D. M., Gill, S. R., Melkonian, K. A., Bingham, J. B., Goodson, H. V., Heuser, J. E., & Schroer, T. A. (1999). Analysis of dynactin subcomplexes reveals a novel actin-related protein associated with the arp1 minifilament pointed end. *The Journal of Cell Biology*, *147*(2), 307-320.

[25] Ellis, T. P., Lukins, H. B., Nagley, P., & Corner, B. E. (1999). Suppression of a nuclear aep2 mutation in Saccharomyces cerevisiae by a base

substitution in the 5'-untranslated region of the mitochondrial oli1 gene encoding subunit 9 of ATP synthase. *Genetics*, *151*(4), 1353-1363.

[26] Everett, R. D., Earnshaw, W. C., Findlay, J., & Lomonte, P. (1999). Specific destruction of kinetochore protein CENP-C and disruption of cell division by herpes simplex virus immediate-early protein Vmw110. *The EMBO Journal*, *18*(6), 1526-1538. doi:10.1093/emboj/18.6.1526

[27] Ferguson, C. A., Tucker, A. S., Christensen, L., Lau, A. L., Matzuk, M. M., & Sharpe, P. T. (1998). Activin is an essential early mesenchymal signal in tooth development that is required for patterning of the murine dentition. *Genes & Development*, *12*(16), 2636-2649.

[28] Fischer, E. G., Riewald, M., Huang, H. Y., Miyagi, Y., Kubota, Y., Mueller, B. M., & Ruf, W. (1999). Tumor cell adhesion and migration supported by interaction of a receptor-protease complex with its inhibitor. *The Journal of Clinical Investigation*, *104*(9), 1213-1221. doi:10.1172/JCI7750

[29] Fisk, D. G., Walker, M. B., & Barkan, A. (1999). Molecular cloning of the maize gene crp1 reveals similarity between regulators of mitochondrial and chloroplast gene expression. *The EMBO Journal*, *18*(9), 2621-2630. doi:10.1093/emboj/18.9.2621

[30] Fletcher, C. M., Pestova, T. V., Hellen, C. U., & Wagner, G. (1999). Structure and interactions of the translation initiation factor eIF1. *The EMBO Journal*, *18*(9), 2631-2637. doi:10.1093/emboj/18.9.2631

[31] Forrester, W. C., Perens, E., Zallen, J. A., & Garriga, G. (1998). Identification of Caenorhabditis elegans genes required for neuronal differentiation and migration. *Genetics*, *148*(1), 151-165.

[32] Gant, T. M., Harris, C. A., & Wilson, K. L. (1999). Roles of LAP2 proteins in nuclear assembly and DNA replication: truncated LAP2beta proteins alter lamina assembly, envelope formation, nuclear size, and DNA replication efficiency in Xenopus laevis extracts. *The Journal of Cell Biology*, *144*(6), 1083-1096.

[33] Gladwin, M. T., Schechter, A. N., Shelhamer, J. H., Pannell, L. K., Conway, D. A., Hrinczenko, B. W., Nichols, J. S., et al. (1999). Inhaled nitric oxide augments nitric oxide transport on sickle cell hemoglobin without affecting oxygen affinity. *The Journal of Clinical Investigation*, *104*(7), 937-945. doi:10.1172/JCI7637

[34] Gollob, J. A., Schnipper, C. P., Orsini, E., Murphy, E., Daley, J. F., Lazo, S. B., Frank, D. A., et al. (1998). Characterization of a novel subset of CD8(+) T cells that expands in patients receiving interleukin-12. *The Journal of Clinical Investigation*, *102*(3), 561-575. doi:10.1172/JCI3861

[35] Goodman, J. L., Nelson, C. M., Klein, M. B., Hayes, S. F., & Weston, B. W. (1999). Leukocyte infection by the granulocytic ehrlichiosis agent is linked to expression of a selectin ligand. *The Journal of Clinical Investigation*, *103*(3), 407-412. doi:10.1172/JCI4230

[36] Harmon, F. G., & Kowalczykowski, S. C. (1998). RecQ helicase, in concert with RecA and SSB proteins, initiates and disrupts DNA recombination. *Genes & Development*, *12*(8), 1134-1144.

[37] Harrison, D. A., McCoon, P. E., Binari, R., Gilman, M., & Perrimon, N. (1998). Drosophila unpaired encodes a secreted protein that activates the JAK signaling pathway. *Genes & Development*, *12*(20), 3252-3263.

[38] Haynes, B. F., Hale, L. P., Weinhold, K. J., Patel, D. D., Liao, H. X., Bressler, P. B., Jones, D. M., et al. (1999). Analysis of the adult thymus in reconstitution of T lymphocytes in HIV-1 infection. *The Journal of Clinical Investigation*, *103*(4), 453-460. doi:10.1172/JCI5201

[39] Hodge, C. A., Colot, H. V., Stafford, P., & Cole, C. N. (1999). Rat8p/Dbp5p is a shuttling transport factor that interacts with Rat7p/Nup159p and Gle1p and suppresses the mRNA export defect of xpo1-1 cells. *The EMBO Journal*, *18*(20), 5778-5788. doi:10.1093/emboj/18.20.5778

[40] Hodges, D., Cripps, R. M., O'Connor, M. E., & Bernstein, S. I. (1999). The role of evolutionarily conserved sequences in alternative splicing at the 3' end of Drosophila melanogaster myosin heavy chain RNA. *Genetics*, *151*(1), 263-276.

[41] Holland, E. C., Hively, W. P., DePinho, R. A., & Varmus, H. E. (1998). A constitutively active epidermal growth factor receptor cooperates with disruption of G1 cell-cycle arrest pathways to induce glioma-like lesions in mice. *Genes & Development*, *12*(23), 3675-3685.

[42] Hosfield, C. M., Elce, J. S., Davies, P. L., & Jia, Z. (1999). Crystal structure of calpain reveals the structural basis for Ca(2+)-dependent protease activity and a novel mode of enzyme activation. *The EMBO Journal*, *18*(24), 6880-6889. doi:10.1093/emboj/18.24.6880

[43] Howden, R., Park, S. K., Moore, J. M., Orme, J., Grossniklaus, U., & Twell, D. (1998). Selection of T-DNA-tagged male and female gametophytic mutants by segregation distortion in Arabidopsis. *Genetics*, *149*(2), 621-631.

[44] Huttley, G. A., Smith, M. W., Carrington, M., & O'Brien, S. J. (1999). A scan for linkage disequilibrium across the human genome. *Genetics*, *152*(4), 1711-1722.

[45] Ilgoutz, S. C., Mullin, K. A., Southwell, B. R., & McConville, M. J. (1999). Glycosylphosphatidylinositol biosynthetic enzymes are localized to a stable tubular subcompartment of the endoplasmic reticulum in Leishmania mexicana. *The EMBO Journal*, *18*(13), 3643-3654. doi:10.1093/emboj/18.13.3643

[46] Janscak, P., MacWilliams, M. P., Sandmeier, U., Nagaraja, V., & Bickle, T. A. (1999). DNA translocation blockage, a general mechanism of cleavage site selection by type I restriction enzymes. *The EMBO Journal*, *18*(9), 2638-2647. doi:10.1093/emboj/18.9.2638

[47] Jeddeloh, J. A., Bender, J., & Richards, E. J. (1998). The DNA methylation locus DDM1 is required for maintenance of gene silencing in Arabidopsis. *Genes & Development*, *12*(11), 1714-1725.

[48] Kopin, A. S., Mathes, W. F., McBride, E. W., Nguyen, M., Al-Haider, W., Schmitz, F., Bonner-Weir, S., et al. (1999). The cholecystokinin-A receptor mediates inhibition of food intake yet is not essential for the maintenance of body weight. *The Journal of Clinical Investigation*, *103*(3), 383-391. doi:10.1172/JCI4901

[49] Lal, A. S., Parker, P. J., & Segal, A. W. (1999). Characterization and partial purification of a novel neutrophil membrane-associated kinase capable of phosphorylating the respiratory burst component p47phox. *The Biochemical Journal*, *338 (Pt 2)*, 359-366.

[50] Leone, G., DeGregori, J., Yan, Z., Jakoi, L., Ishida, S., Williams, R. S., & Nevins, J. R. (1998). E2F3 activity is regulated during the cell cycle and is required for the induction of S phase. *Genes & Development*, *12*(14), 2120-2130.

[51] Linder, J. U., Engel, P., Reimer, A., Krüger, T., Plattner, H., Schultz, A., & Schultz, J. E. (1999). Guanylyl cyclases with the topology of mammalian adenylyl cyclases and an N-terminal P-type ATPase-like domain in Paramecium, Tetrahymena and Plasmodium. *The EMBO Journal*, *18*(15), 4222-4232. doi:10.1093/emboj/18.15.4222

[52] Lloyd, J. R., Landschütze, V., & Kossmann, J. (1999). Simultaneous antisense inhibition of two starch-synthase isoforms in potato tubers leads to accumulation of grossly modified amylopectin. *The Biochemical Journal*, *338 (Pt 2)*, 515-521

[53] MacPhee, C. H., Moores, K. E., Boyd, H. F., Dhanak, D., Ife, R. J., Leach, C. A., Leake, D. S., et al. (1999). Lipoprotein-associated phospholipase A2, platelet-activating factor acetylhydrolase, generates two bioactive products during the oxidation of low-density lipoprotein: use of a novel inhibitor. *The Biochemical Journal*, *338 (Pt 2)*, 479-487.

[54] Magnan, C., Collins, S., Berthault, M. F., Kassis, N., Vincent, M., Gilbert, M., Pénicaud, L., et al. (1999). Lipid infusion lowers sympathetic nervous activity and leads to increased beta-cell responsiveness to glucose. *The Journal of Clinical Investigation*, *103*(3), 413-419. doi:10.1172/JCI3883

[55] Manson, J. C., Jamieson, E., Baybutt, H., Tuzi, N. L., Barron, R., McConnell, I., Somerville, R., et al. (1999). A single amino acid alteration (101L) introduced into murine PrP dramatically alters incubation time of

transmissible spongiform encephalopathy. *The EMBO Journal*, *18*(23), 6855-6864. doi:10.1093/emboj/18.23.6855

[56] Marston, A. L., Thomaides, H. B., Edwards, D. H., Sharpe, M. E., & Errington, J. (1998). Polar localization of the MinD protein of Bacillus subtilis and its role in selection of the mid-cell division site. *Genes & Development*, *12*(21), 3419-3430.

[57] Mayer, K. M., & Forney, J. D. (1999). A mutation in the flanking 5'-TA-3' dinucleotide prevents excision of an internal eliminated sequence from the Paramecium tetraurelia genome. *Genetics*, *151*(2), 597-604.

[58] McCulloch, R., & Barry, J. D. (1999). A role for RAD51 and homologous recombination in Trypanosoma brucei antigenic variation. *Genes & Development*, *13*(21), 2875-2888.

[59] McKee, B. D., Wilhelm, K., Merrill, C., & Ren, X. (1998). Male sterility and meiotic drive associated with sex chromosome rearrangements in Drosophila. Role of X-Y pairing. *Genetics*, *149*(1), 143-155.

[60] McLemore, M. L., Poursine-Laurent, J., & Link, D. C. (1998). Increased granulocyte colony-stimulating factor responsiveness but normal resting granulopoiesis in mice carrying a targeted granulocyte colony-stimulating factor receptor mutation derived from a patient with severe congenital neutropenia. *The Journal of Clinical Investigation*, *102*(3), 483-492. doi:10.1172/JCI3216

[61] McLure, K. G., & Lee, P. W. (1999). p53 DNA binding can be modulated by factors that alter the conformational equilibrium. *The EMBO Journal*, *18*(3), 763-770. doi:10.1093/emboj/18.3.763

[62] Miller, R. K., Matheos, D., & Rose, M. D. (1999). The cortical localization of the microtubule orientation protein, Kar9p, is dependent upon actin and proteins required for polarization. *The Journal of Cell Biology*, *144*(5), 963-975.

[63] Mintz, P. J., Patterson, S. D., Neuwald, A. F., Spahr, C. S., & Spector, D. L. (1999). Purification and biochemical characterization of interchromatin granule clusters. *The EMBO Journal*, *18*(15), 4308-4320. doi:10.1093/emboj/18.15.4308

[64] Morel, J. B., & Dangl, J. L. (1999). Suppressors of the arabidopsis lsd5 cell death mutation identify genes involved in regulating disease resistance responses. *Genetics*, *151*(1), 305-319.

[65] Mosi, R., & Withers, S. G. (1999). Synthesis and kinetic evaluation of 4-deoxymaltopentaose and 4-deoxymaltohexaose as inhibitors of muscle and potato alpha-glucan phosphorylases. *The Biochemical Journal*, *338 (Pt 2)*, 251-256.

[66] Moy, T. I., & Silver, P. A. (1999). Nuclear export of the small ribosomal subunit requires the ran-GTPase cycle and certain nucleoporins. *Genes & Development*, *13*(16), 2118-2133.

[67] Nichols, M. D., DeAngelis, K., Keck, J. L., & Berger, J. M. (1999). Structure and function of an archaeal topoisomerase VI subunit with homology to the meiotic recombination factor Spo11. *The EMBO Journal*, *18*(21), 6177-6188. doi:10.1093/emboj/18.21.6177

[68] Ollmann, M. M., Lamoreux, M. L., Wilson, B. D., & Barsh, G. S. (1998). Interaction of Agouti protein with the melanocortin 1 receptor in vitro and in vivo. *Genes & Development*, *12*(3), 316-330.

[69] Orford, K., Orford, C. C., & Byers, S. W. (1999). Exogenous expression of beta-catenin regulates contact inhibition, anchorage-independent growth, anoikis, and radiation-induced cell cycle arrest. *The Journal of Cell Biology*, *146*(4), 855-868.

[70] Parks, E. J., Krauss, R. M., Christiansen, M. P., Neese, R. A., & Hellerstein, M. K. (1999). Effects of a low-fat, high-carbohydrate diet on VLDL-triglyceride assembly, production, and clearance. *The Journal of Clinical Investigation*, *104*(8), 1087-1096. doi:10.1172/JCI6572

[71] Peakman, M., Stevens, E. J., Lohmann, T., Narendran, P., Dromey, J., Alexander, A., Tomlinson, A. J., et al. (1999). Naturally processed and presented epitopes of the islet cell autoantigen IA-2 eluted from HLA-DR4. *The Journal of Clinical Investigation*, *104*(10), 1449-1457. doi:10.1172/JCI7936

[72] Phillips, J. W., & Berry, M. N. (1999). Long-term maintenance of low concentrations of fructose for the study of hepatic glucose phosphorylation. *The Biochemical Journal*, *337 (Pt 3)*, 497-501.

[73] Plotkin, L. I., Weinstein, R. S., Parfitt, A. M., Roberson, P. K., Manolagas, S. C., & Bellido, T. (1999). Prevention of osteocyte and osteoblast apoptosis by bisphosphonates and calcitonin. *The Journal of Clinical Investigation*, *104*(10), 1363-1374. doi:10.1172/JCI6800

[74] Pryde, F. E., & Louis, E. J. (1999). Limitations of silencing at native yeast telomeres. *The EMBO Journal*, *18*(9), 2538-2550. doi:10.1093/emboj/18.9.2538

[75] Rasnick, D., & Duesberg, P. H. (1999). How aneuploidy affects metabolic control and causes cancer. *The Biochemical Journal*, *340 (Pt 3)*, 621-630.

[76] Reckless, J., & Grainger, D. J. (1999). Identification of oligopeptide sequences which inhibit migration induced by a wide range of chemokines. *The Biochemical Journal*, *340 (Pt 3)*, 803-811.

[77] Rieseberg, L. H., Whitton, J., & Gardner, K. (1999). Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species. *Genetics*, *152*(2), 713-727.

[78] Ritchie, P. J., Decout, A., Amey, J., Mann, C. J., Read, J., Rosseneu, M., Scott, J., et al. (1999). Baculovirus expression and biochemical characterization of the human microsomal triglyceride transfer protein. *The Biochemical Journal*, *338 (Pt 2)*, 305-310.

[79] Rowland-Jones, S. L., Dong, T., Fowke, K. R., Kimani, J., Krausa, P., Newell, H., Blanchard, T., et al. (1998). Cytotoxic T cell responses to multiple conserved HIV epitopes in HIV-resistant prostitutes in Nairobi. *The Journal of Clinical Investigation*, *102*(9), 1758-1765. doi:10.1172/JCI4314

[80] Ruegger, M., Dewey, E., Gray, W. M., Hobbie, L., Turner, J., & Estelle, M. (1998). The TIR1 protein of Arabidopsis functions in auxin response and is related to human SKP2 and yeast grr1p. *Genes & Development*, *12*(2), 198-207.

[81] Russell, I. D., Grancell, A. S., & Sorger, P. K. (1999). The unstable F-box protein p58-Ctf13 forms the structural core of the CBF3 kinetochore complex. *The Journal of Cell Biology*, *145*(5), 933-950.

[82] Saphire, A. C., Bobardt, M. D., & Gallay, P. A. (1999). Host cyclophilin A mediates HIV-1 attachment to target cells via heparans. *The EMBO Journal*, *18*(23), 6771-6785. doi:10.1093/emboj/18.23.6771

[83] Sewell, M. M., Sherman, B. K., & Neale, D. B. (1999). A consensus map for loblolly pine (Pinus taeda L.). I. Construction and integration of individual

linkage maps from two outbred three-generation pedigrees. *Genetics*, 151(1), 321-330.

[84] Schmoll, D., Wasner, C., Hinds, C. J., Allan, B. B., Walther, R., & Burchell, A. (1999). Identification of a cAMP response element within the glucose- 6-phosphatase hydrolytic subunit gene promoter which is involved in the transcriptional regulation by cAMP and glucocorticoids in H4IIE hepatoma cells. *The Biochemical Journal*, *338 (Pt 2)*, 457-463.

[85] Sharp, D. J., McDonald, K. L., Brown, H. M., Matthies, H. J., Walczak, C., Vale, R. D., Mitchison, T. J., et al. (1999). The bipolar kinesin, KLP61F, cross-links microtubules within interpolar microtubule bundles of Drosophila embryonic mitotic spindles. *The Journal of Cell Biology*, *144*(1), 125-138.

[86] Sheehan, J. K., Howard, M., Richardson, P. S., Longwill, T., & Thornton, D. J. (1999). Physical characterization of a low-charge glycoform of the MUC5B mucin comprising the gel-phase of an asthmatic respiratory mucous plug. *The Biochemical Journal*, *338 (Pt 2)*, 507-513.

[87] Shuster, C. B., & Burgess, D. R. (1999). Parameters that specify the timing of cytokinesis. *The Journal of Cell Biology*, *146*(5), 981-992.

[88] Silliman, C. C., Voelkel, N. F., Allard, J. D., Elzi, D. J., Tuder, R. M., Johnson, J. L., & Ambruso, D. R. (1998). Plasma and lipids from stored packed red blood cells cause acute lung injury in an animal model. *The Journal of Clinical Investigation*, *101*(7), 1458-1467. doi:10.1172/JCI1841

[89] Smalley, M. J., Sara, E., Paterson, H., Naylor, S., Cook, D., Jayatilake, H., Fryer, L. G., et al. (1999). Interaction of axin and Dvl-2 proteins regulates

Dvl-2-stimulated TCF-dependent transcription. *The EMBO Journal*, *18*(10), 2823-2835. doi:10.1093/emboj/18.10.2823

[90] Smith, N. G., & Hurst, L. D. (1999). The causes of synonymous rate variation in the rodent genome. Can substitution rates be used to estimate the sex bias in mutation rate? *Genetics*, *152*(2), 661-673.

[91] Smith, T. K., Sharma, D. K., Crossman, A., Brimacombe, J. S., & Ferguson, M. A. (1999). Selective inhibitors of the glycosylphosphatidylinositol biosynthetic pathway of Trypanosoma brucei. *The EMBO Journal*, *18*(21), 5922-5930. doi:10.1093/emboj/18.21.5922

[92] Smolka, A. J., Larsen, K. A., Schweinfest, C. W., & Hammond, C. E. (1999). H,K-ATPase alpha subunit C-terminal membrane topology: epitope tags in the insect cell expression system. *The Biochemical Journal*, *340 (Pt 3)*, 601-611.

[93] Sparks, C. A., Morphew, M., & McCollum, D. (1999). Sid2p, a spindle pole body kinase that regulates the onset of cytokinesis. *The Journal of Cell Biology*, *146*(4), 777-790.

[94] Stark, M. R., Escher, D., & Johnson, A. D. (1999). A trans-acting peptide activates the yeast a1 repressor by raising its DNA-binding affinity. *The EMBO Journal*, *18*(6), 1621-1629. doi:10.1093/emboj/18.6.1621

[95] Taylor, A. F., & Smith, G. R. (1999). Regulation of homologous recombination: Chi inactivates RecBCD enzyme by disassembly of the three subunits. *Genes & Development*, *13*(7), 890-900.

[96] Thomson, S. C., Bachmann, S., Bostanjoglo, M., Ecelbarger, C. A., Peterson, O. W., Schwartz, D., Bao, D., et al. (1999). Temporal adjustment of the juxtaglomerular apparatus during sustained inhibition of proximal

reabsorption. *The Journal of Clinical Investigation*, *104*(8), 1149-1158. doi:10.1172/JCI5156

[97] Thress, K., Evans, E. K., & Kornbluth, S. (1999). Reaper-induced dissociation of a Scythe-sequestered cytochrome c-releasing activity. *The EMBO Journal*, *18*(20), 5486-5493. doi:10.1093/emboj/18.20.5486

[98] Tipnis, S. R., Blake, D. G., Shepherd, A. G., & McLellan, L. I. (1999). Overexpression of the regulatory subunit of gamma-glutamylcysteine synthetase in HeLa cells increases gamma-glutamylcysteine synthetase activity and confers drug resistance. *The Biochemical Journal*, *337 (Pt 3)*, 559-566.

[99] Trial, J., Baughn, R. E., Wygant, J. N., McIntyre, B. W., Birdsall, H. H., Youker, K. A., Evans, A., et al. (1999). Fibronectin fragments modulate monocyte VLA-5 expression and monocyte migration. *The Journal of Clinical Investigation*, *104*(4), 419-430. doi:10.1172/JCI4824

[100] Tumova, S., Hatch, B. A., Law, D. J., & Bame, K. J. (1999). Basic fibroblast growth factor does not prevent heparan sulphate proteoglycan catabolism in intact cells, but it alters the distribution of the glycosaminoglycan degradation products. *The Biochemical Journal*, *337 (Pt 3)*, 471-481.

[101] Turner, J., Hingorani, M. M., Kelman, Z., & O'Donnell, M. (1999). The internal workings of a DNA polymerase clamp-loading machine. *The EMBO Journal*, *18*(3), 771-783. doi:10.1093/emboj/18.3.771

[102] Viner, R. I., Ferrington, D. A., Williams, T. D., Bigelow, D. J., & Schöneich, C. (1999). Protein modification during biological aging: selective tyrosine nitration of the SERCA2a isoform of the sarcoplasmic reticulum

Ca2+-ATPase in skeletal muscle. *The Biochemical Journal*, *340 (Pt 3)*, 657-669.

[103] Walker, M. B., Roy, L. M., Coleman, E., Voelker, R., & Barkan, A. (1999). The maize tha4 gene functions in sec-independent protein transport in chloroplasts and is related to hcf106, tatA, and tatB. *The Journal of Cell Biology*, *147*(2), 267-276.

[104] Waltzer, L., & Bienz, M. (1999). A function of CBP as a transcriptional co-activator during Dpp signalling. *The EMBO Journal*, *18*(6), 1630-1641. doi:10.1093/emboj/18.6.1630

[105] Wardrop, A. J., Wicks, R. E., & Entsch, B. (1999). Occurrence and expression of members of the ferritin gene family in cowpeas. *The Biochemical Journal*, *337 (Pt 3)*, 523-530.

[106] Watts, A. D., Hunt, N. H., Wanigasekara, Y., Bloomfield, G., Wallach, D., Roufogalis, B. D., & Chaudhri, G. (1999). A casein kinase I motif present in the cytoplasmic domain of members of the tumour necrosis factor ligand family is implicated in "reverse signalling." *The EMBO Journal*, *18*(8), 2119-2126. doi:10.1093/emboj/18.8.2119

[107] Watts, B. A., 3rd, & Good, D. W. (1999). Hyposmolality stimulates apical membrane Na(+)/H(+) exchange and HCO(3)(-) absorption in renal thick ascending limb. *The Journal of Clinical Investigation*, *104*(11), 1593-1602. doi:10.1172/JCI7332

[108] Waxman, D., & Peck, J. R. (1999). Sex and adaptation in a changing environment. *Genetics*, *153*(2), 1041-1053.

[109] Webster, P., Ijdo, J. W., Chicoine, L. M., & Fikrig, E. (1998). The agent of Human Granulocytic Ehrlichiosis resides in an endosomal compartment. *The Journal of Clinical Investigation*, *101*(9), 1932-1941. doi:10.1172/JCI1544

[110] Wendland, B., Steece, K. E., & Emr, S. D. (1999). Yeast epsins contain an essential N-terminal ENTH domain, bind clathrin and are required for endocytosis. *The EMBO Journal*, *18*(16), 4383-4393. doi:10.1093/emboj/18.16.4383

[111] Wisdom, R., Johnson, R. S., & Moore, C. (1999). c-Jun regulates cell cycle progression and apoptosis by distinct mechanisms. *The EMBO Journal*, *18*(1), 188-197. doi:10.1093/emboj/18.1.188

[112] Xu, F. Y., Kelly, S. L., & Hatch, G. M. (1999). N-Acetylsphingosine stimulates phosphatidylglycerolphosphate synthase activity in H9c2 cardiac cells. *The Biochemical Journal*, *337 (Pt 3)*, 483-490.

[113] Yeager, T. R., DeVries, S., Jarrard, D. F., Kao, C., Nakada, S. Y., Moon, T. D., Bruskewitz, R., et al. (1998). Overcoming cellular senescence in human cancer pathogenesis. *Genes & Development*, *12*(2), 163-174.

[114] Young, J. C., DeWitt, N. D., & Sussman, M. R. (1998). A transgene encoding a plasma membrane H+-ATPase that confers acid resistance in Arabidopsis thaliana seedlings. *Genetics*, *149*(2), 501-507.

