English for Academic Purposes

English for Academic Purposes

Approaches and Implications

Edited by

Paul Thompson and Giuliana Diani

Cambridge
Scholars
Publishing

English for Academic Purposes: Approaches and Implications

Edited by Paul Thompson and Giuliana Diani

This book first published 2015

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Copyright © 2015 by Paul Thompson, Giuliana Diani and contributors

All rights for this book reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN (10): 1-4438-7439-6
ISBN (13): 978-1-4438-7439-7

TABLE OF CONTENTS

INTRODUCTION

PAUL THOMPSON
UNIVERSITY OF BIRMINGHAM, UK
AND GIULIANA DIANI
UNIVERSITÀ DI MODENA E REGGIO EMILIA, ITALY

Over the last two decades, there has been a prolific increase in scholarly activity in the field of English for Academic Purposes (EAP). In this growth, the notions of corpus and genre have played a central role, with important repercussions for teaching approaches. These notions derive from two approaches to the investigation of academic English: "genre analysis" and "corpus linguistics".

Genre analysis has predominantly focused on genre as text, with the aim of exploring the lexico-grammatical and discursive patterns of particular genres to identify their recognizable structural identity, or what Bhatia (1999, 22) calls "generic integrity". As Hyland observes (2012, 415), "analysing this kind of patterning has yielded useful information about the ways in which texts are constructed and the rhetorical contexts in which such patterns are used, as well as provided valuable input for genre-based teaching".

Within EAP, most genre research has used the move-analysis approach developed by Swales (1990), "which seeks to identify the recognizable stages of particular institutional genres and the constraints on typical move sequences" (Hyland 2012, 415). Substantial work has been devoted to the study of academic research genres, such as research articles, abstracts, textbooks, book reviews, book review articles, and PhD theses (see e.g. Swales 1990, 2004; Myers 1992; Bhatia 1993; Motta-Roth 1998; Bunton 2002; Lorés Sanz 2004; Kwan 2006; Diani 2012; Thompson 2012).

Genre analysis is thus largely an attempt to identify the common traits of academic language in different domains. Comparative studies have gradually shown how disciplinary domains differ from each other not only as regards specialised topics and specialist vocabulary but also in terms of lexico-grammatical characteristics and rhetorical and argumentative structures (e.g. Hyland and Bondi 2006). Even items belonging to what is

known as 'general academic vocabulary' (e.g. verbs such as *note*, *claim* or *suggest*) may be seen to vary in usage and meaning according to the specific disciplinary domain or cultural context in which they are used, thus pointing to the particular ethos of the academic community under scrutiny.

Genre analysts have been greatly assisted in these descriptions by the compilation and investigation of language corpora. The ability to analyse large quantities of data has made it possible to study the particular characteristics of different discourse domains and to investigate variation phenomena.

A significant development in these two traditions of investigating academic English is the view that genre and corpus approaches should not be considered, to borrow Charles, Pecorari and Hunston's (2009, 3) words "as opposing ideas", but rather "as constituting a continuum from top-down to bottom-up", along which researchers "situate their individual studies" (e.g. Baker 2006; Biber et al. 2007; Ädel and Reppen 2008; Charles, Pecorari and Hunston 2009; Gotti and Giannoni 2014).

The integration of genre analysis and corpus-based investigations has had a major impact on EAP pedagogy (see, for example, Weber 2001; Flowerdew 2005; Hyland 2006; Charles 2007, this volume; Cortes 2007), with important implications for teaching academic writing. This is because genre descriptions, on the one hand, support learners by providing "an explicit understanding of how target texts are structured and why they are written in the ways they are" (Hyland 2007, 151). Corpus analyses, on the other hand, encourage learners to understand academic language use, and to see the connections between language and its contexts of use. The increased familiarity of students with electronic tools for corpus analysis has contributed to the development of their language awareness (e.g. Bondi 1999) and promoted learner autonomy (e.g. Lynch 2001).

Contents of the volume

Many of the issues outlined above are investigated in the chapters of this volume. The contributions are arranged into four parts, which highlight how corpus linguistics and genre analysis can work as complementary approaches. Pedagogical implications are also discussed in some detail as the research described here not only aims to investigate features of EAP but also to translate them into classroom applications. The first part presents corpus-based research into EAP at the lexicogrammatical and genre levels, with papers whose focuses range from issues related to patterning and phraseology through to papers focusing on

the language practices of specific disciplines and research genres. The second part is devoted to intercultural EAP research. The third part includes research on English as a Lingua Franca in academic communication. Finally, the last part addresses the relationships between corpus, genre and pedagogy in EAP, with an emphasis on implications and applications.

Overview of the chapters

The first two chapters of Part I "**Corpus, Genre and Disciplinary Discourses**", focus on lexico-grammatical patterning in written academic discourse. The opening article, by **Christopher Gledhill**, investigates how collocation and phraseology, from a lexicogrammar perspective, are relevant to EAP. He sets out a method of textual analysis which exploits the phraseological behaviour of grammatical signs. The chapter provides evidence that grammatical items can be shown to be stable elements in relatively predictable but also productive cascades of expression. His findings suggest that the identification of such extended lexicogrammatical patterns are a key feature in the systematic analysis of EAP texts.

The second chapter, by **Geneviève Bordet**, also treats the phenomenon of collocation as fundamental to the study of language, and genre analysis in particular. Her analysis provides evidence that each genre highlights specific disciplinary strategies based on the use of textual collocational variations, or "collocational chains". Her results show that these chains contribute to the perception of the text as both coherent and persuasive.

