Linguistic Variation in Research Articles

When discipline tells only part of the story

Bethany Gray

Studies in Corpus Linguistics 71

JOHN BENJAMINS PUBLISHING COMPANY

Linguistic Variation in Research Articles

Studies in Corpus Linguistics (SCL)

ISSN 1388-0373

SCL focuses on the use of corpora throughout language study, the development of a quantitative approach to linguistics, the design and use of new tools for processing language texts, and the theoretical implications of a data-rich discipline.

For an overview of all books published in this series, please see
http://benjamins.com/catalog/books/scl

Volume 71

Linguistic Variation in Research Articles
When discipline tells only part of the story
by Bethany Gray

Linguistic Variation in Research Articles

When discipline tells only part of the story

Bethany Gray
Iowa State University

John Benjamins Publishing Company
Amsterdam / Philadelphia

The paper used in this publication meets the minimum requirements of the American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48-1984.

Cover design: Françoise Berserik
Cover illustration from original painting *Random Order*
by Lorenzo Pezzatini, Florence, 1996.

DOI 10.1075/scl.71

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2015026291 (PRINT) / 2015028545 (E-BOOK)

ISBN 978 90 272 0379 3 (HB)
ISBN 978 90 272 6804 4 (E-BOOK)

John Benjamins Publishing Co. · https://benjamins.com

Table of contents

Acknowledgements

The research reported on in this book has benefited from the support of many individuals. Doug Biber provided limitless support, encouragement, and guidance throughout the project, helping me to focus the analyses and keep the scope of the project within reason, and encouraging me to carry the project through to publication. I also wish to thank Viviana Cortes, Randi Reppen, Susan Conrad, and William Grabe for their advice and feedback on the project from the start. I have learned so much from all of you that goes well beyond conducting linguistic research.

Several faculty members at Northern Arizona University deserve particular thanks, as they served as disciplinary experts at various stages in the project: Dr. William Grabe (Applied Linguistics), Dr. Ron Gunderson (Economics), Dr. George Lubick (History), Dr. Timothy Porter (Physics), Dr. George Rudebusch I (Philosophy), Dr. Maribeth Watwood (Biology), and Dr. Stephen Wright (Political Science). These individuals spent valuable time helping me to understand more fully the nature of their respective disciplines and publication practices within the field, and informed many stages of the analysis – from corpus design to interpretation.

I could not have accomplished this book without the support of my family. Chris, Noah, and Ellie – thank you for your patience, your comic relief, and your confidence in me. Chris, thank you for talking through programming logic and linguistic variables with me, helping me through the growing pains of learning to program.

I would also like to thank my colleagues at Iowa State University for their encouragement and support, and an anonymous reviewer who provided detailed feedback on a draft of the manuscript.

List of tables

List of figures

CHAPTER 7

CHAPTER 1

Introduction

1.1 Academic research writing: One register or many?

Applied linguists have long been fascinated with the written language of academia. This interest has developed and expanded over the past few decades, in part due to the premise that much can be learned about disciplinary practices and cultures by examining academic writing: the primary means of the transmission of knowledge in academic fields. The ability to use a single, overarching term like 'academic writing' belies the complexity and range of text types that fall within this label. Academic writing is not a monolithic construct, despite the ease with which we refer to 'academic writing' in general terms. Instead, academic writing is widely regarded as a register exhibiting inherent variation. We easily recognize different types of academic writing, commonly making distinctions between writing produced by students versus professional academics, between L1 and L2 writers, and between texts produced for different purposes. For example, we recognize a range of sub-registers within the domain of 'academic writing', such as textbooks, lab reports, research monographs, conference abstracts, argumentative essays, book reviews, and research articles, to name a few. A great deal of research has been devoted to describing the language of these different texts – from vocabulary use to phraseological patterns, and from grammatical characteristics to discourse structure – with the understanding that these language features are used in distinct ways in different types of academic texts.

Likewise, there is wide recognition that written academic language varies according to discipline – that disciplines utilize linguistic resources in varied ways to construct meaning and build knowledge within their disciplinary communities. The assumption is that the language used by these disciplinary communities is distinct, just as disciplines differ in their epistemological beliefs, research practices, and knowledge structures. Such variation across disciplines is perhaps never more evident than when reading texts from an unfamiliar discipline, as even readers well-versed in research writing in their own fields may be challenged to parse exact meanings out of the words, phrases, and clauses on the page. Consider, for

example, the following excerpts. Both excerpts come from research articles published in academic journals, and even come from the same part of the text – the abstract. Is understanding the content difficult? Can you guess the disciplines to which these excerpts belong?

Excerpt 1

> This paper argues that expressivism faces serious difficulties giving an adequate account of univocal moral disagreements. Expressivist accounts of moral discourse understand moral judgments in terms of various noncognitive mental states, and they interpret moral disagreements as clashes between competing (and incompatible) attitudes. I argue that, for various reasons, expressivists must specify just what mental states are involved in moral judgment. If they do not, we lack a way of distinguishing moral judgments from other sorts of assessment and thus for identifying narrowly moral disagreements. If they do, we can construct cases of intuitively real dispute that do not rest on the theory's preferred mental states. This strategy is possible because our intuitions about moral concept-ascription do not track speakers' noncognitive states. I discuss various ways of developing this basic argument, then apply it to the work of the two most sophisticated proponents of expressivism, Allan Gibbard and Simon Blackburn. I argue that neither is successful in meeting the challenge. [Philosophy abstract; Merli 2008]

Excerpt 2

> During early embryogenesis in Caenorhabditis elegans, the ATL-1CHK-1 (ataxia telangiectasia mutated and Rad relatedChk) checkpoint controls the timing of cell division in the future germ line, or P lineage, of the animal. Activation of the CHK-1 pathway by its canonical stimulus DNA damage is actively suppressed in early embryos so that P lineage cell divisions may occur on schedule. We recently found that the rad-2 mutation alleviates this checkpoint silent DNA damage response and, by doing so, causes damage-dependent delays in early embryonic cell cycle progression and subsequent lethality. In this study, we report that mutations in the smk-1 gene cause the rad-2 phenotype. SMK-1 is a regulatory subunit of the PPH-4.1 (protein phosphatase 4) protein phosphatase, and we show that SMK-1 recruits PPH-4.1 to replicating chromatin, where it silences the CHK-1 response to DNA damage. These results identify the SMK-1PPH-4.1 complex as a critical regulator of the CHK-1 pathway in a developmentally relevant context. [Biology abstract; Kim et al. 2007]

Now, think about what features of the excerpts lead you to make that determination. Perhaps the first feature that comes to mind is content or vocabulary. Both excerpts make extensive use of lexical items that are specific to the areas of inquiry in the respective disciplines: *expressivism, moral disagreements, noncognitive mental states*, and *moral concept-asscription* for the philosophy excerpt, and *embryongenesis, DNA, embryonic cell cycle progression*, and *protein phosphatase* for biology.

Even without knowing the exact meanings of all of these items, you were likely able to make a guess about the disciplines of these two abstracts.

However, the linguistic differences between the texts go beyond the content expressed by lexical items. Perhaps you also noticed the use of personal pronouns in the texts. The philosophy abstract uses many more personal pronouns generally, and uses both third person and first person pronouns to refer to the author and some unspecified individuals – the "expressivists". The biology excerpt uses a few first person pronouns, but the reference is restricted to the authors of the texts. Or, perhaps you recognized the use of 'argue' to describe the main rhetorical function in the philosophy excerpt, contrasted with the focus on reported empirical research results in the biology text (e.g., *We recently found that…, We report that…, We show that…*). Even impressionistically, we can recognize differences in the linguistic resources that are utilized in the two texts. We recognize the simple fact that language use varies between disciplines, and this acknowledgment has provided the motivation for a wealth of research into disciplinary communication practices.

According to Bazerman (1994: 104), the underlying belief of investigations into language use in the disciplines is that "the primary product of most disciplines, and a secondary product of all, are published texts, which are taken to constitute the knowledge of the disciplines". Written discourse is so integrally connected to disciplinary knowledge that Turner (2006) actually uses the presence of disciplinary discourse in his definition of a discipline: a discipline is socially constructed, has regulatory practices, and its members practice a "rhetoric of competence".

The centrality of writing to academic culture, practice, and knowledge building has led to a great deal of research in several fields, including rhetoric and composition, applied linguistics, and English for Academic Purposes (EAP). Often, studies investigating academic writing are motivated by the desire to inform the teaching of writing to native and non-native English-speaking students, through both descriptions of professional academic writing as well as through comparisons of novice writer (native and non-native English-speaking) and expert production. However, while learning about academic writing to better inform teaching content and practices is an important aim, Bazerman (1994) points out that understanding language use in the disciplines also helps us to use language more effectively, can guide editors as they work with contributor texts, and helps provide non-specialist readers with access to the discourse of the disciplines. Describing and understanding patterns of language use in academic prose allows us to understand the disciplinary cultures and practices that they embody for a variety of purposes. To do so, academic research articles have been the focus much of the research with this goal (although disciplinary differences in student writing is gaining in focus, particularly with the development of the Michigan Corpus of Upper-Level Student Papers and the British Academic Written

English corpus; for examples, see Hardy & Römer 2013; Römer & Swales 2010; Nesi & Gardner 2012; Gardner & Nesi 2012).

Disciplinary variation in academic writing is widely recognized, but we know less about the actual patterns of linguistic variation across disciplines than we know about variation across more broadly defined registers of academic writing. Becher (1981: 110) claims that "It is fairly obvious that disciplines differ from one another, but not so obvious what the differences comprise". Although Becher was not discussing the language patterns of different disciplines specifically, the research synthesis in Sections 1.3 and 1.4 below illustrate that the sentiment also applies to documenting the linguistic characteristics of texts across disciplines. Thus, one goal of this book is to elucidate systematic patterns of variation in the linguistic structure of research articles across disciplines.

However, the aims of the present research go well beyond establishing disciplinary differences in research writing. While the focus remains on research articles, part of my goal is to problematize the classification of 'research articles' as a wholly adequate register distinction for published research writing. This goal is partially motivated by the desire to acknowledge variation *within* academic disciplines, in addition to variation that corresponds to discipline-specific norms. Disciplines clearly differ in their basic characteristics, including data sources, areas of inquiry, research methods, and epistemological belief systems. However, variation with respect to these parameters also exists *within* disciplines. For example, many disciplines in the social sciences rely on both quantitative and qualitative research paradigms to conduct academic inquiry. Little research has systematically considered how differences in research paradigm (even within a discipline) might relate to distinct patterns in the linguistic structure of texts. Cao and Hu (2014: 17) claim that "the epistemological assumptions associated with quantitative and qualitative paradigms are believed to….shape the discourse and rhetorical conventions in which empirical research is presented". Their analysis of quantitative and qualitative research articles in three disciplines showed patterns of variation both across research paradigms and across disciplines – at least for the specific feature of interest, interactive metadiscourse markers. However, it is also likely that factors beyond discipline (such as research paradigm) contribute to variation in the use of many types of linguistic features in published research articles.

For illustration, consider the following two excerpts; both are from the results section of research articles in applied linguistics (both articles were published in the same journal, to avoid any differences due to the source journal). However, excerpt 3 reports on a *qualitative* research study, while excerpt 4 reports on a *quantitative* research study. Both are empirical studies, but rely on very different research methodologies and research belief systems.

Excerpt 3

> Ultimately, both the stimulated recall sessions and the talk-in-interaction sessions of both Groups 1 and 2 indicated that all students used the L to determine the meaning of the targeted forms…Groups 1 and 2 revealed four important differences (see Table 1). First, there was a difference in the fluidity of their interactions. Pairs in Group 1 engaged in smooth, continuous interaction. They talked while reading and reviewing the passage, and while discussing the target structures. By contrast, the interactions of pairs in Group 2 were characterized by frequent pauses and fragmented interaction. The students in Group 2 often laughed nervously and looked out the window during pauses, some of which went on for nearly 2 minutes. Although the majority of the students in Group 1 verbalized their thoughts in the L, the students in Group 2 who were trying to use only the L had to translate their L thoughts into the L in order to be able to share them with their conversation partners. [Qualitative applied linguistics, reporting results; Scott & de la Fuente 2010]

Excerpt 4

> The criterion for identification of linguistic variables was based on the components of the functional trisection: function, content, and accuracy (Higgs & Clifford, 1982). Table 2 and Figure 1 show results for the comparison of means t-tests for groups of nullgainers and gainers for three linguistic variables-grammar, accuracy, and vocabulary-in addition to two metalinguistic variables-self-corrected errors and sentence repairs. The score for the norm-referenced ACTR Qualifying Grammar Test, administered prior to the students' departure for Russia, is a classic variable based on domain-specific knowledge. The score reveals the percentage of correctly answered questions on the test. The scores for the group of nullgainers ranged from 36 to 78, and for the group of gainers, scores ranged from 47 to 80. Means for the nullgainers and gainers were 53.3 and 64, respectively. The result for the test of equality of group means was statistically significant at p =.042. [Quantitative applied linguistics, reporting results; Golonka 2006]

Despite coming from the same discipline, the same source journal, and even the same section of the research article, there are marked differences in the way that the outcomes of the research are being reported. The use of third person pronouns, past tense, and relative clauses create a rich, narrative-like description in excerpt 3. In contrast, excerpt 4 uses a mix of past and present tense, relatively few verbs, and many complex noun phrases – which all create a sense of succinct, procedural discourse. The potential for linguistic variation associated with factors such as research paradigm is quite likely, but as of yet these variables are largely unexplored as productive explanations of *within* discipline variation.

At the same time, very different disciplines may use similar research methodologies to explore phenomena in their respective fields. For example, both physics and philosophy rely on theoretical research to develop hypotheses and theories,

and build knowledge within the disciplines. While there are likely substantial linguistic differences due to the nature of the two disciplines, might there also be similarities due to the theoretical nature of the arguments being made? Thus, recognition of multiple types of research reports within disciplines, and how they relate to the range of research types in other disciplines, provides an opportunity to better explain the similarities and differences that can be observed *across* disciplines.

Thus, although substantial research has focused on disciplinary differences across research articles, little attention has been paid to the possibility that research articles themselves are not a monolithic concept. That is, substantial linguistic variation may exist even within the register of *research article* due to situational factors not previously considered, including the research design/methodology being reported, the nature of data, and the role of discipline. Thus, the primary goal of this book is to investigate the linguistic characteristics of registers published in academic journals both *across* and *within* disciplines, while taking into account the varied realizations of research articles (what I will call 'registers' or 'academic journal registers') in fundamentally diverse disciplines.

1.1.1 A note on 'register'

Terms like 'register,' 'genre,' 'text type,' and 'style' have long been used with a variety of meanings to categorize texts (for a comprehensive review on the uses of these terms, see Lee 2001; see also Biber 2006: 10–12). That is, linguists have sought for a way to group texts of a similar nature, and these terms have been used to describe categories of texts. The ability to create such groupings of texts has become particularly important as the field of corpus linguistics has grown. With increasing concern and awareness of the linguistic differences that exist between varieties of language has come the desire to systematically study that linguistic variation. This, in turn, results in the need for the ability to design both balanced and representative corpora that accurately characterize the language varieties under investigation.

While the term 'register' has been used widely by research coming from a variety of different theoretical frameworks, I use 'register' in the sense defined by Biber & Conrad (2009: see Chapter 1; also see Biber 1994, Biber & Finegan 1994: Chapter 1). That is, a register is a variety of language that can be characterized by the situations in which it is used. We can consider such situational factors like mode (written versus spoken), purpose (e.g., argumentative versus informative versus aesthetic), participants and the relationship between participants in a communicative event, and so on. Such factors are characteristics of the texts that are separate from the linguistic structure of those texts; however, the premise behind a

register perspective on language use is that these non-linguistic characteristics are correlated with, or associated with, the linguistic structure of those texts.

Registers can be defined at varying levels of specificity. For example, 'academic writing' is a commonly investigated register. However, within that broad register of academic writing, both textbooks and research articles can be identified as more narrowly defined registers (sometimes called 'sub-registers') under the umbrella term *academic writing*. The two registers share the overarching informational purpose of academic writing, but other non-linguistic features distinguish them from each other. For example, textbooks are intended to introduce the non-expert reader to a field, topic, or discipline, and typically cover a wide range of topics, and summarize established knowledge in the field. Research articles, on the other hand, present more focused, specialized information about new, developing knowledge in the field to expert readers.

Likewise, there is variation in the situational characteristics of texts *within* the register of 'research articles'. Research articles differ in terms of the discipline within which they fall, the journals they occur in, who their authors are, the types of methods that are used to conduct the research being reported on, and so on (see Chapters 2 and 3 for elaboration). Because of the ability to describe differences in the situational characteristics of different types of research articles, I use the terms 'register', 'sub-register', or 'academic journal register' to refer to the range of research article types. For ease of use, the majority of the time I will simply use the term 'register'. My motivation for taking this fine-grained approach to research articles is that in discussions to date (see discussion in Sections 1.2 and 1.3), research articles are generally grouped into a single register that is defined as texts which report on research and are published in academic journals. This rather coarse definition means that similarities across articles which report on distinct research methodologies and disciplines are assumed but not often investigated. In reality, the differing nature of research may be leading to substantial linguistic variation not being captured by the studies that have been conducted to date, as illustrated by the text excerpts presented above. In this book, I identify and examine sub-registers within the broader register of published academic research articles.

1.1.2 Goal of the present book

Thus, the goal of this book is to investigate the linguistic characteristics of registers published in academic journals, taking into account the varied realizations of research articles in fundamentally diverse disciplines. That is, the study seeks to go beyond the traditional and often one-dimensional analysis of a generically-defined research article to first distinguish between different types of articles (or registers)

within and across disciplines, and then to describe those registers according to their characteristic linguistic and non-linguistic features.

The research in this book is based on the Academic Journal Register Corpus, a corpus of 270 research articles from six disciplines: philosophy, history, political science, applied linguistics, biology, and physics. Research articles within these disciplines are further categorized by journal register: theoretical, qualitative, and quantitative research articles. This volume presents analyses of both the non-linguistic and linguistic characteristics of the corpus. In the analysis of the non-linguistic features of these texts, a framework for describing the situational characteristics is first proposed and then applied to each text in the corpus. The linguistic analysis relies on quantitative and qualitative analyses of data extracted through several specialized computer programs, and includes a grammatical survey of the distributions of core grammatical features, a description of grammatical features which 'elaborate' and 'compress' discourse, and a statistical analysis that identifies co-occurrence patterns of 70 linguistic features.

1.2 The linguistic characteristics of academic writing

There is a general consensus, even outside the academic community, that academic writing has distinct characteristics that set it apart from other types of language. Much research (e.g., Biber 1988; Biber et al. 1999; Biber & Gray 2010, 2016; Halliday 2004; Banks 2005, 2008; Fang, Schleppegrell & Cox 2006) has focused on describing one of the defining characteristics of academic prose: its dense reliance on nouns and noun phrase structures. This nominal style contrasts with the structure of, for example, conversation, which relies on the use of more verbs and clausal structures (see Biber 1988; Biber & Gray 2010, 2016). Halliday's work on scientific writing has focused on describing 'grammatical metaphor' (see Halliday 2004 for a collection of key works on nominalization and grammatical metaphor in science writing), whereby processes and actions typically expressed with verbs are nominalized, that is, verbs are changed to nouns through a process of grammatical metaphor.

Much of Biber's work has empirically documented the nominal style of academic writing using large-scale corpus analyses that compare academic writing to other general registers like conversation, fiction, and newspaper writing. In a multi-dimensional analysis using factor analysis to identify how 67 linguistic features co-occur on a statistical basis, Biber (1988) shows that academic writing commonly uses features associated with an 'informational' purpose, such as nouns, prepositions, and attributive adjectives, while relying much less on linguistic features associated with the 'involved' nature of conversation such as private

verbs, *that*-deletions, personal pronouns, WH-questions, modals, and WH-clauses among others (see Biber 1988: Chapter 6).[1]

In perhaps the most comprehensive descriptive reference grammar to date, Biber et al. (1999) describe the distributions of a full range of lexical and grammatical structures in English in *The Longman Grammar of Spoken and Written English (LGSWE)*, comparing academic writing, conversation, newspaper writing, and fiction. In the *LGSWE*, academic prose is represented by extracts from academic books and research articles in 13 disciplines ranging from agriculture to computing to mathematics to sociology (see Biber et al. 1999: 32–33). Although a full summary of the characteristics of academic prose is not possible here, Biber (2006: Chapter 1) provides a comprehensive summary of grammatical features that occur particularly frequently in academic prose based on the *LGSWE*. In particular, Biber (2006: 15–18) notes quite a few grammatical features associated with noun phrases that are either most common or very common in academic prose when compared with other registers, including the overall use of nouns, nouns as pre-modifiers, nominalizations, adjectives, prepositions, *of*-phrases, relative clauses with *which*, and noun post-modifying participle clauses.