Appendix 1

Original list of lexical bundles extracted by *Collocate*

N	Mutual Inf.	Bundle	N	Mutual Inf.	Bundle
906	8.518913	the presence of	31	11.369719	and stored at
632	8.646156	in the presence	31	11.306004	large number of
625	15.556469	data not shown	31	11.141007	be able to
541	13.109891	in the presence of	31	10.961011	that do not
495	8.907142	in the absence	31	10.9444	is not known
481	8.218921	the absence of	31	10.80622	the location of the
387	13.240078	in the absence of	31	10.049087	not shown these
360	15.934436	materials and methods	31	9.534187	been shown that
307	14.240235	as well as	31	9.500299	were identified by
273	7.14912	the number of	31	9.077895	was performed with
259	6.858231	the effect of	31	8.970642	as a single
244	15.403582	as described previously	31	8.890332	was required for
237	7.730166	the ability of	31	8.608501	a portion of
227	10.177912	as described in	31	8.520431	may be a
216	10.021748	shown in figure	31	7.60026	the course of
212	9.372453	consistent with the	31	7.141794	the samples were
209	11.443076	been shown to	31	6.932774	decrease in the
203	6.676684	the addition of	31	6.929324	proportion of the
195	5.469964	the expression of	31	6.929042	the same as
194	11.402583	is required for	31	6.546725	a loss of
190	9.596848	was used to	31	6.319044	determined by the
189	9.46708	in response to	31	6.114871	role for the
183	8.239267	a number of	31	5.857001	by the presence
180	13.490686	results not shown	31	5.831379	the stimulation of
176	7.03375	the effects of	31	5.233924	to have a
173	14.31053	for 30 min	31	5.222478	content of the
172	6.044131	region of the	31	4.801063	of the second
169	5.263514	expression of the	31	4.760724	the time of
168	12.796726	for 10 min	31	3.911549	levels of the
168	7.466129	the level of	30	27.912335	little is known about
165	14.306728	it is possible	30	21.641929	would be expected to
164	15.343361	to determine whether	30	20.974654	these data indicate that
164	6.491655	the role of	30	17.461612	carried out using
161	5.597374	one of the	30	14.581846	with the exception of
158	10.366571	the fact that	30	14.291275	these data indicate
156	14.604337	has been shown	30	14.256518	could be detected
154	11.591088	is consistent with	30	13.327734	have shown that the
154	10.649726	for 1 h	30	13.205487	may play a
154	8.558108	in addition to	30	12.871886	be noted that
154	8.021226	the amount of	30	12.217107	20 min at
153	6.968485	present in the	30	12.132765	activity was measured
152	5.645064	analysis of the	30	11.966266	two copies of
149	6.72299	the formation of	30	11.935256	have also been
148	10.799778	in this study	30	11.923179	in conjunction with
147	14.589377	it has been	30	11.634777	it may be
146	20.813609	it is possible that	30	10.973418	the majority of the
146	18.976404	at room temperature	30	10.327546	were transferred to
146	11.778793	is possible that	30	10.302538	to be involved
145	4.660801	the activity of	30	9.597991	are known to
144	10.970233	was added to	30	9.585737	led to a
144	5.118288	in which the	30	9.199847	were detected by
143	9.830042	the possibility that	30	9.046094	explanation for the
142	6.836724	the rate of	30	8.605915	evidence for a
139	8.326431	the basis of	30	8.034198	due to a
137	16.903517	for review see	30	7.993526	the exception of
137	8.680423	associated with the	30	7.810479	in contrast with
136	10.896266	were incubated with	30	7.611057	tip of the
132	11.636866	on the basis	30	6.983058	result in a

131	5.089786	all of the	30	5.936251	in a similar
130	12.172597	we found that	29	33.811544	it should be noted that
130	7.086636	end of the	29	24.776728	should be noted that
129	16.29173	on the basis of	29	23.184465	performed as described previously
129	4.759286	of the two	29	21.470911	it is not clear
128	10.124116	in order to	29	16.701764	is not required for
126	11.192163	have shown that	29	16.67996	has been implicated
126	5.992579	described in the	29	14.913475	are shown in figure
124	12.172034	the present study	29	14.668916	this suggests that the
122	5.101921	the binding of	29	14.280907	together these results
122	4.411629	activity of the	29	14.209559	results are means
121	6.192634	structure of the	29	13.707829	in the absence of the
119	11.0729	was determined by	29	13.327034	in some cases
119	9.70822	shown to be	29	13.149454	but does not
119	8.662623	suggest that the	29	13.094347	in the presence of the
118	17.079535	were carried out	29	12.87643	was purchased from
118	10.669763	based on the	29	11.884964	with the use of
117	7.620022	involved in the	29	11.837515	inserted into the
116	6.625662	in the same	29	11.10698	is an important
115	8.109195	to determine the	29	10.448816	by the presence of
113	8.323654	as shown in	29	10.345282	this is consistent
113	7.94308	required for the	29	9.815086	in 50 mm
113	3.693238	to that of	29	9.369857	released from the
112	11.206109	an increase in	29	9.30839	to be determined
112	8.557439	are shown in	29	7.911087	was added and
112	7.246018	the use of	29	7.773763	lead to the
112	6.518452	in the present	29	7.528436	implicated in the
112	4.959036	each of the	29	7.512259	added to a
111	10.289522	a variety of	29	7.482539	a set of
110	8.847331	suggesting that the	29	7.428461	and characterization of
110	8.241462	due to the	29	7.382313	was present in
109	11.207677	for 5 min	29	7.196933	with the use
109	8.628752	the majority of	29	6.992128	in support of
108	13.764302	for 15 min	29	6.889801	evidence for the
107	8.652743	were used to	29	6.682257	the medium was
107	6.955978	the regulation of	29	6.670881	reduction in the
106	24.610113	see materials and methods	29	6.667262	in a single
106	13.576438	see materials and	29	6.562147	modification of the
106	7.860096	relative to the	29	6.456281	a fraction of
105	14.287511	no effect on	29	6.314236	it is a
105	8.86862	in contrast to	29	5.264132	case of the
104	19.858479	has been shown to	29	4.883083	by using the
104	13.257763	as described in the	29	4.361803	formation of the
104	5.659265	the activation of	28	26.163907	expressed as a percentage of
101	14.946081	as described above	28	21.475876	expressed as a percentage
101	9.00203	similar to that	28	19.36569	results are consistent with
101	8.717828	suggests that the	28	18.827431	data not shown this
101	8.106348	a role in	28	16.882975	this is consistent with
101	5.353757	presence of the	28	14.356509	significantly different from
101	4.860459	sequence of the	28	14.331764	extracts were prepared
100	12.029767	likely to be	28	13.623435	directed against the
100	5.913965	most of the	28	13.435628	carried out in
96	9.170596	according to the	28	12.873359	we have identified
96	8.811367	effect on the	28	12.77737	results are consistent
96	7.778514	members of the	28	12.600215	see table 1
96	3.240658	cells in the	28	12.163258	not shown thus
96	1.824877	that of the	28	12.111424	can be used
95	10.194058	it is not	28	11.717481	the tip of the
95	4.45752	the results of	28	11.371248	used to determine
94	16.867197	was carried out	28	10.945364	small number of
94	7.607569	in the case	28	10.713625	in this report
94	7.350548	the production of	28	10.46153	was prepared from
94	5.179453	function of the	28	10.449649	for at least
93	12.36017	we show that	28	10.411291	the notion that
93	11.766364	are consistent with	28	10.312546	this result is
93	7.461606	part of the	28	10.299332	was subjected to
93	7.386339	is shown in	28	10.045704	at the restrictive
93	6.843965	increase in the	28	10.033743	an average of
93	6.464483	the loss of	28	9.972419	are associated with
92	12.128537	this suggests that	28	9.953102	are representative of
92	9.800269	responsible for the	28	9.802976	was prepared by

92	9.351441	a role for	28	8.675527	we tested the
92	7.397226	not shown the	28	8.595551	is important to
91	9.4093	the presence of the	28	8.416434	and transferred to
91	8.107317	compared with the	28	8.17825	in the dark
91	6.913945	the case of	28	8.027439	4 h in
90	13.384828	results suggest that	28	7.638179	the function of the
90	12.232864	in the case of	28	7.460102	linked to the
90	11.354294	were treated with	28	7.397007	part of a
90	5.116717	the function of	28	7.214504	was found in
89	6.155405	the localization of	28	6.224054	defects in the
89	5.434559	activation of the	28	5.843365	the range of
88	11.420898	were obtained from	28	5.82287	figure 4 the
88	9.20395	to bind to	28	5.795622	the results are
88	7.386959	in figure 1	28	5.793204	figure 3 the
88	6.300854	the position of	28	5.316261	figure 5 the
88	5.416784	the levels of	28	4.704166	the products of
87	9.646587	a series of	28	3.858327	only in the
86	16.978962	in the present study	28	3.818703	addition of the
86	7.162954	changes in the	28	1.33931	and at the
85	7.627997	by the addition	27	24.69406	washed three times with
84	12.298954	by the addition of	27	17.929432	are likely to be
84	7.37939	added to the	27	17.130247	a wide range
84	6.03128	the concentration of	27	16.97986	a large number of
83	11.080614	are required for	27	16.944923	three independent experiments
83	10.614944	found to be	27	15.740882	previous studies have
83	9.614297	there is a	27	15.426757	does not contain
83	7.367536	the ability to	27	14.701858	in the case of the
82	9.268945	was found to	27	14.658458	in the presence of a
82	8.934008	indicating that the	27	13.852251	be involved in the
81	9.467843	by use of	27	13.359602	an increase in the
81	6.541549	results in a	27	13.028417	with 1 ml of
81	6.120782	role in the	27	12.676638	results demonstrate that
81	0.635117	and in the	27	12.291829	a large number
80	10.172949	for 2 h	27	11.854774	which has been
80	10.05184	was used as	27	11.836482	was supported by
80	8.706089	between the two	27	11.738795	it is important
80	6.916711	the accumulation of	27	11.221776	is based on
80	5.728322	observed in the	27	11.05658	depends on the
79	16.393296	had no effect	27	10.896369	the indicated times
79	14.623057	presence or absence	27	10.813315	in a number of
79	12.626405	appear to be	27	10.560858	is unlikely to
78	13.405048	it is likely	27	10.474525	as measured by
78	12.571346	appears to be	27	10.420889	there is an
77	18.792023	the presence or absence	27	10.193975	at a density
77	13.068312	have been shown	27	10.051325	5 min at
77	10.866613	for 4 h	27	10.050142	not due to
77	9.519501	the observation that	27	9.793155	that has been
77	8.03953	the presence or	27	9.701414	not bind to
77	7.604492	corresponding to the	27	9.480273	by treatment with
77	7.591968	a total of	27	9.366998	the case of the
77	6.133186	similar to the	27	9.184614	to demonstrate that
77	5.540557	the structure of	27	9.146208	also observed in
77	4.976299	used in the	27	9.029516	the conclusion that
76	8.514396	that it is	27	8.275637	on the surface
76	6.049161	regions of the	27	7.845756	to estimate the
75	10.140612	as described by	27	7.625127	was performed in
75	8.483633	or presence of	27	7.487415	were detected in
75	7.914029	1 ml of	27	7.45116	a change in
75	5.070241	effect of the	27	7.124721	to changes in
74	15.131722	have been identified	27	7.067607	fragment from the
74	14.846236	these results suggest	27	6.823437	in fig 1
74	10.27182	were determined by	27	6.519595	the efficiency of
74	9.07775	or absence of	27	6.439425	the behavior of
74	7.91013	by addition of	27	6.197667	the isolation of
74	7.518482	side of the	27	6.125284	in a number
74	6.050876	position of the	27	6.076539	defect in the
73	9.540532	the requirement for	27	6.034168	the detection of
73	8.444697	used in this	27	5.992456	for the first
73	5.951958	the result of	27	5.955858	in the top
72	12.239797	with respect to	27	5.671565	used as the
72	10.068542	we examined the	27	4.543938	it is the

72	9.700905	were grown in	27	4.094289	of the purified
72	5.763839	found in the	27	3.854305	the presence and
72	5.290777	of the same	26	27.068927	had no effect on the
72	4.951476	the control of	26	23.931538	min at 30 8c
71	19.157054	presence or absence of	26	22.236854	here we show that
71	18.790377	have been shown to	26	19.976137	an important role in
71	14.680008	is consistent with the	26	19.064177	not appear to be
71	11.191658	is essential for	26	18.725259	for 20 min at
71	7.966994	such as the	26	18.490457	we were unable to
71	7.504579	the percentage of	26	17.523452	it is important to
71	6.512498	presence of a	26	17.047297	for reviews see
70	15.938168	as shown in figure	26	16.089414	as a consequence of
70	14.498498	we conclude that	26	15.565401	carried out at
70	10.06437	were incubated for	26	15.553475	here we show
70	6.858367	the distribution of	26	14.976272	with respect to the
70	5.78726	of the total	26	14.64186	summarized in table
70	1.773069	and that the	26	13.396462	it will be
69	24.271052	had no effect on	26	13.272257	it is clear
69	23.321792	the presence or absence of	26	12.998139	tested for their
69	13.229701	their ability to	26	12.954065	were then washed
69	12.504526	has not been	26	12.890741	ability to bind
69	10.746974	of this article	26	12.651353	we were unable
69	10.032458	is likely to	26	12.437632	we do not
69	8.692406	used as a	26	12.433287	min at 30
69	6.629754	in contrast the	26	12.072203	we have found
69	6.593124	components of the	26	11.968387	unlikely to be
69	6.543084	the positions of	26	11.939201	been proposed to
69	6.244292	the surface of	26	11.841004	one copy of
68	21.407625	these results suggest that	26	11.676229	important role in
68	12.540734	for 20 min	26	11.187802	we have used
68	12.032724	we have shown	26	10.383253	the same time
68	8.775573	in table 1	26	10.31151	that at least
68	8.332945	indicate that the	26	10.077426	the formation of a
68	4.28971	the sequence of	26	10.033632	were exposed to
67	18.149512	been shown to be	26	10.026209	was analyzed by
67	13.936232	performed as described	26	9.961427	presence of 30
67	10.6348	the presence of a	26	9.7819	not shown we
67	9.054561	the hypothesis that	26	9.331149	model in which
67	7.975288	possible that the	26	9.170894	been observed in
67	7.461698	in figure 2	26	8.55753	in comparison with
67	6.358158	a function of	26	8.387952	respect to the
67	5.724266	in addition the	26	8.305883	some of these
66	19.847418	it is likely that	26	8.149513	are similar to
66	12.305748	10 min at	26	8.018189	are indicated in
66	10.812602	is likely that	26	8.004247	a combination of
65	8.009468	portion of the	26	7.950035	associated with a
65	7.451706	a result of	26	7.764344	as shown by
65	7.051435	change in the	26	7.575135	of a novel
65	6.086636	the end of	26	7.307294	fig 1 a
64	13.472845	as previously described	26	6.826246	alignment of the
64	6.916711	the method of	26	6.822012	both of these
64	6.650817	specificity of the	26	6.434559	identity of the
64	5.622528	the interaction of	26	6.334369	in the bottom
64	5.384216	some of the	26	6.223256	the interaction with
64	4.065854	of the other	26	5.806528	the release of
63	12.182487	that had been	26	5.68754	bottom of the
63	6.010184	the development of	26	5.623263	and the resulting
62	11.970227	not appear to	26	5.449772	version of the
62	8.340423	the absence or	26	5.38501	figure 1 and
62	7.567411	show that the	26	5.332336	the introduction of
62	5.676717	to be a	26	5.29065	of the various
62	4.081917	activity in the	26	5.209067	effect of a
61	16.405454	data not shown the	26	4.12882	in the control
61	13.651824	in the absence or	26	3.983519	and analysis of
61	11.00792	was obtained from	26	3.965135	in the region
61	10.822141	be involved in	26	3.957806	growth of the
61	10.529077	in this case	26	3.413123	in the other
61	10.21407	as a result	26	3.374052	of the complex
61	10.192897	is associated with	26	3.095328	in the two
61	8.780334	the existence of	25	24.067236	it has been suggested
61	7.935106	at the same	25	23.102869	results are means s

61	7.5028	nature of the	25	22.65284	it is likely that the
61	7.5028	the nature of	25	17.219031	we conclude that the
61	7.381574	on the other	25	16.300382	test this hypothesis
61	6.159392	the size of	25	14.770975	at a density of
61	5.926593	expressed in the	25	14.371566	increasing amounts of
60	30.278179	the materials and methods section	25	14.336707	together these data
60	28.455018	in the absence or presence of	25	14.198345	a single copy
60	26.072219	materials and methods section	25	13.886726	able to bind
60	23.766987	in the absence or presence	25	13.618024	is likely that the
60	23.120158	the absence or presence of	25	13.253518	high degree of
60	18.914198	absence or presence of	25	13.21763	as opposed to
60	18.432127	the absence or presence	25	13.174734	the crystal structure
60	17.555433	the materials and methods	25	13.143353	decapping in vivo
60	14.417758	and methods section	25	12.469051	c p m
60	14.226167	absence or presence	25	12.461185	it appears that
60	13.069789	data suggest that	25	12.387101	0 5 µg
60	12.825804	there is no	25	11.847822	ligated into the
60	9.307561	resulted in a	25	11.527813	may also be
60	6.521759	the materials and	25	11.094154	several lines of
59	16.777406	washed twice with	25	11.029529	fig 2 b
59	12.841728	its ability to	25	10.851006	activity was determined
59	12.817143	similar to those	25	10.845586	agouti protein and
59	12.441739	could not be	25	10.827771	the x chromosome
59	5.398573	shown that the	25	10.769527	be important for
58	16.812469	for 1 h at	25	10.612601	to account for
58	8.248847	is present in	25	10.548358	is regulated by
58	8.063842	localized to the	25	10.540769	as has been
58	7.086636	the lack of	25	10.535734	were removed by
58	3.913296	to be the	25	10.329196	the results presented
57	16.654869	has been proposed	25	10.18772	under the same
57	13.564428	final concentration of	25	10.097133	the difference between
57	8.083025	none of the	25	9.877766	is composed of
57	7.556956	the extent of	25	9.771572	that they are
57	5.141918	absence of the	25	9.661768	a requirement for
57	4.614441	control of the	25	9.421595	and analysed by
56	11.183616	were subjected to	25	9.103958	localizes to the
56	10.997246	consistent with this	25	9.039482	was associated with
56	10.534339	to interact with	25	8.917193	was due to
56	9.119092	consistent with a	25	8.83354	at a concentration
56	9.01953	to examine the	25	8.68509	the results obtained
56	5.842363	detected in the	25	8.639799	were obtained with
55	15.96546	as well as the	25	8.50983	in addition we
55	11.692788	high levels of	25	8.180295	parts of the
55	10.447151	in combination with	25	7.995598	characterization of a
55	9.210786	is involved in	25	7.978606	except that the
55	8.68385	was used for	25	7.820443	are found in
55	8.423436	well as the	25	7.633682	for the initial
55	7.191535	component of the	25	7.574596	used in these
55	5.917127	surface of the	25	7.476842	resulting in a
55	4.702371	of the three	25	7.441325	suggested by the
54	23.614984	in the presence or absence	25	7.440993	targeted to the
54	21.327791	mm tris hcl	25	7.337231	were expressed in
54	15.054236	the experimental section	25	7.333901	at the time
54	13.874102	at least two	25	7.213203	the other two
54	13.657621	were purchased from	25	7.111582	the intensity of
54	12.862491	in the presence or	25	7.052472	were present in
54	12.107552	determine whether the	25	6.568427	a family of
54	11.468191	with or without	25	6.474756	for binding to
54	11.282193	were separated by	25	6.463577	recovered in the
54	7.400951	the location of	25	6.377976	of the entire
54	7.382092	half of the	25	6.107597	to those of
54	6.582789	comparison of the	25	5.998688	percentage of the
54	5.59631	ability of the	25	5.858367	the value of
54	5.100951	sites in the	25	5.626828	that this is
54	4.786241	because of the	25	5.197997	top of the
53	17.91969	to determine whether the	25	4.87395	of these cells
53	13.273696	has also been	25	4.727275	map of the
53	12.653239	is dependent on	25	4.637938	and methods the
53	12.466673	results were obtained	25	4.266384	of the interaction
53	11.277291	in the regulation of	25	4.266384	interaction of the
53	10.554759	are likely to	25	4.234066	phase of the

53	9.775303	the position of the	25	4.041992	the study of
53	9.57487	a consequence of	25	2.531521	results of the
53	9.353297	derived from the	24	23.129949	are means s e
53	8.16248	member of the	24	22.725508	at 37 8c for
53	8.123409	5 ml of	24	22.483972	several lines of evidence
53	7.991869	obtained from the	24	22.059944	remains to be determined
53	6.58926	in the regulation	24	21.648352	a wide range of
52	19.051293	washed three times	24	19.923537	were prepared as described
52	15.675107	has been reported	24	19.240561	at restrictive temperatures
52	15.556397	min at room	24	15.31547	to be involved in
52	15.073182	this article has	24	15.054575	medium supplemented with
52	14.617112	a final concentration	24	14.944617	shown in figure 2
52	13.749126	to determine if	24	14.897928	has been demonstrated
52	13.706708	was added to the	24	14.759736	extracts prepared from
52	13.433287	30 min at	24	14.543042	by the fact that
52	13.071092	results indicate that	24	13.586645	it is unlikely
52	11.949588	was confirmed by	24	13.483335	early and late
52	11.308832	was performed on	24	13.386024	for 48 h
52	11.302288	be due to	24	12.876435	but did not
52	10.401352	as determined by	24	12.764452	a previous study
52	10.03277	are involved in	24	12.694802	be the result of
52	8.496117	were found to	24	12.667997	been proposed that
52	8.221942	adjacent to the	24	12.103216	for 16 h
52	7.497021	showed that the	24	12.095524	for the production of
52	6.862892	of a single	24	11.919347	is not yet
52	5.380111	localization of the	24	11.308908	high concentrations of
52	4.911436	of the first	24	11.012386	to associate with
52	4.509622	of the human	24	10.69279	also required for
51	27.186358	min at room temperature	24	10.523846	predicted to be
51	25.452704	has been cited by	24	10.389454	the same conditions
51	21.438349	on the other hand	24	10.327432	fig 1 c
51	19.277129	a final concentration of	24	10.272605	to act as
51	17.133323	amino acid residues	24	10.239059	a role in the
51	13.681056	is required for the	24	10.139459	was not detected
51	13.365336	the other hand	24	9.996561	of a number of
51	11.048687	were unable to	24	9.822044	to note that
51	10.937073	be required for	24	9.778023	the bottom of the
51	10.731468	to test this	24	9.706816	agreement with the
51	9.070903	and analyzed by	24	9.280986	also present in
51	7.281014	the identification of	24	9.015631	was performed by
51	7.157316	was shown to	24	8.986171	to understand the
51	6.614858	found that the	24	8.974112	be required to
51	5.671598	any of the	24	8.862781	been used to
51	4.806528	role of the	24	8.716158	were collected and
51	2.417814	as in the	24	8.542211	are shown as
50	18.421299	in materials and methods	24	8.420574	correspond to the
50	14.936023	at least three	24	8.249556	that we have
50	13.477949	as a function of	24	8.231026	the remainder of
50	13.29327	have been described	24	8.221993	in this model
50	12.675705	similar to that of	24	8.006772	be the result
50	9.979456	account for the	24	7.859663	by the fact
50	9.01658	there was a	24	7.841619	1 h with
50	8.789919	as a function	24	7.731025	function as a
50	8.737696	interact with the	24	7.573434	hypothesis that the
50	8.632713	a defect in	24	7.544263	effects on the
50	7.387624	in materials and	24	7.407493	for the production
50	6.680492	bound to the	24	7.394053	was expressed in
50	6.365974	for the presence	24	7.267682	figure 4 a
50	5.501682	incubated in the	24	7.112565	followed by the
49	18.269175	taken together these	24	6.947105	essential for the
49	13.971315	three times with	24	6.924364	the organization of
49	11.408806	is thought to	24	6.764708	consequence of the
49	9.873339	the interaction between	24	6.748188	this region of
49	9.662093	the ability of the	24	6.684537	determination of the
49	9.392339	in these experiments	24	6.460102	binds to the
49	7.260773	location of the	24	6.404691	a deletion of
49	7.021741	were used in	24	6.397862	the reduction in
49	5.843365	size of the	24	6.310108	of a large
49	4.149204	results in the	24	6.157025	stability of the
49	3.052322	is that the	24	5.993657	is the first
48	14.49167	these data suggest	24	5.955391	of the native