The second trend of Part I is represented by four chapters centring on language variation across disciplines in different academic genres: research articles, book review articles, abstracts, conference posters. **Šarolta Godnič Vičič** and **Mojca Jarc** explore language variation in the genre of research articles within sociology. Three functional key words, *among*, *between* and *these*, were selected for analysis as the most salient in all the corpora analysed. Their study demonstrates that there is intradisciplinary variation in the sociological research articles that is not due to the stylistic flexibility with which authors use language. They suggest that the differences between the preferred meanings and values found in the corpora may be attributed to differences in the research focus and theoretical positionings of the authors, the methodologies they use, and also the niches occupied by the journals.

Michele Sala's chapter examines the way knowledge is conceptualized and grammatically constructed in research article abstracts covering four

disciplinary areas (linguistics, law, economics and medicine). The analysis focuses on linguistic items which are used to portray the construction of disciplinary knowledge, to introduce concepts and methods, and to represent evidence and its interpretation. His study shows a consistent use of *research*, *cognition* and *discourse acts* in the disciplinary corpora analysed.

Stefania M. Maci investigates how conditional constructions are employed in the discourse of medical posters. Her results show that, in this genre, the uses of conditionals reflect the reasoning process of the hard sciences: they can either convey 'facticity' or 'refocusing'. Through the analysis, her study reveals that in the case of 'facticity', facts and results are reported according to the conditions they are associated with and are expressed in the 'Method' and 'Results' sections. As regards factual *if* clauses, their pragmatic role is indicated by the information ordering structuring: prediction seems to be realised with the fronting of protasis, whereas the *if and only if* condition appears to be constructed by means of delaying.

In the next chapter, **Giuliana Diani** explores reflexive phraseology across academic genres and disciplines. Employing a corpus-based approach, the study focuses on how metatextual phraseological units vary across academic research articles and book review articles and academic disciplines (business and economics). Her study shows that phraseological units can be very helpful signals for the analysis of generic and argumentative structure of academic writing. Her findings also demonstrate that convergences and divergences between closely related disciplines and genres help to differentiate different forms of disciplinary discourse.

With Part II of the volume, **"Contrastive EAP rhetoric"**, our attention turns to an investigation of intercultural EAP research. **Rosa Lorés Sanz** illustrates a methodological approach for intercultural research in the use of written academic genres in English by non-native (Spanish) academics, which involves corpus analysis, genre analysis and intercultural rhetoric as central methodological frameworks. She presents some of the findings that have resulted from its application to the exploration of non-native (Spanish) use of EAP. Special emphasis is also made on the advantages and the limitations faced by this methodological approach. The application of a cross-cultural approach to the design of teaching materials and the implementation of EAP courses is also discussed.

Part III "**English as lingua franca in academic settings**" is devoted to research on English as Lingua Franca (ELF) in academic communication. **Laurie Anderson** investigates the pragmatics of academic ELF

communication by examining the role that the thematization of self and other identities in terms pertinent to membership in an international community of scholars plays in peer-to-peer interaction among academics from different national backgrounds. Her study shows that in peer-to-peer interaction with colleagues from different national and linguistic backgrounds, scholars exhibit a particular understanding of international academia, reflecting both the geographical/geopolitical and institutional characteristics of the setting in which the presentations were made. The chapter concludes with a discussion on the extent to which the thematization of 'internationality' is rooted in the specific aims of the genre analysed and the extent to which it is instead a more pervasive aspect of academic communication in ELF settings.

In the next chapter, **Adriano Ferraresi** and **Silvia Bernardini** report on an ongoing project focusing on institutional (vs. disciplinary) academic English as a Lingua Franca as it is used in university websites worldwide to present degree programme descriptions and syllabi and to provide information on a wide range of administrative and organisational matters. They investigate the use of phraseology in native and ELF varieties represented in a 90-million word corpus of institutional academic texts in English. Their findings reveal that ELF university homepages display (phraseological) patterns which are only partially consistent with previous studies of non-native language, and that these deviations might or might not derive from conscious strategies to target an international audience.

In the final chapter of this section, **Giuseppe Palumbo** investigates features of comparable sets of texts written in ELF by universities and in two national varieties of English, with a view to identifying the way the texts construct their respective profiles at the morpho-syntactic level and realize their main, shared function. His study shows that the non-native, ELF texts present similarities and differences from comparable native texts. Some differences between the ELF set and the two native sets concern the use of verbs, while similarities regard the purely structural aspects (such as the generalised tendency to use premodification in noun phrases) and the use of patterns pointing to the adoption of similar signals of stance or engagement, such as the heavy personalization of the discourse through the use of pronouns ("we"/"you" as opposed to "the university"/"students"). His analysis highlights a certain homogeneity between the non-native and native sets with regard to their structural make-up.

Part IV **"Pedagogical implications in EAP"** turns attention to EAP pedagogy. **Maggie Charles** discusses how corpora can be used to enrich EAP pedagogy by facilitating the study of genre and discourse features in

academic writing. She illustrates this through two approaches. The first uses traditional paper-based materials derived from prior analysis of a corpus and shows how pedagogical tasks can contribute to raising student/learner awareness of the variability of genres. The second approach uses a hands-on method, in which the students/learners are responsible for building and investigating their own corpora and shows how they can make use of their corpora to examine discourse functions in their own discipline.

Maria Freddi's contribution focuses on EAP reading pedagogy. The chapter reports on the research informing the development of a course taught by the author at her home university, aimed at undergraduate Humanities students and entitled *Reading Skills in English for the Humanities*. It explores ways in which insights from corpus approaches to academic English combined with genre theory can be brought to bear on the design of the course syllabus and argues for a pedagogically targeted mix of the various paradigms under consideration.