More recently, Biber and colleagues have focused on expanding this analysis of features which contribute to the nominal style of academic writing, both synchronically and diachronically (see Biber & Gray 2010, 2016; Biber, Gray & Poonpon 2011; Biber et al. 2011). In particular, this research has documented the prevalence of nouns and phrasal modifiers in academic writing, such as attributive adjectives, nouns as nominal premodifiers, and prepositional phrases as post-modifiers. Consider the following two examples from a quantitative biology article and a quantitative research article in applied linguistics (head nouns with pre- or post-modification are **bolded**, phrasal post-modifiers are underlined, and adjectives and nouns as nominal premodifiers are *italicized)*:

1.1 Given their **importance** in the **functioning** of *arid* and *semiarid* **ecosystems,** restoring these crusts may contribute to the **recovery** of *ecosystem* **functionality** in *degraded* **areas**. [BIO-QT-12]

1.2 The *main* **aim** of conducting this study was to investigate the *foreign language learning* **needs**, **wants** and **desires** of *undergraduate medical sciences* **students** studying in **faculties** of nursing and midwifery in *various* **universities** in Iran. [AL-QT-10]

1. In fact, 'Dimension 1' of the 1988 MD analysis showed a clear cline of variation that distinguished between written and spoken registers generally. However, academic prose had one of the lowest dimensions scores, indicating its high use of informational features and low use of interactional features.

These two sentences illustrate the dense embedding of nominal modifiers within noun phrases, resulting in condensed structures in which a head noun is modified, often multiple times, in order to pack a great deal of information into a few noun phrases. This style is primarily restricted to academic prose, and Biber and colleagues connect this nominal style to the informational purpose and highly specialized readership of academic prose.

While the research briefly summarized above has focused on describing academic writing without direct consideration of disciplinary differences, a great deal of research has also concentrated on describing the linguistic characteristics of writing in the disciplines, summarized in the next section.

1.3 Linguistic variation and disciplinary writing

Across language-related fields, increasing attention has been being paid to language use in relation to specific disciplinary practices. Movements such as Writing across the Curriculum (WAC; see Russell 1991 for a comprehensive history) have brought attention to the importance of teaching writing within specific curricular areas and disciplines, rather than independent of specific content areas and discourse communities. Alongside the rise of WAC, awareness that language use varies in meaningful ways across disciplines has also grown. Because of this awareness, studies investigating disciplinary language use have flourished, and a look at any journal focusing on English for Academic Purposes (EAP) will reveal a large body of research about language use in the disciplines from a variety of perspectives and research methodologies.

One approach to studying the linguistic characteristics of writing in specific disciplines is to focus on one register in one discipline – to provide detailed descriptions of a particular type of text within the context of a specific disciplinary community. However, the majority of studies concerned with discipline-specific language take a comparative approach to describing the linguistic characteristics of disciplines and/or registers. These studies can be categorized into two major types: (1) those that compare two registers within a discipline or disciplines, and (2) those that compare the same register in multiple disciplines.

As with the treatment of discipline in academic writing research, the registers under investigation also vary in level of specificity. Some studies group several registers together to represent academic language use. For example, Fuertes-Olivera (2007) examines lexical gender in a variety of registers in business, including research articles, product descriptions, political speeches, and governmental reports. Others compare and contrast two or more registers as Koutsantoni (2006) does in an investigation of hedges in research articles and student theses. A great

majority of investigations focus on one particular register. Although research articles are by far the most widely-studied register, other written academic registers that have been studied include textbooks (Biber, Conrad & Cortes 2004; Conrad 1996a; Freddi 2005; Moore 2002), Ph.D. or master's theses/dissertations (Bunton 2005; Charles 2006a, 2006b; Samraj 2008), peer review reports (Fortanet 2008), 'comment' articles (Lewin 2005), book reviews (Groom 2005), and business reports (Yeung 2007).

Many studies also explore the use of linguistic features within specific sections of the target registers. For example, Chen and Ge (2007) examine the distributions of Academic Word List (AWL) words in different sections of research articles. Martínez (2003, 2005) investigates thematic structure and 1st person pronouns in a corpus of biology articles, comparing uses across different sections in the articles. Samraj (2005) looks at the rhetorical structure of research article abstracts and introductions.

While a great deal of the research on academic writing focuses on describing linguistic phenomenon within a genre or single discipline, a smaller body of literature investigates differences in similar genres across disciplines. Since the purpose of this book is to describe disciplinary differences, I will explore these studies in a bit more detail and examine the varying types of linguistic structures that are focused on in these studies.

Aspects of lexis are one feature that is investigated in academic writing, including lexical bundles, keyword analysis, and collocational analysis. Cortes (2004) examines four-word lexical bundles in research articles from history and biology, finding that history articles employed bundles out of the two structural groups noun phrases and prepositional phrases while biology articles employed a much wider range of structures. Cortes also finds that despite using different lexical bundles, both disciplines employed bundles for similar functions.

In her analysis of keywords in introduction chapters in applied linguistics textbooks, Freddi (2005) finds that when compared to a reference corpus, applied linguistics textbook introductions use words that represent processes, logical connections, and interpersonal relationships. Freddi interprets this to be a way that the authors create relationships with the readers of the textbooks and position themselves as a teacher or researcher and those reading the textbook as students.

A second type of analysis investigates grammatical structures. Groom (2005) is an example of such a study, investigating extraposed adjective-controlled *to*- and *that*-clauses in research articles and book reviews in history and literary criticism. Groom finds that the phraseological patterns of these structures differ across the two genres and disciplines. In fact, *that* complement clauses are an often-studied grammatical structure in academic prose (Parkinson 2013; Charles 2006a, 2006b, 2007; Hyland & Tse 2005). *That* complement clauses are productive structures in

academic prose, as they are integral to citing others' work and presenting claims, and are often indicators of an author's stance. Charles (2006b) examines how writers of theses in politics and materials science use verb-controlled *that* clauses to report information, considering the subjects of the clauses (including *it*) and the use of passive voice. Charles's study illustrates a key trend in the investigation of grammatical constructs in academic prose: when investigating grammatical features, researchers typically consider many aspects of the linguistic context and make connections to the specific purposes that the grammatical structures fulfill in academic prose.

In fact, relatively few studies examine single grammatical features. Rather, researchers focus on a collection of lexical and grammatical features that work together to create some type of functional result. For example, stance (as marked by a collection of lexical and grammatical markers) is a widely-researched topic in academic writing. Hyland (1998) compares stance markers in eight disciplines, finding that disciplines in the humanities/social sciences exhibited nearly 2.5 times as many stance markers than the sciences.

Afros and Schryer (2009) also use a combination of lexical and grammatical features to investigate a single construct: self-promotion through the use of lexical items, coordination, comment clauses, personal pronouns, and lexical cohesion. Afros and Schryer found that literary scholars relied on intensifiers to persuade readers to believe their claims, while linguistics scholars relied on self-citation.

Hyland (2002a) and Swales, Ahmad, Chang, Chavez, Dressen, and Seymour (1998) study the use of commands (termed directives and imperatives respectively) in academic research articles. While Swales et al. identify commands based on the criteria that a main verb or emphatic *do* occurs in the base form with no modals and no surface subject, Hyland identifies commands based on a set of 80 lexical search terms (it should be noted that Hyland uses the results from Swales et al. to choose his search terms). Both Hyland and Swales et al. find discipline-specific differences in the use of commands. For example, Hyland finds that 'hard' disciplines (such as mechanical engineering and physics) used many more directives that were intended to direct the reader through procedures and conclusions than the 'soft' disciplines (such as philosophy). Swales et al. find that fields relying on mathematical reasoning employ more commands. These two studies illustrate that, to this point, much of the research that has investigated more than one or two disciplines has interpreted results in terms of categories of disciplines, such as 'hard' or 'soft' disciplines, and disciplinary variation seems to follow generally along those lines.

A final thread found in research on disciplinary writing examines the organizational structure of genres, many times by employing a move analysis. Because of the primarily qualitative nature of move analysis, many studies focus on only

one discipline (e.g., Nwogu's 1997 description of medical research articles), and others investigate only a specific section of the texts. For example, Ozturk (2007) uses Swales's (1990) CARS model to describe introductions in research articles in applied linguistics, while Basturkmen (2012) analyzes discussion sections of dentistry research articles.

However, a few of these move analysis studies do compare disciplines. For example, Stoller and Robinson (2013) compare the rhetorical structure of chemistry and biochemistry articles. Samraj (2005) conducts a move analysis of abstracts and introductions in conservation biology and wildlife behavior, finding that while abstracts and introductions in conservation biology share similar functions and organizations, the two genres in wildlife behavior are not as similar. Holmes (1997) compares discussion sections in articles in history, political science, and sociology. He finds that these three social sciences have similarities and differences when compared to previous findings about discussion sections in natural science research articles, and that the three social sciences disciplines varied amongst themselves as well. Holmes concludes, however, that the political science and sociology discussion sections were sufficiently similar to natural science discussion sections to call them the same genre, with history texts showing much more variation (such as being brief and not containing a cyclical structure). Because of the qualitative nature of move analysis, it has not typically been applied to large-scale corpus-based studies until a recent volume by Biber, Connor, and Upton (2007) and the dissertation by Kanoksilapatham (2005b). However, corpus-based research on move structure is on the rise, and researchers are increasingly paying attention to specific lexical and grammatical features as they are associated with specific rhetorical moves (e.g., Cortes 2013 on lexical bundles associated with moves in RA introductions).

As this summary of research shows, research articles are perhaps the most commonly researched register within academic writing (although there has also been a great deal of research into general academic writing by L2 speakers, as well as novice L1 writers. For examples, see Parkinson & Musgrave 2014; Callies 2013; Grant & Ginther 2000; Green, Christopher & Mei 2000; Hardy & Römer 2013; Hinkel 2003; Jarvis, Grant, Bikowski & Ferris 2003; Schleppegrell 1996; Spycher 2007; Altenberg & Granger 2001; Archer 2008; Flowerdew 2006; Loudermilk 2007). Table 1.1 below summarizes studies on academic research articles, along with the linguistic features investigated in the studies.[2] Table 1.1 illustrates

2. Because the current study focuses on the distribution of lexical and grammatical features of research articles, I have excluded the large body of research using a Swales-inspired moves analysis from this summary table. Interested readers, however, can see the following

that a wide variety of linguistic features are investigated in these studies, and research articles are often compared to either other registers within the realm of academic writing, to novice writer academic writing, or to other general registers like conversation.

As can be seen in the brief literature review above and in Table 1.1, academic writing has been widely researched. In this broad research base, research articles are often compared to student-produced registers (e.g., theses and dissertations written by advanced graduate students, 2nd language writers at a variety of levels), as well as to other sub-registers within academic writing such as textbooks and academic lectures (and less frequently to registers like book reviews and editorials). About as often, however, research articles are not compared to other registers at all.

In terms of the linguistic features that are focused on this research, Table 1.1 documents that research articles are often analyzed for their use of functionally-related lexical and grammatical features. This is illustrated by the fact that much of the terminology used to describe the linguistic features of interest represents functional constructs rather than specific lexical or grammatical structures. For example, a great deal of research has focused on the ways in which authors encode values and judgments in their texts under terms such as hedging, stance, boosting, appraisal, and evaluation. Other examples that illustrate this focus on functional constructs include metadiscourse, citation, self-mention, argument structure, naming conventions, new knowledge claims, and so on.

1.4 Trends and gaps in the study of disciplinary writing

The literature review presented in Section 1.3 illustrates the diverse nature of linguistic studies of academic writing, and of academic research articles more specifically. Despite the large foundation of research into disciplinary writing, it is difficult to arrive at a comprehensive description of disciplinary variation. More specifically, the following trends emerge from this review. First, most large-scale investigations that consider a wide range of linguistic features focus on comparing academic prose in general with other broadly-defined registers like conversation or news. While these comparisons are inherently interesting and useful, providing

---

for moves analyses of research articles: Koutsantoni (2006), Basturkmen (2009), Bhatia (1997), Brett (1994), Bruce (2008, 2009), Bunton (2005), Holmes (1997), Kanoksilapatham (2005a,b), Lim (2006), Lin & Evans (2012); Ozturk (2007), Ruiying & Allison (2004), Samraj (2002, 2004, 2005), Stoller & Robinson (2013).

Table 1.1. Summary of studies investigating language use in academic research articles

Study	Disciplines	Registers Compared to Research Articles	Linguistic Features
Afros & Schryer (2009)	language and literary studies	–	promotional metadiscourse (moves, lexico-grammatical markers)
Aktas & Cortes (2008)	art & design, computer science, economics, environmental engineering, physics, astronomy	L2 graduate-level academic writing	shell nouns
Biber & Gray (2010)	biology, medicine, ecology, physiology, education, psychology, history	conversation	grammatical features of complexity and elaboration
Biber & Gray (2013)	science vs. humanities	–	grammatical features of complexity and elaboration
Biber, Csomay, Jones & Keck (2004)	various	academic lectures, university textbooks	vocabulary-based discourse units
Chen & Ge (2007)	medicine	–	AWL word families
Conrad (1996b)	ecology	composition textbooks, ecology textbooks	various
Cortes (2013)	13 disciplines	RA introductions	associating lexical bundles with discourse moves
Dahl (2008)	economics, linguistics	–	new knowledge claims
Diani (2008)	linguistics, history, economics	academic lectures, book reviews	emphasizer *really*
Dueñas (2007)	business management	–	self-mentions and citations

(*Continued*)

Table 1.1. (Continued) Summary of studies investigating language use in academic research articles

Study	Disciplines	Registers Compared to Research Articles	Linguistic Features
Fang, Schleppegrell & Cox (2006)	various	various	nouns
Feng & Hu (2014)	applied linguistics, education, psychology	quantitative vs. qualitative RAs (post-method sections)	metadiscourse
Gillaerts & Van de Velde (2010)	applied linguistics (pragmatics)	**RA abstracts only	interactional metadiscourse (hedges, boosters, attitude markers)
Gosden (1992)	science	–	marked themes
Gray (2010)	education, sociology	–	demonstrative pronouns and determiners; shell nouns
Groom (2005)	history, literary criticism	book reviews	it + V-ling + ADJ + that/to
Harwood (2005a)	computer science	student writing (project reports, MA theses)	first person pronouns
Harwood (2005b)	business management, computing science, economics, physics	–	first person pronouns
Hemais (2001)	marketing	–	sentence subjects, citations, reporting verbs
Hewings & Hewings (2002)	business	MA dissertations	anticipatory it, extraposed subjects
Hewings, Lillis & Vladimirou (2010)	psychology	–	citations
Hyland & Tse (2007)	sciences, engineering, social sciences	textbooks, book reviews, letters, MA & Ph.D. theses, student projects	academic vocabulary (AWL)
Hyland (1996)	molecular biology	–	hedging
Hyland (1999a)	microbiology, marketing, applied linguistics	textbooks	metadiscourse

(*Continued*)

Table 1.1. (Continued)

Study	Disciplines	Registers Compared to Research Articles	Linguistic Features
Hyland (1999b)	philosophy, sociology, applied linguistics, physics, electrical engineering, marketing, mechanical engineering, biology	–	citation & attribution
Hyland (2001a)	8 disciplines (see Hyland 1999b)	–	2nd person pronouns, inclusive pronouns, questions, directives, etc.
Hyland (2001b)	8 disciplines (ibid.)	–	self-mention
Hyland (2002a)	8 disciplines (ibid.)	–	directives
Hyland (2002b)	8 disciplines (ibid.)	L2 academic writing	authorial identity
Hyland (2007)	8 disciplines (ibid.)	–	exemplifying and reformulating (elaboration)
Hyland (2008)	electrical engineering, biology, business, applied linguistics	MA & Ph.D. theses	lexical bundles
Hyland (2010)	science and engineering	popular science articles	proximity, argument structure
Khedri, Heng, & Ebrahimi (2013)	applied linguistics, economics	research article abstracts only	metadiscourse
Koutsantoni (2004)	electrical engineering	–	appraisal
Koutsantoni (2006)	electrical engineering	research theses	stance
Kuo (1999)	science	–	personal pronouns
Kwan (2012)	2 areas within Information Systems	–	strategies for making evaluations within specific moves in the CARS model
Lee & Chen (2009)	linguistics/applied linguistics	dissertations, student written assignments	keywords
Marco (2000)	medicine	–	collocational frameworks
Martínez (2003)	physical science, biological science, social science	book chapters	finite clauses, transitivity structures

(*Continued*)

Table 1.1. (Continued) Summary of studies investigating language use in academic research articles

Study	Disciplines	Registers Compared to Research Articles	Linguistic Features
Martínez (2005)	biology	NNES manuscripts	first person pronouns
Martínez, Beck & Panza (2009)	agriculture	–	vocabulary
Norman (2003)	bio-medical	RA abstracts only	naming conventions
Parkinson (2013)	social sciences	across sections of RAs	*that*-complement clauses
Peacock (2006)	business, language and linguistics, physics, administration, law, environmental science	–	boosting
Salager-Meyer (1994)	medicine	–	hedges
Stotesbury (2003)	humanities, social sciences, natural sciences	RA abstracts only	evaluation
Swales et al. (1998)	philosophy, sociology, applied linguistics, physics, electrical engineering, marketing, mechanical engineering, biology	–	imperatives
Tarone, Dwyer, Gillette, Icke (1998)	astrophysics	–	passive and active voice
Thomas & Hawes (1994)	medicine	–	reporting verbs
Tucker (2003)	art history	–	evaluation
Vongpumivitch, Huang & Change (2009)	applied linguistics	–	use of AWL words
Warchal (2010)	linguistics	–	conditional clauses
Webber (1994)	medicine	editorials, letters	questions
Williams (1996)	medicine	–	lexical verb use

much-needed accounts of variation in language use broadly, they do not consider disciplinary differences in their research design. As a result, these studies cannot inform discipline-specific language instruction or help us explore the inner workings of disciplinary cultures.

Second, smaller-scale investigations into single disciplines are common. While useful for describing particular disciplines, investigations of single disciplines provide little contrastive information about how the features of interest might vary across disciplines. Also common are studies comparing a small number of disciplines; the comparative approach taken in these studies allows for more contrastive information about how disciplines might vary. However, with both of these lines of inquiry, it's difficult to arrive at a comprehensive picture of disciplinary variation, as replication studies that employ directly comparable methodologies in different disciplines are carried out relatively infrequently.

Third, while an abundance of literature describes research articles, two main problems arise from this trend. Giving preference to research articles above other types of articles published in academic journals ignores many other modes of the transmission of knowledge in academic disciplines and limits our knowledge of the discourse practices within disciplines. This is an important consideration, and one that will need to be further addressed in linguistic research of academic writing in the future, particularly as these studies of research articles often serve the basis for corpus-informed writing instruction.

For the purposes of the present book, a second problem that arises from this focus on research articles is of interest: research articles are generally not a finely-defined register in that similarities across articles which report on distinct research methodologies and disciplines are assumed but not often investigated. In reality, the differing nature of research may be leading to substantial linguistic variation not being captured by the studies that have been conducted to date.

A limited body of research has acknowledged the presence of registers other than research articles in academic journals: Magnet & Carnet (2006), Flowerdew & Dudley-Evans (2002), and Giannoni (2008) on editorials/letters to the editors in academic journals; Fortanet (2008) on evaluative language in peer review referee reports in applied linguistics and business. However, these studies consider types of publications in journals that are not primarily research reports; very little research considers differences in *types* of research reports. Instead, analyses of research reports generally consider all research reports to be a single register, declining to distinguish between articles which report on, for example, qualitative versus quantitative research, or case study research versus survey research, and so on.

Despite the lack of research investigating the linguistic differences in these areas, there is an implicit recognition that such differences exist, and that novice writers may struggle with the characteristics of a particular register. For example, Belcher and Hirvela (2005) focus on the writers of qualitative doctoral dissertations, looking at their motivations and commitments to what is seen as a challenging register for L2 writers, and implicitly acknowledging that

writing up qualitative research is inherently different than writing up quantitative research. Yet, there has been little systematic inquiry into these differences, leading to a lack of empirical evidence of the linguistic characteristics of these various registers.

A few studies have acknowledged different types of research articles, but have primarily used these distinctions to restrict the analysis that is undertaken, or to inform corpus design. For example, Williams (1996) analyzes the use of lexical verbs in clinical and experimental medical research articles. Despite a general claim that descriptions of different types of medical reports are needed for comprehensively informing ESP materials, Williams does not offer explanations of how the two types of reports differ in terms of their non-linguistic characteristics. Likewise, Vande Kopple (1994) restricts his study of complex subjects to experimental science articles, acknowledging that the same patterns may not be found in theoretical science articles (but does not carry out an empirical comparison to test that hypothesis).