48	13.238171	referred to as	24	5.882609	even in the
48	11.486371	be expected to	24	5.751399	which is a
48	10.327546	were able to	24	5.46917	of which are
48	8.711164	contribute to the	24	5.30853	of a number
48	8.596604	led to the	24	5.014647	in the reaction
48	8.54412	and resuspended in	24	4.771594	of the corresponding
48	8.381544	demonstrate that the	24	4.735614	in the formation
48	8.288177	a range of	24	4.672688	in activation of
48	8.170521	suggested that the	24	4.429105	of the small
48	7.130225	the ratio of	24	4.223925	concentration of the
48	6.606102	with the same	24	3.842329	of the growth
48	5.889768	the increase in	23	32.508231	the absence or presence of 30
48	5.856888	to the same	23	32.417319	are means s e m
48	3.943678	the site of	23	28.30227	absence or presence of 30
47	35.260737	in the materials and methods section	23	26.509156	taken together these results
47	22.537991	in the materials and methods	23	24.380357	it has been proposed
47	14.747933	play a role	23	17.595805	provided by dr
47	13.289105	been implicated in	23	17.549778	or presence of 30
47	11.963957	low levels of	23	16.172105	materials and methods the
47	11.504317	in the materials and	23	15.638835	closely related to
47	10.751397	was measured by	23	15.622496	results suggest that the
47	10.631992	was performed as	23	15.441119	has recently been
47	9.529863	is indicated by	23	14.242238	not yet been
47	8.857339	as a control	23	14.197035	was added to a
47	8.402832	was detected in	23	12.876515	been shown previously
47	8.277725	close to the	23	12.587559	then treated with
47	7.210175	the degree of	23	12.572557	min followed by
47	7.192395	in figure 5	23	12.383827	a portion of the
47	7.117663	the action of	23	12.34538	there are no
47	6.412414	in the materials	23	12.173181	been suggested that
47	6.244008	the length of	23	12.119112	known about the
46	28.112939	described in materials and methods	23	12.082942	been identified as
46	28.071689	in the presence or absence of	23	12.012966	its interaction with
46	19.971012	for 30 min at	23	11.961118	we have previously
46	17.079264	described in materials and	23	11.732476	have demonstrated that
46	14.494925	as a result of	23	11.3563	it can be
46	13.690385	has been described	23	11.22859	were separated on
46	11.987362	described in materials	23	11.169955	this work was
46	10.721141	were isolated from	23	11.163402	to ensure that
46	10.401741	are indicated by	23	11.116698	were collected from
46	9.976233	a subset of	23	11.106039	which have been
46	9.769112	dependent on the	23	11.063442	this indicates that
46	9.551005	shown in table	23	11.020472	other members of
46	9.196913	to a final	23	10.911263	two types of
46	8.723137	to investigate the	23	10.784978	are unable to
46	8.199187	is expressed in	23	10.772662	is difficult to
46	8.13215	face of the	23	10.441154	fig 1 b
46	7.896471	each of these	23	10.326608	min at 4
46	5.423199	the mechanism of	23	10.198276	separated on a
46	2.492155	or in the	23	9.987168	is caused by
45	23.321598	it is possible that the	23	9.982652	is activated by
45	17.806625	cells were treated with	23	9.763636	differences between the
45	16.661463	did not affect	23	9.600186	but it is
45	15.880416	has been suggested	23	9.439997	for 3 min
45	15.384447	under the control of	23	9.418942	the length of the
45	14.286781	is possible that the	23	9.409192	is localized to
45	12.805871	one or more	23	9.400002	were harvested and
45	12.747926	thought to be	23	9.395422	in this process
45	12.28854	was performed using	23	9.305756	as a probe
45	11.547796	similar results were	23	9.260472	were washed with
45	11.370445	have not been	23	9.086307	in 20 mm
45	11.25952	were grown at	23	8.940518	the product of the
45	11.226036	may not be	23	8.913903	transferred to a
45	10.902002	for the presence of	23	8.890821	30 min the
45	10.696417	under the control	23	8.83259	the vicinity of
45	10.586718	were generated by	23	8.817248	in the activation of
45	10.45354	were performed as	23	8.645416	interacts with the
45	10.35839	were tested for	23	8.588819	concentration of 0
45	10.26583	it is also	23	8.303841	the effects of the
45	9.68666	followed by a	23	8.281231	it was not
45	8.971584	the structure of the	23	8.258489	affected by the

45	7.606602	in figure 3	23	8.076774	attached to the
45	6.955858	the difference in	23	8.025235	portions of the
45	6.699343	basis of the	23	7.854381	leading to the
45	6.480388	with the indicated	23	7.816709	for a further
45	5.926412	positions of the	23	7.774211	to explain the
45	5.349679	in the first	23	7.698262	leads to the
45	4.109719	the region of	23	7.645403	presence of 0
44	19.987636	play a role in	23	7.53491	the inability of
44	18.555904	used in this study	23	7.363274	of each of the
44	18.088071	we have shown that	23	7.264846	sensitive to the
44	17.309423	does not require	23	7.211672	table 1 in
44	16.718531	was found to be	23	7.118963	all of which
44	16.669354	were washed twice	23	7.029395	on the same
44	12.197815	results show that	23	6.958121	understanding of the
44	11.245125	are expressed as	23	6.736517	with those of
44	11.136425	to confirm that	23	6.637861	presence of 1
44	10.772726	was isolated from	23	6.631989	to inhibit the
44	10.669485	were analyzed by	23	6.572063	the yield of
44	9.144023	were added to	23	6.507487	important for the
44	8.753202	are present in	23	6.349288	of one or
44	8.246206	were used for	23	6.215869	by using a
44	8.151176	for example the	23	6.160165	the combination of
44	8.005601	is similar to	23	6.119292	and 1 mm
44	7.465577	related to the	23	5.784681	found in a
44	7.219038	not shown and	23	5.429126	shown on the
44	5.621853	addition to the	23	5.319521	production of the
44	5.599372	in the medium	23	5.183886	for each of
44	4.552522	sequences of the	23	5.077703	the top of
44	4.384756	domain of the	23	4.763433	added and the
44	3.818147	site of the	23	4.69632	this is a
44	1.346175	that in the	23	4.623262	and used to
43	21.016351	these data suggest that	23	4.58783	out of the
43	16.322198	would be expected	23	4.580566	performed in the
43	14.425083	it should be	23	4.129217	in the activation
43	14.097135	we propose that	23	3.727875	however in the
43	13.894043	we find that	23	1.768039	not in the
43	13.198965	experiments were performed	22	35.975142	according to the manufacturer's instructions
43	13.004378	is not clear	22	30.999606	it has been proposed that
43	12.300193	remains to be	22	25.106804	carried out as described
43	12.015579	were analysed by	22	24.713764	were washed three times
43	11.786432	the relationship between	22	23.097118	to the manufacturer's instructions
43	10.635759	these results are	22	22.836459	an equal volume of
43	9.834938	was detected by	22	21.967782	2 5 lg ml
43	9.072047	a decrease in	22	21.964789	has been proposed that
43	8.180788	were performed in	22	21.616271	has been implicated in
43	7.625606	shows that the	22	21.416388	data not shown thus
43	7.500077	copies of the	22	21.374967	under the same conditions
43	7.15973	resulted in the	22	18.940711	is known about the
43	6.885627	the frequency of	22	18.671969	we asked whether
43	6.50986	indicated that the	22	18.601234	is thought to be
43	5.640775	regulation of the	22	18.461745	has been shown that
43	5.027934	form of the	22	18.148428	an equal volume
43	5.000583	effects of the	22	17.713793	at the same time
42	13.827413	prepared as described	22	15.820091	data not shown and
42	13.723827	increasing concentrations of	22	15.772432	as well as in
42	12.038252	mechanism by which	22	15.730639	did not appear
42	11.367919	we suggest that	22	15.698867	a conformational change
42	10.845594	difference between the	22	15.446925	is a member of
42	10.674662	were stained with	22	14.53655	equal volume of
42	10.64249	known to be	22	14.430691	for 1 h with
42	10.162955	is sufficient to	22	14.394118	for 4 h in
42	8.220219	the onset of	22	13.983517	the permissive temperature
42	7.946458	the importance of	22	13.974507	be explained by
42	7.902278	demonstrated that the	22	13.904994	may be due
42	7.880002	one of these	22	13.796757	there are several
42	6.277898	than that of	22	13.744875	reactions were performed
42	5.481573	used for the	22	13.700184	consistent with previous
42	3.939491	in both the	22	13.626859	were obtained from the
41	17.683295	cells were transfected with	22	13.61486	used to amplify
41	17.136777	no effect on the	22	13.414718	insight into the
41	14.960629	was used as a	22	12.983668	out as described

41	12.998128	data indicate that	22	12.972498	were washed three
41	12.292707	a gift from	22	12.970321	there may be
41	12.026773	this study we	22	12.668212	on the surface of
41	11.135575	the nature of the	22	12.320468	has been used
41	10.941591	however it is	22	12.265526	used to identify
41	10.896011	were prepared from	22	12.104791	at the level of
41	10.315606	not required for	22	12.057016	various concentrations of
41	10.082386	is able to	22	11.877884	be responsible for
41	9.87841	that have been	22	11.838112	no evidence for
41	8.971748	were used as	22	11.743943	have suggested that
41	8.379589	a percentage of	22	11.669424	very similar to
41	7.75969	the context of	22	11.51919	by virtue of
41	7.46342	use of a	22	11.242501	to address this
41	6.037224	the process of	22	10.908837	acts as a
41	5.851248	in the second	22	10.882125	therefore it is
41	4.530444	studies of the	22	10.861687	for 2 hr
40	15.894374	under these conditions	22	10.78146	only a single
40	13.469789	studies have shown	22	10.758894	is a member
40	12.463129	in all cases	22	10.730721	thus it is
40	11.735682	in this paper	22	10.411276	total number of
40	10.451108	is not required	22	10.404247	are essential for
40	10.091291	by incubation with	22	10.399511	the identity of the
40	9.423692	a member of	22	10.100817	have found that
40	9.32991	were performed with	22	9.848885	indicated by an
40	8.93765	of at least	22	9.621651	been found to
40	8.925865	note that the	22	9.310044	interactions between the
40	8.460547	a model for	22	9.298113	in the formation of
40	7.85398	but not in	22	9.284249	was determined as
40	7.730135	the sequence of the	22	9.228604	alone or in
40	7.410367	likely that the	22	9.099951	the positions of the
40	7.233392	in the experimental	22	9.045591	1 min at
40	7.00878	the activity of the	22	8.893709	in the presence and
40	6.958531	copy of the	22	8.287919	the interaction of the
40	6.785933	located in the	22	8.230408	well as in
40	5.253023	the fraction of	22	8.199851	in a manner
40	4.593621	function in the	22	8.160124	located at the
40	3.791654	and the other	22	8.114159	probed with the
39	32.080739	described in the materials and methods	22	7.915046	we used a
39	21.047065	described in the materials and	22	7.904983	transformed with the
39	20.116642	studies have shown that	22	7.884391	we compared the
39	17.557042	is likely to be	22	7.57418	described in figure
39	17.0088	described in the experimental	22	7.478716	at the surface
39	15.955162	described in the materials	22	7.41676	at the level
39	15.849463	as a percentage of	22	7.331055	examination of the
39	12.497097	we did not	22	7.107158	the column was
39	12.337781	there was no	22	7.057404	are described in
39	11.966373	were performed using	22	7.000252	in the upper
39	11.161433	as a percentage	22	6.954538	a comparison of
39	11.022838	away from the	22	6.904585	and table 1
39	10.698852	the basis of the	22	6.893991	removal of the
39	10.220989	as compared with	22	6.652983	features of the
39	10.143702	was able to	22	6.624354	treated with the
39	10.057554	fragment containing the	22	6.608695	to be an
39	10.017695	has shown that	22	6.553056	the results of the
39	9.704425	not shown this	22	6.431442	in a total
39	9.450312	in terms of	22	6.393345	figure 2 a
39	8.212574	is required to	22	6.360805	3 and 5
39	7.893991	the appearance of	22	6.358462	may be the
39	7.791308	to identify the	22	6.096035	orientation of the
39	7.785462	isolated from the	22	5.982179	was used in
39	7.26053	the proportion of	22	5.918103	residue in the
39	5.724066	use of the	22	5.837734	present in a
39	3.033461	two of the	22	5.710233	the differences in
38	28.551614	as described in the materials and	22	5.701664	site at the
38	26.753673	has been shown to be	22	5.625502	amounts of the
38	24.51335	as described in the experimental	22	5.612083	figure 2 and
38	23.459712	as described in the materials	22	5.458605	the association of
38	22.145471	similar results were obtained	22	5.435665	absence of a
38	17.252008	in this study we	22	5.292893	grown in the
38	13.035543	on ice for	22	5.292726	of the full-length
38	12.516678	the fact that the	22	5.290177	localization to the

38	12.295187	appeared to be	22	5.268482	to increase the
38	12.252306	we demonstrate that	22	5.245568	observed for the
38	11.548193	for an additional	22	5.134254	the possibility of
38	11.134014	is necessary for	22	5.072652	of the up
38	10.234852	with the exception	22	4.747586	identified in the
38	10.018661	were resuspended in	22	4.229301	concentrations of the
38	9.727875	the idea that	22	4.221565	result of the
38	9.542619	version of this	22	3.162624	in all the
38	9.302899	on the left	21	27.429447	it has been shown that
38	9.118152	expressed as a	21	27.377839	these results are consistent with
38	8.762916	the absence of the	21	25.848147	at a flow rate of
38	8.371897	by the method	21	25.545614	min at 37 8c
38	8.310718	fact that the	21	24.89218	it seems likely that
38	8.242449	indicates that the	21	21.887948	to test this hypothesis
38	7.360559	the evolution of	21	20.78952	these results are consistent
38	7.248523	evidence that the	21	19.345132	added to a final
38	7.219691	at the indicated	21	18.649446	have been identified in
38	7.217213	of these two	21	18.208802	it seems likely
38	6.932465	characterization of the	21	15.857364	seems likely that
38	6.719342	model for the	21	15.758744	carried out on
38	6.498729	differences in the	21	15.471113	shown in figure 3
38	6.326946	to that of the	21	15.431162	exclude the possibility
38	6.230327	seen in the	21	15.40135	at various times
38	6.065103	the assembly of	21	15.288601	excess of unlabelled
38	4.507107	residues of the	21	14.982594	we tested whether
37	29.655597	described in the experimental section	21	14.419288	we show that the
37	19.843662	in the experimental section	21	14.090449	was introduced into
37	18.444821	should be noted	21	14.047362	min at 37
37	16.480661	does not appear	21	13.902001	h at room
37	15.150291	an important role	21	13.453056	truncated form of
37	15.101027	shown in figure 1	21	13.126576	this implies that
37	14.642619	are consistent with the	21	12.978994	arrows indicate the
37	13.847066	at least one	21	12.48135	total volume of
37	10.706735	in addition to the	21	12.32819	was used as the
37	10.420034	was generated by	21	12.056379	are summarized in
37	9.750124	and probed with	21	12.027704	is involved in the
37	9.628083	was obtained by	21	11.967116	results are expressed
37	9.512366	were obtained by	21	11.872036	were as follows
37	9.477813	supported by the	21	11.586749	the hypothesis that the
37	9.472005	at least in	21	11.495463	three times in
37	8.774138	and subjected to	21	11.381125	can also be
37	8.680222	and stained with	21	11.3179	be caused by
37	8.611591	the timing of	21	11.187937	two or more
37	8.564957	be used to	21	11.035384	containing 0 5
37	8.255736	is independent of	21	10.992672	was based on
37	7.998947	presence of an	21	10.892676	see figure 2
37	7.429082	was observed in	21	10.873298	which is consistent
37	7.196117	from the same	21	10.752485	relationship between the
37	6.781516	the stability of	21	10.74111	although it is
37	5.908687	the activities of	21	10.712577	the presence of 1
37	5.758663	to study the	21	10.629724	rather than the
37	5.115462	residues in the	21	10.609527	distance between the
37	4.95286	that the two	21	10.579118	mg ml in
37	4.739292	expression of a	21	10.523136	in the production of
37	4.398599	and that this	21	10.424595	see figure 1
36	37.158093	as described in the experimental section	21	10.026376	were allowed to
36	35.301326	as described in materials and methods	21	10.003648	suggesting that this
36	27.440109	it should be noted	21	9.884295	was unable to
36	24.267652	as described in materials and	21	9.816046	were made by
36	24.136379	in the present study we	21	9.659343	was induced by
36	22.048619	according to the manufacturer's	21	9.585703	was examined by
36	19.925702	to a final concentration	21	9.344693	that there are
36	19.175749	as described in materials	21	9.07292	it was shown
36	18.801519	the present study we	21	8.940457	in patients with
36	14.880644	little or no	21	8.869224	is predicted to
36	14.595559	present study we	21	8.710781	as seen in
36	14.526114	we found that the	21	8.650829	as part of
36	14.370766	been described previously	21	8.534731	to produce a
36	14.322515	is shown in figure	21	8.35638	to that seen
36	13.036284	15 min at	21	8.332112	the rest of
36	12.981925	by the method of	21	8.267457	localize to the

36	12.596021	when compared with	21	7.645032	observation that the
36	12.165379	the presence of an	21	7.638676	to show that
36	11.882621	was digested with	21	7.618722	on the ability
36	11.870869	as a consequence	21	7.478958	in figure 7
36	11.863935	depending on the	21	7.468836	were prepared and
36	11.612091	in each case	21	7.282141	and can be
36	10.188374	was purified from	21	7.120507	comparison with the
36	9.440153	the end of the	21	6.899126	not shown to
36	9.309796	is due to	21	6.81533	to the right
36	9.170596	to the manufacturer's	21	6.780375	to that observed
36	8.857438	to assess the	21	6.403904	by binding to
36	8.577807	shown in fig	21	6.020248	occur in the
36	8.326116	in table 2	21	6.010193	that the interaction
36	8.247305	a component of	21	5.835105	in the production
36	8.11771	at the end	21	5.828562	that activation of
36	7.840095	possibility that the	21	5.63455	at the site
36	7.613068	a response to	21	5.538115	the rates of
36	7.283944	the reaction was	21	5.524757	the average of
36	7.276117	included in the	21	5.45976	forms of the
36	6.573077	but not the	21	5.350919	such that the
36	6.340952	formation of a	21	5.338543	is one of
36	6.047107	purification of the	21	5.237156	as in figure
36	3.675478	this is the	21	5.091551	activities of the
36	3.337336	shown in the	21	4.609246	specific to the
35	21.483034	it has been shown	21	4.587288	and absence of
35	18.615746	on the basis of the	21	4.495145	of the four
35	12.866281	in the context of	21	4.09551	in the number
35	11.826284	are thought to	21	4.073606	and is not
35	11.380037	in agreement with	21	4.02677	this is in
35	11.085759	is responsible for	21	3.672899	in each of
35	11.063458	for 1 hr	21	3.657013	region in the
35	10.009188	were prepared by	21	3.57758	in all of
35	9.563898	the size of the	21	3.480564	of the indicated
35	9.518432	to be required	21	3.447308	than in the
35	9.178755	that there is	21	3.414056	region and the
35	8.231047	a mixture of	21	3.323035	min in the
35	8.17825	in the context	21	2.292662	both of the
35	8.116794	the control of the	21	2.292662	of both the
35	6.605421	the generation of	20	29.743619	tested for their ability to
35	6.423361	properties of the	20	26.3846	reactions were carried out
35	6.154757	contrast to the	20	25.489587	h at room temperature
35	5.946458	assembly of the	20	23.904514	tested for their ability
35	5.818703	length of the	20	22.090368	were carried out at
35	5.728653	the pattern of	20	21.432241	did not appear to
35	5.395991	figure 2 the	20	21.39387	has been suggested that
35	5.375815	many of the	20	18.484191	at 4 8c with
35	5.340279	product of the	20	17.391227	which is consistent with
35	5.060378	fraction of the	20	16.138005	are shown in table
35	4.959879	those of the	20	15.977047	in the presence of 1
35	4.927911	figure 1 the	20	15.710881	that are required for
35	3.541877	as in a	20	15.626498	have been implicated
34	22.197775	does not appear to	20	15.587263	two copies of the
34	19.976801	was performed as described	20	15.465311	were found to be
34	17.886044	the manufacturer's instructions	20	15.024798	other members of the
34	14.544721	have been reported	20	14.876868	reactions were carried
34	14.201923	these results indicate	20	14.640983	an essential role
34	12.723279	at the end of	20	14.479732	the total number of
34	12.463444	analysis was performed	20	14.207395	in support of this
34	11.963593	the possibility that the	20	13.965816	in the vicinity of
34	10.988871	expected to be	20	13.602617	as reported previously
34	10.809165	act as a	20	13.578275	to distinguish between
34	10.437402	possibility is that	20	13.199643	at a concentration of
34	9.751157	is important for	20	13.187698	a critical role
34	9.693577	to each other	20	13.04469	the existence of a
34	9.502212	encoded by the	20	13.008997	for 60 min
34	9.368751	not affect the	20	12.50372	not shown suggesting
34	9.346092	interaction between the	20	12.288991	consistent with our
34	9.264217	be detected in	20	12.173951	the remainder of the
34	9.248881	conclude that the	20	12.127591	for 24 h
34	9.122687	were grown to	20	12.029243	at the surface of
34	8.965574	the finding that	20	11.940506	compared with control

34	8.567567	a reduction in	20	11.785676	in concert with
34	7.8713	caused by the	20	11.782448	there are two
34	7.689584	prior to the	20	11.747972	4 h in the
34	7.546352	together with the	20	11.645072	a small number
34	6.39363	a concentration of	20	11.549925	a percentage of the
34	6.179889	specific for the	20	11.418404	is also possible
34	5.816547	distribution of the	20	11.380713	was dependent on
34	4.870318	of the reaction	20	11.358561	would result in
34	4.53437	fragment of the	20	11.309624	were crossed to
34	3.432168	of expression of	20	11.272407	to bind to the
33	20.842233	these results indicate that	20	11.206825	introduced into the
33	19.818016	were performed as described	20	11.181308	are responsible for
33	19.625356	for 15 min at	20	11.091021	was dissolved in
33	19.299501	kindly provided by	20	10.927316	between these two
33	19.142948	under the control of the	20	10.911499	the present work
33	16.214004	does not affect	20	10.895615	were processed for
33	15.319713	is known about	20	10.866029	was determined using
33	13.877339	supplemented with 10	20	10.769571	was mixed with
33	13.673959	can be detected	20	10.73747	determine if the
33	13.280182	to test whether	20	10.557332	at this time
33	11.923179	in accordance with	20	10.538967	is subject to
33	11.875075	lines of evidence	20	10.411228	in the amount of
33	11.435843	been reported to	20	10.108206	be consistent with
33	10.832087	loss of function	20	10.088239	shown previously that
33	10.791989	is inhibited by	20	10.046065	be associated with
33	10.39661	been identified in	20	9.949666	are able to
33	10.066721	we have not	20	9.94227	were pooled and
33	9.614555	for 3 h	20	9.818087	30 min in
33	9.386122	the surface of the	20	9.791701	the total number
33	9.072658	from a single	20	9.742981	is sensitive to
33	7.599965	a density of	20	9.326893	present in all
33	7.464342	to form a	20	9.310653	were treated for
33	7.321984	all of these	20	9.295041	not shown figure
33	6.904962	majority of the	20	9.277786	in the vicinity
33	6.778514	the identity of	20	9.236518	along with a
33	6.616457	disruption of the	20	9.208378	was resuspended in
33	6.56721	interaction with the	20	9.116207	associates with the
33	6.558077	to test the	20	8.996279	that are required
33	6.155679	incubated with the	20	8.953498	as indicated by
33	6.031494	the bottom of	20	8.884461	is capable of
33	5.506945	is not a	20	8.872535	support of this
33	5.453357	result in the	20	8.83873	existence of a
33	5.258905	the degradation of	20	8.819957	a function of the
33	5.25539	the product of	20	8.737267	and do not
33	3.678146	of each of	20	8.713151	in the number of
33	2.666667	with that of	20	8.572063	sides of the
32	22.798241	were washed twice with	20	8.566044	not result in
32	18.83578	for their ability to	20	8.484452	identified as a
32	16.603617	data not shown in	20	8.287969	as a model
32	15.679216	little is known	20	8.230032	only one of
32	15.639713	essentially as described	20	8.20756	the localization of the
32	13.603061	may contribute to	20	8.19726	in table 3
32	13.557917	has been observed	20	8.174841	removed from the
32	13.307724	a member of the	20	8.053034	as described for
32	12.996675	for their ability	20	7.967991	remainder of the
32	12.173017	was assessed by	20	7.805379	were washed in
32	11.585146	was replaced with	20	7.724066	diagram of the
32	11.360334	in contrast to the	20	7.715328	with the appropriate
32	11.181106	is mediated by	20	7.692537	shown to have
32	10.789524	to this article	20	7.687902	except for the
32	10.52664	were prepared as	20	7.624064	min after the
32	9.881608	in this experiment	20	7.405927	this type of
32	9.321297	response to this	20	7.138721	5 min and
32	9.272937	for up to	20	7.134709	al 1991 the
32	9.161523	were fixed in	20	7.064412	were obtained in
32	8.788196	is known to	20	7.045994	the question of
32	8.585499	with 1 ml	20	7.019951	at the 5'
32	8.578529	are expressed in	20	7.001563	was similar to
32	8.501337	mediated by the	20	6.9871	the beginning of
32	8.340063	the role of the	20	6.959692	the reactions were
32	8.206317	necessary for the	20	6.921248	so that the

32	8.047383	the effect of the	20	6.912138	the significance of
32	7.810055	in this region	20	6.756487	the removal of
32	7.704166	the tip of	20	6.662634	this is not
32	7.665337	revealed that the	20	6.527142	purified from the
32	7.273683	is found in	20	6.436727	interactions with the
32	7.027291	table 2 the	20	6.406144	sites on the
32	6.821063	to the left	20	6.345554	the incorporation of
32	6.295505	indicated by the	20	6.117887	the origin of
32	5.674993	and it is	20	6.085053	and purification of
32	5.471955	1 2 and	20	6.026095	with the following
32	4.925324	loss of the	20	5.966113	identical to the
32	4.005545	site in the	20	5.932011	in the initial
32	3.587797	from that of	20	5.867191	ends of the
32	3.397137	the analysis of	20	5.789654	recognition of the
31	23.455378	at room temperature for	20	5.787761	figure 1 a
31	16.376765	for 2 h at	20	5.732103	of the central
31	16.057948	to be required for	20	5.723197	in the amount
31	15.883831	room temperature for	20	5.616006	the properties of
31	15.789634	carried out as	20	5.172236	compared to the
31	14.835929	carried out with	20	5.097866	response to the
31	14.531152	can be seen	20	5.011348	of the five
31	14.354612	more than one	20	4.840879	of the resulting
31	14.069582	as judged by	20	4.688442	independent of the
31	13.182224	have been found	20	3.720064	study of the
31	12.786951	it does not	20	3.579294	the growth of
31	11.902325	is supported by	20	3.421346	and the presence
31	11.62512	only a small	20	3.117274	is not the

Appendix 2

Lexical bundles deleted after application of exclusion criteria

Frequency rank	**Bundle**	**Frequency rank**	**Bundle**
1	in the presence	483	the exception of
2	in the absence	484	tip of the
3	materials and methods	485	result in a
4	consistent with the	486	should be noted that
5	the expression of	487	this suggests that the
6	for 30 min	488	results are means
7	region of the	489	in the absence of the
8	expression of the	490	but does not
9	for 10 min	491	in the presence of the
10	one of the	492	inserted into the
11	for 1 h	493	this is consistent
12	present in the	494	in 50 mm
13	analysis of the	495	released from the
14	it has been	496	was added and
15	is possible that	497	lead to the
16	in which the	498	implicated in the
17	associated with the	499	added to a
18	on the basis	500	and characterization of
19	all of the	501	with the use
20	end of the	502	evidence for the
21	of the two	503	the medium was
22	described in the	504	reduction in the
23	the binding of	505	in a single
24	activity of the	506	modification of the
25	structure of the	507	it is a
26	suggest that the	508	case of the
27	based on the	509	by using the
28	involved in the	510	formation of the
29	to determine the	511	expressed as a percentage
30	required for the	512	data not shown this
31	to that of	513	directed against the
32	each of the	514	results are consistent
33	suggesting that the	515	not shown thus
34	due to the	516	the tip of the
35	for 5 min	517	for at least
36	for 15 min	518	this result is
37	the regulation of	519	at the restrictive
38	see materials and	520	we tested the
39	relative to the	521	is important to
40	as described in the	522	and transferred to
41	the activation of	523	4 h in
42	suggests that the	524	the function of the
43	presence of the	525	linked to the
44	sequence of the	526	part of a
45	most of the	527	defects in the
46	according to the	528	figure 4 the
47	effect on the	529	the results are
48	members of the	530	figure 3 the
49	cells in the	531	figure 5 the
50	that of the	532	only in the
51	it is not	533	addition of the
52	in the case	534	and at the
53	function of the	535	washed three times with
54	part of the	536	a wide range