In the last chapter of the volume, **Paul Thompson** explores the lexis of academic lectures through analyses of frequency and range of items within a corpus. He tests the coverage of the Academic Word List (Coxhead 2000) and General Service List against three other options and concludes that Paul Nation's 2K word list, based on BNC frequencies, provides a better indication of the most frequent items in academic lectures than does the General Service List. He then develops a specialised academic lecture listening word list made up of 200 items. Finally, he presents some corpus exploration activities, based around the new word list, which access the British Academic Spoken English (BASE) corpus in the open Sketch Engine interface, that can be used with learners preparing for the challenges of listening to lectures in English.

The various analyses collected in this volume provide a rich overview of the methods of investigation of EAP, the tools and the approaches, bringing together, to differing degrees, two complementary strands of linguistic investigation – corpus analysis and genre analysis. They demonstrate how the wealth of data made available through corpus compilation and searchable through query tools have enabled scholars to identify and give clear descriptions and examples of central concepts in EAP research.

References

Ädel, Annelie, and Randi Reppen. 2008. *Corpora and discourse: The challenges of different settings*. Amsterdam: John Benjamins.

Baker, Paul. 2006. *Using corpora in discourse analysis*. London: Continuum.

Bhatia, Vijay K. 1993. *Analysing genre. Language use in professional settings.* London: Longman.

—. 1999. Integrating products, processes, processes and participants inprofessional writing. In *Writing: Texts, processes and practices*, ed. Christopher N. Candlin and Ken Hyland, 21-39. London: Longman.

Biber, Douglas, Ulla Connor, and Thomas Upton. 2007. *Discourse on the move*. Amsterdam: John Benjamins.

Bondi, Marina. 1999. Language awareness and EFL teacher education. In *English teacher education in Europe: New trends and developments*, ed. Pamela Faber, Wolf Gewehr, Manuel Jiménez Raya and Antony J. Peck, 91-107. Frankfurt: Peter Lang.

Bunton, David. 2002. Generic moves in PhD thesis introductions. In *Academic discourse*, ed. John Flowerdew, 57-75. London: Longman.

Charles, Maggie. 2007. Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes* 6(4): 289-302.

—. this volume. Genre, corpus and discourse: Enrich EAP pedagogy.

Charles, Maggie, Diane Pecorari, and Susan Hunston. 2009. *Academic writing: At the interface of corpus and discourse*. London: Continuum.

Cortes, Viviana. 2007. Genre and corpora in the English for academic writing class. *ORTESOL Journal* 25: 9-16

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.

Diani, Giuliana 2012. *Reviewing academic research in the disciplines: Insights into the book review article in English*. Rome: Officina Edizioni.

Flowerdew, Lynne. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Counting criticisms against corpus-based methodologies. *English for Specific Purposes* 24(3): 321-332.

Gotti, Maurizio, and Davide S. Giannoni. 2014. *Corpus analysis for descriptive and pedagogical purposes: ESP perspectives*. Bern: Peter Lang.

Hyland, Ken. 2006. *English for academic purposes: An advanced resource book*. London: Routledge.

—. 2007. Genre pedagogy: Language, literacy and L2 writing instruction. *Journal of Second Language Writing* 16: 148–164.

—. 2012. English for academic purposes and discourse analysis. In *The Routledge handbook of discourse analysis*, ed. James Paul Gee and Michael Handford, 412-423. London: Routledge.

Hyland, Ken, and Marina Bondi. 2006. *Academic discourse across disciplines*. Bern: Peter Lang.

Kwan, Becky S.C. 2006. The schematic structure of literature reviews in doctoral theses of applied linguistics. *English for Specific Purposes* 25(1): 30-55.

Lorés Sanz, Rosa. 2004. On RA abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes* 23(3): 280-302.

Lynch, Tony. 2001. Promoting EAP learner autonomy in a second language university context. In *Research perspectives on English for academic purposes*, ed. John Flowerdew and Mathew Peacock, 390-403. Cambridge: Cambridge University Press.

Motta-Roth, Désirée 1998. Discourse analysis and academic book reviews: A study of text and disciplinary cultures. In *Genre studies in English for academic purposes*, ed. Inmaculada Fortanet, Santiago Posteguillo, Juan C. Palmer and Juan F. Coll, 29-58. Castelló de la Plana: Universitat Jaume I.

Myers, Gregory. 1992. Textbooks and the sociology of scientific knowledge. *English for Specific Purposes* 11(1): 3-17.

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

—. 2004. *Research genres. Explorations and applications*. Cambridge: Cambridge University Press.

Thompson, Paul. 2012. Thesis and dissertation writing. In *Blackwell handbook of English for specific purposes*, ed. Brian Paltridge and Sue Startfield, 283-300. Oxford: Wiley-Blackwell.

Weber, Jean-Jacques. 2001. A concordance- and genre-informed approach to ESP essay writing. *English Language Teaching Journal* 55(1): 14-20.

PART I

CORPUS, GENRE AND DISCIPLINARY DISCOURSES

CHAPTER ONE

ON THE PHRASEOLOGY OF GRAMMATICAL ITEMS IN LEXICO-GRAMMATICAL PATTERNS AND SCIENCE WRITING

CHRISTOPHER GLEDHILL
UNIVERSITÉ PARIS DIDEROT, FRANCE

1. Introduction

In this chapter I examine the role of grammatical items in lexico-grammatical patterns (or 'LG patterns', for short). In previous work I examined the collocational patterns of individual grammatical items in a particular genre (the cancer research article, Gledhill 1995, 2000a, 2000b). In these studies, I demonstrated that individual functional words (such as 'and' in Titles, 'but' in Abstracts, 'to' in Introductions and so on) have a non-random distribution in these texts, since these words are 'statistically salient' or 'key' in these different parts of the research article. I then went on to examine in detail the phraseological behaviour of these items in each subsection, arguing that, contrary to what one might think, each grammatical item enters into a very restricted, predictable set of phraseological patterns, according to the type of text being analysed. It is often thought that grammatical items do not enter into collocational relations, since they can 'be used anywhere' and thus can 'collocate with anything'. However, one of the findings of my work has been to argue that grammatical items have a highly restricted phraseology in specialised discourse, a feature which makes them ideal targets for analysis, since an analysis of the distribution and behaviour of function words can in effect be seen as a preliminary analysis of the fundamental stylistic and phraseological properties of a particular text type.