There are two recent exceptions to this trend, and the results of both studies indicate that making distinctions between types of research reports is an area where more research is warranted. Cao and Hu (2014) compare "post-method" sections of research articles in three disciplines: applied linguistics, education, and psychology, directly contrasting quantitative and qualitative research articles in the three disciplines. Cao and Hu (2014: 16) find that the metadiscourse markers employed in the research articles in their corpora vary in ways that can be associated with both disciplinary factors, as well as with "the contrasting epistemologies underlying the qualitative and quantitative research paradigms".

Kwan, Chan, and Lam (2012) also suggest that research paradigm correlates to differences in the discourse style of published academic research articles. Taking a slightly different approach, Kwan et al. contrast two sub-disciplines within the field of Information Systems, comparing journals which publish research following a behavioral science research paradigm with a design science research paradigm. In their analysis of the rhetorical strategies that authors use to evaluate previous research while carrying out moves from Swales' CARS model, Kwan et al. find that authors of research in the two paradigms rely on different types of evaluation strategies. They interpret these patterns as reflecting the epistemological orientations of the two different paradigms. In summary, while few studies have considered research method or epistemologies as influences on language variation in research article, what little research exists has uncovered notable patterns in the language and discourse styles.

A final comment that can be made based on the summary of diverse studies above in Table 1.1 is that the research agenda into academic research articles has been rather piecemeal, or has tended to focus on a few narrow areas

(e.g., metadiscourse, stance, moves or rhetorical structure). It is difficult to synthesize large-scale findings on disciplinary variation based on this research, and little research has focused on whether there is variation in the core lexical and grammatical structure of writing across disciplines. In other words, while this research has provided valuable and detailed insights into particular registers and particular disciplines, it is difficult to get an overall picture of the wide range of linguistic variation that is occurring in research articles in different academic disciplines. Thus, the goal of the present book is to take a wide variety of core linguistic features and analyze them systematically across 6 disciplines (philosophy, history, political science, applied linguistics, biology, and physics). To address the three major gaps presented here, I first develop a framework for identifying different academic journal registers, including different types of research articles, across many disciplines. I then carry out a large-scale linguistic analysis to compare and contrast research articles within and across disciplines, creating rich descriptions of a variety of registers and disciplines within academic journal writing.

1.5 Overview of the book: Applying corpus analytical approaches to disciplinary register variation

Corpus linguistics methodologies are useful for investigating differences across disciplines and registers, in part because of the relative ease with which comparisons across varieties can be made using the quantitative data that result from corpus analysis. The research undertaken in this book aims to reveal patterns of linguistic variation within and across disciplines through a consideration of the varied nature of research articles in six disciplines. However, before any corpus-based linguistic analyses can be carried out, work has to be done to identify the range of possible journal registers, select disciplines to be included in the analysis, and determine criteria by which texts can be reliably classified into sub-corpora representing the various registers. That is, non-linguistic analyses must be carried out to inform the design of a corpus that can be used to reliably represent the varieties of interest. This type of non-linguistic analysis is a crucial stage in addressing the issue of external or situational representativeness (McEnery et al. 2006; Biber 1993) in corpus design: the extent to which a corpus is designed to represent "the range of text types in the target population" (Biber 1993: 243).

Furthermore, the registers under investigation need to be described in detail for their purposes, topics, authorship, and so on in order to help explain the patterns of language use that are uncovered during the analysis stage. All of these tasks can be accomplished by undertaking analyses of the non-linguistic, or

situational characteristics, of (a) the target registers at the corpus design stage, and (b) the texts in the resulting corpus. While the situational analysis of the target register can inform the corpus design and increase the probability that external representativeness can be achieved, a situational analysis of the texts in the resulting corpus enables the empirical evaluation of that representativeness, and is needed for interpreting the patterns of variation that are discovered in that corpus.

As a result, the book includes two main types of analyses: (1) analyses of the non-linguistic characteristics of the broad domain of academic journal writing, and of the texts in the corpus more specifically; and (2) linguistic analyses of the patterns of variation in the corpora. Figure 1.1 visually represents the various components of the research being reported in the present book, illustrating the relationship between the individual analyses and the sources of information for these analyses. In the figure, squares represent non-linguistic analyses or sources of information, with sources differentiated from analyses with a dashed outline. Ovals indicate the linguistic analyses carried out in the study. Dashed arrows indicate where information from an analysis/source has informed another component of the study in some way.

Figure 1.1 shows that the study is divided between analyses and information that comes from the world external to the corpus, and analyses that are based on the corpus collected for the study. The analyses of and sources from the broader context outside the corpus all inform the design of the corpus in some way. The main analysis here is a survey of 11 disciplines that identifies ten journal registers and documents how frequently the 11 disciplines publish those registers. This survey, detailed in Chapter 2, is used to inform the choice of registers and disciplines for the Academic Journal Register Corpus, as well as to form the basis for the creation of operational definitions that can be applied to classify texts into the journal registers. However, this survey is not the only criteria in determining the corpus design; meetings with experts within the fields of interest play an important role in refining the general operational definitions for use during corpus collection.

The placement of the circle encompassing corpus design is symbolic of the bridging role that a corpus plays in linguistic research. While the corpus is the basis for the linguistic analyses, it is also intended as a representation of the larger world outside the corpus. That is, it is intended to reflect a type of language that exists in the larger context.

Once the corpus is constructed based on input from the various external analyses and sources, corpus-based research can take place. Going back to Figure 1.1, we can see that the study encompasses one corpus-based situational analysis. The corpus-based situational analysis draws on the analyses and sources from the wider situational context to identify specific features that can be used to characterize the texts (and in turn the registers) in the corpus. Two analyses of the situational

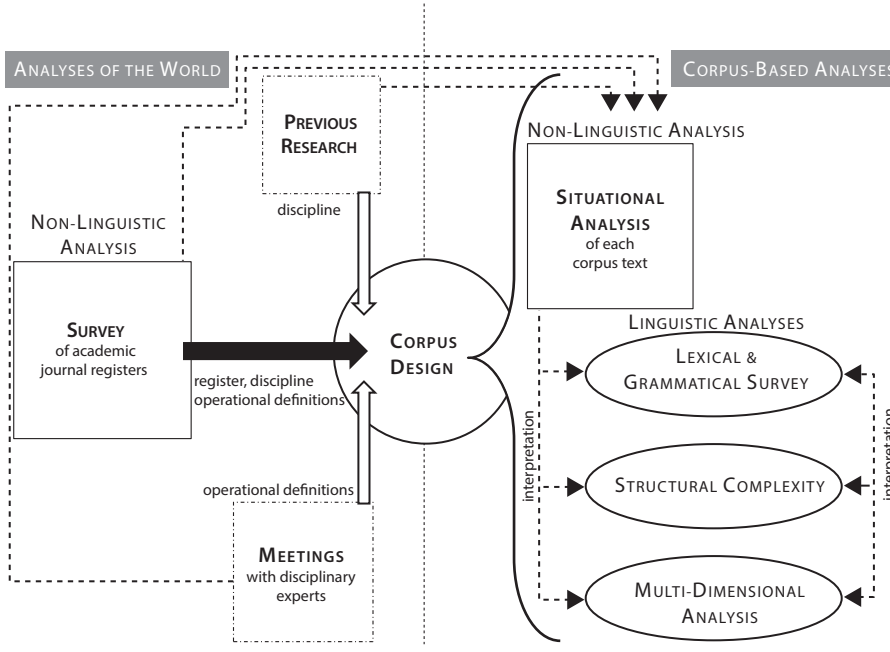


Figure 1.1. Overview of the study: Relationships between and among linguistic and non-linguistic analyses

characteristics of the registers take very different approaches. In the first approach (the survey reported in Chapter 2), the purpose is to describe the broad domain of academic journal publishing with the intent of informing the corpus design and understanding how the registers differ. This approach uses a survey method that focuses on the holistic context of journal publishing. In the second approach, the purpose is to describe the non-linguistic characteristics of the corpus texts themselves. The second approach accordingly employs a corpus-based method in which a framework for describing these non-linguistic (or situational) characteristics is applied to each text in the corpus. It is important to note that these two analyses are not completely separate; rather, information gained during the survey of the domain of academic journal writing informs the development of the framework used in the corpus-based situational analysis of the corpus texts (described in Chapter 4).The situational analyses reported in this book are conducted systematically and empirically, and serve as the foundation for the corpus compilation and linguistic analyses. Many register-based studies, and studies based on academic writing more specifically, begin with pre-conceived, broad definitions of the registers of interest. Some studies survey the field to inform corpus composition, while others analyze certain aspects of the corpus after compilation is complete (or even

after analyses are completed, as a way of interpreting the linguistic results). However, very few studies conduct both types of analyses. Moreover, situational analyses are often applied on the 'register' level. That is, a target register is described in terms of its typical characteristics, with little attention being paid to the individual texts and what their non-linguistic characteristics are. This typically results in registers being described fairly broadly (e.g., informational purpose of academic writing vs. interpersonal relationship management for conversation).

Thus, the use of two situational analyses, one to inform corpus design and one to describe the corpus composition, distinguishes the research reported on in the present book. On one hand, these situational analyses and sources of information help ensure and document that the corpus has 'external' or 'situational' representativeness – that the corpus texts represent the types of texts found in the target domain (see McEnery et al. 2006; Biber 1993). On the other hand, the situational analyses provide crucial information needed in order to interpret the patterns of linguistic variation that are subsequently uncovered in the corpus.

Figure 1.1 illustrates this relationship between situational analyses and corpus design and linguistic analysis. The rich description of the texts that results from the situational analysis is applied within the linguistic analyses to help interpret and explain the uses and distributions of the linguistic devices of interest, as indicated by the dashed arrows in Figure 1.3. Three linguistic analyses are carried out in Chapters 5 through 7.

Chapter 5 provides a descriptive overview of the general grammatical characteristics of academic writing across disciplinary journal registers. The perspective is strictly lexical and grammatical, focusing on the distribution of different word classes and the most commonly occurring classes of lexical items. Chapter 6 moves on to an examination of structural complexity, with features being linked to functions of compressing and elaborating information (based on Biber & Gray 2010; Biber, Gray & Poonpon 2011). This analysis is lexico-grammatical in nature, as the units being analyzed are identified based on a combination of grammatical patterns and frequent lexical items that occur in those grammatical contexts.

Chapter 7 presents a comprehensive register description. Using the multidimensional analysis methodology developed by Biber (1988, 1995), this study considers the co-occurrence patterns of 70 lexical and grammatical features through the statistical technique of exploratory factor analysis. Once the patterns in the use of linguistic features have been determined statistically, those patterns are interpreted according to the functions that those groupings of linguistic features carry out. Biber (2010) describes MD studies as uncovering 'dimensions' of variation as previously-unrecognized linguistic constructs that emerge from the inductive analysis of quantitative patterns in the corpus (see Biber 2010: 179).

In addition to using the situational analyses to interpret the quantitative trends found in the linguistic analyses, the findings from the linguistic analyses themselves can complement one another. In particular, the more narrowly-focused analyses on specific features in Chapters 5 and 6 are particularly influential in interpreting the co-occurrence patterns of those same features as they are uncovered in the multi-dimensional analysis (Chapter 7). Chapter 8 then brings together the results of both the situational and linguistic analyses, to synthesize the major patterns of disciplinary and register variation, and to discuss the implications of these findings for research and disciplinary variation in academic writing.

CHAPTER 2

Describing the domain of academic journal writing

2.1 Introduction

A great deal of research on academic writing acknowledges that different disciplines use language in different ways. This is evidenced by the abundance of research that compares disciplines (as in many of the studies mentioned in Chapter 2), and by the many "how to" books like Zeiger's (1999) *Essentials of Writing Biomedical Research Papers* or Robinson, Stoller, Costanza-Robinson, and Jones's (2008) volume *Write like a Chemist* that focus on offering detailed descriptions within a single discipline. This type of research operates on the belief that each discipline follows its own discourse conventions and patterns.

However, the summary of literature focusing specifically on research articles published in academic journals (Chapter 1) has shown that very little research accounts for variation in the *types* of journal articles published both within and across disciplines. As a consequence, almost no research has documented linguistic variation that occurs due to differences in article type.[1] In using the term 'article type,' I put forth the claim that research articles can be described according to situational characteristics that go beyond the principal features of being written by professionals and experts within a field, published in professional journals, with an informational purpose of relating the results of research. For example, we can also characterize research articles based on research methodologies and ask questions that link these specific characteristics of articles with linguistic patterns, such as 'do empirical studies which report on qualitative research use the same linguistic features to the same extent as studies which report on quantitative research?' By explicitly considering the types of articles that are produced within and across

1. This is not to say that this body of research does not recognize different registers within an academic discipline. For example, *Write like a Chemist* considers the written registers of journal articles, conference abstracts, posters, and research proposals. Rather, these research studies and how-to volumes typically do not consider differences in the *types* of journal articles, or variation within journal articles.

disciplines, linguistic analyses can both describe important patterns of variation that have been unidentified in the research to date, as well as offer more comprehensive explanations of variation that go beyond simple accounts of disciplinary differences.

However, in order to analyze linguistic variation across article types, the first step is to carry out an analysis of the target domain, to identify the range of possible article types and describe the non-linguistic (i.e., situational) features that distinguish different types of research articles. The basic premise here is that article types differ in terms of certain characteristics, and these differences allow us to consider sub-registers within the texts published in academic journals. Because the goal of this study is to break away from the traditional and somewhat monolithic term 'research article,' the purpose at this stage of the research is to survey academic journals in many disciplines in order to document the full range of publication types (including non-empirical research) that occur in these journals, along with the basic characteristics that define these article types. The idea is to situate the often-studied 'research article' within the larger context of academic journal publishing, as well as identify systematic categories within 'research articles'. To do so, a careful consideration of what features identify a particular article as belonging to a certain type is also needed, as the ultimate goal is to represent these article types in a corpus. Thus, the first objective of this chapter (Section 2.2) is to establish a taxonomy that identifies the range of possible text types (registers), and enables the categorization of academic journal articles into different registers according to situational characteristics (such as the type of research that is being reported on).

This target domain description serves as the foundation for corpus design, helping to ensure that a corpus is representative. Two types of corpus representativeness are important relative to corpus design and the interpretation of corpus-based findings. External or situational representativeness is the most commonly addressed type. The second type is internal or linguistic representativeness (McEnery et al. 2006; Biber 1993), and refers to the stability of the linguistic findings from the corpus: whether the corpus represents the range of linguistic variability in the target domain (Biber 1993: 243). Biber (1993) argues that internal/linguistic reliability only occurs if external representativeness has been achieved; thus, this type of situational analysis that describes the target register and informs corpus design is a much-needed step in attaining corpus representativeness – both external/situational and internal/linguistic representativeness.[2]

After the range of publication types have been identified and described in a way that allows for the categorization of an individual article within the larger

2. It is important to note that achieving external/situational representativess does not necessarily mean that a corpus is internally/linguistically representative.

framework (i.e., to identify the register that a particular text belongs to), disciplines can be described in terms of their publication patterns for these registers. In turn, such a description can inform the design of a corpus that can be used to investigate disciplinary and register influences on linguistic variation. Thus, the second objective (Section 2.3) is to identify how the use of these various article types varies across disciplines. This analysis can then inform the design and construction of the corpus that will serve as the foundation for the research.

2.2 Surveying the domain of disciplinary journal writing

In order to develop a taxonomy of research article types, a survey of journals in 11 varied disciplines was undertaken to develop descriptions of article types and identify the use of these article types across a range of disciplines. As mentioned above, the motivation for this type of taxonomy is two-fold. First, if different types of articles published in academic journals are to be considered distinct registers, then there must be aspects of the situational characteristics of these articles that differ, and a taxonomy allows us to systematically describe these differences. The second motivation is methodological: if the goal of the research is to empirically investigate linguistic variation across these registers, then a taxonomy based on non-linguistic features of these texts is needed in order to design and build a corpus to represent a particular register. That is, corpus builders need a reliable method for categorizing texts into register categories so that the corpus is a suitable representation of that register. In this section I describe the process used to develop a taxonomy for defining registers published in academic journals.

2.2.1 Procedures

The first step in the development of the taxonomy was to develop a list of possible article types based on my own background knowledge as a reader of academic journals and as a trained researcher. Along with this list, a preliminary or draft inventory of distinguishing characteristics that an article of a particular type might exhibit (such as presence or absence of observed data, type of data, and so on) was compiled. These lists served as starting points for the analysis in that they allowed texts to be grouped into broad categories that could then be refined based on the second stage of taxonomy development: an inductive survey of journals in a range of disciplines.

The inductive survey involved reading carefully but widely in four disciplines (chemistry, economics, sociology, and philosophy), categorizing as many articles as possible in four to seven journals in each discipline. Journals were identified based on their placement in topic groupings in the periodical section of the

library, and were selected from those published in 2001 or later. Journal descriptions enabled the sample to be restricted to peer-reviewed journals, and to ensure the inclusion of both 'general' and more specialized journals within each field. The 'general' journals were included systematically because they covered a wide range of topics within a field. Other journals examined during the survey represented publications focused on a range of more specific sub-topics or areas. The journal description also helped to identify journals which might publish different types of articles. For example, most journal descriptions included a statement of the types of research that they accept for publication.

These selection guidelines were chosen based on a desire to describe the current state of respectable academic research, and to capture as many different types of writing as possible. For example, in sociology, the journals *Theory and Society* and *Qualitative Sociology* were reviewed because their titles and journal descriptors indicated the presence of theoretical and qualitative research articles respectively. In economics, the journal *Quarterly Journal of Economics* was included in the survey based on its description as covering "all aspects of the field," while *Computational Economics* was selected based on its focus on a subfield not often mentioned in other journal descriptions, and the statement in its Aims and Scopes that it publishes three specific types of articles: state of the art reports, brief reports, and critical reviews.

All journals examined in each discipline are listed in Appendix A. For this initial survey, every article in one issue of each journal was examined (however, "editorials" written by the editor of the journal, book announcements, or obituaries were not included). The following aspects/portions of the text served as the basis for the analysis:

1. the title
2. the abstract, if present
3. a goal/purpose statement, typically located in the first few pages of the article
4. internal headings
5. descriptions of data and/or procedures, if present
6. textual aspects like the presence/absence of data tables and figures, formulas
7. labels assigned to the text by the journal itself (e.g., *article, commentary, book review, note,* etc.)
8. key words throughout the article that described the study as dealing with data and methods, such as *experiment, treatment, survey, interviews, case study, observational,* and so on.

Based on the reading at this stage, the taxonomy categories and operational definitions were revised to include additional categories and descriptors. The second stage involved reading in 2–4 journals in additional disciplines (geology, physics,

applied linguistics, psychology, political science, pediatric medicine, and general and civil engineering). I applied the revised framework to articles in those journals in order to evaluate the applicability of the taxonomy to disciplines not included in the initial survey. The journals examined in this stage are also listed in the Appendix A. The next section contains the revised general taxonomy for article types along with the operational definitions, as well as an illustration of how the taxonomy was applied to categorizing texts.

2.2.2 A taxonomy of academic journal registers

The taxonomy of published journal writing developed for this study has three 'meta' article types: empirical research reports, theoretical articles, and evaluative documents. Empirical research reports are those that analyze some type of observed data. The term 'empirical' is used here in a traditional sense, so that it encompasses both quantitative data and qualitative data. However, as I will discuss a bit later, a major question in categorizing texts are distinct disciplinary conceptualizations of what constitutes 'observed data.' Theoretical articles are those that discuss matters of theory within the discipline, and a key feature of these is that they do not analyze any empirical data. Rather, they focus on explicating and extending key premises in the discipline. Evaluative texts are those whose primary purpose is to offer critique or summary of the state of the field, a particular article, book or product. Each of these three types contains subsets of texts within them, and each subtype is described in terms of its distinguishing characteristics (or operational definition) in Table 2.1 below.

Empirical research reports include quantitative, qualitative, and mixed methods research. Quantitative research includes experimental/quasi-experimental research as well as observational research. Qualitative research analyzes any observed data which is not quantitative in nature and encompasses a variety of research methods and data types, such as ethnographies, case studies, focus groups, interviews, and field observations. Mixed methods research is a term reserved for studies which use a nearly equal focus on quantitative and qualitative data analysis, and only occurs in disciplines where both quantitative and qualitative research is already conducted.

Theoretical articles can also be one of three subtypes. Some articles focus on explicating details of a famous scholar's work and ideas on a theory, and this constitutes what I term 'author interpretation' articles. Other theoretical articles are logic-based and rely on formal statements of logic to parse through theoretical constructs. All other theoretical articles are grouped under a 'general theoretical' article subtype.

Evaluative texts can be of three main types. The first type, which is fairly common, is the book or product review. These texts are typically shorter than full

articles, and focus on summarizing a book or product while incorporating a degree of evaluation/critique. The 'commentary/forum' evaluative texts focus on critique or put forward an argument on an issue in the field, and they often take the form of a forum where scholars with opposing viewpoints each present a critique/argument. The third type is a review article which synthesizes current research in a particular area of the field without presenting new data analysis. In this type of article the focus is on the synthesis and not necessarily on evaluation.