55	increase in the	537	three independent experiments
56	responsible for the	538	in the case of the
57	not shown the	539	in the presence of a
58	the presence of the	540	be involved in the
59	compared with the	541	an increase in the
60	the case of	542	with 1 ml of
61	activation of the	543	a large number
62	to bind to	544	which has been
63	changes in the	545	it is important
64	by the addition	546	depends on the
65	added to the	547	there is an
66	the concentration of	548	at a density
67	there is a	549	5 min at
68	indicating that the	550	that has been
69	results in a	551	not bind to
70	role in the	552	the case of the
71	and in the	553	to estimate the
72	for 2 h	554	to changes in
73	between the two	555	fragment from the
74	observed in the	556	in a number
75	presence or absence	557	defect in the
76	the presence or absence	558	for the first
77	have been shown	559	used as the
78	for 4 h	560	it is the
79	the presence or	561	of the purified
80	corresponding to the	562	the presence and
81	similar to the	563	had no effect on the
82	used in the	564	min at 30 8c
83	that it is	565	for 20 min at
84	regions of the	566	here we show
85	or presence of	567	with respect to the
86	1 ml of	568	it will be
87	effect of the	569	tested for their
88	or absence of	570	were then washed
89	side of the	571	ability to bind
90	position of the	572	we were unable
91	used in this	573	we do not
92	we examined the	574	min at 30
93	found in the	575	one copy of
94	of the same	576	that at least
95	presence or absence of	577	the formation of a
96	is consistent with the	578	presence of 30
97	such as the	579	not shown we
98	presence of a	580	respect to the
99	and that the	581	some of these
100	the presence or absence of	582	associated with a
101	has not been	583	of a novel
102	of this article	584	fig 1 a
103	used as a	585	alignment of the
104	in contrast the	586	both of these
105	components of the	587	identity of the
106	for 20 min	588	bottom of the
107	indicate that the	589	and the resulting
108	been shown to be	590	version of the
109	the presence of a	591	figure 1 and
110	possible that the	592	of the various
111	in addition the	593	effect of a
112	10 min at	594	and analysis of
113	is likely that	595	growth of the
114	portion of the	596	in the other
115	change in the	597	of the complex
116	specificity of the	598	in the two
117	some of the	599	results are means s
118	of the other	600	it is likely that the
119	that had been	601	we conclude that the
120	the absence or	602	test this hypothesis
121	show that the	603	a single copy
122	to be a	604	able to bind
123	activity in the	605	is likely that the
124	data not shown the	606	the crystal structure
125	in the absence or	607	decapping in vivo

126	nature of the	608	c p m
127	on the other	609	0 5 μg
128	expressed in the	610	ligated into the
129	the materials and methods section	611	may also be
130	materials and methods section	612	several lines of
131	in the absence or presence	613	fig 2 b
132	the absence or presence of	614	agouti protein and
133	absence or presence of	615	the x chromosome
134	the absence or presence	616	is regulated by
135	the materials and methods	617	as has been
136	and methods section	618	under the same
137	absence or presence	619	that they are
138	there is no	620	and analysed by
139	resulted in a	621	localizes to the
140	the materials and	622	at a concentration
141	washed twice with	623	in addition we
142	could not be	624	parts of the
143	shown that the	625	characterization of a
144	for 1 h at	626	except that the
145	localized to the	627	for the initial
146	to be the	628	used in these
147	final concentration of	629	resulting in a
148	none of the	630	suggested by the
149	absence of the	631	targeted to the
150	control of the	632	were expressed in
151	consistent with a	633	the other two
152	to examine the	634	for binding to
153	detected in the	635	recovered in the
154	as well as the	636	of the entire
155	well as the	637	to those of
156	component of the	638	percentage of the
157	surface of the	639	that this is
158	of the three	640	top of the
159	in the presence or absence	641	of these cells
160	mm tris hcl	642	map of the
161	the experimental section	643	and methods the
162	at least two	644	of the interaction
163	in the presence or	645	interaction of the
164	determine whether the	646	phase of the
165	with or without	647	results of the
166	half of the	648	are means s e
167	comparison of the	649	at 37 8c for
168	ability of the	650	at restrictive temperatures
169	sites in the	651	extracts prepared from
170	because of the	652	early and late
171	to determine whether the	653	for 48 h
172	has also been	654	but did not
173	the position of the	655	for 16 h
174	derived from the	656	is not yet
175	member of the	657	high concentrations of
176	5 ml of	658	the same conditions
177	obtained from the	659	fig 1 c
178	in the regulation	660	a role in the
179	washed three times	661	the bottom of the
180	min at room	662	agreement with the
181	this article has	663	to understand the
182	a final concentration	664	were collected and
183	was added to the	665	correspond to the
184	30 min at	666	that we have
185	adjacent to the	667	be the result
186	showed that the	668	by the fact
187	of a single	669	1 h with
188	localization of the	670	function as a
189	of the first	671	hypothesis that the
190	of the human	672	effects on the
191	min at room temperature	673	for the production
192	has been cited by	674	was expressed in
193	a final concentration of	675	figure 4 a
194	amino acid residues	676	followed by the
195	is required for the	677	essential for the
196	the other hand	678	consequence of the

197	and analyzed by	679	determination of the
198	found that the	680	binds to the
199	any of the	681	stability of the
200	role of the	682	is the first
201	as in the	683	of the native
202	in materials and methods	684	even in the
203	at least three	685	which is a
204	account for the	686	of which are
205	there was a	687	of a number
206	as a function	688	in the reaction
207	interact with the	689	of the corresponding
208	in materials and	690	in the formation
209	bound to the	691	in activation of
210	for the presence	692	of the small
211	incubated in the	693	concentration of the
212	three times with	694	of the growth
213	the ability of the	695	the absence or presence of 30
214	location of the	696	are means s e m
215	size of the	697	absence or presence of 30
216	results in the	698	it has been proposed
217	is that the	699	provided by dr
218	contribute to the	700	or presence of 30
219	led to the	701	materials and methods the
220	and resuspended in	702	results suggest that the
221	demonstrate that the	703	has recently been
222	suggested that the	704	not yet been
223	with the same	705	was added to a
224	in the materials and methods	706	then treated with
225	in the materials and	707	min followed by
226	close to the	708	a portion of the
227	in the materials	709	there are no
228	described in materials and methods	710	been suggested that
229	for 30 min at	711	known about the
230	described in materials and	712	we have previously
231	described in materials	713	it can be
232	dependent on the	714	which have been
233	to a final	715	fig 1 b
234	to investigate the	716	min at 4
235	is expressed in	717	separated on a
236	face of the	718	is activated by
237	each of these	719	differences between the
238	or in the	720	but it is
239	it is possible that the	721	for 3 min
240	cells were treated with	722	the length of the
241	is possible that the	723	were harvested and
242	one or more	724	as a probe
243	have not been	725	in 20 mm
244	may not be	726	the product of the
245	under the control	727	transferred to a
246	it is also	728	30 min the
247	followed by a	729	the vicinity of
248	the structure of the	730	in the activation of
249	basis of the	731	interacts with the
250	with the indicated	732	concentration of 0
251	positions of the	733	the effects of the
252	in the first	734	it was not
253	were washed twice	735	affected by the
254	for example the	736	attached to the
255	related to the	737	portions of the
256	not shown and	738	leading to the
257	addition to the	739	for a further
258	in the medium	740	to explain the
259	sequences of the	741	leads to the
260	domain of the	742	presence of 0
261	site of the	743	of each of the
262	that in the	744	sensitive to the
263	it should be	745	table 1 in
264	is not clear	746	all of which
265	these results are	747	on the same
266	shows that the	748	understanding of the
267	copies of the	749	with those of

268	resulted in the	750	presence of 1
269	indicated that the	751	to inhibit the
270	regulation of the	752	important for the
271	form of the	753	of one or
272	effects of the	754	by using a
273	increasing concentrations of	755	and 1 mm
274	difference between the	756	found in a
275	demonstrated that the	757	shown on the
276	one of these	758	production of the
277	than that of	759	for each of
278	used for the	760	added and the
279	in both the	761	this is a
280	cells were transfected with	762	and used to
281	no effect on the	763	out of the
282	was used as a	764	performed in the
283	this study we	765	in the activation
284	the nature of the	766	however in the
285	however it is	767	not in the
286	that have been	768	to the manufacturer's instructions
287	use of a	769	2 5 lg ml
288	in the second	770	has been proposed that
289	studies of the	771	data not shown thus
290	studies have shown	772	is known about the
291	by incubation with	773	has been shown that
292	of at least	774	data not shown and
293	note that the	775	a conformational change
294	but not in	776	for 1 h with
295	the sequence of the	777	for 4 h in
296	likely that the	778	the permissive temperature
297	in the experimental	779	reactions were performed
298	the activity of the	780	were obtained from the
299	copy of the	781	insight into the
300	located in the	782	out as described
301	function in the	783	were washed three
302	and the other	784	there may be
303	described in the materials and methods	785	various concentrations of
304	described in the materials and	786	acts as a
305	described in the experimental	787	therefore it is
306	described in the materials	788	for 2 hr
307	we did not	789	only a single
308	there was no	790	is a member
309	as a percentage	791	thus it is
310	away from the	792	the identity of the
311	the basis of the	793	indicated by an
312	fragment containing the	794	interactions between the
313	not shown this	795	the positions of the
314	to identify the	796	1 min at
315	isolated from the	797	in the presence and
316	use of the	798	the interaction of the
317	two of the	799	well as in
318	as described in the materials and	800	located at the
319	as described in the experimental	801	probed with the
320	as described in the materials	802	we used a
321	the fact that the	803	transformed with the
322	version of this	804	we compared the
323	expressed as a	805	at the level
324	the absence of the	806	examination of the
325	fact that the	807	the column was
326	indicates that the	808	and table 1
327	evidence that the	809	removal of the
328	of these two	810	features of the
329	characterization of the	811	treated with the
330	model for the	812	to be an
331	differences in the	813	the results of the
332	to that of the	814	figure 2 a
333	seen in the	815	3 and 5
334	residues of the	816	may be the
335	should be noted	817	orientation of the
336	does not appear	818	residue in the
337	are consistent with the	819	present in a
338	at least one	820	site at the

339	in addition to the	821	amounts of the
340	and probed with	822	figure 2 and
341	supported by the	823	absence of a
342	at least in	824	grown in the
343	and subjected to	825	of the full-length
344	and stained with	826	localization to the
345	presence of an	827	to increase the
346	from the same	828	observed for the
347	to study the	829	of the up
348	residues in the	830	identified in the
349	that the two	831	concentrations of the
350	expression of a	832	result of the
351	and that this	833	in all the
352	as described in materials and	834	min at 37 8c
353	to a final concentration	835	these results are consistent
354	as described in materials	836	added to a final
355	the present study we	837	it seems likely
356	present study we	838	seems likely that
357	we found that the	839	carried out on
358	15 min at	840	excess of unlabelled
359	the presence of an	841	we show that the
360	depending on the	842	min at 37
361	the end of the	843	h at room
362	to the manufacturer's	844	truncated form of
363	to assess the	845	arrows indicate the
364	at the end	846	was used as the
365	possibility that the	847	is involved in the
366	the reaction was	848	the hypothesis that the
367	included in the	849	three times in
368	but not the	850	can also be
369	formation of a	851	two or more
370	purification of the	852	containing 0 5
371	this is the	853	which is consistent
372	shown in the	854	relationship between the
373	on the basis of the	855	although it is
374	for 1 hr	856	the presence of 1
375	the size of the	857	rather than the
376	that there is	858	distance between the
377	in the context	859	mg ml in
378	the control of the	860	that there are
379	properties of the	861	in patients with
380	contrast to the	862	to produce a
381	assembly of the	863	localize to the
382	length of the	864	observation that the
383	figure 2 the	865	on the ability
384	many of the	866	were prepared and
385	product of the	867	and can be
386	fraction of the	868	comparison with the
387	those of the	869	not shown to
388	figure 1 the	870	by binding to
389	as in a	871	occur in the
390	these results indicate	872	that the interaction
391	the possibility that the	873	in the production
392	act as a	874	that activation of
393	to each other	875	forms of the
394	encoded by the	876	such that the
395	not affect the	877	is one of
396	interaction between the	878	activities of the
397	conclude that the	879	specific to the
398	caused by the	880	and absence of
399	prior to the	881	of the four
400	together with the	882	in the number
401	a concentration of	883	and is not
402	specific for the	884	this is in
403	distribution of the	885	in each of
404	of the reaction	886	region in the
405	fragment of the	887	in all of
406	of expression of	888	than in the
407	for 15 min at	889	region and the
408	under the control of the	890	min in the
409	supplemented with 10	891	both of the

410	loss of function	892	of both the
411	is inhibited by	893	reactions were carried out
412	we have not	894	h at room temperature
413	for 3 h	895	tested for their ability
414	the surface of the	896	at 4 8c with
415	from a single	897	in the presence of 1
416	to form a	898	two copies of the
417	all of these	899	other members of the
418	majority of the	900	reactions were carried
419	disruption of the	901	at a concentration of
420	interaction with the	902	the existence of a
421	to test the	903	for 60 min
422	incubated with the	904	not shown suggesting
423	is not a	905	the remainder of the
424	result in the	906	for 24 h
425	the degradation of	907	there are two
426	of each of	908	4 h in the
427	with that of	909	a percentage of the
428	a member of the	910	were crossed to
429	for their ability	911	to bind to the
430	in contrast to the	912	introduced into the
431	to this article	913	between these two
432	response to this	914	determine if the
433	with 1 ml	915	were pooled and
434	are expressed in	916	30 min in
435	mediated by the	917	the total number
436	the role of the	918	present in all
437	necessary for the	919	not shown figure
438	the effect of the	920	in the vicinity
439	revealed that the	921	along with a
440	table 2 the	922	associates with the
441	indicated by the	923	that are required
442	and it is	924	support of this
443	1 2 and	925	existence of a
444	loss of the	926	a function of the
445	site in the	927	and do not
446	from that of	928	sides of the
447	for 2 h at	929	identified as a
448	room temperature for	930	only one of
449	more than one	931	the localization of the
450	it does not	932	removed from the
451	and stored at	933	remainder of the
452	that do not	934	diagram of the
453	the location of the	935	with the appropriate
454	not shown these	936	shown to have
455	been shown that	937	except for the
456	as a single	938	min after the
457	may be a	939	5 min and
458	the samples were	940	al 1991 the
459	decrease in the	941	at the 5'
460	proportion of the	942	the reactions were
461	determined by the	943	so that the
462	role for the	944	this is not
463	by the presence	945	purified from the
464	the stimulation of	946	interactions with the
465	to have a	947	sites on the
466	content of the	948	and purification of
467	of the second	949	identical to the
468	levels of the	950	in the initial
469	these data indicate	951	ends of the
470	have shown that the	952	recognition of the
471	may play a	953	figure 1 a
472	be noted that	954	of the central
473	20 min at	955	in the amount
474	two copies of	956	compared to the
475	have also been	957	response to the
476	it may be	958	of the five
477	the majority of the	959	of the resulting
478	to be involved	960	independent of the
479	led to a	961	study of the
480	explanation for the	962	and the presence

481	evidence for a	963	is not the
482	due to a		

Appendix 3

List of target bundles after application of exclusion criteria

N	Mutual Inf.	Bundle	N	Mutual Inf.	Bundle
906	8.518913	the presence of	33	6.778514	the identity of
625	15.556469	data not shown	33	6.031494	the bottom of
541	13.109891	in the presence of	33	5.25539	the product of
481	8.218921	the absence of	32	22.798241	were washed twice with
387	13.240078	in the absence of	32	18.83578	for their ability to
307	14.240235	as well as	32	16.603617	data not shown in
273	7.14912	the number of	32	15.679216	little is known
259	6.858231	the effect of	32	15.639713	essentially as described
244	15.403582	as described previously	32	13.603061	may contribute to
237	7.730166	the ability of	32	13.557917	has been observed
227	10.177912	as described in	32	12.173017	was assessed by
216	10.021748	shown in figure	32	11.585146	was replaced with
209	11.443076	been shown to	32	11.181106	is mediated by
203	6.676684	the addition of	32	10.52664	were prepared as
194	11.402583	is required for	32	9.881608	in this experiment
190	9.596848	was used to	32	9.272937	for up to
189	9.46708	in response to	32	9.161523	were fixed in
183	8.239267	a number of	32	8.788196	is known to
180	13.490686	results not shown	32	7.810055	in this region
176	7.03375	the effects of	32	7.704166	the tip of
168	7.466129	the level of	32	7.273683	is found in
165	14.306728	it is possible	32	6.821063	to the left
164	15.343361	to determine whether	32	3.397137	the analysis of
164	6.491655	the role of	31	23.455378	at room temperature for
158	10.366571	the fact that	31	16.057948	to be required for
156	14.604337	has been shown	31	15.789634	carried out as
154	11.591088	is consistent with	31	14.835929	carried out with
154	8.558108	in addition to	31	14.531152	can be seen
154	8.021226	the amount of	31	14.069582	as judged by
149	6.72299	the formation of	31	13.182224	have been found
148	10.799778	in this study	31	11.902325	is supported by
146	20.813609	it is possible that	31	11.62512	only a small
146	18.976404	at room temperature	31	11.306004	large number of
145	4.660801	the activity of	31	11.141007	be able to
144	10.970233	was added to	31	10.9444	is not known
143	9.830042	the possibility that	31	9.500299	were identified by
142	6.836724	the rate of	31	9.077895	was performed with
139	8.326431	the basis of	31	8.890332	was required for
137	16.903517	for review see	31	8.608501	a portion of
136	10.896266	were incubated with	31	7.60026	the course of
130	12.172597	we found that	31	6.929042	the same as
129	16.29173	on the basis of	31	6.546725	a loss of
128	10.124116	in order to	31	4.760724	the time of
126	11.192163	have shown that	30	27.912335	little is known about
124	12.172034	the present study	30	21.641929	would be expected to
119	11.0729	was determined by	30	20.974654	these data indicate that
119	9.70822	shown to be	30	17.461612	carried out using
118	17.079535	were carried out	30	14.581846	with the exception of
116	6.625662	in the same	30	14.256518	could be detected
113	8.323654	as shown in	30	12.132765	activity was measured
112	11.206109	an increase in	30	11.923179	in conjunction with
112	8.557439	are shown in	30	10.327546	were transferred to
112	7.246018	the use of	30	9.597991	are known to
112	6.518452	in the present	30	9.199847	were detected by
111	10.289522	a variety of	30	7.810479	in contrast with

109	8.628752	the majority of	30	5.936251	in a similar
107	8.652743	were used to	29	33.811544	it should be noted that
106	24.610113	see materials and methods	29	23.184465	performed as described previously
105	14.287511	no effect on	29	21.470911	it is not clear
105	8.86862	in contrast to	29	16.701764	is not required for
104	19.858479	has been shown to	29	16.67996	has been implicated
101	14.946081	as described above	29	14.913475	are shown in figure
101	9.00203	similar to that	29	14.280907	together these results
101	8.106348	a role in	29	13.327034	in some cases
100	12.029767	likely to be	29	12.87643	was purchased from
95	4.45752	the results of	29	11.884964	with the use of
94	16.867197	was carried out	29	11.10698	is an important
94	7.350548	the production of	29	10.448816	by the presence of
93	12.36017	we show that	29	9.30839	to be determined
93	11.766364	are consistent with	29	7.482539	a set of
93	7.386339	is shown in	29	7.382313	was present in
93	6.464483	the loss of	29	6.992128	in support of
92	12.128537	this suggests that	29	6.456281	a fraction of
92	9.351441	a role for	28	26.163907	expressed as a percentage of
90	13.384828	results suggest that	28	19.36569	results are consistent with
90	12.232864	in the case of	28	16.882975	this is consistent with
90	11.354294	were treated with	28	14.356509	significantly different from
90	5.116717	the function of	28	14.331764	extracts were prepared
89	6.155405	the localization of	28	13.435628	carried out in
88	11.420898	were obtained from	28	12.873359	we have identified
88	7.386959	in figure 1	28	12.600215	see table 1
88	6.300854	the position of	28	12.111424	can be used
88	5.416784	the levels of	28	11.371248	used to determine
87	9.646587	a series of	28	10.945364	small number of
86	16.978962	in the present study	28	10.713625	in this report
84	12.298954	by the addition of	28	10.46153	was prepared from
83	11.080614	are required for	28	10.411291	the notion that
83	10.614944	found to be	28	10.299332	was subjected to
83	7.367536	the ability to	28	10.033743	an average of
82	9.268945	was found to	28	9.972419	are associated with
81	9.467843	by use of	28	9.953102	are representative of
80	10.05184	was used as	28	9.802976	was prepared by
80	6.916711	the accumulation of	28	8.17825	in the dark
79	16.393296	had no effect	28	7.214504	was found in
79	12.626405	appear to be	28	5.843365	the range of
78	13.405048	it is likely	28	4.704166	the products of
78	12.571346	appears to be	27	17.929432	are likely to be
77	9.519501	the observation that	27	16.97986	a large number of
77	7.591968	a total of	27	15.740882	previous studies have
77	5.540557	the structure of	27	15.426757	does not contain
75	10.140612	as described by	27	12.676638	results demonstrate that
74	15.131722	have been identified	27	11.836482	was supported by
74	14.846236	these results suggest	27	11.221776	is based on
74	10.27182	were determined by	27	10.896369	the indicated times
74	7.91013	by addition of	27	10.813315	in a number of
73	9.540532	the requirement for	27	10.560858	is unlikely to
73	5.951958	the result of	27	10.474525	as measured by
72	12.239797	with respect to	27	10.050142	not due to
72	9.700905	were grown in	27	9.480273	by treatment with
72	4.951476	the control of	27	9.184614	to demonstrate that
71	18.790377	have been shown to	27	9.146208	also observed in
71	11.191658	is essential for	27	9.029516	the conclusion that
71	7.504579	the percentage of	27	8.275637	on the surface
70	15.938168	as shown in figure	27	7.625127	was performed in
70	14.498498	we conclude that	27	7.487415	were detected in
70	10.06437	were incubated for	27	7.45116	a change in
70	6.858367	the distribution of	27	6.823437	in fig 1
70	5.78726	of the total	27	6.519595	the efficiency of
69	24.271052	had no effect on	27	6.439425	the behavior of
69	13.229701	their ability to	27	6.197667	the isolation of
69	10.032458	is likely to	27	6.034168	the detection of
69	6.543084	the positions of	27	5.955858	in the top
69	6.244292	the surface of	26	22.236854	here we show that
68	21.407625	these results suggest that	26	19.976137	an important role in
68	12.032724	we have shown	26	19.064177	not appear to be
68	8.775573	in table 1	26	18.490457	we were unable to

68	4.28971	the sequence of	26	17.523452	it is important to
67	13.936232	performed as described	26	17.047297	for reviews see
67	9.054561	the hypothesis that	26	16.089414	as a consequence of
67	7.461698	in figure 2	26	15.565401	carried out at
67	6.358158	a function of	26	14.64186	summarized in table
66	19.847418	it is likely that	26	13.272257	it is clear
65	7.451706	a result of	26	12.072203	we have found
65	6.086636	the end of	26	11.968387	unlikely to be
64	13.472845	as previously described	26	11.939201	been proposed to
64	6.916711	the method of	26	11.676229	important role in
64	5.622528	the interaction of	26	11.187802	we have used
63	6.010184	the development of	26	10.383253	the same time
62	11.970227	not appear to	26	10.033632	were exposed to
61	11.00792	was obtained from	26	10.026209	was analyzed by
61	10.822141	be involved in	26	9.331149	model in which
61	10.529077	in this case	26	9.170894	been observed in
61	10.21407	as a result	26	8.55753	in comparison with
61	10.192897	is associated with	26	8.149513	are similar to
61	8.780334	the existence of	26	8.018189	are indicated in
61	7.935106	at the same	26	8.004247	a combination of
61	7.5028	the nature of	26	7.764344	as shown by
61	6.159392	the size of	26	6.334369	in the bottom
60	28.455018	in the absence or presence of	26	6.223256	the interaction with
60	13.069789	data suggest that	26	5.806528	the release of
59	12.841728	its ability to	26	5.332336	the introduction of
59	12.817143	similar to those	26	4.12882	in the control
58	8.248847	is present in	26	3.965135	in the region
58	7.086636	the lack of	25	24.067236	it has been suggested
57	16.654869	has been proposed	25	14.770975	at a density of
57	7.556956	the extent of	25	14.371566	increasing amounts of
56	11.183616	were subjected to	25	14.336707	together these data
56	10.997246	consistent with this	25	13.253518	high degree of
56	10.534339	to interact with	25	13.21763	as opposed to
55	11.692788	high levels of	25	12.461185	it appears that
55	10.447151	in combination with	25	10.851006	activity was determined
55	9.210786	is involved in	25	10.769527	be important for
55	8.68385	was used for	25	10.612601	to account for
54	13.657621	were purchased from	25	10.535734	were removed by
54	11.282193	were separated by	25	10.329196	the results presented
54	7.400951	the location of	25	10.097133	the difference between
53	12.653239	is dependent on	25	9.877766	is composed of
53	12.466673	results were obtained	25	9.661768	a requirement for
53	11.277291	in the regulation of	25	9.039482	was associated with
53	10.554759	are likely to	25	8.917193	was due to
53	9.57487	a consequence of	25	8.68509	the results obtained
52	15.675107	has been reported	25	8.639799	were obtained with
52	13.749126	to determine if	25	7.820443	are found in
52	13.071092	results indicate that	25	7.333901	at the time
52	11.949588	was confirmed by	25	7.111582	the intensity of
52	11.308832	was performed on	25	7.052472	were present in
52	11.302288	be due to	25	6.568427	a family of
52	10.401352	as determined by	25	5.858367	the value of
52	10.03277	are involved in	25	4.041992	the study of
52	8.496117	were found to	24	22.483972	several lines of evidence
51	21.438349	on the other hand	24	22.059944	remains to be determined
51	11.048687	were unable to	24	21.648352	a wide range of
51	10.937073	be required for	24	19.923537	were prepared as described
51	10.731468	to test this	24	15.31547	to be involved in
51	7.281014	the identification of	24	15.054575	medium supplemented with
51	7.157316	was shown to	24	14.944617	shown in figure 2
50	13.477949	as a function of	24	14.897928	has been demonstrated
50	13.29327	have been described	24	14.543042	by the fact that
50	12.675705	similar to that of	24	13.586645	it is unlikely
50	8.632713	a defect in	24	12.764452	a previous study
49	18.269175	taken together these	24	12.694802	be the result of
49	11.408806	is thought to	24	12.667997	been proposed that
49	9.873339	the interaction between	24	12.095524	for the production of
49	9.392339	in these experiments	24	11.012386	to associate with
49	7.021741	were used in	24	10.69279	also required for
48	14.49167	these data suggest	24	10.523846	predicted to be
48	13.238171	referred to as	24	10.272605	to act as