In this chapter, I use similar methods and I make similar claims. However, my focus here is somewhat different. In this study, I concentrate on longer stretches of wording, and in particular I am interested here in examining discontinuous stretches of text or what Renouf and Sinclair (1991) have called "collocational frameworks". A discontinuous stretch of text is a short sequence of words such as *a(n) * *-ed in* in which only a few selected kinds of lexical item (marked by *) can fit meaningfully into the lexical pattern (in this case '*A substance found in... A cytokine implicated in'...* etc.). What I find interesting about such sequences is that they often correspond to short extracts of very highly specialised discourse. The main hypothesis I wish to test here is that by searching for a given sequence of grammatical signs, it is possible to identify a regular pattern of discourse, provided that this pattern is looked for in a relatively coherent body of texts (i.e. an electronic archive or corpus). Furthermore, I would suggest that by looking at discontinuous sequences in this way, the corpus analyst should be able to identify some of the most characteristic features and functions of that genre in an efficient and systematic manner. In this chapter, I look at examples taken from a corpus of research articles and their corresponding abstracts (referred to below generally as RAs), as well as 'journalistic accounts' of the same research (referred to as JA). However, as I point out in the data analysis below, although many discontinuous stretches are highly regular and recurrent in the particular texts I am interested in here, some kinds of writing (notably scientific journalism) also involve 'hybrid' patterns, i.e. an original blending of two or more patterns that are commonly found in other types of discourse.

In this contribution I use the term "grammatical item" (also known as closed-class item, function word, small word, stop-word, etc.) in contrast to "lexical item", and I assume that the difference between the two is that grammatical items belong to a relatively closed class of high-frequency, polyvalent words with relatively abstract meanings, such as *auxiliaries, conjunctions, determiners, grammatical adverbs, prepositions* and *pronouns*. Until recently, grammatical items as an entire class have not received much attention in English for Specific Purposes and corpus-based genre analysis. Indeed, it has often been supposed that grammatical items are of little interest in text linguistics because they can "go with anything" (i.e. collocate with any lexical item, appear in a vast range of grammatical structures, occur in a uniform way across almost all text types, etc.). Thus in the early days of lexicometrics and Natural Language Processing (NLP), specific procedures were developed to filter these words out from the analysis (for example, Smadja 1993 introduced a well-known method of extracting collocations from the Hansard bilingual corpus, but only after

a process of automatically filtering out the function words). To a certain extent the concept of the stop-word is still widespread, and entirely understandable: analysts are interested in getting at what they consider to be important data (equating this with lexical items, content-words, terminology), while the large quantities of apparently ubiquitous grammatical items which fill most types of text appear to be redundant 'noise'.

However, since the advent of corpus linguistics (and particularly the use of corpora by grammarians), there has been a growing body of evidence to suggest that grammatical items enter into collocational relations that are just as interesting and revealing as lexical items, and thus have an important role to play in the phraseological patterning of texts. Renouf and Sinclair (1991) and Renouf (1992) most notably argued that grammatical items are the building blocks of idiomatic language, and coined the term "collocational framework" to refer to such productive sequences as *a(n) X of (a [dash, handful, smattering] of)*. More recently, and in a way that mirrors the current move to rehabilitate 'junk DNA',[1] some researchers in NLP now recognise the importance of functional words in automatic text recognition, terminology extraction and other corpus-based applications (Meyer 1988; Riloff 1995; Vergne 2004). Similarly, collocational frameworks and related notions such as "bundles", "clusters", "n-grams" and so on, have become an accepted part of the descriptive apparatus of corpus-based applied linguistics, and many researchers have examined the distribution and collocational behaviour of specific grammatical items as they occur in specialised corpora (Luzón Marco 1999, 2000; van der Wouden 2001, 2007; Biber et al. 2004; Cheng et al. 2008; Hyland 2008; Bordet 2011) as well as the role of grammatical sequences and frameworks in discourse analysis, language learning and evaluation (Hasselgren 2002; Groom 2005, 2010; Scott and Tribble 2006; Biber and Barbieri 2007; Lee et al. 2008).

But although the study of function words has now become an accepted part of the corpus-based approach, the notion that grammatical items can be the focus of phraseological analysis still requires some theorisation within a broader analytical framework. In addition, it seems to me that for most observers, the idea that grammatical items can be the starting point for textual analysis is still not obvious. Therefore, before embarking on an analysis of discontinuous lexico-grammatical patterns, I set out in the first half of this chapter some of the arguments for studying grammatical items from the point of view of corpus-based genre analysis.