The operational definitions presented in Table 2.1 were formed based on prior knowledge and on the analysis of the texts in the journals listed in Appendix A. In order to illustrate how this worked, several examples are provided below. As mentioned above in the methodology section, many pieces of evidence in each text were examined, from the title to internal headings, to the presence of tables or formulas in the text, to purpose statements, and so on.

The first piece of evidence examined was how a text was labeled (if labeled by the journal) in the table of contents. For those labeled "commentaries," "notes," and "book reviews" (labels that clearly fit within the evaluative type of article), several additional features from the operational definitions were confirmed before the texts were labeled with the corresponding sub-type of evaluative texts. For example, book reviews are a very straight forward type of text to determine. Not only are they always labeled 'book review', but across disciplines, they typically begin with a citation to the book or product being reviewed (rather than a creative title), they are usually 2–3 pages long, and have a focus on summary that is illustrated by the fact that the text usually contains a lot of markers like "in Chapter 4" and "the focus of the second part of X's book is….".

For texts labeled as "article," the analysis was a bit more complex. Often the title of the text gave a first indication of the type of article. For example, the article "Wittegenstein on Metaphysical/Everyday Use" (Baker 2002) follows a typical pattern for a theoretical article title that is an author interpretation. That is, the title follows the pattern "X on Y." In addition, the statement of the paper's aim or goal explicitly identifies the purpose of the text as an interpretation of Wittegenstein's claims: the author states that he "shall make a case for a very different reading of this remark" (Baker 2002: 289). A further look through the article reveals a heavy focus on the author's ideas. All of this evidence confirms that Baker's article should be labeled 'theoretical: author interpretation.'

Empirical articles were identified by the presence of observed data, and a data mention could appear in the abstract or in a section labeled "Data" or something similar. Take, for example, the article "The long arm of the law: Effects of labeling on employment" (Davis & Tanner 2003). Here, the title indicates that it is probably a quantitative analysis (the use of "effects of X on Y"). Within the abstract, the authors state that they will use "the National Longitudinal Survey of Youth, a large and nationally representative sample, to examine …" (Davis & Tanner 2003: 385).

Table 2.1. Operational definitions for the text taxonomy

	Article Type	Content Operational Definition	Textual/Genre Features Operational Definition
Empirical *Analyzes observed data*	*Quantitative: Experimental*	– analyzes numerically-based data – object of study is manipulated in some way, either physically or through a 'treatment' – includes quasi-experimental – usually includes comparison of groups	– generally includes a 'experimental' or 'procedure' section – usually includes tables, figures illustrating quantitative analysis – key words: *experiment, control, group, laboratory setting, treatment, procedure*
	Quantitative: Observational	– analyzes numerically-based data – object of study is not submitted to any type of treatment or manipulation – data comes from a variety of sources depending on discipline, e.g., survey or demographic data, test scores, observations from nature	– usually includes a description of the data – usually includes tables, figures illustrating quantitative analysis – key words: *survey, demographic, measure*
	Qualitative	– observational in nature – does not analyze quantitative data – typically does not manipulate/apply treatment conditions to get data	– usually includes description of data – only limited statement of 'procedure' or methodology – key words: *ethnography, longitudinal, interviews, focus groups*
	Mixed Methods	– uses both quantitative and qualitative methods with an equal focus	– has features of both qualitative and quantitative research reports – key words: *qualitative/quantitative or mixed methods*

Table 2.1. (Continued) Operational definitions for the text taxonomy

(*Continued*)

	Article Type	Content Operational Definition	Textual/Genre Features Operational Definition
Theoretical *No observed data is analyzed. Advances theory within field.*	*General*	– discusses/advances a theoretical aspect of the field	– lack of distinguishing features of other types (e.g., not labeled 'review,' no data description or methods section)
	Author Interpretation	– comprehensive and in-depth description/explication of one author's ideas/theories on a particular issue	– title usually includes "X on Y" – internal headings often include author's name, or first sentences if no headings
	Logic-based	– uses formulas to advance logic, but no data	– includes a progression of formulas within text of analysis
Evaluative *Offers critique of state of field, issue, article, book, or product*	*Commentary/ Forum*	– Presents critique/evaluation of state of field, issue within the field, or a particular article – Focus is on critique with less summary	– typically labeled 'commentary', 'discussion note' – fewer references than a theoretical article – begins with statement introducing article to be critiqued – initial critiques have descriptive, clear title, e.g., "A critique of X" – response typically titled "A Reply to X"
	Synthesis/ Review	– focus is on synthesizing what is known in the field or recent research on a particular area	– often termed 'review articles'
	Book/Product Review	– offers summary and evaluative comments regarding a book or product – focus is on summary, while critique is there but often backgrounded	– typically shorter (2–3 pages) than articles – title usually a citation to reviewed book – usually at end of journal issue – labeled "Book Review" by journal/title – typically have an internal structure of summary marked by adverbial phrases like "In Chapter 2..." or "Ch. 2 deals with..."

Also indicative that this text is an empirical article is the internal headings "Research Questions and Methods" and "Measures" (Davis & Tanner 2003: 391). The presence of a large table showing descriptive statistics for their measures further confirms that it is a quantitative study. Because the data is described as a survey, this study is labeled 'observational'.

The examples that I have presented above represent fairly straightforward applications of the taxonomy. However, in practice, applying the taxonomy is much more complicated. In what follows, two of the most problematic issues that I encountered in applying the taxonomy are discussed: (1) that the lines between certain pairs of registers are not as transparent as others, and (2) that some types of articles bridge registers.

2.2.3 Some issues in applying a taxonomy of research articles

The first issue in applying this taxonomy is that without familiarity in a discipline, it can be difficult to distinguish between certain registers. At the heart of this issue is what is considered 'observed data' in a discipline. One of the most difficult distinctions to make is between qualitative research and theoretical research. For example, a study reporting on an ethnography in which focus groups, observations, and interviews were conducted is clearly qualitative research, in part based on the fact that the article calls itself 'ethnographic' (a widely-acknowledged qualitative methodology), as well as meta-language in the article that labels and describes data. However, it is not as clear (at least to an outsider of the field) whether a political science article that provides an analysis based on legislative decisions and court records is also qualitative (i.e., empirical). An article such as the latter typically does not have a section in which the data is systematically described, but it is also not purely theoretical. While I have considered an article such as this qualitative research in my taxonomy, an important consideration is the perspective from inside the discipline. Taking into account what members of the discipline consider data is key to understanding the discipline and the writing that takes place in the discipline.

Related to this issue is the fact that as an outsider of these disciplines, it can be difficult to distinguish between data types. For example, in chemistry and physics, despite careful reading, it was often difficult to determine if research was based on experimental or observational data. Thus, for registers and disciplines represented in a corpus of journal registers, discussions with disciplinary informants are important in writing operational definitions that can be applied on a more discipline-specific level.

The second main issue is that some articles do not clearly fit into any one category. For example, some articles make a theoretical or methodological argument, and then present a brief data analysis to support that argument. The data analysis is not the focus of the article, but rather is used for illustrative purposes. While the article may have some characteristics of an empirical article, such as quantitative

data, tables and figures, and so on, they less often contain key sections such as a description of procedures. This type of hybrid article is not accounted for in the current taxonomy (and is also not included in the corpus designed for this study).

Furthermore, a type of article that I am not distinguishing in this analysis is that of the brief report. Brief reports carry the same features as empirical articles, but are labeled by journals as 'brief reports' or something similar. One journal described brief reports as reports of research that are much abbreviated, or reports of research still in progress. Interestingly, primarily the 'hard' disciplines tended to publish these (e.g., chemistry, pediatrics, physics, and engineering).

In sum, the major consideration for most articles is the presence or absence of data, and if present, the nature of that data and the methods through which the data was analyzed. In the next section, I present the results of my analysis of eleven disciplines across a range of academic areas, and discuss the variation that I observed within.

2.3 Journal registers in the disciplines

While carrying out the taxonomy development reported on in Section 2.2, I also analyzed the extent to which each discipline surveyed publishes each article type. This analysis is presented in Table 2.2, where ++ indicates that this type occurs frequently in journals in that discipline, + indicates that it occurs on a regular basis, +– indicates that a few examples of that type of article were found but that they occurred rarely, and – indicates that no articles of this type were observed in the discipline. It should be noted here that these estimations are just that – estimates based on a survey of a limited number of journals in these disciplines. For example, although a register may be labeled with a – in Table 2.2, it is possible that the register would occur in other journals or sub-disciplines not considered in this survey.

Several interesting trends emerge from this analysis. First, the evaluative types of articles, particularly book reviews, show a decreasing trend as disciplines move (in traditional terms) from soft to hard disciplines. Likewise, commentary/forum articles are nearly non-existent along that same parameter. A second trend is that theoretical articles are more frequent in philosophy and political science. Empirical research is generally not present in philosophy. As we move into disciplines which fall more in the middle between hard and soft disciplines (such as sociology and economics), quantitative observational studies become more frequent, and as we move further into the 'hard' realm, experimental research studies become more common. In addition, all theoretical article types decrease as the disciplines approach the 'hard' side of the continuum. Physics is an exception to this rule, however, because of the sub-discipline of theoretical physics, which one of the journals surveyed in this task represented.

Table 2.2. Types of articles by discipline

Discipline	Empirical				Theoretical			Evaluative	
	Quantitative		Qualitative	Mixed	General	Author Interp.	Logic-based	Commentary, Forum	Book/ Product Review
	Exp.	Obs.							
*Chemistry†		++	–	–	+–	–	–	–	–
Physics†		++	–	–	+	–	++	–	–
Medicine (Pediatrics)	+	++	+–	–	-	–	–	+–	+–
Geology	+	++	–	–	+–	–	–	–	+–
Engineering (General & Civil)†		++	–	–	+–	–	–	–	+–
*Economics	+	++	–	–	+	+	+	+–	+
*Sociology	+–	++	++	+–	+	–	–	+–	+
Psychology	++	++	–	–	+	–	–	+	+
Applied Linguistics	+	++	++	+	+	–	–	+	+
Political Science	–	++	++	–	+	+	+–	+	+
*Philosophy	+–	–	–	–	++	++	+	+–	+

Key: ++ frequently occurs, + occurs with regularity, +– occurs rarely, – not found
*Discipline investigated in more detail as part of initial taxonomy formation
† A field in which as a non-specialist, I could not reliably distinguish between experimental and observational research. However, my belief is that these are primarily experimental research designs, particularly in engineering and chemistry.

The case of physics brings up an important point about variation across sub-disciplines and in variation across journals. For several of the disciplines, one sub-discipline is theoretically-oriented. Thus, journals within that sub-discipline publish only theoretical (and sometimes evaluative) articles. While some variation occurred across sub-disciplines in all the disciplines, psychology was perhaps the most varied, and this is most likely because psychology is a very broad and diverse discipline with a wide range of sub-disciplines.

In addition, this analysis reveals much more variation due to the effect of journal than expected. For example, in each of the disciplines, relatively few journals published evaluative texts, particularly forums. Some journals did not publish book reviews, while one journal only published reviews. In particular, I noted that most journals publish primarily either theoretical work or empirical work (there are a few exceptions to this), even if their descriptions state that they publish both. The journals that do publish both types tend to publish primarily empirical work, and a theoretical article may appear once in a while.

While there are some journals that are more general in nature, other journals can be associated with a particular sub-discipline, and thus publish more of a certain type of article. Some journals have distinct article types that were not represented elsewhere. For example, the journal *Philosophy and Phenomenological Research* contained a section of articles labeled "Book Symposium" in which an author of a book writes an article introducing the book, which is then followed by several reviews of the book by other scholars, and concluded with a response from the reviewed book's author.

Two major points can be summarized from this analysis. First, there is wide within-discipline variation that often follows along sub-disciplinary lines, and which is reflected in journals that are aimed at those specific sub-disciplinary areas of inquiry. Thus, when selecting disciplines to represent registers, care will need to be taken to select disciplines in which sampling from a wide range of journals is possible in order to reach a desired number of texts for inclusion in the corpus.

Second, and perhaps most notably, there are no disciplines in which all journal registers were found. That is, disciplines typically published a small number of registers with greater frequency, rather than a broad range of registers. The implications for these two trends are discussed in Section 2.4, as well as a description of how the results of this survey and taxonomy development have been applied to the design of the corpus that serves as the foundation for the analyses contained in this book.

2.4 Implications for corpus design

The premise behind creating a general taxonomy of published journal article types was to identify potential disciplinary differences in the types of journal articles which are published, and thus aid in the selection of the disciplines and journal registers to be represented in the corpus. This analysis has revealed that each discipline publishes a variety of texts, but that most disciplines do not publish *each* type of text. Therefore, two principles for corpus design are apparent. First, in order to investigate across-discipline variation, two disciplines should represent each type of article included in the corpus. Second, in order to investigate the possibility of within-discipline variation due to register differences, disciplines should be represented by at least two article registers whenever possible.

The first way in which the survey and taxonomy development (Section 2.3) is applied to corpus design is in the selection of journal *registers* to be included in the present study. While my original conceptualization for this project had been to include both research reports and evaluative texts, the findings displayed in Table 2.2 above illustrate that these evaluative texts are published much less frequently than research reports. In addition, I found that commentaries and forums, in which academics engage in more interactive scholarly discussion, were much

less frequent than even book reviews (the second type of evaluative texts). In fact, the publication of forums and commentaries is highly dependent upon journal – that is, only a few journals ever publish this type of text. Furthermore, evaluative texts as identified by this taxonomy are largely absent from the natural sciences (e.g., chemistry, physics, geology, etc., see Table 2.2). A further complicating factor is that these evaluative texts are not usually published in each issue or even volume of the journal. It would be difficult to sample these evaluative texts in a manner consistent with the sampling of theoretical or empirical articles, which occur most frequently in all disciplines. Therefore, evaluative registers have been excluded from the corpus design for this study.

The corpus for the present study has been limited to the primary article types of empirical and theoretical research. Within empirical research, the analysis shows that both quantitative and qualitative research is common, with great variation across disciplines. That is, research is exclusively quantitative in the hard sciences, while social sciences often publish both types of research. Mixed methods research is less common overall (Table 2.2) and is limited to disciplines which publish both quantitative and qualitative research. Thus, within empirical registers, I have chosen to represent quantitative and qualitative research in the corpus. Within theoretical research, the two specific types of theoretical research (logic-based and author interpretation) are generally less frequently published. Based on the decision to sample from physics and philosophy (described next) to represent theoretical articles, these specific types of theoretical articles have been collapsed into one general 'theoretical' register.

The second way in which the taxonomy development and survey has informed the corpus design is in the selection of disciplines to be studied in the project. In fact, the selection of disciplines has been influenced by several factors: (a) quantitative trends in the types of research published in disciplines, as reported in Table 2.2, (b) the desire to represent each of the selected registers (theoretical, quantitative, and qualitative research) with at least two disciplines, (c) the desire to represent disciplines by more than one register whenever possible, (d) the desire to include disciplines from a range of academic areas, and (e) the benefits of including disciplines that have been analyzed in previous linguistic research. Consequently, six disciplines were selected.

To represent theoretical articles, physics and philosophy showed a frequent use of this register, as well as representing disciplines clearly situated on opposite ends of the 'hard' and 'soft' continuum. Quantitative research reports are also readily available in physics. Two social science disciplines, political science and applied linguistics, exhibited strong publication rates for both quantitative and qualitative research, and thus have been selected for inclusion in the corpus. Finally, two disciplines, biology and history, which were not included in the survey summarized in Section 2.3, were selected for inclusion in the corpus based on their frequent presence in

research on disciplinary variation and because they are disciplines that are characterized by their quantitative and qualitative research methods respectively. Table 2.3 summarizes the registers and disciplines selected for the present study.

Table 2.3. Disciplines and registers represented in the corpus

Discipline	Theoretical	Qualitative	Quantitative
Philosophy	✓		
History		✓	
Political Science		✓	✓
Applied Linguistics		✓	✓
Biology			✓
Physics	✓		✓

To the extent possible, two registers were chosen to represent each discipline. However, for philosophy, history, and biology, it was not possible to sample more than one journal register because the discipline relied primarily upon one type of article. Although differences surely exist within these disciplines, making those highly fine-grained distinctions would have been unreliable within the scope of the present project.

The six disciplines that have been identified in this chapter represent a range of fields along the 'hard' and 'soft' parameter of disciplinary variation (see Becher, 1994), and are capable of representing the three journal registers selected for the study. As a result, the corpus design specified here allows me to investigate linguistic variation within and across both discipline and register. More specifically, however, several comparisons for describing variation are possible:

1. comparisons across all disciplines and registers,
2. comparisons across discipline for a single register type (e.g., quantitative research reports in political science, applied linguistics, biology, and physics), and
3. comparisons across register type within a discipline (e.g., theoretical versus quantitative research in physics).

In order to build a corpus of these disciplines and registers and enable these comparisons, the taxonomy presented above in Table 2.1 was refined on a disciplinary basis in consultation with expert informants from the discipline. The process undertaken to revise the taxonomy and the resulting operational definitions that were used to construct the corpus are detailed in Chapter 3, along with the general analytical methods employed in the study.

CHAPTER 3

Building and analyzing the Academic Journal Register Corpus

3.1 Introduction

In Chapter 2, I described the broader situational domain of academic journal writing based on a survey of eleven different disciplines. At the end of that chapter, I proposed a corpus design which would allow me to investigate linguistic variation within and across academic disciplines and journal article registers. This corpus design includes three journal registers (theoretical, quantitative, and qualitative research) and six disciplines (philosophy, history, political science, applied linguistics, biology, and physics). In this chapter, the focus is on the methodological procedures undertaken to build, annotate, and analyze this corpus, called the Academic Journal Register Corpus.

3.2 Corpus collection procedures

One difficulty in applying a general taxonomy of academic journal registers to specific disciplines is that the perspective of a disciplinary insider is sometimes necessary to fully understand and categorize journal articles according to the taxonomy. As one purpose of building a taxonomy is to create operational definitions by which individual texts can be categorized into registers, it is important that the categorizations be both reliable and valid. One approach to accomplishing this reliability and validity is to take into account the beliefs of disciplinary insiders, verifying how a journal article register would be characterized within the disciplinary knowledge community. In Section 3.2.1, I describe the approach taken to incorporate discipline-specific information into corpus compilation, followed by descriptions of the corpus collection and annotation procedures.

3.2.1 Formation of operational definitions for journal registers in specific disciplines

The first step in building a corpus of disciplinary writing in philosophy, history, political science, applied linguistics, biology, and physics was to revise the general operational definitions that I reported on in Chapter 2 (see Table 2.1) to account for discipline-specific characteristics. First, the journal survey process described in Chapter 2 was replicated for each of the six disciplines to be included in the corpus. That is, a range of journals and articles in each discipline were analyzed, this time with a focus on the three journal registers of theoretical, quantitative, and qualitative research. During this survey, features which seemed to characterize articles within the specific disciplines were identified, and questions regarding how to categorize example articles in each discipline were compiled. I then consulted with disciplinary experts in each of the disciplines in order to validate and refine these operational definitions.

The first meeting with each expert included a discussion of the general nature of the discipline, and the expert informant examined Table 2.2 in Chapter 2 (displaying the frequency and types of articles by discipline). The expert evaluated the accuracy of my estimation of the types of research that are common in the discipline. In all cases, the trends were confirmed, and the experts explicated on other types of research that are less commonly included in their discipline. Also in this first meeting, the informants and I discussed the operational definition drafts, confirming and adding to these definitions. The expert informant also helped generate a list of high quality academic journals within the field which could serve as sources for the corpus texts. More specifically, we focused on identifying high quality journals which were considered 'generalist' journals (meaning that they published research from many different topic areas and/or sub-disciplines) as well as more specialized journals across a range of sub-disciplines. Journals which would frequently publish the specific registers of interest for that discipline were also identified.

After the first meeting, each discipline-specific operational definition was revised and applied to several articles. When needed, a second meeting with the disciplinary expert allowed me to validate those operational definitions. This validation process included clarifying any remaining points of confusion (e.g., is an analysis of financial records in political science considered quantitative research?) as well as the two of us looking at several articles together to see if we agreed on the register it would be classified in. While I had expected that the resulting operational definitions would vary to a certain degree, in fact, the general operational definitions formed in the initial stages of research were fairly reliable. Rather, the main factor that varied across disciplines concerned the nature of data in distinguishing between quantitative versus qualitative research, and between qualitative and theoretical research. Instances where the nature of data affected how a text was categorized are discussed in Chapter 4.