48	11.486371	be expected to	24	10.139459	was not detected
48	10.327546	were able to	24	9.996561	of a number of
48	8.288177	a range of	24	9.822044	to note that
48	7.130225	the ratio of	24	9.280986	also present in
48	5.889768	the increase in	24	9.015631	was performed by
48	5.856888	to the same	24	8.974112	be required to
48	3.943678	the site of	24	8.862781	been used to
47	35.260737	in the materials and methods section	24	8.542211	are shown as
47	14.747933	play a role	24	8.231026	the remainder of
47	13.289105	been implicated in	24	8.221993	in this model
47	11.963957	low levels of	24	6.924364	the organization of
47	10.751397	was measured by	24	6.748188	this region of
47	10.631992	was performed as	24	6.404691	a deletion of
47	9.529863	is indicated by	24	6.397862	the reduction in
47	8.857339	as a control	24	6.310108	of a large
47	8.402832	was detected in	23	26.509156	taken together these results
47	7.210175	the degree of	23	15.638835	closely related to
47	7.192395	in figure 5	23	12.876515	been shown previously
47	7.117663	the action of	23	12.082942	been identified as
47	6.244008	the length of	23	12.012966	its interaction with
46	28.071689	in the presence or absence of	23	11.732476	have demonstrated that
46	14.494925	as a result of	23	11.22859	were separated on
46	13.690385	has been described	23	11.169955	this work was
46	10.721141	were isolated from	23	11.163402	to ensure that
46	10.401741	are indicated by	23	11.116698	were collected from
46	9.976233	a subset of	23	11.063442	this indicates that
46	9.551005	shown in table	23	11.020472	other members of
46	5.423199	the mechanism of	23	10.911263	two types of
45	16.661463	did not affect	23	10.784978	are unable to
45	15.880416	has been suggested	23	10.772662	is difficult to
45	15.384447	under the control of	23	9.987168	is caused by
45	12.747926	thought to be	23	9.409192	is localized to
45	12.28854	was performed using	23	9.395422	in this process
45	11.547796	similar results were	23	9.260472	were washed with
45	11.25952	were grown at	23	7.53491	the inability of
45	10.902002	for the presence of	23	6.572063	the yield of
45	10.586718	were generated by	23	6.160165	the combination of
45	10.45354	were performed as	23	5.077703	the top of
45	10.35839	were tested for	22	35.975142	according to the manufacturer's instructions
45	7.606602	in figure 3	22	30.999606	it has been proposed that
45	6.955858	the difference in	22	25.106804	carried out as described
45	4.109719	the region of	22	24.713764	were washed three times
44	19.987636	play a role in	22	22.836459	an equal volume of
44	18.555904	used in this study	22	21.616271	has been implicated in
44	18.088071	we have shown that	22	21.374967	under the same conditions
44	17.309423	does not require	22	18.671969	we asked whether
44	16.718531	was found to be	22	18.601234	is thought to be
44	12.197815	results show that	22	18.148428	an equal volume
44	11.245125	are expressed as	22	17.713793	at the same time
44	11.136425	to confirm that	22	15.772432	as well as in
44	10.772726	was isolated from	22	15.730639	did not appear
44	10.669485	were analyzed by	22	15.446925	is a member of
44	9.144023	were added to	22	14.53655	equal volume of
44	8.753202	are present in	22	13.974507	be explained by
44	8.246206	were used for	22	13.904994	may be due
44	8.005601	is similar to	22	13.796757	there are several
43	21.016351	these data suggest that	22	13.700184	consistent with previous
43	16.322198	would be expected	22	13.61486	used to amplify
43	14.097135	we propose that	22	12.668212	on the surface of
43	13.894043	we find that	22	12.320468	has been used
43	13.198965	experiments were performed	22	12.265526	used to identify
43	12.300193	remains to be	22	12.104791	at the level of
43	12.015579	were analysed by	22	11.877884	be responsible for
43	11.786432	the relationship between	22	11.838112	no evidence for
43	9.834938	was detected by	22	11.743943	have suggested that
43	9.072047	a decrease in	22	11.669424	very similar to
43	8.180788	were performed in	22	11.51919	by virtue of
43	6.885627	the frequency of	22	11.242501	to address this
42	13.827413	prepared as described	22	10.411276	total number of
42	12.038252	mechanism by which	22	10.404247	are essential for
42	11.367919	we suggest that	22	10.100817	have found that

42	10.674662	were stained with	22	9.621651	been found to
42	10.64249	known to be	22	9.298113	in the formation of
42	10.162955	is sufficient to	22	9.284249	was determined as
42	8.220219	the onset of	22	9.228604	alone or in
42	7.946458	the importance of	22	8.199851	in a manner
41	12.998128	data indicate that	22	7.57418	described in figure
41	12.292707	a gift from	22	7.478716	at the surface
41	10.896011	were prepared from	22	7.057404	are described in
41	10.315606	not required for	22	7.000252	in the upper
41	10.082386	is able to	22	6.954538	a comparison of
41	8.971748	were used as	22	6.431442	in a total
41	8.379589	a percentage of	22	5.982179	was used in
41	7.75969	the context of	22	5.710233	the differences in
41	6.037224	the process of	22	5.458605	the association of
40	15.894374	under these conditions	22	5.134254	the possibility of
40	12.463129	in all cases	21	27.429447	it has been shown that
40	11.735682	in this paper	21	27.377839	these results are consistent with
40	10.451108	is not required	21	25.848147	at a flow rate of
40	9.423692	a member of	21	24.89218	it seems likely that
40	9.32991	were performed with	21	21.887948	to test this hypothesis
40	8.460547	a model for	21	18.649446	have been identified in
40	5.253023	the fraction of	21	15.471113	shown in figure 3
39	20.116642	studies have shown that	21	15.431162	exclude the possibility
39	17.557042	is likely to be	21	15.40135	at various times
39	15.849463	as a percentage of	21	14.982594	we tested whether
39	11.966373	were performed using	21	14.090449	was introduced into
39	10.220989	as compared with	21	13.126576	this implies that
39	10.143702	was able to	21	12.48135	total volume of
39	10.017695	has shown that	21	12.056379	are summarized in
39	9.450312	in terms of	21	11.967116	results are expressed
39	8.212574	is required to	21	11.872036	were as follows
39	7.893991	the appearance of	21	11.3179	be caused by
39	7.26053	the proportion of	21	10.992672	was based on
38	26.753673	has been shown to be	21	10.892676	see figure 2
38	22.145471	similar results were obtained	21	10.523136	in the production of
38	17.252008	in this study we	21	10.424595	see figure 1
38	13.035543	on ice for	21	10.026376	were allowed to
38	12.295187	appeared to be	21	10.003648	suggesting that this
38	12.252306	we demonstrate that	21	9.884295	was unable to
38	11.548193	for an additional	21	9.816046	were made by
38	11.134014	is necessary for	21	9.659343	was induced by
38	10.234852	with the exception	21	9.585703	was examined by
38	10.018661	were resuspended in	21	9.07292	it was shown
38	9.727875	the idea that	21	8.869224	is predicted to
38	9.302899	on the left	21	8.710781	as seen in
38	8.371897	by the method	21	8.650829	as part of
38	7.360559	the evolution of	21	8.35638	to that seen
38	7.219691	at the indicated	21	8.332112	the rest of
38	6.065103	the assembly of	21	7.638676	to show that
37	29.655597	described in the experimental section	21	7.478958	in figure 7
37	19.843662	in the experimental section	21	6.81533	to the right
37	15.150291	an important role	21	6.780375	to that observed
37	15.101027	shown in figure 1	21	5.63455	at the site
37	10.420034	was generated by	21	5.538115	the rates of
37	9.628083	was obtained by	21	5.524757	the average of
37	9.512366	were obtained by	21	5.237156	as in figure
37	8.611591	the timing of	21	3.480564	of the indicated
37	8.564957	be used to	20	29.743619	tested for their ability to
37	8.255736	is independent of	20	22.090368	were carried out at
37	7.429082	was observed in	20	21.432241	did not appear to
37	6.781516	the stability of	20	21.39387	has been suggested that
37	5.908687	the activities of	20	17.391227	which is consistent with
36	37.158093	as described in the experimental section	20	16.138005	are shown in table
36	35.301326	as described in materials and methods	20	15.710881	that are required for
36	27.440109	it should be noted	20	15.626498	have been implicated
36	24.136379	in the present study we	20	15.465311	were found to be
36	22.048619	according to the manufacturer's	20	14.640983	an essential role
36	14.880644	little or no	20	14.479732	the total number of
36	14.370766	been described previously	20	14.207395	in support of this
36	14.322515	is shown in figure	20	13.965816	in the vicinity of
36	12.981925	by the method of	20	13.602617	as reported previously

36	12.596021	when compared with	20	13.578275	to distinguish between
36	11.882621	was digested with	20	13.187698	a critical role
36	11.870869	as a consequence	20	12.288991	consistent with our
36	11.612091	in each case	20	12.029243	at the surface of
36	10.188374	was purified from	20	11.940506	compared with control
36	9.309796	is due to	20	11.785676	in concert with
36	8.577807	shown in fig	20	11.645072	a small number
36	8.326116	in table 2	20	11.418404	is also possible
36	8.247305	a component of	20	11.380713	was dependent on
36	7.613068	a response to	20	11.358561	would result in
35	21.483034	it has been shown	20	11.181308	are responsible for
35	12.866281	in the context of	20	11.091021	was dissolved in
35	11.826284	are thought to	20	10.911499	the present work
35	11.380037	in agreement with	20	10.895615	were processed for
35	11.085759	is responsible for	20	10.866029	was determined using
35	10.009188	were prepared by	20	10.769571	was mixed with
35	9.518432	to be required	20	10.557332	at this time
35	8.231047	a mixture of	20	10.538967	is subject to
35	6.605421	the generation of	20	10.411228	in the amount of
35	5.728653	the pattern of	20	10.108206	be consistent with
34	22.197775	does not appear to	20	10.088239	shown previously that
34	19.976801	was performed as described	20	10.046065	be associated with
34	17.886044	the manufacturer's instructions	20	9.949666	are able to
34	14.544721	have been reported	20	9.742981	is sensitive to
34	12.723279	at the end of	20	9.310653	were treated for
34	12.463444	analysis was performed	20	9.208378	was resuspended in
34	10.988871	expected to be	20	8.953498	as indicated by
34	10.437402	possibility is that	20	8.884461	is capable of
34	9.751157	is important for	20	8.713151	in the number of
34	9.264217	be detected in	20	8.566044	not result in
34	9.122687	were grown to	20	8.287969	as a model
34	8.965574	the finding that	20	8.19726	in table 3
34	8.567567	a reduction in	20	8.053034	as described for
33	20.842233	these results indicate that	20	7.805379	were washed in
33	19.818016	were performed as described	20	7.405927	this type of
33	19.299501	kindly provided by	20	7.064412	were obtained in
33	16.214004	does not affect	20	7.045994	the question of
33	15.319713	is known about	20	7.001563	was similar to
33	13.673959	can be detected	20	6.9871	the beginning of
33	13.280182	to test whether	20	6.912138	the significance of
33	11.923179	in accordance with	20	6.756487	the removal of
33	11.875075	lines of evidence	20	6.345554	the incorporation of
33	11.435843	been reported to	20	6.117887	the origin of
33	10.39661	been identified in	20	6.026095	with the following
33	7.599965	a density of	20	5.616006	the properties of
			20	3.579294	the growth of

Appendix 4

Complete list of target bundles

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
7.730166	237	**the ability of**	ability	NP+of	description		
7.367536	83	**the ability to**	ability	other NP	description		
13.229701	69	their ability to	ability	other NP	description		
18.83578	32	for their ability to	ability	other PP	description		
12.841728	59	its ability to	ability	other NP	description		
10.082386	41	**is able to**	able	V/A+to	description		
9.949666	20	are able to	able	V/A+to	description		
11.141007	31	be able to	able	V/A+to	description		
10.143702	39	was able to	able	V/A+to	description		
10.327546	48	**were able to**	able	V/A+to	inferential		(we) were able to [demonstrate, detect, identify]
8.218921	481	**the absence of**	absence	NP+of	description		
13.240078	387	**in the absence of**	absence	PP+of	framing		(occur) in the (complete) absence of
28.455018	60	**in the absence or presence of**	absence	PP+of	framing		(in the) absence or presence of
11.923179	33	**in accordance with**	accordance	other PP	framing	citation	
10.612601	25	**to account for**	account	V/A+to	objective	inferential	
6.916711	80	**the accumulation of**	accumulation	NP+of	procedure		
10.272605	24	**to act as**	act	V/A+to	description		
7.117663	47	**the action of**	action	NP+of	procedure		
4.660801	145	**the activity of**	activity	NP+of	procedure		
5.908687	37	the activities of	activity	NP+of	procedure		
10.970233	144	**was added to**	add	passive+PP	procedure		
9.144023	44	were added to	add	passive+PP	procedure		
6.676684	203	**the addition of**	addition	NP+of	procedure		
12.298954	84	**by the addition of**	addition	PP+of	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
7.91013	74	by addition of	addition	PP+of	procedure		
8.558108	154	**in addition to**	addition	other PP	additive		
11.548193	38	**for an additional**	additional	other PP	quantification		
11.242501	22	**to address this**	address	V/A+to	objective		
16.661463	45	**did not affect**	affect	other V fragment	causative		
16.214004	33	does not affect	affect	other V fragment	causative		
11.380037	35	**in agreement with**	agreement	other PP	comparative	citation	in (good) agreement with
10.026376	21	**were allowed to**	allow	other passive	procedure		
10.411228	20	**in the amount of**	amount	PP+of	quantification		
8.021226	154	**the amount of**	amount	NP+of	quantification		
3.397137	32	**the analysis of**	analysis	NP+of	procedure		
10.669485	44	**were analyzed by**	analyze	passive+PP	procedure		
10.026209	26	was analyzed by	analyze	passive+PP	procedure		
12.015579	43	were analysed by	analyze	passive+PP	procedure		
12.461185	25	**it appears that**	appear	anticipatory it	inferential	stance	(thus) it [appears, would appear] that
12.626405	79	**appear to be**	appear	V/A+to	inferential	stance	[appear, appears, appeared] to be
12.571346	78	appears to be	appear	V/A+to	inferential	stance	
12.295187	38	appeared to be	appear	V/A+to	inferential	stance	
11.970227	62	**not appear to**	appear	V/A+to	inferential	stance	[does, did] not appear to [affect, be, contain, have, involve]
22.197775	34	does not appear to	appear	V/A+to	inferential	stance	
19.064177	26	not appear to be	appear	V/A+to	inferential	stance	
15.730639	22	did not appear	appear	other V fragment	inferential	stance	
21.432241	20	did not appear to	appear	V/A+to	inferential	stance	
7.893991	39	**the appearance of**	appearance	NP+of	description		
18.671969	22	**we asked whether**	ask	we+V	objective		
6.065103	38	**the assembly of**	assembly	NP+of	procedure		
12.173017	32	**was assessed by**	assess	passive+PP	procedure		
10.192897	61	**is associated with**	associate	passive+PP	inferential	stance	[is, are, was [can, could, may, might be]] (closely, significantly, strongly, tightly) associated with
9.972419	28	are associated with	associate	passive+PP	inferential	stance	
9.039482	25	was associated with	associate	passive+PP	inferential	stance	
10.046065	20	be associated with	associate	passive+PP	inferential	stance	
11.012386	24	**to associate with**	associate	V/A+to	inferential	stance	
5.458605	22	**the association of**	association	NP+of	inferential	stance	

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
10.033743	28	**an average of**	average	NP+of	quantification		
5.524757	21	the average of	average	NP+of	quantification		
11.221776	27	**is based on**	base	passive+PP	framing		[is, was] (largely, mainly) based on
10.992672	21	was based on	base	passive+PP	framing		
8.326431	139	**the basis of**	basis	NP+of	framing		
16.29173	129	**on the basis of**	basis	PP+of	framing		
8.326431	139	the basis of	basis	NP+of	framing		
6.9871	20	**the beginning of**	beginning	NP+of	procedure		
6.439425	27	**the behavior of**	behavior	NP+of	description		
6.334369	26	**in the bottom**	bottom	other PP	location		
6.031494	33	**the bottom of**	bottom	NP+of	location		
8.884461	20	**is capable of**	capable	be+AP	description		
15.565401	26	**carried out at**	carry out	passive+PP	procedure		
13.435628	28	**carried out in**	carry out	passive+PP	procedure		
14.835929	31	**carried out with**	carry out	passive+PP	procedure		
17.461612	30	carried out using	carry out	other passive	procedure		
17.079535	118	**were carried out**	carry out	other passive	procedure		
16.867197	94	was carried out	carry out	other passive	procedure		
22.090368	20	were carried out at	carry out	passive+PP	procedure		
12.232864	90	**in the case of**	case	PP+of	framing		
10.529077	61	**in this case**	case	other PP	framing		
12.463129	40	**in all cases**	case	other PP	framing		
11.612091	36	**in each case**	case	other PP	framing		
13.327034	29	**in some cases**	case	other PP	framing		
9.987168	23	**is caused by**	cause	passive+PP	causative		[is, [could, may] be] caused by
11.3179	21	be caused by	cause	passive+PP	causative	stance	
7.45116	27	**a change in**	change	NP+other	procedure		
13.272257	26	**it is clear**	clear	anticipatory it	stance		it is clear (from) (that)
21.470911	29	**it is not clear**	clear	anticipatory it	stance		it is [not clear, unclear] [how, if, what, whether, which, why]
11.116698	23	**were collected from**	collect	passive+PP	procedure		
8.004247	26	**a combination of**	combination	NP+of	grouping		
6.160165	23	the combination of	combination	NP+of	grouping		
10.447151	55	**in combination with**	combination	other PP	additive	framing	(alone or) in combination with

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
9.228604	22	alone or in	combination	other AP	additive	framing	
9.228604	22	alone or in	combination	other AP	additive	framing	
11.940506	20	**compared with control**	compare	passive+PP	procedure		
10.220989	39	**as compared with**	compare	as+V	comparative		[as, when] compared [to, with]
12.596021	36	when compared with	compare	passive+PP	comparative		
6.954538	22	**a comparison of**	comparison	NP+of	procedure		
8.55753	26	**in comparison with**	comparison	other PP	comparative		
8.247305	36	**a component of**	component	NP+of	grouping		
9.877766	25	**is composed of**	compose	passive+PP	framing		is composed (entirely, largely, mainly, predominantly) of
11.785676	20	**in concert with**	concert	other PP	additive	framing	(alone or) in concert with
14.498498	70	**we conclude that**	conclude	we+V	inferential	stance	(therefore) we conclude that
9.029516	27	**the conclusion that**	conclusion	V/N+that cl	inferential		
15.894374	40	**under these conditions**	condition	other PP	framing		under [these, the] conditions (used)
21.374967	22	**under the same conditions**	condition	other PP	framing		
11.949588	52	**was confirmed by**	confirm	passive+PP	procedure		
11.136425	44	**to confirm that**	confirm	V/A+to	objective		
11.923179	30	**in conjunction with**	conjunction	other PP	additive	framing	
11.870869	36	**as a consequence**	consequence	other PP	causative		
16.089414	26	**as a consequence of**	consequence	PP+of	causative		
9.57487	53	a consequence of	consequence	NP+of	causative		
9.57487	53	**a consequence of**	consequence	NP+of	causative		
11.591088	154	**is consistent with**	consistent	be+AP	comparative	citation	(this) [result, conclusion, finding, hypothesis, idea, this] is consistent with [[this, our, the] (previous) [data, hypothesis, idea, observations, notion, reports, results, studies, work]
10.997246	56	consistent with this	consistent	other AP	comparative	citation	
16.882975	28	this is consistent with	consistent	others	comparative	citation	
13.700184	22	consistent with previous	consistent	other AP	comparative	citation	
12.288991	20	consistent with our	consistent	other AP	comparative	citation	
17.391227	20	which is consistent with	consistent	be+AP	comparative	citation	
11.766364	93	**are consistent with**	consistent	be+AP	comparative	citation	[these, our] [results, data, findings, observations, studies] are consistent with [[this, our, the] (previous) [data, idea, hypothesis, observations, notion, reports, results, studies, work]
10.997246	56	consistent with this	consistent	other AP	comparative	citation	
19.36569	28	results are consistent with	consistent	others	comparative	citation	
13.700184	22	consistent with previous	consistent	other AP	comparative	citation	