2. Why study grammatical items?

2.1. Grammatical items have collocations

Not all analysts accept that grammatical items enter into collocational relations. For example, in one well-known British dictionary of linguistics, "collocation" is defined as the co-occurrence of lexical items, while grammatical items are explicitly stated as having no collocational relations:

> collocation, n. A term used in lexicology by some (especially Firthian linguists) to refer to the habitual co-occurrence of individual lexical items […] Some words have no specific collocational restrictions - grammatical words such as *the, of, after, in* [...] Another important feature of collocations is that they are formal (not semantic) statements of co-occurrence [...] (Crystal 2008, 86-87)

This definition seems to represent a commonly-held view among linguists. However, I feel that Crystal rather misrepresents the way Firth (1957) would have understood the term, or at least as Firth's successors understand it. From a Firthian point of view, every single sign in the language (whether lexeme or morpheme) has a consistent and contrastive context of use. In other words, each sign is used in a consistent and contrastive set of linguistic co-texts (e.g. the preposition *of* typically has as a complement the noun *course*) and each sign is used in a consistent and contrastive set of situational contexts (e.g. *of course* tends to be used as an informal, concessive adjunct). The term "context" is clearly central to this approach, and in typical Firthian fashion it is used to indicate three things at the same time: a) the "the co-occurrence of forms within the same stretch of text" (usually what is meant by "co-text"), b) the immediate "context of situation", and c) the broader "context of culture". The point about co-text and context is that they essentially shape the meaning of the linguistic sign, since signs are mutually dependent on their typical collocational partners in discourse, as Firth puts it:

> The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. The words are mutually expectant and mutually prehended. (Firth 1957 [1951], 181)

This kind of definition sets itself against an "essentialist" or "semantic trait" approach to meaning. Thus, it is claimed that the meaning of a word such as *of* can only be seen as a composite of its particular uses, which

depend partly on its co-occurrence with *course* in *of course* and partly on other uses, such as its co-occurrence with a nominal post-modifier referring to people in terms of subjective, usually high-mindedly positive qualities (*a man / woman of [action, honour, humility*, *steel, quality]*) and so on. Sinclair (1991) argued that these and the other typical lexico-grammatical patterns associated with *of* contribute to our general understanding of this rather idiosyncratic preposition. As he puts it:

> 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 text (Sinclair 1991, 108).

2.2. Grammatical items have different distributions in different text types

Many linguists would agree that different varieties of language exploit different lexical and grammatical resources, an assumption which lies behind the multi-factorial method of register analysis developed by Biber et al. (2002). Yet it is surprising to see how few studies of a particular text type begin by setting out the relative distribution of grammatical forms, especially grammatical items. In this chapter, I examine the role of grammatical items in their local contexts (as the most recurrent, pivotal items in lexico-grammatical patterns). But before examining their local role, it is important to realise that grammatical items (indeed all items, whether functional or lexical) have a particular distribution in different texts, and even within different subsections of texts. The reason for this is that that, as shall be shown in the following sections, if the occurrence of a particular grammatical item is statistically more 'salient' or 'key' in a particular text type or subsection of a text, this is because this item is pivotal in those lexical patterns which have an important role or discourse function to play in that text. One example of this will suffice here: in Gledhill (2000) I showed that the word 'to' is an statistically significant item in cancer research article introductions. One reason for this, it would seem, is that 'to' is a pivotal element in post-posed attributive clauses such as *'it is (important, necessary, possible) to (assess the cell differentiation at this stage, construct a series of structures, identify TAAs, repeat measurements)...*' but also in passive non-finite projecting clauses such as '*(HPV 16 E6, hyperphasis, metabolic inc-cells, is (known, likely, thought...) to (be involved, be a major factor, determine celle cycle) in' etc*.' (from the Pharmaceutical Sciences Corpus, Gledhill 2000, 151). Examples such as these demonstrate three principles: 1) grammatical items collocate with other lexical and grammatical items (note the discontinuous

sequence *it is X to* in the first pattern and *X is Y-ed to Z in* in the second pattern), 2) each lexico-grammatical pattern expresses a specific discourse function (in the first case, 'strongly stating the case for a clinical methodology' and in the second case 'tentatively proposing a biochemical explanation'), and 3) the most systematic way of identifying these patterns is, in my view, to compare the relative distribution of grammatical items across different text corpora.

As mentioned in the introduction, in Gledhill (1995, 2000a, 2000b) I attempted to show that grammatical items have a particular distribution across the different rhetorical sections (Titles, Abstracts, Introductions, etc.) of 150 research articles (RAs) in the field of cancer research (the Pharmaceutical Sciences Corpus, PSC). At the time, I did not have access to a specialised reference corpus in English but later I used the British National Corpus (BNC) as a reference corpus. For example, the following tables (1, 2 and 3) present the results of a "keyword" comparison using AntConc (Antony 2002). In order to obtain these results, AntConc first creates a word list for each corpus (the BNC has 100,520,565 tokens with 448,005 types, and the PSC has 1896869 tokens with 48,537 types).[2] AntConc then calculates a score for each word by comparing a particular item's percentage chance of occurring in the study corpus as opposed to its chances of occurring in the reference corpus. For example, in the PSC *were* occurs 17,968 times divided by 1,896,869 tokens (=0,0961 or roughly 0.96%), whereas in the BNC *were* occurs 306,801 times divided by 100,520,565 tokens (=0,00305 or roughly 0.30%). Such a large percentage difference shows the extent to which certain function words, such as *were,* can have consistently different distributions across text types. It is of course important to be able to judge what is meant by "large percentage difference", and the "Keywords" module of AntConc is an important stage in this analysis. Keywords compares the relative percentages of items as they occur in the study corpus and reference corpus, and then assigns a rank to each item according to a statistical test (for details see Scott and Tribble 2006).[3]

The tables set out the first 20 keywords in the PSC (ranked by decreasing keyword score) as compared with the BNC (Table 1) and the first 20 key grammatical items in the PSC as compared with the BNC (Table 2).