3.2.2 Journal and article selection

For each discipline, eight to ten peer-reviewed journals were selected based on the input of disciplinary informants and on information about the journals themselves. After the initial meeting with the disciplinary experts, each journal suggested by the experts was further researched to determine suitability for the corpus; the journal descriptions (published on the journals' own websites) were examined to verify the topics and article types published in each journal. ISI ratings were also examined for physics, based on a suggestion by the disciplinary informant who emphasized the importance of these ratings for physics journals (T. Porter, personal communication, May 6, 2009).

The goal in selecting journals was to sample from reputable journals in many areas of the discipline (limited to journals for which electronic versions of articles were available). For each discipline, I included as many 'generalist' journals as possible. 'Generalist' journals are those which do not focus on a specific topic or sub-discipline, but rather publish from the range of topics in a discipline. For example, in philosophy I included *Philosophical Quarterly, Philosophy,* and *Journal of Philosophy* as 'generalist' journals because the journals' own descriptions and input from the disciplinary expert indicated that they published articles on all areas of philosophical inquiry.

After identifying as many 'generalist' journals as possible, the sampling frame was filled out with journals focused on a specific sub-discipline/topic, making sure to include a variety of journals that covered a wide range of topics within the discipline. For example, in applied linguistics, more specialized journals like *Language Learning & Technology* and *World Englishes* (focused on computer-assisted language learning and sociolinguistics respectively) were included in addition to the more generalist journal of *Applied Linguistics*. A complete listing of journals sampled for each discipline appears below in Table 3.1. In order to avoid confounding register with source journal in disciplines for which two registers were being sampled, I first focused on journals which published both registers, and then supplemented with journals which primarily published one type of research.

For each discipline and register combination, 30 texts were selected to make up the sub-corpus. To select articles, one article was randomly selected from three different issues of a journal, one each from 2006, 2007, and 2008. Each article was categorized using the discipline-specific operational definitions. If the article matched the operational definition for a theoretical article (physics and philosophy), a quantitative research report (political science, applied linguistics, biology, physics), or qualitative research report (history, political science, applied linguistics), it was included in the corpus. If it did not, then a different article was randomly selected from that issue and categorized; this process was repeated until the target number of texts was reached.

Table 3.1. Journals represented in the Academic Journal Register Corpus

Philosophy	**History**	**Political Science**
1. Philosophical Quarterly*	1. American Historical Review*	1. American Journal of Political Science*
2. Philosophy*	2. Journal of World History	2. Third World Quarterly
3. Journal of Philosophy*	3. Historical Research (British)	3. International Studies Perspectives
4. Inquiry*	4. The Journal of American History	4. American Politics Research
5. Ethics	5. Journal of Women's History	5. Journal of International Development Perspectives on Politics*
6. Law and Philosophy	6. Journal of Colonialism and Colonial History	6. Political Quarterly
7. Journal of Ethics	7. Journal of Urban History	7. Policy Studies Journal
8. Philosophy of Science	8. Journal of Contemporary History (European)	8. Foreign Affairs
	9. The Western Historical Quarterly	9. Politics & Policy*
Applied Linguistics	**Biology**	**Physics**
1. Applied Linguistics*	1. PNAS	1. Physical Review B: Condensed Matter
2. TESOL Quarterly	2. Journal of Natural History	2. Journal of Applied Physics
3. Journal of English for Academic Purposes	3. Applied & Environmental Microbiology	3. New Journal of Physics*
4. Language Learning & Technology	4. Microbial Ecology	4. Nuclear Physics A and B
5. World Englishes	5. Journal of Cell Biology	5. Annals of Physics*
6. Language Teaching Research	6. American Journal of Physiology	6. Journal of Physical Chemistry B
7. Modern Language Journal	7. Ecology	7. Astrophysical Journal
8. International Journal of Applied Linguistics*	8. Evolution	8. European Physical Journal C. Particles and Fields
9. Journal of Second Language Writing	9. Conservation Biology	9. Journal of Geophysical Research – Atmospheres
10. Canadian Modern Language Review	10. Oikos	10. Journal of Physics B

*indicates a 'generalist' journal

3.2.3 File conversion and clean-up

After all corpus files were obtained, each file was converted to plain text, and a standardized header was added to the beginning of the file. The header contained the full bibliographic information for the research article. Because the conversion to plain text is problematic in terms of page layout, each text was manually edited.

In this editing process, all page headers and footers (typically containing the journal, article or author name and a page number) were deleted, as they are not a part of the language of the article itself. For articles containing footnotes within the text, each footnote was located and moved to the end of the file. Because the notes often contain substantial material about background information, claims and counterclaims, and even data (as Conrad 1996a: 142 mentions for history articles specifically), the notes were retained but moved so that the main prose of the text was not interrupted by footnotes. All reference lists were deleted from the files, along with figures and tables. In some cases (particularly in history and political science), references were cited in full in footnotes. In this case, the footnotes were removed from the files.

In addition, the physics texts contained many formulas and special symbols throughout the texts. Each text was edited using the following principle: formulas set apart from the text of the article on its own line were removed. If, however, a short formula or symbol was embedded in the prose of the article, then it was retained. All other aspects of the text prose, for example, headings, were retained in their entirety. All text files were given a descriptive filename containing indicators of the discipline, the register (qualitative, quantitative, or theoretical), a unique identification number, and the original journal in which the article was published.

3.3 Corpus description: The Academic Journal Register Corpus

The final Academic Journal Register Corpus is composed of 270 research articles from 56 academic journals, and contains about 2 million words. Each discipline/register combination is represented by 30 research articles (Table 3.2). Table 3.3 describes the corpus in terms of the number of words per discipline and register.

Table 3.2. Corpus description in number of texts

	Theoretical	Qualitative	Quantitative	Total
Philosophy	30	-	-	30
History	-	30	-	30
Political Science	-	30	30	60
Applied Linguistics	-	30	30	60
Biology	-	-	30	30
Physics	30	-	30	60
Total	*60*	*90*	*120*	*270*

Table 3.3. Corpus description in number of words

	Theoretical	Qualitative	Quantitative	Total
Philosophy	280,826	-	-	280,826
History	-	282,898	-	282,898
Political Science	-	191,791	230,386	422,177
Applied Linguistics	-	237,089	202,871	439,960
Biology	-	-	154,824	154,824
Physics	194,029	-	183,279	377,308
Total	*474,855*	*711,778*	*771,360*	*1,957,993*

3.4 Corpus annotation

To enable many of the automatic analyses available with corpus linguistics techniques, the Academic Journal Register Corpus was automatically 'tagged' to annotate each word in each text with grammatical information. The tagging process was followed by procedures to evaluate and improve the accuracy of the automatic annotation.

3.4.1 'Tagging': Part of speech annotation

All texts in the corpus were 'tagged' using the Biber tagger, which is available at Northern Arizona University. The tagger is a computer program developed by Biber (see Biber 1988; Biber et al. 1999) to assign 'tags' indicating grammatical information for each word in a text. The tagger uses large dictionaries, lexical information, probabilistic information, and contextual rules to assign tags for the major part of speech (e.g., noun, verb, adjective, preposition), verb tense, aspect and voice (e.g., active vs. passive, perfect aspect, modality), and syntactic structures (e.g., *that*-clauses, *to*-clauses, conditional clauses). The use of an automatic tagger allows for a great deal of language to be annotated with detailed grammatical and lexical information. However, as with any automatic tool, an important concern is the degree to which that tool performs accurately. In the next section, I describe the process undertaken to evaluate the accuracy of the automatic tagging.

3.4.2 Accuracy of automatic tagging

In order to evaluate the accuracy of the automatic tags produced by the Biber tagger, a subsample of files from the corpus were hand-coded for tagging errors. Using a randomly selected sample of 15 research articles representing all disciplines and registers in the corpus, I first extracted excerpts of the texts to obtain a coding sample which was broad yet small enough to make hand-coding feasible.

Each text excerpt constituted exactly one-third of the original file, and an equal number of excerpts were taken from the beginning, middle or end of the file (that is, five excerpts were composed of the first one-third of the text, five came from the middle third of the text, and five came from the final third of the text).

Each word in these 15 excerpts was hand-coded to identify automatic tagging errors, focusing on the major parts of speech (noun, verb, adjective and so on), verb tense, aspect and voice, demonstrative pronouns versus determiners, and so on. A complete list of the features for which accuracy measures were computed can be seen in Appendix B. A second computer program was then written to count and classify the errors. This program performed two tasks, which are illustrated here using demonstrative pronouns as an example. First, the program computed error rates for each linguistic feature being evaluated. These error rates are based on three components:

1. the number of times a word was automatically tagged as a demonstrative pronoun
2. the number of times the demonstrative pronoun tag was assigned correctly by the tagger
3. the number of actual occurrences of demonstrative pronouns (indicated by a special tag added during the hand-coding process)

Using these three counts, two measures of reliability were calculated: precision and recall. In addition, a third measure was developed and applied to obtain an overall accuracy rate for each feature. These three measures give complementary information about the accuracy and reliability of the automatic tagging process. The first two measures, precision and recall, quantify the extent to which automatic tags are applied accurately (i.e., that the tag assigned to any individual word is accurate). The third measure, overall accuracy rate, quantifies the accuracy of quantitative data that results from the automatic tagging process. Each of these measures is discussed in turn.

The first reliability measure, *precision,* is represented in Equation 1 below. Precision is an estimate of how often, for example, a word that has been tagged as a demonstrative pronoun is in fact a demonstrative pronoun. The result of this equation is a proportion; thus, precision reflects the proportion of all words tagged as a feature that are true instances of that feature.

Equation 1

$$Precision = \frac{\#\ Correctly\ Tagged\ as\ X}{\#\ Automatically\ Tagged\ as\ X}$$

On the other hand, the second reliability measure, *recall*, is a measure of whether or not all instances of a linguistic feature are being tagged with the

correct tag (Equation 2). Continuing with our example of demonstrative pronouns, the recall measure compares the number of times a word is correctly tagged as a demonstrative pronoun to the number of actual demonstrative pronouns occur in the sample. Thus, recall measures the extent to which words that are demonstrative pronouns are in fact tagged as demonstrative pronouns. This is measured by dividing the number of words correctly tagged as a demonstrative pronoun with the number of actual occurrences of demonstrative pronouns. Again, the result is a proportion. In this case, recall tells us the proportion of all occurrences of a feature that have been correctly identified as that feature.

Equation 2

$$Recall = \frac{\#\ Correctly\ Tagged\ as\ X}{\#\ Actual\ Occurrences\ of\ X}$$

Precision and recall measures give us estimates regarding the accuracy of tags, specifically for determining whether we can trust that words tagged in a certain way are actually instances of that particular feature. The third measure, which I call *overall accuracy rate*, is slightly different. Rather than considering the accuracy of tags in combination with the instances of actual words, *overall accuracy rate* is a strictly quantitative measure that tells us how accurate the counts resulting from tags are. Equation 3 displays how an overall accuracy rate is calculated. This measure is by no means a replacement for measures of precision and recall, which tell us the true accuracy of the automatic tagging process. However, overall accuracy rate is an *additional* measure by which we can evaluate how much we trust the rates of occurrence of a linguistic feature that we have calculated based on tagged information. That is, overall accuracy rate measures the accuracy of the *number* that results when we count all instances of a linguistic feature based on the automatically assigned tag.

Returning to our demonstrative pronoun example, the overall accuracy rate can be calculated by first subtracting the *actual* number of occurrences of demonstrative pronouns from the number of words *automatically tagged* as demonstrative pronouns. This gives the difference between the true number of instances of the feature and the number of times a word has been tagged as that feature (note that this number can be positive or negative, which only reflects whether counts based on the automatic tags are over- or under-representations of the frequency of that feature). The next step in calculating overall accuracy rate is to divide the absolute value of this difference by the *actual* number of occurrences of demonstrative pronouns. This gives us a proportion between the difference between the two counts and the true number of times the feature occurs. By itself, this gives us an *error* rate; that is, it tells what percentage of the counts are likely errors. In order to make the interpretation of this measure parallel with the measures of

precision and recall, we then subtract the resulting proportion from 1.00 to get the proportion of the count which is accurate. Thus, the overall accuracy rate measure will not indicate whether the tags are correctly being assigned to the right words, but it will indicate whether the rates of occurrence calculated from tags accurately reflect the real rate of occurrence of a particular feature.

Equation 3

$$\text{Accuracy Rate} = 1.00 - \frac{|\text{\# Automatically Tagged X} - \text{\# Actual Occurrences of X}|}{\text{\# Actual Occurrences of X}}$$

For all three measures of reliability, the result is a proportion, which can be multiplied by 100 to get a percentage. In general, rates greater than 95% can be considered good, and greater than 90% can be considered acceptable. To see how this works, let's return to our example of demonstrative pronouns. Table 3.4 below gives the results of the hand-coded error analysis for demonstrative pronouns, showing that 107 demonstrative pronouns were automatically tagged, and 101 of those instances were correctly tagged. Meanwhile, the actual number of demonstrative pronouns in the sample was 107 (meaning that there were 6 demonstrative pronouns that were not automatically tagged). The precision measure indicates that 94% of the time, a word tagged as a demonstrative pronoun was in fact a demonstrative pronoun. Likewise, the recall measure indicates that 94% of the occurrences of demonstrative pronouns were correctly tagged as demonstrative pronouns.

Table 3.4. Reliability rates (precision and recall) for demonstrative determiners

Feature	*Automatically Tagged*	*Correctly Tagged*	*Actual Occurrences*	*Precision*	*Recall*
Demonstrative pronouns	107	101	107	$P = \frac{101}{107} = .94 = 94\%$	$R = \frac{101}{107} = .94 = 94\%$

Ninety-four percent precision and recall rates are acceptable for the study, and tell us about the accuracy of the demonstrative pronoun tag. However, we can also consider the overall error rate of the tag. Doing so, we get:

$$\text{Overall Accuracy Rate} = 1 - \frac{|107 - 107|}{107} = 1 - 0 = 1 = 100\%$$

This shows that in terms of the quantitative counts only, there is no difference between the actual number of occurrences of a feature and the number of times

demonstrative pronouns are tagged. However, this measure should be used with caution. Overall reliability is useful when the primary focus of the analysis is the quantitative trend. A high overall reliability but low rates of precision and recall are not valuable particularly if the research requires analysis other than a rate of occurrence. For example, if the purpose of a study is to describe rates of occurrences for noun + *that* complement clauses and includes a description of the common nouns that control the *that*-clauses, high levels of both precision and recall are needed to create and accurate and reliable analysis; overall reliability should be used cautiously.

Measures of precision, recall, and overall accuracy rate were calculated for all of the features listed in Appendix B. After the initial round of tag-checking (coding errors; calculation of precision, recall, and overall reliability), any feature with a precision or recall rate below 95% was further evaluated. Above, I mentioned that the error analysis program performed two tasks. First, it calculated the rates of precision and recall. Second, however, it created key-word-in-context (KWIC) lines for each instance of an error for each feature. The KWIC lines allowed for an analysis of all errors to determine the reason for the errors, and identify any systematic errors that could be easily corrected with further computer programs. Automatic scripts to correct for the most frequent errors were developed and tested. First, the scripts were run on the error-coded files and then examined to verify that the scripts made the intended corrections and did not result in any unintended changes. Then, final error rates were calculated on the sample. These error rates are provided in Appendix B. The scripts program was then run on all corpus files to ensure the highest degree of tagging accuracy possible with automatic tools.

3.5 Overview: Procedures for quantitative corpus analysis

While Chapter 4 presents a detailed situational (or non-linguistic) analysis of the Academic Journal Register Corpus, Chapters 5–7 of this book present a series of studies that investigate the linguistic characteristics of the journal registers and disciplines represented in the Academic Journal Register Corpus. These analysis chapters contain quantitative analyses paired with functional interpretations of the patterns observed in the corpus. Each of these chapters contains a brief method section detailing the procedures undertaken to carry out the analysis reported on in the chapter, along with information about the specialized computer programs utilized for the analyses. Because of the use of these specialized computer programs, it's useful to give an overview of the nature of the quantitative data reported in Chapters 5–7.

Discussing the unit of analysis in corpus-based studies, and the resulting types of analyses that are possible, Biber and Jones (2009) identify three types of corpus studies. Type A studies as those which count each occurrence of a linguistic item as an observation in the analysis. Data from this type of study allow the researcher to describe the variants of a linguistic feature, and the data is in the form of frequencies of the variants under investigation. Because this data represents frequencies of nominal categories, cross-tabulation tables and chi-squared tests can be used to interpret the quantitative data (see Biber & Jones 2009).

Biber and Jones's Type B studies, on the other hand, treat an individual text or text sample as the unit of analysis, and quantitative data takes the form of rates of occurrence of a particular feature in the text. This type of study allows the researcher to describe differences between texts or text varieties. Because rate of occurrence is an interval data type, inferential statistics can be used to test the significance of differences observed between texts or text varieties.

Finally, Biber and Jones identify Type C studies, which treat an entire corpus or sub-corpus as the unit of analysis. The resulting data is a frequency for the feature in the sub-corpus. Type C studies can be used to either describe a linguistic feature and its variants or to describe the differences between genres. However, like Type A studies, the lack of data for each text in the corpus means that inferential statistics, which rely on means and standard deviations, are not possible with Type C studies.

For corpus studies which rely on concordancers and commercially/publicly available software packages for the primary analysis, the corpus is generally considered the unit of analysis. That is, the frequency of linguistic items is reported in terms of how often it occurs in a corpus or sub-corpus (the linguistic feature can also be considered the unit of analysis in studies using ready-made corpus analysis programs). However, one strength of specialized programs developed by the researcher is the ability to treat each text in the corpus as an observation, enabling the calculation of rates of occurrence for each text in a corpus. Biber and Jones (2009) explain how treating the text as an observation opens up many possibilities for how the researcher analyzes the quantitative data, and allows for the use of inferential statistics.

Thus, all analyses in this book use the text as the observation (Type B). The programs described in Chapter 5–7 all produce rates of occurrence for the target linguistic features for each text in the corpus. These rates of occurrence are calculated by 'norming' raw frequencies of features. To calculate a normed (also called 'normalized') rate of occurrence, the total raw frequency of a feature is divided by the total number of words in the text or corpus (depending on the unit of observation), and then multiplied by a norming number, typically 1000 or 1 million (see Equation 4, also see Biber, Conrad & Reppen 1998: 263–264).

Equation 4. Calculating normed rates of occurrence (per text)

$$Normed\ rate\ of\ occurrence = \frac{raw\ frequency\ of\ Feature\ X\ in\ Text\ A}{total\ number\ of\ words\ in\ Text\ A} \times norming\ number$$

In order to report rates of occurrence for a particular register using per-text counts, the normed rates of occurrence for all of the texts representing that register can be averaged. A mean rate of occurrence for a linguistic feature in a register is beneficial in several ways. First, because per-text counts are used, the standard deviation can be calculated along with the mean to describe the variability within the texts representing a register. In addition, a mean rate of occurrence minimizes the effect of a few texts that use a linguistic feature in a markedly different way (in quantitative terms).

In the remainder of the book, I present the results of my linguistic and non-linguistic analyses of three journal registers in six disciplines. First, in Chapter 4 I describe the non-linguistic, situational characteristics of the texts in the corpus.

CHAPTER 4

The situational characteristics of the Academic Journal Register Corpus

4.1 Introduction

Registers are defined based on the external characteristics of the larger context in which they are used (i.e., situational characteristics), rather than on internal, linguistic characteristics (see Biber & Conrad 2009). Thus, the goal of corpus-based studies of register variation is to identify a group of texts which are systematically similar based on situational characteristics, and then to investigate the linguistic variation that occurs within these groupings. The general taxonomy developed in Chapter 2 provided a means for designing and collecting corpora that represent various journal registers. One of the key characteristics of studies using corpus linguistics methodologies is the ability to locate quantitative trends in the use of linguistic features, and then link those trends in meaningful ways to analyses of the functions of the linguistic features. These functional analyses require a consideration of the non-linguistic, or situational, characteristics of registers in order to help us understand how and why linguistic features are used in texts. The purpose of this chapter is to present a framework for analyzing the specific situational characteristics of different journal registers, and apply that framework to the non-linguistic analysis of the Academic Journal Register Corpus.

4.2 Motivating a new situational framework for journal registers

Biber (1994), Conrad (1996a), and Biber & Conrad (2009) offer frameworks for describing the situational characteristics of registers, building upon scholarship in a variety of research traditions (e.g., Biber 1988; Crystal & Davy 1969; Halliday 1978; Hymes 1974; Basso 1974; see Biber 1994 and Conrad 1996a for further discussion). These frameworks include ways to categorize the characteristics of participants and the relationships between these participants, the setting of the communicative event, including whether time and place are shared by partici-

pants, the channel and mode of the linguistic message, the purpose of the event, and the specific topic or subject of the event (See Biber 1994: 40–41). The frameworks presented in Biber (1994) and Biber & Conrad (2009) are valuable for distinguishing between registers with broad differences, such as conversation and academic writing (and many registers in between). Conrad's (1996a) framework is more specific to academic writing and is used to describe research articles, textbooks, and student writing. However, because the situational characteristics of different journal registers vary in more restricted ways, a framework that can capture these subtle differences is needed.