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
27.377839	21	these results are consistent with	consistent	others	comparative	citation	
12.288991	20	consistent with our	consistent	other AP	comparative	citation	
10.108206	20	**be consistent with**	consistent	be+AP	comparative	citation	
15.426757	27	**does not contain**	contain	other V fragment	description		
7.75969	41	**the context of**	context	NP+of	framing		
12.866281	35	**in the context of**	context	PP+of	framing		[in, within] the context of
7.75969	41	the context of	contrast	NP+of	framing		
8.86862	105	**in contrast to**	contrast	other PP	comparative		in contrast, in contrast [to, with]
7.810479	30	in contrast with	contrast	other PP	comparative		
13.603061	32	**may contribute to**	contribute	other V fragment	causative	stance	
4.951476	72	**the control of**	control	NP+of	procedure		
8.857339	47	**as a control**	control	other PP	procedure		
4.12882	26	**in the control**	control	other PP	procedure		
15.384447	45	**under the control of**	control	other PP	framing		
7.60026	31	**the course of**	course	NP+of	framing		
8.17825	28	**in the dark**	dark	other PP	location		
9.072047	43	**a decrease in**	decrease	NP+other	quantification		
8.632713	50	**a defect in**	defect	NP+other	description		
7.210175	47	**the degree of**	degree	NP+of	description		
13.253518	25	high degree of	degree	NP+of	description		
6.404691	24	**a deletion of**	deletion	NP+of	procedure		
12.676638	27	**results demonstrate that**	demonstrate	V/N+that cl	inferential		(these) *(our)* [data, results] demonstrate that
12.252306	38	**we demonstrate that**	demonstrate	we+V	inferential	stance	we [demonstrate, have demonstrated] that
11.732476	23	have demonstrated that	demonstrate	V/N+that cl	inferential	stance	
14.897928	24	**has been demonstrated**	demonstrate	other passive	citation	inferential	(it) has been demonstrated (that)
9.184614	27	**to demonstrate that**	demonstrate	V/A+to	objective		
7.599965	33	**a density of**	density	NP+of	quantification		
14.770975	25	**at a density of**	density	PP+of	quantification		
12.653239	53	**is dependent on**	dependent	be+AP	framing		
11.380713	20	was dependent on	dependent	be+AP	framing		
15.403582	244	**as described previously**	describe	as+V	structuring		([was, were] carried out, performed, prepared) (essentially) as (previously) described (previously) (above, in the experimental section, in materials and methods)
10.177912	227	as described in	describe	as+V	structuring		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
14.946081	101	as described above	describe	as+V	structuring		
13.936232	67	performed as described	describe	other passive	structuring		
13.472845	64	as previously described	describe	as+V	structuring		
10.631992	47	was performed as	describe	passive+PP	structuring		
10.45354	45	were performed as	describe	passive+PP	structuring		
13.827413	42	prepared as described	describe	passive+PP	structuring		
29.655597	37	described in the experimental section	describe	passive+PP	structuring		
37.158093	36	as described in the experimental section	describe	as+V	structuring		
35.301326	36	as described in materials and methods	describe	as+V	structuring		
19.976801	34	was performed as described	describe	passive+PP	structuring		
19.818016	33	were performed as described	describe	passive+PP	structuring		
15.639713	32	essentially as described	describe	as+V	structuring		
15.789634	31	carried out as	describe	passive+PP	structuring		
23.184465	29	performed as described previously	describe	passive+PP	structuring		
19.923537	24	were prepared as described	describe	passive+PP	structuring		
25.106804	22	carried out as described	describe	passive+PP	structuring		
10.140612	75	**as described by**	describe	as+V	citation		([was, were] carried out, performed, prepared) (essentially) as described [by, for, in]
10.177912	227	as described in	describe	as+V	citation		
13.936232	67	performed as described	describe	other passive	citation		
10.631992	47	was performed as	describe	passive+PP	citation		
10.45354	45	were performed as	describe	passive+PP	citation		
13.827413	42	prepared as described	describe	passive+PP	citation		
19.976801	34	was performed as described	describe	passive+PP	citation		
19.818016	33	were performed as described	describe	passive+PP	citation		
15.639713	32	essentially as described	describe	as+V	citation		
15.789634	31	carried out as	describe	passive+PP	citation		
23.184465	29	performed as described previously	describe	passive+PP	citation		
19.923537	24	were prepared as described	describe	passive+PP	citation		
25.106804	22	carried out as described	describe	passive+PP	citation		
14.370766	36	**been described previously**	describe	other passive	citation		[has, have] been described (previously)
13.29327	50	have been described	describe	other passive	citation		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
13.690385	46	has been described	describe	other passive	citation		
7.057404	22	**are described in**	describe	passive+PP	structuring	citation	
8.053034	20	**as described for**	describe	as+V	structuring	citation	
9.834938	43	**was detected by**	detect	passive+PP	procedure		
9.199847	30	were detected by	detect	passive+PP	procedure		
8.402832	47	**was detected in**	detect	passive+PP	inferential		[was, were [can, could] be] detected (in)
9.264217	34	be detected in	detect	passive+PP	inferential	stance	
13.673959	33	can be detected	detect	other passive	inferential	stance	
14.256518	30	could be detected	detect	other passive	inferential	stance	
7.487415	27	were detected in	detect	passive+PP	inferential		
10.139459	24	**was not detected**	detect	other passive	inferential		
6.034168	27	**the detection of**	detection	NP+of	procedure		
9.284249	22	**was determined as**	determine	passive+PP	procedure		
11.0729	119	**was determined by**	determine	passive+PP	procedure		
10.866029	20	was determined using	determine	other passive	procedure		
10.27182	74	were determined by	determine	passive+PP	procedure		
10.851006	25	activity was determined	determine	other passive	procedure		
10.401352	52	**as determined by**	determine	as+V	inferential		
15.343361	164	**to determine whether**	determine	V/A+to	objective		
13.749126	52	to determine if	determine	V/A+to	objective		
6.010184	63	**the development of**	development	NP+of	procedure		
6.955858	45	**the difference in**	difference	NP+other	comparative		the [difference, differences] in
5.710233	22	the differences in	difference	NP+other	comparative		
10.097133	25	**the difference between**	difference	NP+other	comparative		
14.356509	28	**significantly different from**	different	other AP	comparative		
10.772662	23	**is difficult to**	difficult	be+AP	engagement		[it] is difficult to
11.882621	36	**was digested with**	digest	passive+PP	procedure		
11.091021	20	**was dissolved in**	dissolve	passive+PP	procedure		
13.578275	20	**to distinguish between**	distinguish	V/A+to	objective		
6.858367	70	**the distribution of**	distribution	NP+of	grouping		
9.309796	36	**is due to**	due to	be+AP	causative		[is, was, [could, may, might] be] (likely, mainly, possibly, presumably, probably) due to
11.302288	52	be due to	due to	be+AP	causative	stance	
8.917193	25	was due to	due to	be+AP	causative		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
13.904994	22	may be due	due to	be+AP	causative	stance	
10.050142	27	**not due to**	due to	other AP	causative	stance	
14.287511	105	**no effect on**	effect	NP+other	causative		had no (detectable, detrimental, significant, similar) [effect, effects] on
16.393296	79	had no effect	effect	other V fragment	causative		
24.271052	69	had no effect on	effect	other V fragment	causative		
6.858231	259	**the effect of**	effect	NP+of	causative		
7.03375	176	the effects of	effect	NP+of	causative		
6.519595	27	**the efficiency of**	efficiency	NP+of	quantification		
12.723279	34	**at the end of**	end	PP+of	location		
6.086636	65	**the end of**	end	NP+of	location		
11.163402	23	**to ensure that**	ensure	V/A+to	objective		
11.191658	71	**is essential for**	essential	be+AP	engagement		[is, are] (absolutely) essential for
10.404247	22	are essential for	essential	be+AP	stance		
11.875075	33	**lines of evidence**	evidence	other NP	inferential		(several) lines of evidence
22.483972	24	several lines of evidence	evidence	other NP	inferential		
11.838112	22	**no evidence for**	evidence	NP+other	inferential		
7.360559	38	**the evolution of**	evolution	NP+of	procedure		
9.585703	21	**was examined by**	examine	passive+PP	procedure		
14.581846	30	**with the exception of**	exception	PP+of	framing		with the exception [of, that]
10.234852	38	with the exception	exception	other PP	framing		
15.431162	21	**exclude the possibility**	exclude	other V fragment	inferential	engagement	[one, we, data, results, studies] [cannot, do not] [discount, eliminate, exclude, rule out] the possibility [of, that]
9.830042	143	the possibility that	exclude	V/N+that cl	inferential	engagement	
5.134254	22	the possibility of	exclude	NP+of	inferential	engagement	
8.780334	61	**the existence of**	existence	NP+of	description		
11.486371	48	**be expected to**	expect	V/A+to	inferential	stance	[can, might, would] be expected to (be)
16.322198	43	would be expected	expect	other passive	inferential	stance	
21.641929	30	would be expected to	expect	V/A+to	inferential	stance	
10.988871	34	expected to be	expect	V/A+to	inferential	stance	
10.988871	34	**expected to be**	expect	V/A+to	inferential	stance	[is, are, as] expected to be
9.392339	49	**in these experiments**	experiment	other PP	structuring		
9.881608	32	in this experiment	experiment	other PP	structuring		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
13.974507	22	**be explained by**	explain	passive+PP	causative	inferential	
10.033632	26	**were exposed to**	expose	passive+PP	procedure		
11.245125	44	**are expressed as**	express	passive+PP	structuring		[data, results, values] are expressed as [means, units, as a percentage of]
11.967116	21	results are expressed	express	other passive	structuring		
7.556956	57	**the extent of**	extent	NP+of	description		
10.366571	158	**the fact that**	fact	V/N+that cl	framing		
14.543042	24	by the fact that	fact	V/N+that cl	framing		
6.568427	25	**a family of**	family	NP+of	grouping		
15.938168	70	**as shown in figure**	figure	as+V	structuring		(as) [depicted, described, illustrated, presented, shown] in [fig, figure 1,2,3...]
8.323654	113	as shown in	figure	as+V	structuring		
10.021748	216	shown in figure	figure	passive+PP	structuring		
15.101027	37	shown in figure 1	figure	passive+PP	structuring		
8.577807	36	shown in fig	figure	passive+PP	structuring		
14.944617	24	shown in figure 2	figure	passive+PP	structuring		
7.57418	22	described in figure	figure	passive+PP	structuring		
15.471113	21	shown in figure 3	figure	passive+PP	structuring		
7.386959	88	in figure 1	figure	other PP	structuring		
7.461698	67	in figure 2	figure	other PP	structuring		
7.192395	47	in figure 5	figure	other PP	structuring		
7.606602	45	in figure 3	figure	other PP	structuring		
6.823437	27	in fig 1	figure	other PP	structuring		
7.478958	21	in figure 7	figure	other PP	structuring		
7.386959	88	in figure 1	figure	other PP	structuring		
7.461698	67	in figure 2	figure	other PP	structuring		
7.192395	47	in figure 5	figure	other PP	structuring		
7.606602	45	in figure 3	figure	other PP	structuring		
6.823437	27	in fig 1	figure	other PP	structuring		
7.478958	21	in figure 7	figure	other PP	structuring		
14.322515	36	**is shown in figure**	figure	passive+PP	structuring		[is, are] [depicted, described, illustrated, presented, shown] in [fig, figure 1,2,3...]
14.913475	29	are shown in figure	figure	passive+PP	structuring		
10.021748	216	shown in figure	figure	passive+PP	structuring		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
8.557439	112	are shown in	figure	passive+PP	structuring		
7.386339	93	is shown in	figure	passive+PP	structuring		
15.101027	37	shown in figure 1	figure	passive+PP	structuring		
8.577807	36	shown in fig	figure	passive+PP	structuring		
14.944617	24	shown in figure 2	figure	passive+PP	structuring		
7.57418	22	described in figure	figure	passive+PP	structuring		
15.471113	21	shown in figure 3	figure	passive+PP	structuring		
5.237156	21	**as in figure**	figure	other PP	structuring		as in figure [1,2,3…]
7.386959	88	in figure 1	figure	other PP	structuring		
7.461698	67	in figure 2	figure	other PP	structuring		
7.192395	47	in figure 5	figure	other PP	structuring		
7.606602	45	in figure 3	figure	other PP	structuring		
6.823437	27	in fig 1	figure	other PP	structuring		
7.478958	21	in figure 7	figure	other PP	structuring		
10.614944	83	**found to be**	find	V/A+to	inferential	citation	[have been, was, were] found to (be)
9.268945	82	was found to	find	V/A+to	inferential		
8.496117	52	were found to	find	V/A+to	inferential		
16.718531	44	was found to be	find	V/A+to	inferential		
13.182224	31	have been found	find	other passive	citation	inferential	
9.621651	22	been found to	find	V/A+to	inferential		
15.465311	20	were found to be	find	V/A+to	inferential		
7.273683	32	**is found in**	find	passive+PP	generalization	inferential	[is, are] found in
7.820443	25	are found in	find	passive+PP	generalization	inferential	
7.214504	28	**was found in**	find	passive+PP	inferential		
12.172597	130	**we found that**	find	we+V	inferential	stance	we [find, found, have found] that
13.894043	43	we find that	find	we+V	inferential	stance	
12.072203	26	we have found	find	we+V	inferential	stance	
10.100817	22	have found that	find	V/N+that cl	inferential	stance	
8.965574	34	**the finding that**	finding	V/N+that cl	inferential		
9.161523	32	**were fixed in**	fix	passive+PP	procedure		
11.872036	21	**were as follows**	follow	as+V	structuring		
6.026095	20	**with the following**	following	other PP	structuring		
6.72299	149	**the formation of**	formation	NP+of	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
9.298113	22	in the formation of	formation	PP+of	procedure		
6.456281	29	**a fraction of**	fraction	NP+of	quantification		
5.253023	40	**the fraction of**	fraction	NP+of	quantification		
6.885627	43	**the frequency of**	frequency	NP+of	quantification		
5.116717	90	**the function of**	function	NP+of	description		
6.358158	67	a function of	function	NP+of	description		
13.477949	50	**as a function of**	function	PP+of	framing		
10.586718	45	**were generated by**	generate	passive+PP	procedure		
10.420034	37	was generated by	generate	passive+PP	procedure		
6.605421	35	**the generation of**	generation	NP+of	procedure		
12.292707	41	**a gift from**	gift	NP+other	acknowledgment		
11.25952	45	**were grown at**	grow	passive+PP	procedure		
9.700905	72	**were grown in**	grow	passive+PP	procedure		
9.122687	34	**were grown to**	grow	passive+PP	procedure		
3.579294	20	**the growth of**	growth	NP+of	procedure		
21.438349	51	**on the other hand**	hand	other PP	comparative	additive	
9.054561	67	**the hypothesis that**	hypothesis	V/N+that cl	inferential		
13.035543	38	**on ice for**	ice	other PP	procedure		
9.727875	38	**the idea that**	idea	V/N+that cl	framing		
7.281014	51	**the identification of**	identification	NP+of	procedure		
9.500299	31	**were identified by**	identify	passive+PP	procedure		
18.649446	21	**have been identified in**	identify	passive+PP	citation	inferential	
15.131722	74	have been identified	identify	other passive	citation	inferential	
10.39661	33	been identified in	identify	passive+PP	citation	inferential	
12.082942	23	**been identified as**	identify	passive+PP	citation	inferential	
15.131722	74	have been identified	identify	other passive	citation	inferential	
12.873359	28	**we have identified**	identify	we+V	inferential	stance	
6.778514	33	**the identity of**	identity	NP+of	description		
13.289105	47	**been implicated in**	implicate	passive+PP	citation	inferential	[has, have] been (directly, previously, strongly) implicated [as, in]
16.67996	29	has been implicated	implicate	other passive	citation	inferential	
21.616271	22	has been implicated in	implicate	passive+PP	citation	inferential	
15.626498	20	have been implicated	implicate	other passive	citation	inferential	
13.126576	21	**this implies that**	imply	V/N+that cl	inferential		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
7.946458	42	**the importance of**	importance	NP+of	description		
9.751157	34	**is important for**	important	be+AP	engagement		[is, [may, might, will] be] (critically) important for
10.769527	25	be important for	important	be+AP	stance		
11.10698	29	**is an important**	important	be+AP	engagement		
7.53491	23	**the inability of**	inability	NP+of	description		
6.345554	20	**the incorporation of**	incorporation	NP+of	procedure		
11.206109	112	**an increase in**	increase	NP+other	quantification		
5.889768	48	the increase in	increase	NP+other	quantification		
14.371566	25	**increasing amounts of**	increase	NP+of	quantification		
10.06437	70	**were incubated for**	incubate	passive+PP	procedure		
10.896266	136	**were incubated with**	incubate	passive+PP	procedure		
8.255736	37	**is independent of**	independent	be+AP	framing		
11.063442	23	**this indicates that**	indicate	V/N+that cl	inferential		this (strongly) indicates that
13.071092	52	**results indicate that**	indicate	V/N+that cl	inferential		these [data, findings, results] indicate that
20.842233	33	these results indicate that	indicate	V/N+that cl	inferential		
12.998128	41	data indicate that	indicate	V/N+that cl	inferential		
20.974654	30	these data indicate that	indicate	V/N+that cl	inferential		
9.529863	47	**is indicated by**	indicate	passive+PP	structuring		
10.401741	46	are indicated by	indicate	passive+PP	structuring		
8.018189	26	**are indicated in**	indicate	passive+PP	structuring		
7.219691	38	**at the indicated**	indicate	other passive	structuring		at the indicated [concentrations, doses, intervals, times]
10.896369	27	the indicated times	indicate	other NP	structuring		
3.480564	21	**of the indicated**	indicate	other passive	structuring		
8.953498	20	**as indicated by**	indicate	as+V	inferential	structuring	
9.659343	21	**was induced by**	induce	passive+PP	procedure		
7.111582	25	**the intensity of**	intensity	NP+of	description		
10.534339	56	**to interact with**	interact	V/A+to	procedure		
5.622528	64	**the interaction of**	interaction	NP+of	procedure		
9.873339	49	the interaction between	interaction	NP+other	procedure		
6.223256	26	the interaction with	interaction	NP+other	procedure		
12.012966	23	its interaction with	interaction	NP+other	procedure		
14.090449	21	**was introduced into**	introduce	passive+PP	procedure		
5.332336	26	**the introduction of**	introduction	NP+of	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
10.822141	61	**be involved in**	involve	passive+PP	inferential	causative	[is, are, [could, may, might] be, appear] [known, likely, shown, suggested, thought] to be involved in
9.210786	55	is involved in	involve	passive+PP	inferential	causative	
10.03277	52	are involved in	involve	passive+PP	inferential	causative	
15.31547	24	to be involved in	involve	passive+PP	inferential	causative	
10.721141	46	**were isolated from**	isolate	passive+PP	procedure		
10.772726	44	was isolated from	isolate	passive+PP	procedure		
6.197667	27	**the isolation of**	isolation	passive+PP	procedure		
14.069582	31	**as judged by**	judge	as+V	inferential		
15.319713	33	**is known about**	know	passive+PP	generalization		[less, little, nothing] is known [about, of, regarding]
27.912335	30	little is known about	know	passive+PP	generalization		
15.679216	32	little is known	know	other passive	generalization		
10.64249	42	**known to be**	know	V/A+to	generalization		
8.788196	32	is known to	know	V/A+to	generalization		[is, are] (previously, well) known to (be)
9.597991	30	are known to	know	V/A+to	generalization		
10.9444	31	**is not known**	know	other passive	generalization		
7.086636	58	**the lack of**	lack	NP+of	description		
6.310108	24	**of a large**	large	other PP	quantification		
9.302899	38	**on the left**	left	other PP	location		
6.821063	32	**to the left**	left	other PP	location		
6.244008	47	**the length of**	length	NP+of	quantification		
12.104791	22	**at the level of**	level	NP+of	description		
7.466129	168	**the level of**	level	NP+of	description		
11.692788	55	high levels of	level	NP+of	description		
11.963957	47	low levels of	level	NP+of	description		
5.416784	88	the levels of	level	NP+of	description		
10.032458	69	**is likely to**	likely	V/A+to	stance	inferential	(it) [is, are] (more, most) likely to (be)
12.029767	100	likely to be	likely	V/A+to	stance	inferential	
17.557042	39	is likely to be	likely	V/A+to	stance	inferential	
13.405048	78	it is likely	likely	anticipatory it	stance	inferential	
10.554759	53	are likely to	likely	V/A+to	stance	inferential	
17.929432	27	are likely to be	likely	V/A+to	stance	inferential	
19.847418	66	**it is likely that**	likely	anticipatory it	stance	inferential	it [is, seems] likely that

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
13.405048	78	it is likely	likely	anticipatory it	stance	inferential	
24.89218	21	it seems likely that	likely	anticipatory it	stance	inferential	
14.880644	36	**little or no**	little	other AP	quantification		
6.155405	89	**the localization of**	localization	NP+of	location		
9.409192	23	**is localized to**	localize	passive+pp	location		
7.400951	54	**the location of**	location	NP+of	location		
6.464483	93	**the loss of**	loss	NP+of	procedure		
6.546725	31	a loss of	loss	NP+of	procedure		
8.628752	109	**the majority of**	majority	NP+of	quantification		
9.816046	21	**were made by**	make	passive+PP	procedure		
8.199851	22	**in a manner**	manner	other PP	framing		in a manner [analogous to, similar to, that]
22.048619	36	**according to the manufacturer's**	manufacturer	other NP	procedure		
35.975142	22	according to the manufacturer's instructions	manufacturer	other NP	procedure		
17.886044	34	the manufacturer's instructions	manufacturer	other NP	procedure		
12.132765	30	**activity was measured**	measure	other passive	procedure		
10.474525	27	**as measured by**	measure	as+V	procedure		
10.751397	47	**was measured by**	measure	passive+PP	procedure		
12.038252	42	**mechanism by which**	mechanism	other NP	procedure		
5.423199	46	**the mechanism of**	mechanism	NP+of	procedure		
11.181106	32	**is mediated by**	mediate	passive+PP	procedure		
9.423692	40	**a member of**	member	NP+of	grouping		
15.446925	22	is a member of	member	NP+of	grouping		
11.020472	23	other members of	member	NP+of	grouping		
6.916711	64	**the method of**	method	NP+of	procedure		
8.371897	38	**by the method**	method	other PP	procedure		
12.981925	36	by the method of	method	PP+of	procedure		
10.769571	20	**was mixed with**	mix	passive+PP	procedure		
8.231047	35	**a mixture of**	mixture	NP+of	grouping		
8.460547	40	**a model for**	model	NP+other	framing		
9.331149	26	**model in which**	model	NP+other	framing		
8.287969	20	**as a model**	model	other PP	framing		
8.221993	24	**in this model**	model	other PP	framing		
7.5028	61	**the nature of**	nature	NP+of	description		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
11.134014	38	**is necessary for**	necessary	be+AP	engagement		
27.440109	36	**it should be noted**	note	anticipatory it	engagement	stance	it should be noted (however) (that)
33.811544	29	it should be noted that	note	anticipatory it	engagement	stance	
9.822044	24	**to note that**	note	V/A+to	engagement	stance	it is important to [acknowledge, emphasize, note, stress] (that)
17.523452	26	it is important to	note	anticipatory it	engagement	stance	
10.411291	28	**the notion that**	notion	V/N+that cl	framing		[consistent with] [confirm, support] the notion that
8.239267	183	**a number of**	number	NP+of	quantification		
16.97986	27	**a large number of**	number	NP+of	quantification		
11.306004	31	large number of	number	NP+of	quantification		
11.645072	20	**a small number**	number	other NP	quantification		
10.945364	28	small number of	number	NP+of	quantification		
10.813315	27	**in a number of**	number	PP+of	quantification		
9.996561	24	of a number of	number	PP+of	quantification		
7.14912	273	**the number of**	number	NP+of	quantification		
8.713151	20	in the number of	number	PP+of	quantification		
10.411276	22	**total number of**	number	NP+of	quantification		
14.479732	20	the total number of	number	NP+of	quantification		
9.519501	77	**the observation that**	observation	V/N+that cl	inferential		[consistent with] [supported by] the observation that
7.429082	37	**was observed in**	observe	passive+PP	inferential		[was, has (also) been] observed in
13.557917	32	has been observed	observe	passive+PP	inferential		
9.146208	27	also observed in	observe	passive+PP	inferential		
9.170894	26	been observed in	observe	passive+PP	inferential		
6.780375	21	**to that observed**	observe	other passive	comparative		[equivalent, comparable, similar] to that observed
9.00203	101	similar to that	observe	other AP	comparative		
9.628083	37	**was obtained by**	obtain	passive+PP	procedure		
9.512366	37	were obtained by	obtain	passive+PP	procedure		
11.420898	88	**were obtained from**	obtain	passive+PP	procedure		
11.00792	61	was obtained from	obtain	passive+PP	procedure		
8.220219	42	**the onset of**	onset	NP+of	procedure		
13.21763	25	**as opposed to**	oppose	as+V	comparative		
10.124116	128	**in order to**	order	others	objective		
6.924364	24	**the organization of**	organization	NP+of	procedure		
6.117887	20	**the origin of**	origin	NP+of	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
11.735682	40	**in this paper**	paper	other PP	structuring		
8.650829	21	**as part of**	part	PP+of	grouping		
5.728653	35	**the pattern of**	pattern	NP+of	procedure		
8.379589	41	**a percentage of**	percentage	NP+of	quantification		
15.849463	39	as a percentage of	percentage	PP+of	quantification		
7.504579	71	**the percentage of**	percentage	NP+of	quantification		
26.163907	28	expressed as a percentage of	percentage	passive+PP	structuring		
9.015631	24	**was performed by**	perform	passive+PP	procedure		
8.180788	43	**were performed in**	perform	passive+PP	procedure		
7.625127	27	was performed in	perform	passive+PP	procedure		
11.308832	52	was performed on	perform	passive+PP	procedure		
12.28854	45	**was performed using**	perform	other passive	procedure		
9.077895	31	was performed with	perform	passive+PP	procedure		
11.966373	39	were performed using	perform	passive+PP	procedure		
9.32991	40	were performed with	perform	passive+PP	procedure		
12.463444	34	**analysis was performed**	perform	other passive	procedure		
13.198965	43	experiments were performed	perform	other passive	procedure		
8.608501	31	**a portion of**	portion	NP+of	grouping		
6.300854	88	**the position of**	position	NP+of	location		
6.543084	69	the positions of	position	NP+of	location		
9.830042	143	**the possibility that**	possibility	V/N+that cl	stance	inferential	
10.437402	34	**possibility is that**	possible	V/N+that cl	stance	inferential	[an alternative, another, one, a second, a third] possibility is that
14.306728	165	**it is possible**	possible	anticipatory it	stance	inferential	(therefore) it [is, remains] (also) possible (that)
20.813609	146	it is possible that	possible	anticipatory it	stance	inferential	
11.418404	20	is also possible	possible	be+AP	stance	inferential	
8.869224	21	**is predicted to**	predict	V/A+to	inferential		[is, are] predicted to [be]
10.523846	24	predicted to be	predict	V/A+to	inferential		
10.009188	35	**were prepared by**	prepare	passive+PP	procedure		
9.802976	28	was prepared by	prepare	passive+PP	procedure		
10.46153	28	**was prepared from**	prepare	passive+PP	procedure		
10.896011	41	were prepared from	prepare	passive+PP	procedure		
10.52664	32	**were prepared as**	prepare	passive+PP	procedure		
14.331764	28	extracts were prepared	prepare	other passive	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
8.518913	906	**the presence of**	presence	NP+of	description		
13.109891	541	**in the presence of**	presence	PP+of	framing		(only) in the presence of
28.071689	46	**in the presence or absence of**	presence	PP+of	framing		
10.902002	45	**for the presence of**	presence	PP+of	framing		
10.448816	29	**by the presence of**	presence	PP+of	framing		
8.248847	58	**is present in**	present	other AP	description		
9.280986	24	also present in	present	other AP	description		
8.753202	44	are present in	present	other AP	description		
7.382313	29	was present in	present	other AP	description		
7.052472	25	were present in	present	other AP	description		
6.518452	112	**in the present**	present	other PP	structuring		
16.978962	86	in the present study	present	other PP	structuring		
24.136379	36	in the present study we	present	other PP	structuring		
12.172034	124	the present study	present	other NP	structuring		
10.895615	20	**were processed for**	process	passive+PP	procedure		
6.037224	41	**the process of**	process	NP+of	procedure		
9.395422	23	**in this process**	process	other PP	framing		
5.25539	33	**the product of**	product	NP+of	causative		
4.704166	28	the products of	product	NP+of	causative		
7.350548	94	**the production of**	production	NP+of	procedure		
12.095524	24	for the production of	production	PP+of	procedure		
10.523136	21	in the production of	production	PP+of	procedure		
5.616006	20	**the properties of**	property	NP+of	description		
7.26053	39	**the proportion of**	proportion	NP+of	quantification		
30.999606	22	**it has been proposed that**	propose	anticipatory it	citation		it has been proposed that
16.654869	57	has been proposed	propose	other passive	citation		
12.667997	24	been proposed that	propose	V/N+that cl	citation		
11.939201	26	**been proposed to**	propose	V/A+to	citation		[it] has been proposed to
16.654869	57	has been proposed	propose	other passive	citation		
14.097135	43	**we propose that**	propose	we+V	inferential	stance	
19.299501	33	**kindly provided by**	provide	passive+PP	acknowledgment		
13.657621	54	**were purchased from**	purchase	passive+PP	procedure		
12.87643	29	was purchased from	purchase	passive+PP	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
10.188374	36	**was purified from**	purify	passive+PP	procedure		
7.045994	20	**the question of**	question	NP+of	framing		
8.288177	48	**a range of**	range	NP+of	grouping		
21.648352	24	a wide range of	range	NP+of	grouping		
5.843365	28	**the range of**	range	NP+of	grouping		
6.836724	142	**the rate of**	rate	NP+of	quantification		
5.538115	21	the rates of	rate	NP+of	quantification		
25.848147	21	**at a flow rate of**	rate	PP+of	quantification		
7.130225	48	**the ratio of**	ratio	NP+of	quantification		
8.567567	34	**a reduction in**	reduction	NP+other	quantification		
6.397862	24	the reduction in	reduction	NP+other	quantification		
13.238171	48	**referred to as**	refer	passive+PP	structuring		
4.109719	45	**the region of**	region	NP+of	location		
6.748188	24	this region of	region	NP+of	location		
3.965135	26	**in the region**	region	other PP	location		
7.810055	32	in this region	region	other PP	location		
11.277291	53	**in the regulation of**	regulation	PP+of	procedure		
15.638835	23	**closely related to**	relate	other AP	inferential		
11.786432	43	**the relationship between**	relationship	NP+other	inferential		
5.806528	26	**the release of**	release	NP+of	procedure		
12.300193	43	**remains to be**	remain	V/A+to	objective		remains to be [determined, established, investigated]
22.059944	24	remains to be determined	remain	V/A+to	objective		
9.30839	29	to be determined	remain	other passive	objective		
8.231026	24	**the remainder of**	remainder	NP+of	grouping		
6.756487	20	**the removal of**	removal	NP+of	procedure		
10.535734	25	**were removed by**	remove	passive+PP	procedure		
11.585146	32	**was replaced with**	replace	passive+PP	procedure		
10.713625	28	**in this report**	report	other PP	structuring		
15.675107	52	**has been reported**	report	other passive	citation		[has, have] been reported (to)
14.544721	34	have been reported	report	other passive	citation		
11.435843	33	been reported to	report	V/A+to	citation		
13.602617	20	**as reported previously**	report	as+V	citation		
9.953102	28	**are representative of**	representative	be+AP	inferential		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
11.402583	194	**is required for**	require	passive+PP	framing		(that) [is, are, was, [could, may, might, will, would] be] [[appear, known, seem shown] to be] (also) required for
11.080614	83	are required for	require	passive+PP	framing		
10.937073	51	be required for	require	passive+PP	framing	stance	
9.518432	35	to be required	require	other passive	framing		
16.057948	31	to be required for	require	passive+PP	framing		
8.890332	31	was required for	require	passive+PP	framing		
10.69279	24	also required for	require	passive+PP	framing		
15.710881	20	that are required for	require	passive+PP	framing		
17.309423	44	**does not require**	require	other V fragment	framing		
10.315606	41	**not required for**	require	passive+PP	framing		(is) not required for
10.451108	40	is not required	require	passive+PP	framing		
16.701764	29	is not required for	require	passive+PP	framing		
8.212574	39	**is required to**	require	V/A+to	framing		[is, [may, will, would]] be required to
8.974112	24	be required to	require	V/A+to	framing	stance	
9.540532	73	**the requirement for**	requirement	NP+other	framing		
9.661768	25	a requirement for	requirement	NP+other	framing		
12.239797	72	**with respect to**	respect	other PP	framing		
9.46708	189	**in response to**	response	other PP	causative		
7.613068	36	**a response to**	response	NP+other	causative		
11.085759	35	**is responsible for**	responsible	be+AP	causative		[is, are, was, were, [may, might] be] (directly, largely, primarily) responsible for
11.877884	22	be responsible for	responsible	be+AP	causative	stance	
11.181308	20	are responsible for	responsible	be+AP	causative		
8.332112	21	**the rest of**	rest	NP+of	grouping		
10.21407	61	**as a result**	result	other PP	causative		
5.951958	73	**the result of**	result	NP+of	causative		[is, are, was, were] [may, might be] (likely) [a, the] result of
12.694802	24	be the result of	result	be+AP	causative	stance	[may be] [the, a] result of
4.45752	95	the results of	result	NP+of	causative		
7.451706	65	a result of	result	NP+of	causative		
14.494925	46	**as a result of**	result	PP+of	causative		
7.451706	65	a result of	result	NP+of	causative		
22.145471	38	**similar results were obtained**	result	other passive	comparative		similar results were [found, observed, obtained, seen] [with, in]
11.547796	45	similar results were	result	other NP	comparative		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
12.466673	53	results were obtained	result	other passive	comparative		
8.639799	25	were obtained with	result	passive+PP	comparative		
7.064412	20	were obtained in	result	passive+PP	comparative		
11.358561	20	**would result in**	result	other V fragment	causative	stance	
8.566044	20	**not result in**	result	other V fragment	causative		
10.329196	25	**the results presented**	result	other NP	inferential		the results presented [here, in]
8.68509	25	**the results obtained**	result	other NP	inferential		the results obtained [from, in, with]
10.018661	38	**were resuspended in**	resuspend	passive+PP	procedure		
9.208378	20	was resuspended in	resuspend	passive+PP	procedure		
6.81533	21	**to the right**	right	other PP	location		
6.491655	164	**the role of**	role	NP+of	causative		
8.106348	101	**a role in**	role	NP+other	causative		[(may) play] [a, an] (central, critical, crucial, essential, important, key, major, pivotal, significant) role in
14.747933	47	play a role	role	other V fragment	causative		
19.987636	44	play a role in	role	other V fragment	causative		
15.150291	37	an important role	role	other NP	causative	stance	
19.976137	26	an important role in	role	NP+other	causative	stance	
11.676229	26	important role in	role	NP+other	causative	stance	
14.640983	20	an essential role	role	other NP	causative	stance	
13.187698	20	a critical role	role	other NP	causative	stance	
9.351441	92	**a role for**	role	NP+other	causative		
6.929042	31	**the same as**	same	other AP	comparative		
6.625662	116	**in the same**	same	other PP	comparative		in the same [direction, manner, way]
7.935106	61	**at the same**	same	other PP	comparative		in the same [rate, time]
5.856888	48	**to the same**	same	other PP	comparative		to the same [degree, extent, region]
35.260737	47	**in the materials and methods section**	section	other PP	structuring		in the [experimental, materials and methods] section
19.843662	37	in the experimental section	section	other PP	structuring		
16.903517	137	**for review see**	see	other V fragment	structuring	engagement	for [review, reviews] see
17.047297	26	for reviews see	see	other V fragment	structuring	engagement	
10.424595	21	**see figure 1**	see	other V fragment	structuring	engagement	see figure [1,2…]
10.892676	21	see figure 2	see	other V fragment	structuring	engagement	
12.600215	28	**see table 1**	see	other V fragment	structuring	engagement	see table [1,2…]
24.610113	106	**see materials and methods**	see	other V fragment	inferential	engagement	