Rank	Item	Freq. in PSC	Keyword score(vs. BNC)	Rank	Item	Freq. in PSC	Keyword score (vs. BNC)
1	&	6071	48432.645	11	Table	2231	8976.632
2	patients	8563	36825.897	12	clinical	1812	8445.988
3	et	4540	22956.671	13	cell	2165	8335.861
4	al	4544	21851.859	14	min	1447	7897.745
5	study	5791	19067.584	15	Fig	2126	7643.490
6	cells	3911	16769.641	16	cases	3053	7609.082
7	were	17968	15873.864	17	patient	2249	7373.035
8	results	3664	11613.943	18	studies	2467	7307.815
9	treatment	3212	10486.346	19	significant	2410	6703.943
10	of	82182	9404.549	20	tissue	1373	6402.795

Table 1. First twenty keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.

Rank	Item	Freq. in PSC	Keyword score(vs. BNC)	Rank	Item	Freq. in PSC	Keyword score (vs. BNC)
7	were	17968	15874.251	228	these	3859	1652.307
10	of	82182	9404.238	236	However	1702	1627.852
38	with	21063	5318.315	247	due	1203	1595.017
45	in	47809	4915.890	260	may	4165	1529.965
58	and	61723	3824.316	261	The	16217	1528.736
91	during	2581	2971.304	347	Therefore	461	1250.765
117	between	4148	2521.590	354	or	9940	1230.950
124	In	6005	2464.055	361	after	3479	1213.382
163	vs	396	2114.845	364	was	20876	1208.869
189	versus	456	1888.239	519	both	2313	921.387

Table 2. First twenty grammatical keywords in the Pharmaceutical Sciences Corpus (PSC) vs. the BNC.

Analysts familiar with keyword lists will have little difficulty in interpreting these data. The keywords in Table 1 show some of the major textual features of the PSC (such as the presentation of *data* in *Tables* and *Figures*) as well as the topical preoccupations of the PSC (the nominal expression of material processes, e.g. the *study* of *cells, cases* or *groups* of different *ages* and at different *times*, the *treatment* of *patients*, the *use* of drugs / *treatments* and the verbal expression of communicative or perceptive processes *significant results found* or *reported*). Similar

comments can be made about the key grammatical items in the PSC (Table 2): they are predominantly prepositions and coordinators (*of, and, or,* typically involved in elaborate nominal post-modifiers in the PSC) or prepositions involved in adjuncts / post-modifiers expressing cause (*due to*), accompaniment or manner (*with*), temporal extent (*after, between, during, in*) and comparison (*between, versus, vs.*). Table 2 also shows the typical markers of cohesion which we might expect to find in elaborate written discourse, such as pronouns / determiners (*both, The, these*) and sentence-initial conjuncts (*However*, *Therefore*). Other items are perhaps less obviously typical of written discourse, but Table 2 shows that they are salient in science writing: *may* (the preferred modal verb for "hedging", especially in Discussion sections of the PSC) and *were* (usually an auxiliary expressing the past passive in Methods sections).

As mentioned above, my initial description of the PSC was an intra-varietal analysis, conducted in order to establish the main differences between the different rhetorical sections of the research article and the PSC corpus as a whole (Titles, Abstracts, Introductions, Methods, Results, Discussions). I shall not repeat these data here, but for illustrative purposes, the following Table 3 sets out the main results for the first ten key grammatical items across each sub-section of the PSC:

Rank	Rhetorical Section of the PSC					
	Title	Abstract	Introduction	Methods	Results	Discussion
1	of	but	been	were	no	that
2	for	these	has	was	in	be
3	on	of	have	at	did	may
4	and	there	is	then	not	is
5	in	in	such	for	had	our
6	-	was	can	each	after	in
7	-	that	it	and	there	not
8	-	did	we	from	the	this
9	-	who	of	after	when	we
10	-	both	to	with	all	have

Table 3. First ten grammatical keywords in the six main sub-sections of the PSC.

Although general comparisons (Tables 1 and 2) give a good idea of the general features of the PSC, Table 3 shows the extent to which there is also much variation within the research article genre itself. In fact there is so much internal variation that some items which are statistically salient in their respective sub-sections of the PSC, are also more typical of the

general language (BNC) when compared with the PSC as a whole (in particular: the pronouns *we, our* in Introductions and Discussions, the modal *can* which is only salient in Introductions, the item *to* which is also salient in Introductions, as mentioned above). In Table 3, I have indicated the items which stand out in relation to the other sub-sections of the RA (by underlining). I shall not go into a detailed analysis of these data here. It is sufficient to note that over half the keywords in the Introductions (*been, has, such, can, it, to*) and Methods (*were, at, then, each, from, with*) are only specifically "key" in these sections, a result which suggests that these sections have a specific phraseology which is quite unlike the rest of the research article (although these items are of course not exclusive to these sections).

2.3. Grammatical items are pivotal elements in lexico-grammatical patterns

In the previous section, I showed that grammatical items do not have an even distribution across text types, and that their distribution varies considerably, even within the same text type. In this section I argue that the identification of "key" grammatical items can be seen as a useful first stage in the search for longer stretches of phraseology. Some authors (notably Hunston and Francis 2000) use the term "lexical pattern" to refer to regular multi-word units which do not necessarily correspond to the traditional constituents of the clause. In this chapter, (and elsewhere, Gledhill 2011) I refer to such sequences as "lexico-grammatical" (LG) patterns, in an attempt to make it clear that in any multi-word phrase at least one grammatical item (or grammatical structure) is a permanent or "pivotal" element around which the rest of the phrase is built.