For example, the framework presented in Biber and Conrad (2009: Chapter 2) includes characteristics such as mode (speech, writing, signing) and production circumstances (real time, planned, scripted, revised and edited). These characteristics are useful for describing differences among registers with higher level differences, such as conversation (spoken in real time) and professional academic writing (written over a longer period of time that includes revisions/editing). In contrast, the journal registers studied here do not vary with respect to these parameters. That is, all registers included in the corpus are written texts that have been revised and edited prior to publication. Thus, while such descriptors are useful for describing academic writing as a broader register, they are not as useful for distinguishing between sub-registers within academic journal writing.

Conrad's (1996a) framework is more directly applicable to the present study and has been used as a starting point for the framework I propose here, as Conrad's framework is specific to forms of academic writing. Despite this, the framework still contains a few factors that are not relevant for the present study. For example, Conrad (1996a: Chapter 3) includes parameters such as the level of training in the discipline of the writer and the relationship between the writer and audience. In the registers in her study, these parameters reflected differences – writers of academic textbooks were professionals in the discipline with a high level of training, while writers of the student texts had much less training in the discipline. Although the relationship between the writer and reader of academic research articles were equal (both typically being trained professionals), the relationship between the writers of textbooks and the readers of those textbooks was not equal.

In sum, although the registers studied in this book can be described in important ways by these parameters (and I will include descriptions of some of these characteristics below), these parameters do not identify *differences* across the journal registers that may be related to variations in the use of linguistic feature. Thus, a revised framework is needed that can identify differences in the situational characteristics of registers within published academic journals. In the next section, I present such a framework.

4.3 A framework for the situational characteristics of journal registers

The framework adopted here is informed by four primary sources: Conrad's (1996a) framework, information from the journal taxonomy and operational definitions developed for this study, published research about these disciplines, and aspects of the texts that have noticeably varied as I have been working with the texts in the corpus. Beginning with Conrad's (1996a) framework, I identified characteristics used by Conrad that are also important markers of variation with journal articles specifically. As will be discussed below, I then adapted the coding options included in Conrad (1996a) in order to make the application of those characteristics feasible for a corpus of 270 articles.

In addition, many characteristics of the registers were identified in the survey analysis reported on in Chapter 2 and in the formation of operational definitions in Chapters 2 and 3. At times, I have also relied upon published research that discusses the nature and characteristics of particular disciplines when a broader knowledge base was required in order to describe the registers. Finally, some of the features that I analyze in this study simply stood out as a variable aspect of research articles as I worked with texts during the corpus collection and preparation processes described in Chapter 3. These prominent features often led me to delve into a certain characteristic with more detail. Accordingly, the development of this framework has been cyclical, informed throughout the different stages of this research. The specific ways in which these sources have shaped the framework adopted in this study are described below.

Conrad (1996a: 42) outlines four guidelines for analytical frameworks of situational characteristics of texts. First, the situational analysis should remain independent of linguistic analyses. Second, the framework should be comprehensive, covering all characteristics that may be important (not only those which are determined to be important on an a priori basis). Third, such a framework must be capable of being applied to a large number of texts. Finally, the framework must be effective in comparing texts at a variety of levels. As I have adapted Conrad's analytical framework and applied it to the 270 research articles in my corpus of academic journal registers, I have attempted to follow these guidelines to the extent possible.

Table 4.1 summarizes the situational framework. Overall, there are eight categories of characteristics in the framework: participants, textual layout and organization, setting, subject/topic, purpose, nature of data or evidence, methodology, and explicitness of research design. Each of these categories is discussed in turn in the sections that follow, along with a description of how analyses for each of these characteristics were carried out.

Table 4.1. Framework for describing the situational characteristics of academic journal registers

1. Participants	
A. Writer	1 (single) 2–4 (small group) 5+ (large group)
2. Textual Layout & Organization	
A. Length	Mean page length Standard deviation of page length Page length range
B. Headings	None Un-numbered Numbered
C. Use of Abstracts	Yes No
D. Visual Elements	None Tables Figures Tables & figures Equations
E. Sections / Organization	IMRD IMRD with varied order Other Standardized section headings Variable section headings/names
3. Setting	
A. Nature of Journal	Generalist Specialized
4. Subject/Topic	
A. General Topic of the Discipline	Varied (based on informal survey of a range of sources)
5. Purpose	
A. General Academic Purpose	Varied (based on initial register survey)
6. Nature of Data or Evidence	
A. Presence of Observed Data	Yes No
B. Use of Numerical Data	Yes No
C. Primary Presentation of Evidence	Extensive prose Quantitative displays Mathematical formulas
D. Object of Study	Varied (based on trends in the corpus)

(*Continued*)

Table 4.1. (Continued)

7. Methodology	
A. General Method Type	Observational Experimental n/a (for theoretical articles)
B. Statistical Techniques	n/a (for theoretical and qualitative articles) descriptive statistics statistical difference testing other advanced statistics
8. Explicitness of Research Design	
A. Explicitness of Purpose	Direct statement Indirect / No discernible statement
B. Explicitness of Research Questions	Direct statement Indirect / No discernible statement
C. Explicitness of Citations	Within the text In footnotes/endnotes
D. Explanation of Evidence	Extensive Mention / No discernible statement
E. Explanation of Procedures	Extensive Mention / No discernible statement

4.3.1 Participants

In this framework, the participants of journal registers are analyzed in terms of characteristics of the writers. The analysis is limited to the number of authors (single-authored, small group of 2–4, and large group of 5 or more). Because other characteristics of the writer (e.g., level of training in the field) and of the reader (e.g., level of training in the field) included in Conrad's (1996a) framework remain constant in these journal registers, these parameters have not been included in the current framework. Other characteristics of the writers, such as native language, may be interesting; however, it is generally not feasible to identify this information for a single author, and this becomes even more difficult and possibly even irrelevant for articles authored by multiple scholars.

4.3.2 Textual layout and organization

The second category of characteristics involves the textual layout and organization of texts ('physical layout' in Conrad 1996a). Previous frameworks have identified the characteristics of length and internal structure, and these are included here. Length is characterized by both the mean number of pages per article as well as the range of article lengths. Also included are the use of headings within the article

text, the use of abstracts as research summaries, and the use of visual elements like tables and figures. Headings have been included here because they were a feature that varied markedly throughout the corpus, both in terms of the presence or absence of the headings, the standard (or lack of standard) form of the heading language to refer to sections of the article (i.e., introduction, method, results, and discussion), and whether or not the headings were numbered or un-numbered. While these characteristics have not been systematically described in analytical frameworks before, they varied prominently in the texts in this corpus and so have been included here.

The various sections of research articles, however, have been the focus of much linguistic analysis. For example, abstracts have been analyzed (e.g., Gillaerts & Van de Velde 2010; Stotesbury 2003; Norman 2003), showing variation in the internal structure and linguistic characteristics of abstracts. However, discussions of abstracts mostly assume the use of abstracts across disciplines, an assumption which is problematic for the texts in this corpus. Thus, in this framework, articles and registers are characterized by the general use or non-use of abstracts.

The sections or organization of articles have also received considerable attention in research on academic writing. Science writing in particular is often described as having an IMRD (Introduction, Method, Results, Discussion) organization, and many investigations look at the use of such organizational patterns (e.g., Kanoksilapatham 2005a; Ruiying & Allison 2004; Posteguillo 1999; Li & Ge 2009), examine the way that linguistic features vary across sections in an article (e.g., Martínez 2003, 2005; Biber & Finegan 2001), and offer detailed descriptions of features in particular sections (e.g., Hirano 2009; Dahl 2008; Bhatia 1997; Samraj 2002; Ozturk 2007 on introductions; Harwood 2005a; Bruce 2008; Lim 2006 on methods sections; Bruce 2009; Brett 1994 on results sections; Holmes 1997 on discussion sections).

In this preliminary situational analysis, no attempt was made to segment texts into these sections. Rather, the general structures of the research articles were analyzed by examining how sections of the article were labeled, and if not labeled, by the basic topics covered in different parts of the article. In this taxonomy, three categories of organization were identified: articles with (at least) sections including the content of the traditional IMRD sections, articles with the IMRD sections appearing in a different order, and articles with other sections and/or organizations.

While some of these features (such as the use of numbered or non-numbered headings, the standardized heading titles, etc.) may seem to merely reflect disciplinary convention, Becher (1994: 153) claims that cultural phenomena (like the culture of academic disciplines) "may be best understood in terms of an arbitrary convention". Thus, I have included these types of conventionalized characteristics as possible markers of registers, with the assumption that trends in how various aspects of research articles are conventionalized reflect the disciplinary culture.

4.3.3 Setting

As Conrad (1996a) points out, the setting of a communicative event has traditionally referred to place and time in analytical frameworks. Aspects of the setting that are relevant to journal registers, however, are much more restricted. In this framework, I am only concerned with the status of the journal in which texts were published, and I distinguish here between generalist journals (those that publish in a variety of areas within a discipline) and specialist journals (those which focus on a particular sub-discipline or topic). Each journal was categorized as a 'specialized' or 'generalist' journal based on input from the disciplinary informants and through close readings of the journal descriptors provided in the front cover of journals or in bibliographic entries within online archiving systems.

While Conrad (1996a) also distinguished between journals published in Britain and America, this distinction has not been made in the present framework, as no effort was made to verify the status of authors as speakers of American or British English (or as non-native English speakers) in the corpus collection process for this study, and the corpus design intentionally included international journals.

4.3.4 Subject/topic

In common use, the subject or topic of a text is simply what it is about. In academic writing, what a text is about is commonly thought to be largely determined by discipline. In fact, at the most basic level, disciplines can be characterized by their subject matter, although researchers such as Becher (1981) argue that disciplines are much more than a collection of knowledge about a particular subject matter.[1] Still, the topics that form the basis for disciplinary cultures seem to be a unifying parameter for all disciplines, even those which have more varied subtopics. For example, both biology and applied linguistics have a general topic that binds together scholarship in these disciplines, although biology has a large diversity of subtopics within the discipline (even leading Becher 1981:117 to call biology a 'fragmented' subject). However, the degree to which a discipline is 'fragmented' is difficult to measure empirically or reliably. For example, a field with which a person is very familiar may seem more fragmented than an unknown field, simply due to the fact that this person recognizes many distinct topical divisions where an outsider would not. On the other hand, an outsider could perceive a discipline as more fragmented than it actually is when faced with a detailed listing of possible

1. Specifically, Becher (1981:109) calls disciplines "cultural phenomena" that are "embodied in collections of like-minded people, each with their own codes of conduct, sets of values and distinctive intellectual tasks".

topics within the field if he or she is unable to recognize the relationships between the topics. Thus, synthesizing the topics or a range of texts can be rather challenging for the non-specialist.

Topics can be classified quite broadly in ways that characterize the subject matter of entire disciplines, yet perhaps more common is to describe a discipline by the many subtopics that are considered in that discipline. For example, disciplinary professional associations and handbooks introducing a particular field of study are likely to have an overarching statement of the topic of the discipline, but also focus on identifying a listing of sub-topics. Schmitt (2010), in an introduction to the field of applied linguistics, identifies seven areas of important inquiry in the field, including second language acquisition, psycholinguistics, descriptions of language and language use, and so on. Likewise, the American Political Science Association website lists 42 subtopics related to political science,[2] each falling under their broader definition of political science as the "study of governments, public policies and political processes, systems, and political behavior".[3]

In sum, we can classify subject/topic quite broadly in a way that encompasses the nature of the entire discipline, or we can very specifically state the subject matter of a particular article (which is often indicated in part by listing of key words near the beginning of the article). For the purposes of the present study, however, a broader approach is necessary due to the feasibility of identifying the topics of the 270 research articles in the corpus. While ideally a grouping of topics could be identified inductively from an analysis of each article, this was not feasible for the present study for several reasons. The most influential reason, however, is the fact that when it comes to the more specific sub-topics within a discipline, it is often the case that a single article falls within multiple subject areas, an issue that has previously been raised by Conrad (1996a: 53–54).

Because no attempt was made to categorize the topics of disciplines in either the initial register taxonomy or the formation of operational definitions, the following approach was taken to classify the topics in the various sub-corpora. First, an informal survey was taken to identify how each of the six disciplines are described in textbooks, handbooks introducing the field, previous linguistic and epistemological research on these disciplines, and on websites for professional organizations associated with each discipline. A general statement of topic was formulated based on commonalities found in these sources, and any prominent sub-topics were also noted. This general topic is presented in Table 4.2 below.

2. ⟨http://www.apsanet.org/content_4596.cfm⟩

3. ⟨http://www.apsanet.org/content_9181.cfm?navID=727⟩

Then, each text was examined to verify that the topics covered by the texts in the corpus fall within that general topic area.

4.3.5 Purpose

In studies of register variation, the purpose of different registers is typically seen as one of the most influential situational factors, and discussions of the varying purposes of texts and how those relate to the use of linguistic features often play a dominant role in interpretations of register variation. At the most general level, academic writing is seen as having an informational purpose (Biber 1988, 1995, 1994; Biber & Conrad 2009). Researchers conducting linguistic analyses of academic writing often offer statements of the purpose of academic writing, and these various purposes statements typically lead back to general informational purposes.

Conrad (1996a) points out that the purpose of any given register or text can be analyzed at several different levels of specificity, and defining specific purposes can be difficult, often requiring more subjective interpretations. Taking an extensive look at purposes proposed in the literature for academic writing, Conrad proposes seven general purposes for academic writing, but her analysis of research articles in biology and history classifies the purpose for all the research articles as "contribute new knowledge" (p. 112).[4]

However, we can state additional purposes of academic articles according to different levels of specificity. For example, we can specify a purpose that is related to research type, as identified in the taxonomy development analysis carried out in Chapter 4: the overall purposes for reports of theoretical and empirical research are as follows:

- Theoretical Research → to propose, explore, and advance theoretical arguments
- Empirical Research → to present the analysis of observed data

That is, we can classify the overall purpose of all academic research articles (to contribute new knowledge) and the overall purposes of different article types (to propose, explore and advance theoretical arguments versus to present the analysis of observed data). In this study, I am primarily making a distinction of the overall purpose between theoretical and empirical research articles. While we can also identify more specific purposes in each article, I found it difficult (and thus challenging in the ability to reliably code the purpose) to quantify specific purposes on a per-text basis. Thus, the purposes of disciplines and registers have been more generally characterized based on (a) published statements about the purpose of

4. Conrad (1996) also goes on to analyze the purposes for each section in the research article.

research in particular disciplines, (b) information from the disciplinary informants that I consulted, and (c) an inductive analysis of the explicit statements of purpose, goals, or aims of the texts represented in the Academic Journal Register Corpus.

To analyze the purpose statements posited by the article authors themselves, I first examined abstracts for statement of purpose, followed by a reading of the first several pages of the article to identify whether any statements of purposes were included in the article. Here, I relied upon explicit statements that could be identified more objectively, such as direct statements of the purpose, goals, or aims of the study. I then inductively identified patterns in these purpose statements for each discipline and register. The results of this process, along with comments about purposes gleaned from previous research and disciplinary informants, are presented in the prose discussion of each discipline in Section 4.4 below.

It should be noted, however, that the more specific purposes stated within research articles themselves typically refer more generally to the purpose of the *research,* and not of the rhetorical package (i.e., the article). The more specific purpose statements that I summarize below are reflective more generally of *research* in these disciplines and registers, and less so to the articles themselves (where the true purpose is to convey the entire research process, from motivations to methodology to results).

4.3.6 Nature of data or evidence

In every discipline, the foundation for research is in evidence. However, the nature of that evidence is widely varied, and there is often debate even within a discipline about the nature and role of evidence. For example, Jordanova (2000) illustrates both the fundamental importance of evidence, as well as differing views on the role of evidence in history. Jordanova at once claims that historians' "engagement with their chosen [sources] is so fundamental" (Jordanova 2000: 28) that the issue of evidence deserves a position in the forefront of the discussion of the discipline of history, and yet that "it is potentially misleading to think of historical research in terms of 'data'" (Jordanova 2000: 23). For Jordanova, there is a clear mismatch between what is viewed as 'data' in academia, and what historians themselves use as the foundation for their research. Just as topic or subject matter varies widely across disciplines, so too does the nature of the data and evidence that is used to explore that subject matter. Thus, it is highly likely that the nature of evidence will be illustrative of disciplinary register differences.

In this light, the category of 'Nature of Data or Evidence' is included in the present framework, particularly in terms of three ways of describing that evidence and/or data. At the most fundamental level is the question of whether the research relies on observed data at all (not always a straightforward distinction,

as illustrated by the quote from Jordanova 2000 above). A description of what is considered data in a discipline is discussed in the analysis below.

The second characteristic that I am using to describe the nature of data has strong connections to methodology, and is concerned with whether that data is quantifiable and is then analyzed through quantitative means. There are two pairs of contrasts that can be made with respect to this characteristic: quantitative versus qualitative research, and quantitative versus mathematical research. In qualitative research the data typically does not lend itself to quantification, but rather is interpreted based on common themes found in the data as interpreted by the researcher. Quantitative research, on the other hand, contains rich numerical data that are measurements of some phenomenon, and which can be summarized and/or tested using statistical means (including simple descriptive statistics). Empirical quantitative research, however, differs from purely mathematical research. Mathematical research (such as the research in theoretical physics) contains no observed data, but rather relies on mathematical constructs, formulas, and calculations.

These differences lead to the third characteristic included under the category 'Nature of Evidence': primary presentation of evidence. This characteristic focuses on the way in which evidence is presented in the text: through extensive prose descriptions, through quantitative displays such as tables and figures, or through progressions of mathematical formulas. While all research generally contains varying amounts of prose around the presentation of evidence as that evidence is discussed or analyzed, this characteristic focuses on the primary way in which writers place the evidence in the discourse.

The final characteristic of importance is the actual object of study. That is, what is being examined, and what does it represent? In this analysis, information from published sources about the nature of evidence and an inductive survey of the evidence used in the studies included in the corpus provide the basis for analysis.

4.3.7 Methodology

Related to the nature of data are the procedures that researchers use to make sense of that data. Traditionally, there has been a split between experimental and observational research, and I have maintained this distinction for qualitative and quantitative research. The nature of the research design is typically explicitly stated through descriptions of research protocol, and I have relied upon these author-provided descriptors to make these distinctions in my analysis. When the author does not provide such a label, this coding decision has been based on the description of the research methods.

Also under the category of methodology, I have included a description of how quantitative data is analyzed. Based on an initial survey of the studies represented

in the corpus, and the desire to make reliable decisions about this characteristic, I use a three-way distinction in this analysis: descriptive statistics, methods which test for significant differences between groups, and all other advanced statistical techniques (e.g., regression, modeling, etc.). Although many more sophisticated distinctions are possible here, these three distinctions aptly characterized the research designs represented in the corpus to a degree of specificity that was feasible and reliable to apply.

4.3.8 Explicitness of research design

The final major category in the situational framework has been conceptually adapted from Conrad (1996a), and deals with the explicitness with which aspects of the research are described in the articles. In the present framework, the focus is on how explicitly the following five aspects of the research are addressed: purpose, research questions, citations to previous research, explanation of evidence, and explanation of research methods and procedures.

While Conrad (1996) characterized her 'explicitness' parameters using a three-way descriptor (extensive, mention, and none), I have adopted a two-way descriptor in the present study. For purpose and research questions, the concern is with whether or not there is a direct statement regarding the author's goals/aims in the study, or the specific research questions that the research is intended to address. Explicit statements for both of these features were searched for in (a) the abstract if present, and (b) the first few pages of the article, particularly in the paragraphs preceding any sections on data/methods (if present). Any statement which described what the authors intended to do in the study were counted as explicit statements, although the form of this varied across disciplines (which will be discussed below). Research questions were categorized as directly stated when they appeared in question form.

Citation practices have been investigated in disciplinary discourses before, and Hyland (1999b) offers one of the most comprehensive studies that look at citations from a linguistic point of view. Following Swales (1990), Hyland distinguishes between 'integral' and 'non-integral' citations, defining 'integral' citations as those which use the researchers' names and a reporting verb within the prose of the article, and non-integral citations as those which are presented solely in parentheses within the text or as cited footnotes or endnotes.

The distinction made in the present framework is slightly broader, and is based on preliminary observations about the citation practices represented in the corpus. That is, while many disciplines/registers used both integral and non-integral citations in the sense proposed by Swales (1990) and documented by Hyland (1999b), in the present corpus a more common split seemed to exist between disciplines

which present citations within the text of the article (i.e., either as integrated structures using names and reporting verbs or in a name and date format in parentheses in the text) and those which include only numbered footnotes or endnotes that then contain the referenced material. Thus, the present framework distinguishes between articles whose primary means of citation is through references within the text, and those which use footnotes or endnotes to contain the references. The decision to categorize citation practices in this way was reinforced by the fact that often the articles in the Academic Journal Register Corpus either include both integral and non-integral citations (in Hyland's sense), or references in footnotes (without reference information embedded within sentences).