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
14.531152	31	**can be seen**	see	other passive	inferential	engagement	(it) can be seen
8.710781	21	**as seen in**	see	as+V	inferential	engagement	as (can be) seen in
8.35638	21	**to that seen**	see	other passive	comparative		
9.742981	20	**is sensitive to**	sensitive	be+AP	framing		
11.282193	54	**were separated by**	separate	passive+PP	procedure		
11.22859	23	**were separated on**	separate	passive+PP	procedure		
4.28971	68	**the sequence of**	sequence	NP+of	grouping		
9.646587	87	**a series of**	series	NP+of	grouping		
7.482539	29	**a set of**	set	NP+of	grouping		
13.796757	22	**there are several**	several	others	inferential	framing	there are several [aspects, mechanisms, possible explanations, reasons]
20.116642	39	**studies have shown that**	show	V/N+that cl	citation		(a) (previous, recent) [results, study, studies, work] [has, have] shown (that)
11.192163	126	have shown that	show	V/N+that cl	citation		
10.017695	39	has shown that	show	V/N+that cl	citation		
15.740882	27	previous studies have	show	other NP	citation		
12.764452	24	a previous study	show	other NP	citation		
12.197815	44	**results show that**	show	V/N+that cl	citation		
15.556469	625	**data not shown**	show	other passive	structuring		[data, results] not shown (in)
13.490686	180	results not shown	show	other passive	structuring		
16.603617	32	data not shown in	show	passive+PP	structuring		
11.443076	209	**been shown to**	show	V/A+to	citation	inferential	(it) [[has, have] been, was] (previously) shown to (be)
19.858479	104	has been shown to	show	V/A+to	citation	inferential	
14.604337	156	has been shown	show	other passive	citation	inferential	
9.70822	119	shown to be	show	V/A+to	citation	inferential	
18.790377	71	have been shown to	show	V/A+to	citation	inferential	
7.157316	51	was shown to	show	V/A+to	citation	inferential	
26.753673	38	has been shown to be	show	V/A+to	citation	inferential	
21.483034	35	it has been shown	show	anticipatory it	citation	inferential	
12.876515	23	been shown previously	show	other passive	citation	inferential	
9.07292	21	it was shown	show	anticipatory it	citation	inferential	
27.429447	21	**it has been shown that**	show	anticipatory it	citation	inferential	it [has (recently) been, was] shown (previously) that
21.483034	35	it has been shown	show	anticipatory it	citation	inferential	
9.07292	21	it was shown	show	anticipatory it	citation	inferential	

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
10.088239	20	shown previously that	show	V/N+that cl	citation	inferential	
12.36017	93	**we show that**	show	we+V	inferential	stance	(here) we [show, have shown] (previously) (that)
11.192163	126	have shown that	show	V/N+that cl	inferential		
12.032724	68	we have shown	show	we+V	inferential	stance	
18.088071	44	we have shown that	show	we+V	inferential	stance	
22.236854	26	here we show that	show	we+V	inferential	stance	
10.088239	20	shown previously that	show	V/N+that cl	inferential		
8.323654	113	**as shown in**	show	as+V	structuring		
7.764344	26	**as shown by**	show	as+V	inferential		
7.638676	21	**to show that**	show	V/A+to	objective		
8.542211	24	**are shown as**	show	passive+PP	structuring		
6.912138	20	**the significance of**	significance	NP+of	description		
9.00203	101	**similar to that**	similar	other AP	comparative		[is, are] (very) similar to [that, those] (observed, seen)
12.817143	59	similar to those	similar	other AP	comparative		
12.675705	50	similar to that of	similar	other AP	comparative		
8.005601	44	is similar to	similar	be+AP	comparative		
8.149513	26	are similar to	similar	be+AP	comparative		
11.669424	22	very similar to	similar	other AP	comparative		
7.001563	20	was similar to	similar	be+AP	comparative		
8.35638	21	to that seen	similar	other passive	comparative		
6.780375	21	to that observed	similar	other passive	comparative		
5.936251	30	**in a similar**	similar	other PP	comparative		in a similar [fashion, manner]
3.943678	48	**the site of**	site	NP+of	location		
5.63455	21	**at the site**	site	other PP	location		
6.159392	61	**the size of**	size	NP+of	quantification		
11.62512	31	**only a small**	small	other AP	quantification		
6.781516	37	**the stability of**	stability	NP+of	description		
10.674662	42	**were stained with**	stain	passive+PP	procedure		
5.540557	77	**the structure of**	structure	NP+of	description		
4.041992	25	**the study of**	study	NP+of	procedure		
10.799778	148	**in this study**	study	other PP	structuring		
17.252008	38	in this study we	study	other PP	structuring		
12.172034	124	**the present study**	study	other NP	structuring		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
11.183616	56	**were subjected to**	subject	passive+PP	procedure		
10.299332	28	was subjected to	subject	passive+PP	procedure		
10.538967	20	**is subject to**	subject	be+AP	framing		
9.976233	46	**a subset of**	subset	NP+of	grouping		
10.162955	42	**is sufficient to**	sufficient	be+AP	framing		
12.128537	92	**this suggests that**	suggest	V/N+that cl	inferential	stance	this (strongly) suggests that
13.384828	90	**results suggest that**	suggest	V/N+that cl	inferential	stance	(taken together) [these, our] [data, experiments, findings, observations, results] (strongly) suggest (that)
14.846236	74	these results suggest	suggest	other V fragment	inferential	stance	
21.407625	68	these results suggest that	suggest	V/N+that cl	inferential	stance	
13.069789	60	data suggest that	suggest	V/N+that cl	inferential	stance	
18.269175	49	taken together these	suggest	others	inferential	stance	
14.49167	48	these data suggest	suggest	other V fragment	inferential	stance	
21.016351	43	these data suggest that	suggest	V/N+that cl	inferential	stance	
14.280907	29	together these results	suggest	others	inferential	stance	
14.336707	25	together these data	suggest	others	inferential	stance	
26.509156	23	taken together these results	suggest	others	inferential	stance	
11.367919	42	**we suggest that**	suggest	we+V	inferential	stance	
11.743943	22	**have suggested that**	suggest	V/N+that cl	citation		
24.067236	25	**it has been suggested**	suggest	anticipatory it	citation		it has been suggested that
15.880416	45	has been suggested	suggest	other passive	citation		
21.39387	20	has been suggested that	suggest	V/N+that cl	citation		
10.003648	21	**suggesting that this**	suggest	other V fragment	inferential	stance	
15.054575	24	**medium supplemented with**	supplement	other passive	procedure		
11.902325	31	**is supported by**	support	passive+PP	inferential	acknowledgment	(this [work, study)] [is, was] (further) supported (in part) by
11.836482	27	was supported by	support	passive+PP	inferential	acknowledgment	
11.169955	23	this work was	support	other NP	acknowledgment		
6.992128	29	**in support of**	support	PP+of	inferential		in support of (this)
14.207395	20	in support of this	support	other PP	inferential		
6.244292	69	**the surface of**	surface	NP+of	location		
7.478716	22	**at the surface**	surface	other PP	location		
12.668212	22	**on the surface of**	surface	PP+of	location		
12.029243	20	at the surface of	surface	PP+of	location		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
8.275637	27	on the surface	surface	other PP	location		
9.551005	46	**shown in table**	table	passive+PP	structuring		[is, are] [described, given, listed, presented, shown, summarized] in table [1,2,3…]
8.557439	112	are shown in	table	passive+PP	structuring		
7.386339	93	is shown in	table	passive+PP	structuring		
14.64186	26	summarized in table	table	passive+PP	structuring		
12.056379	21	are summarized in	table	passive+PP	structuring		
16.138005	20	are shown in table	table	passive+PP	structuring		
8.775573	68	in table 1	table	other PP	structuring		
8.326116	36	in table 2	table	other PP	structuring		
8.19726	20	in table 3	table	other PP	structuring		
18.976404	146	**at room temperature**	temperature	other PP	quantification		
23.455378	31	at room temperature for	temperature	other PP	quantification		
9.450312	39	**in terms of**	term	PP+of	framing		
14.982594	21	**we tested whether**	test	we+V	procedure		
10.35839	45	**were tested for**	test	passive+PP	procedure		
29.743619	20	tested for their ability to	test	passive+PP	procedure		
13.280182	33	**to test whether**	test	V/A+to	objective		
10.731468	51	**to test this**	test	V/A+to	objective		to test this [hypothesis, idea, possibility]
21.887948	21	to test this hypothesis	test	V/A+to	objective		
11.408806	49	**is thought to**	think	V/A+to	generalization	inferential	[is, are] (generally, usually) thought to (be)
12.747926	45	thought to be	think	V/A+to	generalization	inferential	
11.826284	35	are thought to	think	V/A+to	generalization	inferential	
18.601234	22	is thought to be	think	V/A+to	generalization	inferential	
4.760724	31	**the time of**	time	NP+of	quantification		
17.713793	22	**at the same time**	time	other PP	framing	additive	at [about, approximately] the same time
10.383253	26	the same time	time	other NP	framing		
7.333901	25	**at the time**	time	other PP	framing		
15.40135	21	**at various times**	time	other PP	framing		
10.557332	20	**at this time**	time	other PP	framing		
8.611591	37	**the timing of**	timing	NP+of	description		
7.704166	32	**the tip of**	tip	NP+of	location		
5.077703	23	**the top of**	top	NP+of	location		
5.955858	27	**in the top**	top	other PP	location		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
7.591968	77	**a total of**	total	NP+of	quantification		
6.431442	22	in a total	total	other PP	quantification		
5.78726	70	**of the total**	total	other PP	quantification		
10.327546	30	**were transferred to**	transfer	passive+PP	procedure		
9.310653	20	**were treated for**	treat	passive+PP	procedure		
11.354294	90	**were treated with**	treat	passive+PP	procedure		
9.480273	27	**by treatment with**	treatment	other PP	procedure		
7.405927	20	**this type of**	type	NP+of	grouping		
10.911263	23	**two types of**	type	NP+of	grouping		
11.048687	51	**were unable to**	unable	V/A+to	description		
10.784978	23	are unable to	unable	V/A+to	description		
9.884295	21	was unable to	unable	V/A+to	description		
18.490457	26	**we were unable to**	unable	we+V	inferential	stance	we [were, have been] unable to [confirm, detect, express, identify]
10.560858	27	**is unlikely to**	unlikely	V/A+to	stance	inferential	is unlikely to (be)
11.968387	26	unlikely to be	unlikely	V/A+to	stance	inferential	
13.586645	24	**it is unlikely**	unlikely	anticipatory it	stance	inferential	it [seems, is] unlikely (that)
9.272937	32	**for up to**	up	other PP	quantification		
7.000252	22	**in the upper**	upper	other PP	location		
9.467843	81	**by use of**	use	PP+of	procedure		
11.884964	29	**with the use of**	use	PP+of	procedure		
7.246018	112	**the use of**	use	NP+of	procedure		
9.596848	190	**was used to**	use	passive+PP	procedure		
8.652743	107	were used to	use	passive+PP	procedure		
8.564957	37	be used to	use	other passive	procedure		
8.862781	24	been used to	use	other passive	procedure		
12.111424	28	can be used	use	other passive	procedure		
12.320468	22	has been used	use	other passive	procedure		
13.61486	22	used to amplify	use	other passive	procedure		
11.371248	28	used to determine	use	other passive	procedure		
12.265526	22	used to identify	use	other passive	procedure		
10.05184	80	**was used as**	use	passive+PP	procedure		
8.971748	41	were used as	use	passive+PP	procedure		
12.111424	28	can be used	use	other passive	procedure		

MI	Freq	Bundle	Keyword	Structure	Function 1	Function 2	Variations
12.320468	22	has been used	use	other passive	procedure		
8.68385	55	**was used for**	use	passive+PP	procedure		
8.246206	44	were used for	use	passive+PP	procedure		
12.111424	28	can be used	use	other passive	procedure		
12.320468	22	has been used	use	other passive	procedure		
7.021741	49	**were used in**	use	passive+PP	procedure		
5.982179	22	was used in	use	passive+PP	procedure		
12.111424	28	can be used	use	other passive	procedure		
12.320468	22	has been used	use	other passive	procedure		
11.187802	26	we have used	use	we+V	procedure		
5.858367	25	**the value of**	value	NP+of	quantification		
10.289522	111	**a variety of**	variety	NP+of	grouping		
13.965816	20	**in the vicinity of**	vicinity	PP+of	location		
11.51919	22	**by virtue of**	virtue	PP+of	causative		
22.836459	22	**an equal volume of**	volume	NP+of	quantification		
18.148428	22	an equal volume	volume	other NP	quantification		
14.53655	22	equal volume of	volume	NP+of	quantification		
12.48135	21	total volume of	volume	NP+of	quantification		
7.805379	20	**were washed in**	wash	passive+PP	procedure		
9.260472	23	**were washed with**	wash	passive+PP	procedure		
24.713764	22	were washed three times	wash	other passive	procedure		
22.798241	32	were washed twice with	wash	passive+PP	procedure		
14.240235	307	**as well as**	well	others	additive		as well as (in)
15.772432	22	as well as in	well	others	additive		
10.911499	20	**the present work**	work	other NP	structuring		
6.572063	23	**the yield of**	yield	NP+of	causative		

LEGEND: Prototypical bundles are in bold and highlighted in gray. NP + of – Noun phrase + *of*-phrase fragment; NP + other – Noun phrase with other post-modifier fragment; other NP – Other noun phrase; passive + PP – Passive + prepositional-phrase fragment; other passive – Other passive fragment; *we* + V – Verb phrase with personal pronoun *we*; other V fragment – Other verbal fragment; PP + *of* – Prepositional phrase + *of;* other PP – Other prepositional phrase (fragment); V/A + *to* – Verb or adjective + *to*-clause fragment; V/N + *that*-cl – Verb phrase or noun phrase + *that*-clause fragment; as + V – Adverbial clause fragment; *be* + AP – Copula *be* + adjective phrase; other AP – Other adjectival phrase; anticipatory it - Anticipatory *it* + verb or adjectival phrase; Others – Other expression

Examples provided in digital version on CD

Appendix 5

List of prototypical target bundles

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
237	11.38	4	3.31	the ability of	ability	NP+of	description	
83	3.99	4	3.31	the ability to	ability	other NP	description	
41	1.97	3	2.49	is able to	able	V/A+to	description	
48	2.31	21	17.40	were able to	able	V/A+to	inferential	
387	18.58	22	18.22	in the absence of	absence	PP+of	framing	
60	2.88	0	0.00	in the absence or presence of	absence	PP+of	framing	
33	1.58	3	2.49	in accordance with	accordance	other PP	framing	citation
25	1.20	3	2.49	to account for	account	V/A+to	objective	inferential
80	3.84	9	7.46	the accumulation of	accumulation	NP+of	procedure	
24	1.15	0	0.00	to act as	act	V/A+to	description	
47	2.26	4	3.31	the action of	action	NP+of	procedure	
145	6.96	11	9.11	the activity of	activity	NP+of	procedure	
144	6.92	9	7.46	was added to	add	passive+PP	procedure	
84	4.03	4	3.31	by the addition of	addition	PP+of	procedure	
154	7.40	6	4.97	in addition to	addition	other PP	additive	
38	1.82	0	0.00	for an additional	additional	other PP	quantification	
22	1.06	0	0.00	to address this	address	V/A+to	objective	
45	2.16	0	0.00	did not affect	affect	other V fragment	causative	
35	1.68	22	18.22	in agreement with	agreement	other PP	comparative	citation
21	1.01	1	0.83	were allowed to	allow	other passive	procedure	
20	0.96	2	1.66	in the amount of	amount	PP+of	quantification	
32	1.54	11	9.11	the analysis of	analysis	NP+of	procedure	
44	2.11	6	4.97	were analyzed by	analyze	passive+PP	procedure	
25	1.20	1	0.83	it appears that	appear	anticipatory it	inferential	stance