In order to illustrate this notion, let us return to the particular case of Abstracts in cancer research articles (in the PSC there are 400 Abstracts = 123,296 words or 6.5% of the corpus). As mentioned above, the first ten grammatical keywords in this sub-section are (in order of rank) *but, these, of, there, in, was, that, did, who, both.* Out of context, it is not clear what patterns of usage these items might represent. An item such as *that* can be used in many different lexico-grammatical contexts (conjunction, pronoun, determiner etc.), and it is therefore necessary to analyse each item separately, not only within Abstracts, but also contrastively, in the rest of the research article. This is not an easy task, not least because the analysis of grammatical items usually generates a vast amount of data. However, I would suggest that the task is simpler when looking at a specialised genre than for the general language. For example, in the 1st edition of the Collins

Cobuild dictionary (Sinclair 1995), there are 19 entries for *of* (not including idiomatic uses), whereas in the PSC (Gledhill 2000,142-149) the number of patterns varies between 3 (Titles, Abstracts) and 5 (Introductions). Even so, it is still difficult to represent this kind of data, and I will not repeat the analysis of each of these patterns here, largely because this type of detailed analysis requires long lists of concordances. However, in Gledhill (1995) I suggested one way of resolving the problem of data representation, which I called the "collocational cascade". An example of this is set out in the following figure:

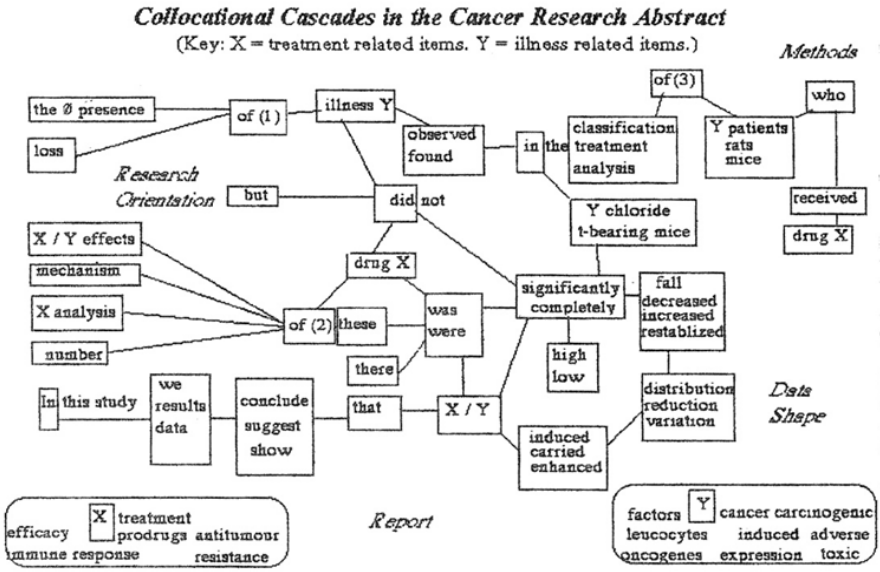

Figure 1. Collocational cascades in Cancer Research Abstracts.[4]

I would argue that "collocational cascades" are an efficient way of summarising the most salient phraseology of a particular text type. Thus in Figure 1 we can see that the outstanding or salient phraseology of Abstracts involves a specification of the general shape of data (as evidenced by lexico-grammatical patterns involving *in, of, that, there*), statements about who or what was affected by various treatments (the LG patterns around *in*, *of, who*) and the extent to which an effect was or was not observed (the LG patterns involving *but, did, not, was, were*). Unlike diagrams representing collocational networks (Williams 1998),

collocational cascades do not represent a formal or statistical relationship between lexical items. Rather, the cascade is broadly meant to be read from left to right, as an informal representation of interlocking lexico-grammatical patterns which, as the cascade metaphor suggests, fall or lead on from one choice of expression into another further on in the clause. In Figure 1, each grammatical item in the cascade (*but, did* + *not, in, of, that, there, these, who,* but not *both*) is linked to one or more of the main patterns as observed in the Abstracts sub-corpus (not counting of course the many sub-patterns or variants of these patterns). We can also see in the diagram that some items appear more than once. Thus, the diagram shows that there are 3 (main) patterns involving the preposition *of* in Abstracts: (1) a quantification (*loss / presence*) of a (usually post-modified) disease-related item (*cancer cells, carcinogenic factors, leucocytes...*), (2) an observation, quantity or facet (*amount*, *analysis*, *effect*) of a treatment-related item *(antitumour response, immune response, prodrug*), and (3) an extended pattern involving a reduced relative clause (expressing a mental / empirical-oriented process: *found*, *observed*) qualified by a complex nominal (expressing a research-oriented process: *in the* + *analysis / classification / treatment of* + *disease-related item*). While these patterns are clearly prevalent in Abstracts, they are also clearly typical of the complex (post-modifying) nominals analysts have come to expect in academic and scientific writing in general. In addition, in the following section, we see that pattern 4 has a slightly different realisation in journalism (examples 4g, 4h and 4i).

One of the defining features of collocational cascades is that the patterns they represent are all related (each grammatical and lexical item is linked indirectly to one or more other items, creating a complex, although sometimes incomplete chain of patterns). I would claim that the grammatical items in the cascade are "pivotal", that is to say they are used consistently in each of these patterns. The lexical items on the other hand represent a "paradigm", in that they usually represent semantic abstractions or families of related lexical items rather than specific examples (as in *of* pattern 1: *loss / presence of* + Y, where Y is the name of a specific disease-related item such as *leucocytes*). In addition, collocational cascades have a certain directionality. What I mean by this is that the cascade as a whole represents the general way in which information is structured within Abstracts. For example, expressions and phrases which are research- or report-oriented (*In this study, we conclude that...*) as well as empirical observations (*loss / presence of item Y*) appear in theme (clause-initial) position in this diagram, whereas clinical methods

(*item Y who received item X*) and results (*did not significantly fall / increase...*) appear in rheme / clause-final position.