Finally, within the category of explicitness, I have included (following Conrad 1996a) categories for the manner in which evidence and methodological procedures are described in articles. Here, the distinction is between articles which contain developed descriptions of data and methods, and those which primarily omit explicit information about these aspects of research.

4.4 The situational characteristics of the Academic Journal Register Corpus

In this section, I present an analysis of the situational characteristics of three journal registers as represented by the 270 texts in the Academic Journal Register Corpus. First, I address a few of the situational characteristics that are included in other analytical frameworks, but which were not included in the present study because of the lack of variation with regard to these characteristics. Second, I turn to a description of each register based on the framework presented above. While Table 4.2 presents exact distributions whenever possible, the discussion below will focus on identifying common trends in the characteristics of each discipline and register. As mentioned above, the description of characteristics such as topic and the nature of evidence have been informed by published sources within the disciplines of interest and through inductive surveys of articles in the corpora. First, though, I begin with a few situational characteristics that are shared across journal registers and disciplines.

4.4.1 Common characteristics across journal registers

As mentioned above, previous frameworks for comparing the situational characteristics of registers have addressed many characteristics that are not included in the framework adapted for this study because of a lack of variation in those characteristics. However, this is not to say that those characteristics should be omitted from discussions as we try to understand these registers. In this section, I very

Table 4.2. The situational characteristics of texts in the Academic Journal Register Corpus

	Philosophy (Theo)	History (Qual)	Political Science (Qual)	Political Science (Quant)	Applied Ling. (Qual)	Applied Ling. (Quant)	Biology (Quant)	Physics (Quant)	Physics (Theo)
Participants									
Writer									
1	26	30	23	9	21	13	1	1	6
2–4	4	0	7	21	9	17	20	14	21
5+	0	0	0	0	0	0	9	15	3
Textual Layout & Organization									
Length									
Mean page length	22.6	24.3	17.7	17.5	20.8	19.9	10.5	13.4	15.7
Standard deviation	5.9	6.7	6.8	6.6	4.9	5.6	2.9	7.3	8.6
Page length range	10–33	8–42	8–39	9–34	14–32	10–32	6–19	5–32	4–37
Headings									
None	0	20	0	0	0	0	0	1	1
Un-numbered	5	10	4	5	22	24	0	3	1
Numbered	25	0	26	25	8	6	30	26	28
Use of Abstracts									
yes	20	15	22	23	30	30	30	30	30
no	10	15	8	7	0	0	0	0	0
Visual Elements									
None	29	21	19	0	17	0	0	0	1
Tables	0	1	4	17	10	17	0	0	0
Figures	1	6	2	0	1	0	6	11	21
Tables & Figures	0	2	5	13	2	13	24	19	8
Equations	0	0	0	8	0	0	14	24	30

(*Continued*)

	Philosophy (Theo)	History (Qual)	Political Science (Qual)	Political Science (Quant)	Applied Ling. (Qual)	Applied Ling. (Quant)	Biology (Quant)	Physics (Quant)	Physics (Theo)
Sections/Organization									
IMRD	0	0	1	26	23	30	25	28	0
IMRD in other order	0	0	0	0	0	0	5	2	0
Other	30	30	29	4	7	0	0	0	30
Standardized section heading	0	0	1	15	5	20	30	25	0
Variable section heading	30	10	29	15	25	10	0	5	30
Setting									
Nature of Journal									
generalist	19	8	8	10	6	6	3	4	3
specialized	11	22	22	20	24	24	27	26	27
Subject/Topic									
General Topic	Human thought, knowledge, and morality	Human & societal actions and events	Public and political processes, systems, and behavior		Structure, use and acquisition of human language		Living organisms	Physical world	
Purpose									
General Academic Purpose									
To propose, explore and advance theoretical arguments	30	0	0	0	0	0	0	0	30
To report on the analysis of observed data	0	30	30	30	30	30	30	30	0

(*Continued*)

Table 4.2. (Continued) The situational characteristics of texts in the Academic Journal Register Corpus

	Philosophy (Theo)	**History (Qual)**	**Political Science (Qual)**	**Political Science (Quant)**	**Applied Ling. (Qual)**	**Applied Ling. (Quant)**	**Biology (Quant)**	**Physics (Quant)**	**Physics (Theo)**
Nature of Data or Evidence									
Presence of Observed Data									
yes	0	30	30	30	30	30	30	30	0
no	30	0	0	0	0	0	0	0	30
Use of Numerical Evidence									
yes	0	0	9	30	0	30	30	30	30
no	30	30	21	0	30	0	0	0	0
Primary Presentation of Evidence									
Prose discussion	30	30	30	0	30	0	0	0	0
Quantitative displays	0	0	0	30	0	30	30	30	0
Mathematical formulas	0	0	0	0	0	0	0	0	30
Object of Study	unreal scenarios; logical progressions	historical documents and artifacts	historical documents and artifacts; survey data	survey data, government statistics	language production; process observation interviews	language production	measures of living organisms	laboratory data; measures. of natural phenomena	numerical models & simulations; mathematical logic formula
Methodology									
observational	–	30	30	27	30	13	15	5	–
experimental	–	0	0	3	0	17	15	25	–

(*Continued*)

	Philosophy (Theo)	History (Qual)	Political Science (Qual)	Political Science (Quant)	Applied Ling. (Qual)	Applied Ling. (Quant)	Biology (Quant)	Physics (Quant)	Physics (Theo)
Statistical Techniques									
None	–	–	–	0	–	0	0	0	–
Descriptive Statistics	–	–	–	6	–	6	2	0	–
Statistical Difference Testing	–	–	–	2	–	24	13	0	–
Other advanced statistics	–	–	–	22	–	0	15	30	–
Explicitness of Research Design									
Explicitness of Purpose									
Direct statement	26	18*	23	30	29	30	29	27	29
Minimal / No statement	4	12	7	0	1	0	1	3	1
Explicitness of RQs	.								
Direct statement	–	1	8	15	7	25	3	1	0
Minimal / No statement	–	29	22	15	23	5	27	29	29
Explicitness of Citations							.		
Within text	15	0	16	25	30	30	17	7	10
In notes	15	30	14	5	0	0	13	23	20
Explanation of Evidence									
extensive	0	0	9	22	27	30	30	30	30
mention / none	30	30*	21	8	3	0	0	0	0
Explanation of Procedures									
extensive	–	0	2	25	23	30	30	30	–
mention / none	–	30	28	5	7	0	0	0	–

briefly discuss some of the characteristics omitted from the formal framework: mode, production circumstances, reader and writer status, and general academic purpose.

All of the texts in the corpus are written texts, produced over a period of time in which they can be extensively revised and edited. These texts are typically written by highly knowledgeable individuals with advanced degrees, and are written for a specialized audience of fellow academics who typically have a high level of shared knowledge with the authors. Thus, although little direct context is shared (e.g., time and place), we can assume a certain degree of shared context in terms of both readers and writers understanding specific content and the more general research contexts (e.g., typical research methods in the field, previous research on the topic, etc.).

While these features are key characteristics of all of the disciplines and registers represented in the corpus, there are many other features that vary. Table 4.2 displays the full situational analysis for the texts in the Academic Journal Register Corpus based on the framework presented in Section 4.3. In the sections that follow, key information from this table is summarized and illustrated with examples from the corpus texts. However, it is important to note that while I have tried to develop a framework which can be comprehensively applied to identify differences and similarities in the situational characteristics of texts across a range of disciplines and registers, the analysis presented here is restricted to the texts that are represented in the corpus. Indeed, it is beyond the scope of the current study to describe the nature of all research articles within a discipline.

4.4.2 Theoretical articles in philosophy

Articles in philosophy are typically single-authored texts that rely on extensive prose (as evidenced by the 2nd highest mean page length, and the lack of space-consuming tables and figures). As texts reporting on theoretical rather than empirical research, there is a lack of observed data, and consequently features associated with the collection and analysis of observed data are absent from these articles (i.e., specific statements of research questions, explanations of research procedures). Likewise, the traditional Introduction-Methods-Results-Discussion structure of texts is not used in theoretical philosophy articles.

As a discipline, the subject matter of philosophy is focused on understanding fundamental problems or workings of the human existence, including the nature of knowledge, morality, truth, human rights, and key aspects of the human psyche such as desire, belief, and trust. Most articles in theoretical philosophy have an

explicit statement of the purpose of the research, and it is not uncommon to find this purpose in an abstract (if present) as well as in the first part of the article. These stated purposes typically reflect the general purpose of theoretical articles (to propose, explore and advance theoretical arguments), and are often described in terms that illustrate the exploratory, logical progression of arguments, as well as the end conclusion to such exploration – an argument for or against a particular concept or theory:

4.1 In this paper I **explore** the tension between the view that art is to be appreciated for its own sake and the apparent fact that much art is made to serve extrinsic functions… [PHIL-TH-03]

4.2 This paper **examines** the relationship between truth and liberal politics via the work of Bernard Williams and Richard Rorty. I **argue** that Williams is right to think that there are positive relations between truth, specifically a realist understanding of truth, and liberal politics that Rorty's abandonment of the realist vocabulary of truth undermines. [PHIL-TH-08]

4.3 I would like to **challenge** the idea that only objective theories have this uncomfortable feature. I will **show** that any plausible theory justifying the defense of others, whether subjectively or objectively, can lead to situations of normative inconsistency. [PHIL-TH-15]

In order to carry out these purposes, researchers in theoretical philosophy rely on descriptions of imagined or unreal situations (excerpt 4.4), as well as progressions of logic (4.5). However, these pieces of evidence, which are often both used within a single article, are not explicitly presented as evidence. That is, the authors of theoretical philosophy articles do not offer meta-talk that describes these forms of evidence but rather use them to illustrate and provide a basis for their conclusions and claims throughout the sections in the article.

4.4 Having a right to do X is normally taken to imply that others must not prevent the doing of X. A right to do wrong may then come about in a situation in which a person has a duty not to do X, but still has a claim against others that they refrain from preventing him or her from doing X. That it is a right to do wrong implies that the agent has a duty to do otherwise. [PHIL-TH-26]

4.5 Bradley's regress is familiar. Suppose a is F. Suppose, for reductio, that it follows that a relation of instantiation holds between a and F, symbolized as Ra F. But now, it seems, R holds between a and F, and there is just as much reason to think that a relation of instantiation must bind R, a and F as there was to think that a relation of instantiation must bind a and F. So a relation holds between R, a and F. [PHIL-TH-22]

4.4.3 Qualitative articles in history

The discipline of history is concerned with describing and analyzing the events of the past, and can include topics focused on particular time periods or events, specific geographic locations or regions, individual people and social groups, and political and institutional entities and processes. The overarching theme, however, is that history is concerned with the analysis and description of human events.[5] Wilson (1999: 29–30) characterizes the goal of historians as follows: "[m]any historians do not want merely to recover the past, they seek to render a meaningful history of the past". Historians come to these descriptions through basic techniques to describe, narrate, and analyze historical events (Tosh 2000: 92).

Although only about two-thirds of the history articles in the corpus included direct statements of the purpose of the article/research, these characteristics are illustrated through the purpose statements that are provided, where verbs and phrases like *explore, examine, discuss, seek an alternative explanation, analyze, trace, uncover what happened, contrasts,* and *look at* reflect the exploratory nature of research, and then lead into the authors' claims or interpretations*:*

4.6 This article will **attempt to present** the Labour party's thinking on the land question. It will **examine** the changing nature of land-related policies brought forward by Labour during the inter-war period and indicate the different, and indeed contradictory, policy positions adopted by the party. It is hoped that by contextualizing Labour's thought in this way it will be possible to **achieve a better understanding** of the Attlee government's policy of not nationalizing the land. [HIST-QL-09]

4.7 This article **explores** the efforts of civic leaders to secure federal slum clearance in these three important cities and **argues** that the failure to embrace urban renewal did not simply stem from conservative leadership, but from a significant shift in political culture. [HIST-QL-27]

4.8 This essay **discusses** the Apaches and Pawnees who joined the post-Civil War U. S. Army as workers and **argues** that they functioned and were used as colonized labor, a special race-based colonial labor system characterized by constant negotiation and tension between integration and exclusion, valuing and othering, and indigenous freedom and colonial control. [HIST-QL-25]

History as a discipline varies in nature when compared to other empirical, scientific disciplines, and in this section I devote a bit of discussion to this difference

---

5. This general topic statement is in line with Conrad (1996a), and reflects various summaries of the content of historical inquiry, namely Jordanova (2000), Wilson (1999), and Tosh (2000).

as many of the situational characteristics in the framework vary in ways that can be better understood by first considering the nature of history from the point of view within history itself. Munslow (1997: 4) summarizes the major difference between history and other scientific, empirical disciplines as follows: "history cannot claim to be straightforwardly scientific in the sense that we understand the physical sciences to be because it does not share the protocol of hypothesis-testing, does not employ deductive reasoning, and neither is it an experimental and objective process producing incontrovertible facts." Wilson (1999: 1) echoes this sentiment, claiming that the "nature of historical evidence is actually quite distinct from scientific evidence because history cannot be repeated in similar conditions." That is, although the interpretations and analyses that historians offer are based on various forms of historical evidence, the nature of the data and the nature of the inquiry are fundamentally different than other empirical research.

Historians rely on a great variety of evidence, typically in the form of written documents (e.g., chronologies, autobiographies, press reportage, official publications and records, and private registers such as diaries and letters; see Tosh 2000).[6] And while the "traces of the past are thus traditionally viewed as empirical objects from which to mine *the* meaning, or as sources out of which social theories of explanation can be constructed" (Munslow 1997: 7; emphasis in original), the terms 'empirical' and 'data' are *not* commonly used within the field (personal communication, G. Lubick, January 26, 2010; Jordanova 2000) to describe the research practices which take place.

Thus, it is not altogether surprising that history articles typically do not directly state research questions, do not explain what they will be using for evidence and how they will go about analyzing it, and are not organized in the traditional IMRD sections (see Table 4.2). Articles in history continue this general trend of non-explicit marking with their infrequent use of headings (20 out of the 30 articles in the corpus do not use any headings) and with their use of references primarily in endnotes or footnotes rather than embedded into the prose of the article.

---

6. This multitude of evidence types exemplifies why no attempt was made to categorize the evidence types in this situational analysis, but rather to offer general statements about the nature of that data. As Tosh (2000: 65) points out, in History research the "procedure is rather to amass as many pieces of evidence as possible from a wide range of sources – preferably from *all* the sources that have a bearing on the problem at hand". This results in articles which use many different types of evidence in a single research endeavor.

4.4.4 Qualitative and quantitative articles in political science

Political science is a broad discipline that is concerned with describing public and political processes, systems, and behavior,[7] with the mission "to elucidate how social power is, can be, and should be exercised and constrained" (Goodin 2009:6). Collier (1993) identifies three major types of research within political science: statistical analyses, experimental research, and historical studies. In broad terms, the first two types of research correspond to quantitative research in political science. Historical studies, the third type, generally correspond to qualitative research (S. Wright, personal communication, September 10, 2009).

Unlike quantitative research in political science, qualitative research has a primary goal to provide an analysis with the intent of explaining the topic of interest (S. Wright, personal communication, September 10, 2009). In fact, all of the qualitative political science articles in the corpus are historical in nature, and share many of the situational characteristics of qualitative history articles, such as this general purpose. Also similarly to qualitative history articles, qualitative political science articles generally do not contain tables and figures, are not organized into traditional IMRD sections, and use a varied range of historical data for analysis, including governmental records. While many qualitative political science articles do have a direct statement of their research purposes (see Examples 4.9–4.10), there is little direct explanation of what the data is or specifically how the data was collected and analyzed. If there is an explicit mention of the data used for the analysis, it is generally quite brief as in 4.11 (data mention is underlined). Rather, qualitative studies jump into the contextualization of the situation and analysis (S. Wright, personal communication, September 10, 2009).

4.9 This paper **examines** the particular political economy of the PTT- the politics, negotiation and contestation constituting its implementation–and **considers** how the programme and later processes of debt forgiveness and parcelisation, shaped conditions of access to resources, livelihood options, and processes of land use for resettled communities. [POLISCI-QL-05]

4.10 This paper seeks to **contextualise** the 1999 Turkish earthquake within the institutional structure of Turkey's development. Particularly **focused** upon the role – and culpability – of the state in the disaster, it **outlines** a number of key continuities within Turkey's political tradition. In all, it **argues** that Ankara's inadequate response can be understood both in terms of the persistence of these older social structures and in a more recent weakening of the public sector. [POLISCI-QL-22]

7. This broad statement is in line with topics and generalizations covered in Goodin (2009), Goodin & Klingermann (1998), Heineman (1995), and the American Political Science Association website.

4.11 This article **investigates** the language adaptations that facilitated the changes in agricultural conservation policy by analyzing the five policy-design elements developed by Ingram and Schneider. [POLISCI-QL-13]

A key characteristic of qualitative political science articles is that some articles (9 out of 30 in the corpus used here) report quantitative data. However, these studies are distinguished from quantitative articles in the purposes for which they use that data. In qualitative studies, the data is used to tell a story, but the data can be considered "passive" in that it is not the subject of direct analysis (S. Wright, personal communication, October 5, 2009), and no statistical analyses are used to analyze that data.

Quantitative articles in political science, on the other hand, are more often authored by multiple researchers, typically use data displays such as tables and figures, and are often organized into the general Introduction-Methods-Results-Discussion format. Evidence in quantitative studies typically comes from large datasets, surveys, and government-level data, and typically use sophisticated statistical techniques to analyze data. While quantitative political science articles more often provide direct statements of data and research methods, these statements are much less detailed than, for example, data and method descriptions in applied linguistics articles. In addition, purpose statements are stated in direct terms what the study did, as in the following example:

4.12 In this article, we **investigate** one highly significant aspect of the role of money in judicial elections: whether campaign spending increases citizen participation in the recruitment and retention of judges. Specifically, **by using** a two-stagemodeling strategy that allows us to separate the effects of challengers from the effects of money, we **assess** whether relatively expensive campaigns improve the chances that citizens will vote in the 260 supreme court elections held from 1990 through 2004 in 18 states using partisan or nonpartisan elections to staff the high court bench.We find that increased spending significantly improves citizen participation in these races. [POLISCI-QT-01]

4.4.5 Qualitative and quantitative articles in applied linguistics

The subject matter of applied linguistics can be summarized as concerned with the structure, use, learning and teaching, and users of language.[8] The knowledge-making processes are often explicitly stated in both qualitative and quantitative articles, particularly in terms of the stated purposes of the articles/research and the

8. This summary statement is based off of trends found in Schmitt (2010), Kaplan (2010), and the American Association for Applied Linguistics strand listings. Available at ⟨www.aaal.org⟩

explanations of data that is used for the analyses. These stated purposes typically focus on the goals of the overall research project:

4.13 This qualitative exploratory study was designed to **provide insight** into the role of the L1 when L2 learners are engaged in consciousness-raising, form-focused tasks. [AL-QL-08]

4.14 The present study aims to **expand our existing knowledge** of task-induced involvement by testing its predictive power on word learning by beginning learners of Spanish and by assessing its impact on both passive (receptive) and active (productive) word knowledge. [AL-QL-16]

Both applied linguistics registers typically directly discuss research procedures and methods (although qualitative articles do this to a slightly lesser extent). One major difference, however, is that quantitative articles typically explicitly state one or more research questions which the study is intended to address, while very few qualitative articles do so.

A further difference between qualitative and quantitative applied linguistics articles is that all of the quantitative articles in the corpus roughly followed an IMRD organization pattern with fairly standardized headings indicating, for example 'methods' and 'participants' and 'results'. Headings in quantitative applied linguistics articles are longer, more descriptive, and more variable than headings in, for example, biology and physics; however, they are relatively standardized in that they contain key terminology to indicate the content of the section (e.g., data, methods, procedures, discussion). In contrast, qualitative articles are a bit more likely to use other organizations, and are much more likely to use highly descriptive headings that varied widely across articles.

In contrast to quantitative political science (and, as we will see, quantitative biology and physics), quantitative applied linguistics articles primarily rely upon statistical techniques that test for significant differences between group means, such as ANOVAs, t-tests, Chi-square tests, and so on, and rely fairly evenly on observational and (quasi-) experimental research designs.

Related to this issue of the application of statistical techniques is the nature of the data being analyzed in applied linguistics. Although both registers within applied linguistics often rely on language data, qualitative research also focuses on characteristics of the language learners, thus analyzing classroom processes and learners' reactions to those processes to a greater extent than quantitative research.