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
53	2.55	1	0.83	appear to be	appear	V/A+to	inferential	stance
36	1.73	0	0.00	not appear to	appear	V/A+to	inferential	stance
39	1.87	2	1.66	the appearance of	appearance	NP+of	description	
22	1.06	0	0.00	we asked whether	ask	we+V	objective	
38	1.82	1	0.83	the assembly of	assembly	NP+of	procedure	
32	1.54	0	0.00	was assessed by	assess	passive+PP	procedure	
61	2.93	3	2.49	is associated with	associate	passive+PP	inferential	stance
24	1.15	0	0.00	to associate with	associate	V/A+to	inferential	stance
22	1.06	0	0.00	the association of	association	NP+of	inferential	stance
28	1.34	0	0.00	an average of	average	NP+of	quantification	
27	1.30	3	2.49	is based on	base	passive+PP	framing	
129	6.19	12	9.94	on the basis of	basis	PP+of	framing	
20	0.96	11	9.11	the beginning of	beginning	NP+of	procedure	
27	1.30	3	2.49	the behavior of	behavior	NP+of	description	
26	1.25	0	0.00	in the bottom	bottom	other PP	location	
33	1.58	0	0.00	the bottom of	bottom	NP+of	location	
20	0.96	0	0.00	is capable of	capable	be+AP	description	
26	1.25	11	9.11	carried out at	carry out	passive+PP	procedure	
28	1.34	11	9.11	carried out in	carry out	passive+PP	procedure	
31	1.49	17	14.08	carried out with	carry out	passive+PP	procedure	
118	5.67	29	24.02	were carried out	carry out	other passive	procedure	
90	4.32	14	11.60	in the case of	case	PP+of	framing	
61	2.93	12	9.94	in this case	case	other PP	framing	
40	1.92	9	7.46	in all cases	case	other PP	framing	
36	1.73	0	0.00	in each case	case	other PP	framing	
29	1.39	7	5.80	in some cases	case	other PP	framing	
23	1.10	1	0.83	is caused by	cause	passive+PP	causative	
27	1.30	3	2.49	a change in	change	NP+other	procedure	
26	1.25	0	0.00	it is clear	clear	anticipatory it	stance	
29	1.39	1	0.83	it is not clear	clear	anticipatory it	stance	
23	1.10	7	5.80	were collected from	collect	passive+PP	procedure	
26	1.25	6	4.97	a combination of	combination	NP+of	grouping	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
55	2.64	0	0.00	in combination with	combination	other PP	additive	framing
20	0.96	0	0.00	compared with control	compare	passive+PP	procedure	
39	1.87	1	0.83	as compared with	compare	as+V	comparative	
22	1.06	0	0.00	a comparison of	comparison	NP+of	procedure	
26	1.25	8	6.63	in comparison with	comparison	other PP	comparative	
36	1.73	0	0.00	a component of	component	NP+of	grouping	
20	0.96	0	0.00	in concert with	concert	other PP	additive	framing
70	3.36	1	0.83	we conclude that	conclude	we+V	inferential	stance
27	1.30	0	0.00	the conclusion that	conclusion	V/N+that cl	inferential	
40	1.92	1	0.83	under these conditions	condition	other PP	framing	
22	1.06	1	0.83	under the same conditions	condition	other PP	framing	
52	2.50	2	1.66	was confirmed by	confirm	passive+PP	procedure	
44	2.11	0	0.00	to confirm that	confirm	V/A+to	objective	
30	1.44	0	0.00	in conjunction with	conjunction	other PP	additive	framing
26	1.25	3	2.49	as a consequence of	consequence	PP+of	causative	
154	7.40	3	2.49	is consistent with	consistent	be+AP	comparative	citation
93	4.47	4	3.31	are consistent with	consistent	be+AP	comparative	citation
20	0.96	0	0.00	be consistent with	consistent	be+AP	comparative	citation
27	1.30	0	0.00	does not contain	contain	other V fragment	description	
35	1.68	2	1.66	in the context of	context	PP+of	framing	
105	5.04	2	1.66	in contrast to	contrast	other PP	comparative	
32	1.54	1	0.83	may contribute to	contribute	other V fragment	causative	stance
72	3.46	2	1.66	the control of	control	NP+of	procedure	
47	2.26	0	0.00	as a control	control	other PP	procedure	
26	1.25	4	3.31	in the control	control	other PP	procedure	
45	2.16	0	0.00	under the control of	control	other PP	framing	
31	1.49	2	1.66	the course of	course	NP+of	framing	
28	1.34	7	5.80	in the dark	dark	other PP	location	
43	2.06	7	5.80	a decrease in	decrease	NP+other	quantification	
50	2.40	0	0.00	a defect in	defect	NP+other	description	
47	2.26	6	4.97	the degree of	degree	NP+of	description	
24	1.15	0	0.00	a deletion of	deletion	NP+of	procedure	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
27	1.30	0	0.00	results demonstrate that	demonstrate	V/N+that cl	inferential	
38	1.82	0	0.00	we demonstrate that	demonstrate	we+V	inferential	stance
24	1.15	2	1.66	has been demonstrated	demonstrate	other passive	citation	inferential
27	1.30	0	0.00	to demonstrate that	demonstrate	V/A+to	objective	
25	1.20	0	0.00	at a density of	density	PP+of	quantification	
53	2.55	2	1.66	is dependent on	dependent	be+AP	framing	
244	11.72	4	3.31	as described previously	describe	as+V	structuring	
75	3.60	5	4.14	as described by	describe	as+V	citation	
36	1.73	0	0.00	been described previously	describe	other passive	structuring	citation
22	1.06	0	0.00	are described in	describe	passive+PP	structuring	citation
20	0.96	0	0.00	as described for	describe	as+V	structuring	citation
43	2.06	2	1.66	was detected by	detect	passive+PP	procedure	
47	2.26	6	4.97	was detected in	detect	passive+PP	inferential	
24	1.15	3	2.49	was not detected	detect	other passive	inferential	
27	1.30	4	3.31	the detection of	detection	NP+of	procedure	
22	1.06	1	0.83	was determined as	determine	passive+PP	procedure	
119	5.71	9	7.46	was determined by	determine	passive+PP	procedure	
52	2.50	5	4.14	as determined by	determine	as+V	inferential	
164	7.88	1	0.83	to determine whether	determine	V/A+to	objective	
63	3.03	8	6.63	the development of	development	NP+of	procedure	
45	2.16	7	5.80	the difference in	difference	NP+other	comparative	
25	1.20	1	0.83	the difference between	difference	NP+other	comparative	
28	1.34	2	1.66	significantly different from	different	other AP	comparative	
23	1.10	3	2.49	is difficult to	difficult	be+AP	stance	
36	1.73	3	2.49	was digested with	digest	passive+PP	procedure	
20	0.96	3	2.49	was dissolved in	dissolve	passive+PP	procedure	
20	0.96	0	0.00	to distinguish between	distinguish	V/A+to	objective	
70	3.36	5	4.14	the distribution of	distribution	NP+of	grouping	
36	1.73	3	2.49	is due to	due to	be+AP	causative	
27	1.30	0	0.00	not due to	due to	other AP	causative	stance
105	5.04	1	0.83	no effect on	effect	NP+other	causative	
259	12.44	37	30.65	the effect of	effect	NP+of	causative	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
27	1.30	6	4.97	the efficiency of	efficiency	NP+of	quantification	
34	1.63	5	4.14	at the end of	end	PP+of	location	
23	1.10	1	0.83	to ensure that	ensure	V/A+to	objective	
71	3.41	2	1.66	is essential for	essential	be+AP	stance	
33	1.58	1	0.83	lines of evidence	evidence	other NP	inferential	
22	1.06	0	0.00	no evidence for	evidence	NP+other	inferential	
38	1.82	1	0.83	the evolution of	evolution	NP+of	procedure	
21	1.01	1	0.83	was examined by	examine	passive+PP	procedure	
30	1.44	6	4.97	with the exception of	exception	PP+of	framing	
21	1.01	0	0.00	exclude the possibility	exclude	other V fragment	inferential	engagement
61	2.93	8	6.63	the existence of	existence	NP+of	description	
34	1.63	1	0.83	expected to be	expect	V/A+to	inferential	stance
49	2.35	2	1.66	in these experiments	experiment	other PP	structuring	
22	1.06	7	5.80	be explained by	explain	passive+PP	causative	inferential
26	1.25	1	0.83	were exposed to	expose	passive+PP	procedure	
44	2.11	6	4.97	are expressed as	express	passive+PP	structuring	
57	2.74	6	4.97	the extent of	extent	NP+of	description	
158	7.59	21	17.40	the fact that	fact	V/N+that cl	framing	
25	1.20	0	0.00	a family of	family	NP+of	grouping	
70	3.36	1	0.83	as shown in figure	figure	as+V	structuring	
36	1.73	2	1.66	is shown in figure	figure	passive+PP	structuring	
21	1.01	0	0.00	as in figure	figure	other PP	structuring	
83	3.99	5	4.14	found to be	find	V/A+to	inferential	citation
32	1.54	4	3.31	is found in	find	passive+PP	generalization	inferential
28	1.34	13	10.77	was found in	find	passive+PP	inferential	
130	6.24	2	1.66	we found that	find	we+V	inferential	stance
34	1.63	0	0.00	the finding that	finding	V/N+that cl	inferential	
32	1.54	2	1.66	were fixed in	fix	passive+PP	procedure	
21	1.01	2	1.66	were as follows	follow	as+V	structuring	
20	0.96	2	1.66	with the following	following	other PP	structuring	
149	7.16	10	8.28	the formation of	formation	NP+of	procedure	
29	1.39	1	0.83	a fraction of	fraction	NP+of	quantification	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
40	1.92	1	0.83	the fraction of	fraction	NP+of	quantification	
43	2.06	9	7.46	the frequency of	frequency	NP+of	quantification	
90	4.32	0	0.00	the function of	function	NP+of	description	
50	2.40	7	5.80	as a function of	function	PP+of	framing	
45	2.16	4	3.31	were generated by	generate	passive+PP	procedure	
35	1.68	13	10.77	the generation of	generation	NP+of	procedure	
41	1.97	1	0.83	a gift from	gift	NP+other	acknowledgment	
45	2.16	2	1.66	were grown at	grow	passive+PP	procedure	
72	3.46	3	2.49	were grown in	grow	passive+PP	procedure	
34	1.63	0	0.00	were grown to	grow	passive+PP	procedure	
20	0.96	8	6.63	the growth of	growth	NP+of	procedure	
51	2.45	30	24.85	on the other hand	hand	other PP	comparative	additive
67	3.22	4	3.31	the hypothesis that	hypothesis	V/N+that cl	inferential	
38	1.82	0	0.00	on ice for	ice	other PP	procedure	
38	1.82	6	4.97	the idea that	idea	V/N+that cl	framing	
51	2.45	6	4.97	the identification of	identification	NP+of	procedure	
31	1.49	0	0.00	were identified by	identify	passive+PP	procedure	
21	1.01	0	0.00	have been identified in	identify	passive+PP	citation	inferential
23	1.10	0	0.00	been identified as	identify	passive+PP	citation	inferential
28	1.34	1	0.83	we have identified	identify	we+V	inferential	stance
33	1.58	0	0.00	the identity of	identity	NP+of	description	
47	2.26	4	3.31	been implicated in	implicate	passive+PP	citation	inferential
21	1.01	2	1.66	this implies that	imply	V/N+that cl	inferential	
42	2.02	9	7.46	the importance of	importance	NP+of	description	
34	1.63	0	0.00	is important for	important	be+AP	stance	
29	1.39	5	4.14	is an important	important	be+AP	stance	
23	1.10	1	0.83	the inability of	inability	NP+of	description	
20	0.96	2	1.66	the incorporation of	incorporation	NP+of	procedure	
112	5.38	19	15.74	an increase in	increase	NP+other	quantification	
25	1.20	0	0.00	increasing amounts of	increase	NP+of	quantification	
70	3.36	5	4.14	were incubated for	incubate	passive+PP	procedure	
136	6.53	2	1.66	were incubated with	incubate	passive+PP	procedure	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
37	1.78	1	0.83	is independent of	independent	be+AP	framing	
23	1.10	1	0.83	this indicates that	indicate	V/N+that cl	inferential	
52	2.50	4	3.31	results indicate that	indicate	V/N+that cl	inferential	
47	2.26	0	0.00	is indicated by	indicate	passive+PP	structuring	
26	1.25	0	0.00	are indicated in	indicate	passive+PP	structuring	
38	1.82	1	0.83	at the indicated	indicate	other passive	structuring	
21	1.01	0	0.00	of the indicated	indicate	other passive	structuring	
20	0.96	1	0.83	as indicated by	indicate	as+V	inferential	structuring
21	1.01	1	0.83	was induced by	induce	passive+PP	procedure	
25	1.20	1	0.83	the intensity of	intensity	NP+of	description	
56	2.69	0	0.00	to interact with	interact	V/A+to	procedure	
64	3.07	6	4.97	the interaction of	interaction	NP+of	procedure	
21	1.01	0	0.00	was introduced into	introduce	passive+PP	procedure	
26	1.25	0	0.00	the introduction of	introduction	NP+of	procedure	
61	2.93	4	3.31	be involved in	involve	passive+PP	inferential	causative
46	2.21	5	4.14	were isolated from	isolate	passive+PP	procedure	
27	1.30	0	0.00	the isolation of	isolation	passive+PP	procedure	
31	1.49	1	0.83	as judged by	judge	as+V	inferential	
33	1.58	1	0.83	is known about	know	passive+PP	generalization	
31	1.49	3	2.49	is not known	know	other passive	generalization	
58	2.79	6	4.97	the lack of	lack	NP+of	description	
24	1.15	0	0.00	of a large	large	other PP	quantification	
38	1.82	0	0.00	on the left	left	other PP	location	
32	1.54	0	0.00	to the left	left	other PP	location	
47	2.26	0	0.00	the length of	length	NP+of	quantification	
22	1.06	0	0.00	at the level of	level	NP+of	description	
168	8.07	7	5.80	the level of	level	NP+of	description	
69	3.31	1	0.83	is likely to	likely	V/A+to	stance	inferential
66	3.17	1	0.83	it is likely that	likely	anticipatory it	stance	inferential
36	1.73	0	0.00	little or no	little	other AP	quantification	
89	4.27	0	0.00	the localization of	localization	NP+of	location	
23	1.10	0	0.00	is localized to	localize	passive+pp	location	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
54	2.59	1	0.83	the location of	location	NP+of	location	
93	4.47	3	2.49	the loss of	loss	NP+of	procedure	
109	5.23	11	9.11	the majority of	majority	NP+of	quantification	
21	1.01	2	1.66	were made by	make	passive+PP	procedure	
22	1.06	0	0.00	in a manner	manner	other PP	framing	
36	1.73	0	0.00	according to the manufacturer's	manufacturer	other NP	procedure	
30	1.44	6	4.97	activity was measured	measure	other passive	procedure	
27	1.30	0	0.00	as measured by	measure	as+V	procedure	
47	2.26	5	4.14	was measured by	measure	passive+PP	procedure	
42	2.02	1	0.83	mechanism by which	mechanism	other NP	procedure	
46	2.21	3	2.49	the mechanism of	mechanism	NP+of	procedure	
32	1.54	0	0.00	is mediated by	mediate	passive+PP	procedure	
40	1.92	1	0.83	a member of	member	NP+of	grouping	
64	3.07	14	11.60	the method of	method	NP+of	procedure	
38	1.82	14	11.60	by the method	method	other PP	procedure	
20	0.96	1	0.83	was mixed with	mix	passive+PP	procedure	
35	1.68	12	9.94	a mixture of	mixture	NP+of	grouping	
40	1.92	0	0.00	a model for	model	NP+other	framing	
26	1.25	1	0.83	model in which	model	NP+other	framing	
20	0.96	0	0.00	as a model	model	other PP	framing	
24	1.15	0	0.00	in this model	model	other PP	framing	
61	2.93	2	1.66	the nature of	nature	NP+of	description	
38	1.82	0	0.00	is necessary for	necessary	be+AP	stance	
36	1.73	1	0.83	it should be noted	note	anticipatory it	engagement	stance
24	1.15	2	1.66	to note that	note	V/A+to	engagement	stance
28	1.34	0	0.00	the notion that	notion	V/N+that cl	framing	
27	1.30	5	4.14	a large number of	number	NP+of	quantification	
20	0.96	1	0.83	a small number	number	other NP	quantification	
27	1.30	2	1.66	in a number of	number	PP+of	quantification	
273	13.11	53	43.90	the number of	number	NP+of	quantification	
22	1.06	12	9.94	total number of	number	NP+of	quantification	
77	3.70	0	0.00	the observation that	observation	V/N+that cl	inferential	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
37	1.78	11	9.11	was observed in	observe	passive+PP	inferential	
21	1.01	0	0.00	to that observed	observe	other passive	comparative	
37	1.78	5	4.14	was obtained by	obtain	passive+PP	procedure	
88	4.23	13	10.77	were obtained from	obtain	passive+PP	procedure	
42	2.02	0	0.00	the onset of	onset	NP+of	procedure	
25	1.20	0	0.00	as opposed to	oppose	as+V	comparative	
128	6.15	54	44.73	in order to	order	others	objective	
24	1.15	0	0.00	the organization of	organization	NP+of	procedure	
20	0.96	0	0.00	the origin of	origin	NP+of	procedure	
40	1.92	6	4.97	in this paper	paper	other PP	structuring	
21	1.01	2	1.66	as part of	part	PP+of	grouping	
35	1.68	2	1.66	the pattern of	pattern	NP+of	procedure	
41	1.97	3	2.49	a percentage of	percentage	NP+of	quantification	
71	3.41	10	8.28	the percentage of	percentage	NP+of	quantification	
24	1.15	5	4.14	was performed by	perform	passive+PP	procedure	
43	2.06	3	2.49	were performed in	perform	passive+PP	procedure	
45	2.16	1	0.83	was performed using	perform	other passive	procedure	
34	1.63	5	4.14	analysis was performed	perform	other passive	procedure	
31	1.49	0	0.00	a portion of	portion	NP+of	grouping	
88	4.23	2	1.66	the position of	position	NP+of	location	
143	6.87	1	0.83	the possibility that	possibility	V/N+that cl	stance	inferential
34	1.63	0	0.00	possibility is that	possible	V/N+that cl	stance	inferential
165	7.92	8	6.63	it is possible	possible	anticipatory it	stance	inferential
21	1.01	0	0.00	is predicted to	predict	V/A+to	inferential	
35	1.68	0	0.00	were prepared by	prepare	passive+PP	procedure	
28	1.34	0	0.00	was prepared from	prepare	passive+PP	procedure	
32	1.54	0	0.00	were prepared as	prepare	passive+PP	procedure	
541	25.98	67	55.50	in the presence of	presence	PP+of	framing	
46	2.21	0	0.00	in the presence or absence of	presence	PP+of	framing	
45	2.16	2	1.66	for the presence of	presence	PP+of	framing	
29	1.39	9	7.46	by the presence of	presence	PP+of	framing	
58	2.79	1	0.83	is present in	present	other AP	description	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
112	5.38	18	14.91	in the present	present	other PP	structuring	
20	0.96	0	0.00	were processed for	process	passive+PP	procedure	
41	1.97	5	4.14	the process of	process	NP+of	procedure	
23	1.10	0	0.00	in this process	process	other PP	framing	
33	1.58	2	1.66	the product of	product	NP+of	causative	
94	4.51	14	11.60	the production of	production	NP+of	procedure	
20	0.96	1	0.83	the properties of	property	NP+of	description	
39	1.87	4	3.31	the proportion of	proportion	NP+of	quantification	
22	1.06	1	0.83	it has been proposed that	propose	anticipatory it	citation	
26	1.25	1	0.83	been proposed to	propose	V/A+to	citation	
43	2.06	1	0.83	we propose that	propose	we+V	inferential	stance
33	1.58	0	0.00	kindly provided by	provide	passive+PP	acknowledgment	
54	2.59	7	5.80	were purchased from	purchase	passive+PP	procedure	
36	1.73	0	0.00	was purified from	purify	passive+PP	procedure	
20	0.96	0	0.00	the question of	question	NP+of	framing	
48	2.31	5	4.14	a range of	range	NP+of	grouping	
28	1.34	8	6.63	the range of	range	NP+of	grouping	
142	6.82	12	9.94	the rate of	rate	NP+of	quantification	
21	1.01	6	4.97	at a flow rate of	rate	PP+of	quantification	
48	2.31	7	5.80	the ratio of	ratio	NP+of	quantification	
34	1.63	4	3.31	a reduction in	reduction	NP+other	quantification	
48	2.31	1	0.83	referred to as	refer	passive+PP	structuring	
45	2.16	1	0.83	the region of	region	NP+of	location	
26	1.25	3	2.49	in the region	region	other PP	location	
53	2.55	1	0.83	in the regulation of	regulation	PP+of	procedure	
23	1.10	1	0.83	closely related to	relate	other AP	inferential	
43	2.06	4	3.31	the relationship between	relationship	NP+other	inferential	
26	1.25	1	0.83	the release of	release	NP+of	procedure	
43	2.06	0	0.00	remains to be	remain	V/A+to	objective	
24	1.15	1	0.83	the remainder of	remainder	NP+of	grouping	
20	0.96	1	0.83	the removal of	removal	NP+of	procedure	
25	1.20	0	0.00	were removed by	remove	passive+PP	procedure	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
32	1.54	1	0.83	was replaced with	replace	passive+PP	procedure	
28	1.34	0	0.00	in this report	report	other PP	structuring	
52	2.50	12	9.94	has been reported	report	other passive	citation	
20	0.96	0	0.00	as reported previously	report	as+V	citation	
28	1.34	0	0.00	are representative of	representative	be+AP	inferential	
194	9.32	2	1.66	is required for	require	passive+PP	framing	
44	2.11	0	0.00	does not require	require	other V fragment	framing	
41	1.97	0	0.00	not required for	require	passive+PP	framing	
39	1.87	0	0.00	is required to	require	V/A+to	framing	
73	3.51	0	0.00	the requirement for	requirement	NP+other	framing	
72	3.46	24	19.88	with respect to	respect	other PP	framing	
189	9.08	19	15.74	in response to	response	other PP	causative	
36	1.73	1	0.83	a response to	response	NP+other	causative	
35	1.68	1	0.83	is responsible for	responsible	be+AP	causative	
21	1.01	3	2.49	the rest of	rest	NP+of	grouping	
73	3.51	7	5.80	the result of	result	NP+of	causative	
46	2.21	3	2.49	as a result of	result	PP+of	causative	
38	1.82	4	3.31	similar results were obtained	result	other passive	comparative	
20	0.96	0	0.00	would result in	result	other V fragment	causative	stance
20	0.96	2	1.66	not result in	result	other V fragment	causative	
25	1.20	1	0.83	the results presented	result	other NP	inferential	
25	1.20	11	9.11	the results obtained	result	other NP	inferential	
21	1.01	0	0.00	to the right	right	other PP	location	
164	7.88	9	7.46	the role of	role	NP+of	causative	
101	4.85	2	1.66	a role in	role	NP+other	causative	
92	4.42	0	0.00	a role for	role	NP+other	causative	
31	1.49	2	1.66	the same as	same	other AP	comparative	
116	5.57	15	12.43	in the same	same	other PP	comparative	
61	2.93	16	13.25	at the same	same	other PP	comparative	
48	2.31	6	4.97	to the same	same	other PP	comparative	
47	2.26	0	0.00	in the materials and methods section	section	other PP	structuring	
137	6.58	0	0.00	for review see	see	other V fragment	structuring	engagement

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
21	1.01	0	0.00	see figure 1	see	other V fragment	structuring	engagement
28	1.34	0	0.00	see table 1	see	other V fragment	structuring	engagement
106	5.09	1	0.83	see materials and methods	see	other V fragment	structuring	engagement
31	1.49	8	6.63	can be seen	see	other passive	inferential	engagement
21	1.01	1	0.83	as seen in	see	as+V	inferential	engagement
21	1.01	0	0.00	to that seen	see	other passive	comparative	
20	0.96	1	0.83	is sensitive to	sensitive	be+AP	framing	
54	2.59	5	4.14	were separated by	separate	passive+PP	procedure	
23	1.10	1	0.83	were separated on	separate	passive+PP	procedure	
68	3.27	3	2.49	the sequence of	sequence	NP+of	grouping	
87	4.18	5	4.14	a series of	series	NP+of	grouping	
29	1.39	1	0.83	a set of	set	NP+of	grouping	
22	1.06	1	0.83	there are several	several	others	inferential	framing
39	1.87	1	0.83	studies have shown that	show	V/N+that cl	citation	
44	2.11	4	3.31	results show that	show	V/N+that cl	citation	
625	30.01	20	16.57	data not shown	show	other passive	structuring	
209	10.04	5	4.14	been shown to	show	V/A+to	citation	inferential
21	1.01	1	0.83	it has been shown that	show	anticipatory it	citation	inferential
93	4.47	3	2.49	we show that	show	we+V	inferential	stance
113	5.43	7	5.80	as shown in	show	as+V	structuring	
26	1.25	0	0.00	as shown by	show	as+V	inferential	
21	1.01	1	0.83	to show that	show	V/A+to	objective	
24	1.15	1	0.83	are shown as	show	passive+PP	structuring	
20	0.96	1	0.83	the significance of	significance	NP+of	description	
101	4.85	9	7.46	similar to that	similar	other AP	comparative	
30	1.44	4	3.31	in a similar	similar	other PP	comparative	
48	2.31	3	2.49	the site of	site	NP+of	location	
21	1.01	2	1.66	at the site	site	other PP	location	
61	2.93	5	4.14	the size of	size	NP+of	quantification	
31	1.49	0	0.00	only a small	small	other AP	quantification	
37	1.78	2	1.66	the stability of	stability	NP+of	description	
42	2.02	3	2.49	were stained with	stain	passive+PP	procedure	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
77	3.70	7	5.80	the structure of	structure	NP+of	description	
25	1.20	6	4.97	the study of	study	NP+of	procedure	
148	7.11	24	19.88	in this study	study	other PP	structuring	
124	5.95	20	16.57	the present study	study	other NP	structuring	
56	2.69	2	1.66	were subjected to	subject	passive+PP	procedure	
20	0.96	1	0.83	is subject to	subject	be+AP	framing	
46	2.21	1	0.83	a subset of	subset	NP+of	grouping	
42	2.02	0	0.00	is sufficient to	sufficient	be+AP	framing	
92	4.42	1	0.83	this suggests that	suggest	V/N+that cl	inferential	stance
90	4.32	4	3.31	results suggest that	suggest	V/N+that cl	inferential	stance
42	2.02	1	0.83	we suggest that	suggest	we+V	inferential	stance
22	1.06	1	0.83	have suggested that	suggest	V/N+that cl	citation	
25	1.20	1	0.83	it has been suggested	suggest	anticipatory it	citation	
21	1.01	1	0.83	suggesting that this	suggest	other V fragment	inferential	stance
24	1.15	0	0.00	medium supplemented with	supplement	other passive	procedure	
31	1.49	1	0.83	is supported by	support	passive+PP	inferential	acknowledgment
29	1.39	2	1.66	in support of	support	PP+of	inferential	
22	1.06	2	1.66	at the surface	surface	other PP	location	
22	1.06	3	2.49	on the surface of	surface	PP+of	location	
46	2.21	12	9.94	shown in table	table	passive+PP	structuring	
146	7.01	8	6.63	at room temperature	temperature	other PP	quantification	
39	1.87	6	4.97	in terms of	term	PP+of	framing	
21	1.01	0	0.00	we tested whether	test	we+V	procedure	
45	2.16	1	0.83	were tested for	test	passive+PP	procedure	
33	1.58	0	0.00	to test whether	test	V/A+to	objective	
51	2.45	0	0.00	to test this	test	V/A+to	objective	
49	2.35	0	0.00	is thought to	think	V/A+to	generalization	inferential
31	1.49	3	2.49	the time of	time	NP+of	quantification	
22	1.06	6	4.97	at the same time	time	other PP	framing	additive
25	1.20	1	0.83	at the time	time	other PP	framing	
21	1.01	1	0.83	at various times	time	other PP	framing	
20	0.96	1	0.83	at this time	time	other PP	framing	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
37	1.78	1	0.83	the timing of	timing	NP+of	description	
32	1.54	2	1.66	the tip of	tip	NP+of	location	
23	1.10	1	0.83	the top of	top	NP+of	location	
27	1.30	0	0.00	in the top	top	other PP	location	
77	3.70	6	4.97	a total of	total	NP+of	quantification	
70	3.36	10	8.28	of the total	total	other PP	quantification	
30	1.44	5	4.14	were transferred to	transfer	passive+PP	procedure	
20	0.96	0	0.00	were treated for	treat	passive+PP	procedure	
90	4.32	4	3.31	were treated with	treat	passive+PP	procedure	
27	1.30	0	0.00	by treatment with	treatment	other PP	procedure	
20	0.96	1	0.83	this type of	type	NP+of	grouping	
23	1.10	1	0.83	two types of	type	NP+of	grouping	
51	2.45	4	3.31	were unable to	unable	V/A+to	description	
26	1.25	1	0.83	we were unable to	unable	we+V	inferential	stance
24	1.15	0	0.00	it is unlikely	unlikely	anticipatory it	stance	inferential
32	1.54	2	1.66	for up to	up	other PP	quantification	
22	1.06	2	1.66	in the upper	upper	other PP	location	
81	3.89	0	0.00	by use of	use	PP+of	procedure	
29	1.39	0	0.00	with the use of	use	PP+of	procedure	
112	5.38	12	9.94	the use of	use	NP+of	procedure	
190	9.12	24	19.88	was used to	use	passive+PP	procedure	
80	3.84	10	8.28	was used as	use	passive+PP	procedure	
55	2.64	8	6.63	was used for	use	passive+PP	procedure	
49	2.35	4	3.31	were used in	use	passive+PP	procedure	
25	1.20	3	2.49	the value of	value	NP+of	quantification	
111	5.33	11	9.11	a variety of	variety	NP+of	grouping	
20	0.96	0	0.00	in the vicinity of	vicinity	PP+of	location	
22	1.06	0	0.00	by virtue of	virtue	PP+of	causative	
22	1.06	0	0.00	an equal volume of	volume	NP+of	quantification	
20	0.96	0	0.00	were washed in	wash	passive+PP	procedure	
23	1.10	1	0.83	were washed with	wash	passive+PP	procedure	
307	14.74	44	36.45	as well as	well	others	additive	

HSC Raw	HSC Norm	NNS Raw	NNS Norm	Bundle	Keyword	Structure	Function 1	Function 2
20	0.96	10	8.28	the present work	work	other NP	structuring	
23	1.10	1	0.83	the yield of	yield	NP+of	causative	

LEGEND: NP + of – Noun phrase + *of*-phrase fragment; NP + other – Noun phrase with other post-modifier fragment; other NP – Other noun phrase; passive + PP – Passive + prepositional-phrase fragment; other passive – Other passive fragment; *we* + V – Verb phrase with personal pronoun *we*; other V fragment – Other verbal fragment; PP + *of* – Prepositional phrase + *of;* other PP – Other prepositional phrase (fragment); V/A + *to* – Verb or adjective + *to*-clause fragment; V/N + *that*-cl – Verb phrase or noun phrase + *that*-clause fragment; as + V – Adverbial clause fragment; *be* + AP – Copula *be* + adjective phrase; other AP – Other adjectival phrase; anticipatory it - Anticipatory *it* + verb or adjectival phrase; Others – Other expression