I make no claim here about the linguistic status of collocational cascades; I primarily see them as a way of depicting the outstanding phraseology of a particular text type. However, I would suggest that this kind of representation does capture something of the social or psychological reality of this kind of formulaic language, in which members of the discourse community recognise that they write in "chunks" and "formulae", and claim to "skim" research articles before deciding whether to pore through them line-by-line. These notions, as well as non-linear processing, predictive text analysis and lexical priming, have become important themes in applied linguistics (de Cock 1998; Simpson 2004; Hoey 2005). However, I will not dwell on these issues here. The more general point I am making is that lexico-grammatical patterns are significantly recognisable within a particular text type, and that LG patterns can be seen as parts of a broader set of interlocking collocational cascades within that particular type of discourse.

3. Extended lexico-grammatical patterns

So far, I have argued the case for seeing grammatical items as key elements in the corpus-based analysis of genres, since these words are the pivotal building blocks of lexico-grammatical patterns. In this section, I attempt to establish whether longer stretches of LG patterns can be identified on the basis of corpus analysis, and my particular focus here is on patterns involving extended (and usually discontinuous) sequences of grammatical items. Previous work on collocational frameworks has usually focused on the collocation of two grammatical items within a pre-defined window of words (Renouf and Sinclair 1991; Cheng et al. 2008; Groom 2010). Here, on the other hand, I am interested in identifying patterns in which at least one grammatical sign is bound between two other grammatical items, such as *the* * * *of* **s* (where * is a wild-card representing one lexical item of any length, * * are contiguous lexical items, and [* *of*] or [**s*] stand for lexical items with an intervening grammatical item or an attached grammatical morpheme). My hypothesis is that discontinuous sequences of grammatical items are particular to specific genres, and that when they can be observed with sufficient regularity, they provide good evidence for the existence of the typical lexico-grammatical patterns of that text type. In other words, given two sequences, such as a) and b) below (both sequences are complete sentences), it should be possible to predict whether they belong more or

less to the typical discourse of a research article (RA) or a journalistic article (JA), and within these sequences it should be possible to identify those patterns that are typical of the genre and those which are 'merely' local innovations:

a) * * * is a * *of * *s and some *s * that *ly * of *s are *ed.
b) *s * a * *er to *ing an * * * and * * of * of the most * and * *s.

In order to test the 'extended pattern' hypothesis, I have examined a sample of complete sentences taken from research articles on cancer cachexia (two of which authored or co-authored by Professor Michael Tisdale, Aston University, and both included in the Pharmaceutical Sciences Corpus, PSC) and from a selection of journalistic articles which all refer to this research as a "breakthrough".[5] In the following sections I look at the initial sentences from two research articles written by Michael Tisdale and his colleagues (RA1: *Trends in Pharmaceutical Sciences* and RA2: *Journal of the National Cancer Institute*) and I then look at the initial sentences from journalistic accounts which refer to this research as a 'breakthrough' (JA1: *The Daily Telegraph,* JA2: *The Independent,* JA3: *The Guardian* and JA4: *The Birmingham Post*). In each case, I attempt to find the sequence in two corpora: the British National Corpus (BNC) for the general language and the Pharmaceutical Sciences Corpus (PSC) for scientific discourse. For each sequence, it is possible to multi-word searches of the form *the * of ** (although AntConc often interprets the * symbol as more than one word). When a sequence turns out to be impossible to find (which is the case for most examples above approximately 10 signs), I then search for increasingly shorter extracts.

(RA1) *Trends in Pharmaceutical Sciences*[6]

The first sentence of this research article reads as follows:

(1)
Progressive weight loss is a characteristic feature of malignant diseases, and some studies suggest that nearly 90% of patients are affected. (RA1)

Here is the sequence of grammatical items used in the search:

* * * is a * *of * *s, and some *s * that *ly * of *s are *ed. (RA1)

I have included the sign *is* in the search sequence, even though it is used here as a lexical, copula verb (the verb *are* in the second half of the

extract is an auxiliary, a more *bona fide* grammatical item). One justification for doing this is that if *is* is not included, a search for the sequence * * * * *a* * **of* produces too many hits and includes many irrelevant patterns. Using *is* to narrow down the search, I find 264 examples of the sequence: *is a * * of *, and* in the BNC. Many of these examples do not include a clause break before *and,* as in *this show is a triumphant affirmation of life and vitality*, and very few involve a complex nominal (as we have at the beginning of extract 1). Only 11 BNC examples are structurally close (but still not exactly matching) extract 1. However, it is interesting to note how similar these examples are topically to extract 1, all involving highly technical subject nouns and an attributive clause which either defines or evaluates the subject as a more general "cause", "source", "method", "product" etc.:

(1a)
COPD **is a leading cause of** morbidity and mortality worldwide, and results in an economic and social burden that is both substantial and increasing. (BNC)

(1b)
Pamidronate, a second-generation bisphosphonate, **is a potent inhibitor of** resorption, and has been successful in the treatment of TIH. (BNC)

(1c)
This dividing technique **is a useful method of** increase, and works well, provided each piece has some root and some dormant buds or young shoots. (BNC)

A similar search in the PSC of course finds RA, plus 10 other examples, this time with complex nominal structures in subject position. In Gledhill (2000) I found that the sequence *is a* is a salient sequence in research article introductions. It seems that this usage is simply one realisation of a more general pattern in academic writing, in which *to be* introduces an attributive complement in the present tense and has the discourse function of expressing explicit evaluation. As can be seen in the following PSC examples, the phraseology of these patterns is similar to that of the BNC, except that the complement typically refers either to a key biochemical agent / participant, or to a source / cause from the point of view of the observer (example 1f):

(1d)
The present inhibition studies show that MAMC **is a competitive inhibitor of** dextromethorphan, and vice versa. (PSC)