4.4.6 Quantitative articles in biology

As noted above, the discipline of biology is highly diverse in terms of the subject matter covered under the larger umbrella of the discipline. In general terms,

however, biology is concerned with life, that is, with the living organisms that occupy the planet (e.g., see the introductory chapter in Reece et al. 2010). As any perusal into an introductory biology textbook shows, the study of life can be applied at many levels, including cells, molecules, plants, animals, and ecological interactions. Although I will not attempt to classify each of the articles to one of these specific subtopics, the articles included in the corpus represent a range of these sub-disciplines, from aquatic beetles to thyroid hormones, from parthenogenetic lizards to microbial resistance in broiler chickens, and so on.

Biology articles are typically written by 2–4 authors, and have the shortest mean length of any register or discipline in the corpus. In addition, biology articles are highly cohesive in terms of their situational characteristics. All biology articles had abstracts and contained sections corresponding to the IMRD format with no variation in the standardized headings of each section. The only difference between biology articles with respect to organization is that articles from one journal had the sections ordered as Introduction-Results-Discussion-Methods. All articles contained extensive descriptions of both data/evidence and the procedures undertaken to collect and/or analyze that data, and all but one contained a direct statement of purpose. However, although specific purpose statements for the research are provided in almost every article, these statements are often embedded in the text of the introduction, are typically quite brief, and are often stated more implicitly than the purpose statements seen in, for example quantitative political science and applied linguistics. For example, abstracts typically contain a statement of what was done in the study, but this is not overtly packaged as a purpose, as in the following statement from the beginning of an abstract:

4.15 Magnetic resonance imaging (MRI) was used to spatially resolve the structure, water diffusion, and copper transport of a phototrophic biofilm and its fate. [BIO-QT-07]

This more general statement of what was done contrasts with the purpose statements packaged more explicitly like those in abstracts in applied linguistics and political science. Excerpt 4.16 below comes from the same article as 4.15, but occurs embedded within the introduction section of the article, and more explicitly packages the content of the experiment as a purpose:

4.16 The key aim of this study was to investigate the ability of MRI to characterize the transport and fate of heavy metal in a natural biofilm, thus demonstrating its potential for probing biofilm-metal interactions in natural systems. [BIO-QT-07]

While purpose statements are always present, it is a bit surprising that research questions are not directly stated in biology articles; however, research questions

can be implicitly stated, as in 4.17 where the research questions are assumed from the statement of the gaps in the previous research:

4.17 However it is rarely clear the extent to which individual reproductive isolating barriers related to habitat differentiation contribute to total isolation. Furthermore, it is often difficult to determine the specific environmental variables that drive the evolution of those ecological barriers, and the geographic scale at which habitat-mediated speciation occurs. Here, **we address these questions** through an analysis of the population structure and reproductive isolation between coastal perennial and inland annual forms of the yellow monkeyflower, Mimulus guttatus. [BIO-QT-25]

4.4.7 Quantitative and theoretical articles in physics

Perhaps even more so than biology, physics is a wide and varied discipline with a great number of sub-disciplines, from astrophysics to quantum physics, from condensed matter physics to nuclear physics, and so on. In all of these fields, however, the overarching theme is the investigation of the natural world. Because of the broad nature of the discipline, and the degree of specialist knowledge required to identify distinctions between categories of physics inquiry, my analysis of the topics represented in the corpus will be restricted to a discussion of the journals from which the texts in the corpus come from. While several more general journals are included, a wide variety of sub-disciplines within physics are also represented. For example, in addition to articles from general journals, the corpus is also composed of articles within condensed matter, nuclear physics, astrophysics, particle physics, geophysics, and atomic and molecular physics; thus, the articles in the corpus reflect a broad range of topics.

Quantitative and theoretical physics registers are included in the corpus, and the two registers have several shared characteristics, as well as some substantial differences in the situational characteristics of the texts. For example, few physics articles are written by single authors; however, half of the quantitative texts have five or more co-authors for a single text, while only three theoretical articles have five or more authors. Both registers use primarily numbered headings and always begin with abstracts. Both registers also rely heavily on visual elements within the texts, such as tables, figures, and substantial equations (especially common in theoretical articles).

Quantitative articles follow a fairly standardized IMRD format where sections are usually labeled with standardized headings. In these sections, data and procedures are described in detail. In contrast, theoretical physics articles do not have these same sections, in part due to the fact that the theoretical articles do not contain data in the same way that empirical articles do.

Like biology, physics articles nearly always state a purpose for the research/article. However, this statement is often less explicitly labeled as a purpose, and is instead focused on a description of the exact research topic. The statement of purpose is typically embedded within the introduction section of the article, as in 4.18 and 4.19 below. While similar information also appears in the abstract, the abstract is framed more procedurally. For example, in 4.19, the introduction presents a goal of finding expecting relationships, while the abstract in contrast focuses on the fact that "[a] theoretical model is developed".

4.18 Measurements of proton-induced fission on 232 Th, 238 U, 237 Np, 239 Pu and 241 Am nuclei at proton energies of 26.5 MeV and 62.9 MeV **were performed** at the Louvain-la-Neuve cyclotron facility. [PHYS-QT-01 from the abstract]
[…]
In this work, we **report** on some of the results of our research program that focuses on measuring the properties of proton-induced fission on target nuclei such as 232 Th, 238 U, 237 Np, 239 Pu, and 241 Am, at proton projectile energies (Ep) of 26.5 and 62.9 MeV. [PHYS-QT-01 from the introduction]

4.19 A theoretical **model is developed** that is applicable to the electric field fluctuations that arise in the polar summer mesosphere as a result of the coupling of the charged species to the neutral air turbulence. [PHYS-TH-17 from the abstract]
[…]
The **goal of this work is to find the expected relationship** between the electric field fluctuations, the charge density fluctuations, and the neutral air turbulence. [PHYS-TH-17 from the introduction]

Twenty-five of the quantitative physics articles are experimental in nature, and contain laboratory data on the measurements of various physical phenomena. A variety of sophisticated measurement and statistical techniques are used. Theoretical articles, on the other hand, base their analyses upon either computer-simulated data or complex progressions of mathematical formulas. The following excerpt from theoretical physics illustrates the use of mathematical formulas, where writers move through a series of formulas, leading the reader through the steps of the formulas:

4.20 Here, a more general version of the result of Ref. 20 will be obtained in a form that is almost equally convenient for large-scale applications although it does require additional parameters – namely dipole matrix elements,
[formula], (6)
and on-site (X' = X) matrix elements of the momentum operator,
[formula], (7)
where α labels an orbital centered on the nucleus whose instantaneous position is X. Recall that ℓ labels both nucleus and orbital, so at a given instant in time
[formula]. (8) [PHYS-TH-22]

4.5 Trends in the situational characteristics of the Academic Journal Register Corpus

The situational analysis of the Academic Journal Register Corpus presented in this chapter has shown variation both within and across disciplines in terms of the non-linguistic characteristics of these texts. While these have been summarized in Table 4.2 above and discussed in Section 4.4, in this section I would like to take note of a few somewhat surprising trends that have come out of this situational analysis.

First, it is interesting to note that the number of authors for a particular text is highly related to the discipline and registers of the texts. That is, philosophy texts are overwhelmingly single-authored, as well as qualitative history. Likewise, within political science and applied linguistics, qualitative texts are much more likely to be single-authored than their quantitative counterparts. In contrast, the hard natural sciences of biology and physics are rarely single-authored, and in fact are the only texts that have five or more co-authors per text. On the one hand, this may be related to disciplinary divisions, where research in the hard sciences (particularly experimental research) requires sophisticated and expensive equipment, involves many technical steps in the analysis, and may therefore require more human power to carry out the research. On the other hand, the prevalence of single-authored articles in qualitative research may be related to the nature of qualitative inquiry, which relies upon the ability of the researcher to observe things in their natural setting and brings the researcher into the research context (Denzin & Lincoln 2000).

A third point of interest is the number of features which appeared to be standardized by journals and were rarely deviated from, such as the use of abstracts, citation styles, numbered or un-numbered headings, and the use of standardized terminology/labels for headings. Despite the fact that these features may indeed be dictated by journal style, we can also argue that the decisions journals make as a set style reflects some sort of value held within the disciplinary community. For example, why do the hard sciences nearly always use an IMRD style, with numbered headings, and standardized section labels? Why don't, for example, articles publishing qualitative research obligate researchers to standardize headings to label portions of the text with general descriptors, but rather allow longer, more descriptive titles?

A final trend, and one which was particularly surprising to me, was the relative lack of explicitly stated research questions in the hard sciences (and quantitative political science as well). One possibility is that the explicit statement of research questions is related to disciplinary epistemologies, with some disciplines requiring

overt research questions, while others rely on implied/implicit research questions (and thus, perhaps my surprise is a reflection of my training in research in the social sciences, which did tend to specify research questions more often than other disciplines). A further possibility, however, is that the coding framework used in this study did not allow for the identification of research questions which were more implicitly embedded into text, like the implied research questions identified in excerpt 4.17 above. Such an analysis was not feasible for the current study, due to the intensive work needed to analyze 270 articles at this level of detail, along with the need for a 2nd rater to enable reliability testing for a category which could not be identified by primarily objective means.

While I do not have the answers to these specific questions and issues raised in this section, these are items worthy of further reflection and study. I will return to these situational characteristics in Chapter 8 as the findings from the linguistic and non-linguistic analyses of the disciplines and registers are synthesized.

CHAPTER 5

A lexical and grammatical survey

5.1 Introduction

Previous research has documented the ways in which academic writing differs linguistically from spoken registers such as conversation (e.g., Biber 1988, 1992; Biber et al. 1999; Biber & Gray 2010; Halliday 1989; Wells 1960), particularly in terms of the grammatical structures that are widely used in academic writing and less frequently relied upon in conversation. Biber (2006) summarizes a comprehensive number of these structures, including many features associated with noun phrases: a dense use of all nouns, plural nouns, nominalizations, definite articles and demonstrative determiners, adjectives (particularly attributive adjectives), nouns as noun pre-modifiers, prepositional phrases as noun post-modifiers, and relative clauses with *which*. Biber (2006) also notes the prevalent use of several aspects of verb phrases, such as the copular verb *be,* existence verbs, verbs with inanimate subjects, simple aspect, present tense, and the use of passive voice verbs (particularly the short-passive). Other features more frequent in academic prose include linking adverbials, extraposed *that*- and *to*-clauses, prepositions, and *of*-phrases (see Biber et al. 1999; also see Biber 2006: 15–18 for a more complete summary).

This research has described academic writing from a more global perspective – that is, including multiple registers (e.g., textbooks, research articles, academic books) and disciplines, and have encompassed a range of grammatical features. Typically, studies on the grammatical characteristics of academic writing have considered academic writing as a whole, represented by a variety of disciplines or in some cases, by science writing specifically (e.g., Halliday's work largely focused on science writing). The possibility that meaningful variation exists in the general grammatical structure of writing in different disciplines and sub-registers within academic prose has been largely disregarded in this previous research.

Thus, it is this possibility that motivates the analysis reported on in this chapter, in which I take a corpus-based approach to analyzing lexical and grammatical variation in academic journal registers and disciplines. In this chapter, I focus on core grammatical categories and their distributions of use across disciplines and registers. In Section 5.2, I briefly summarize previous research documenting dis-

ciplinary differences in the use of these core grammatical concepts. In Section 5.3, I detail the method used to investigate general lexical and grammatical features in the present study. Finally, in Section 5.4, I report the results of this analysis.

5.2 Grammatical variation in academic prose

Previous studies on variation across broad register boundaries have shown that one of the primary differences across registers is the relative use of content word classes, such as nouns, verbs, adjectives and adverbs (for example, see Biber 1988 on a range of spoken and written registers and Biber 2006 on academic spoken and written registers). In these studies, spoken registers and written registers typically show distinct patterns, with written registers exhibiting a much greater reliance on nouns and adjectives, while spoken registers rely on nouns and verbs to similar extents but show a higher use of adverbs (see Biber 2006: Chapter 4). However, this research has focused on describing patterns across broader register categories that have greater differences in the situational characteristics of the registers (such as spoken versus written). In this study the registers under investigation are all written academic language, and thus we can expect less variation in the overall use of these core content word classes. We can also expect that all of the registers will show overall trends that correspond with previous findings that written registers rely on noun and adjectives, and less so on verbs and adverbs.

However, the possibility of disciplinary variation in the use of general grammatical features has been relatively omitted from much previous research. In fact, research into the grammatical characteristics of academic writing can typically be placed into three strands of inquiry which vary with respect to the amount of attention they pay to disciplinary variation and the types of grammatical features that are investigated:

1. research which compares the *basic* grammatical characteristics of academic writing in general with other spoken and written registers; much of this research has disregarded disciplinary differences or has focused on science writing more narrowly as a range of core grammatical structures are investigated
2. research which focuses on a particular lexical or grammatical structure within specific disciplines, typically comparing either a small number of disciplines or different types of writing in the same discipline (see Chapter 1, Table 1.1); while this research does often include a discipline-specific or a comparative approach to identify disciplinary variation, it also typically focuses on a small range of more specialized linguistic features

3. research which investigates a functional construct (e.g., stance) that has been realized through a range of related grammatical features, either for academic writing more generally, or in specific disciplines.

A substantial portion of the research in this third strand has been based on Hyland's corpus of research articles in 8 disciplines and has focused on the marking of stance (e.g., Hyland 1996, 1998; Hyland & Tse 2005) and other interactional features of discourse (e.g., Hyland 1999b on citation and attribution; Hyland 2001a; Hyland 2002a on directives; Hyland 2002b on authorial identity; Hyland 2001b on self-mention; Hyland 2007 on exemplifying and reformulating; Swales et al. 1998 on imperatives).

One of the few studies that do consider disciplinary differences in core grammatical features is Biber (2006: Chapter 4). For example, Biber (2006: 65) found that engineering and natural science disciplines used passives more frequently than other disciplines, while the education and humanities disciplines employed past tense verbs much more frequently than other disciplines, especially engineering (Biber 2006: 61). In addition, Biber (2006: 53–55) found quite a few differences in the use of nouns and verbs in different semantic categories, with mental nouns more prevalent in business and humanities, abstract/process nouns more common in business and engineering, and concrete (but not animate) nouns frequent in engineering. For verbs, he found that natural science used occurrence verbs more frequently than other disciplines, while education relied on communication, mental, and activity verbs (Biber 2006: 60–61). However, the disciplinary comparisons made in Biber (2006) are based on academic textbooks (and classroom teaching) in these disciplines, rather than on research articles. As the situational characteristics of textbooks and research articles differ in important ways, we can expect that the language use of these two academic registers varies as well. Little knowledge currently exists regarding variation in the general grammatical characteristics of research articles across disciplines. In this study, I address this gap by examining a wide range of grammatical features that have been shown to vary across register in previous research. In addition to core grammatical categories, I also offer a brief analysis of typical lexical patterns within some of these grammatical categories.

5.3 Carrying out a lexical and grammatical survey

The analysis in this chapter explores the rates of occurrence for core grammatical categories (nouns, verbs, adjectives, and adverbs) across disciplines and journal registers. This investigation also involves an examination of the semantic

groupings of nouns and verbs, pronoun usage, and aspects of the verb phrase such as tense, aspect, and voice. Table 5.1 lists the features considered in this analysis. The words included in the semantic categories for nouns come from Biber (2006: Appendix A), and lists of all words included in the present study are summarized in Appendix C. The semantic sets of words were compiled by examining the highly frequent nouns in large corpora (the Longman Corpus of Spoken and Written English for verbs, and the T2K-SWAL corpus for nouns), and categorizing them into groups based on the most typical meanings for those words (see Biber 2006: 244–250).

Rates of occurrence were calculated for each of these linguistic features using a specialized computer program called 'Tagcount'. Developed by Biber, Biber's tagcount program relies on the grammatical tags produced by the Biber tagger in combination with lexical information to produce normed frequency counts per text for over 120 grammatical and lexical features, including grammatical classes of words, verb tense/aspect/voice, embedded clauses, stance devices, etc. Biber (2006: Appendices) lists many of the features identified in this tagcount program.

Table 5.1. Grammatical categories included in the lexical and grammatical survey with examples

A. Nouns	
1. all nouns	*book, child, gravel, fish, idea, position*
2. semantic sets of nouns	
– cognition nouns	*ability, decision, concept, idea, knowledge, reason*
– group nouns	*church, committee, government, institute, university*
– animate nouns	*applicant, child, immigrant, patient, owner, president*
– technical nouns	*atom, compound, equation, message, particle, sample*
– other abstract nouns	*advantage, background, culture, equity, quality, setting*
– place nouns	*bench, country, habitat, office, region, store, territory*
– process nouns	*achievement, comparison, formation, process, result*
– quantity nouns	*amount, century, frequency, percentage, volt, weight*
– concrete nouns	*acid, brain, computer, glacier, magnet, radio, statue*
B. Pronouns	
3. pronouns	
– 1st person	*I, we*
– 2nd person	*you*
– 3rd person	*he, she, they*
– 'it'	*it*
– demonstratives	*this, these, that, those*
– nominal	*somebody, anyone*

(*Continued*)

Table 5.1. (Continued)

C. Verbs & Verb Phrases	
1. uninflected present tense, imperative, and third person	*confirm, confirms, expect, expects, focus, focuses, look, looks*
2. past tense	*claimed, concluded, demonstrated, found, reported*
3. perfect aspect	*have argued, have discussed, has shown, had used*
4. pres. progressive aspect	*is becoming, is causing, are seeking, are studying*
5. passive voice verbs	
• agentless	*is attributed to X, have been considered, were examined*
• by-phrase	*are accompanied by, were confirmed by, were provided by*
6. semantic sets of verbs	
• activity verbs	*bring, combine, encounter, obtain, produce, repeat, take*
• communication verbs	*acknowledge, claim, discuss, explain, question, specify*
• existence verbs†	*appear, define, illustrate, indicate, reflect, tend*
• mental verbs	*assess, confirm, discover, find, identify, observe, predict, think*
D. Other Classes	
1. Adjectives	*better, central, different, evident, important, unable*
2. Adverbs	*already, clearly, effectively, often, only, partly, widely*

†Existence verbs "report a state that exists between entities" (Biber et al. 1999: 364).

In order to facilitate a qualitative analysis of the semantic categories of nouns and verbs, a second program was developed in Perl to produce counts for each of the words included in each semantic category per register, as well as to add an additional tag to the annotation to enable concordance searching for all words in a particular semantic category. The purpose of this step in the analysis was not to consider the impact of any particular word (as the sub-corpora are too small to say much about individual lexical items with great reliability), but rather to facilitate a comparison of the uses of these semantic classes of words across registers and disciplines. Thus, the features investigated in this general grammatical and lexical description are identified based on either automatically assigned grammatical tags, or on lexical information.

5.4 Distribution of core grammatical features

As mentioned above, previous research has documented the relative distributions of the four major content word classes for the overall register of academic writing without attending to variation within academic writing. Figure 5.1 shows that the

same established patterns hold across research articles in different disciplines. That is, nouns are by far the most frequently used part of speech in all disciplines and registers, occurring more than twice as frequently as any other content word class. Adjectives are used more frequently than adverbs in all academic journal registers and disciplines, usually occurring even more frequently than verbs (represented here by uninflected verb forms and 3rd person singular).

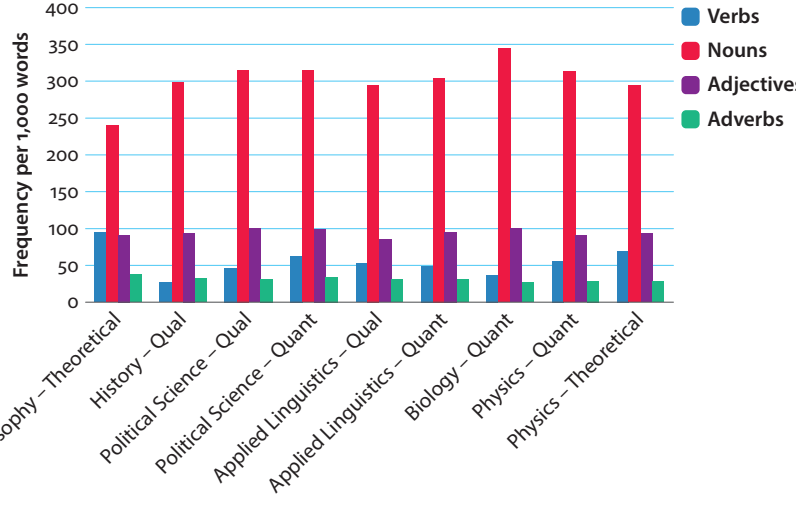


Figure 5.1. Distributions of verbs, nouns, adjectives and adverbs across register

Despite these overall similarities, several differences across disciplines are also shown in Table 5.1. For example, nouns are most frequent in biology and least frequent in philosophy, ranging in frequency from 240 to 344 times per 1,000 words. Philosophy also differs from other disciplines in its reliance on verbs. That is, while all other registers and disciplines use adjectives more frequently than verbs, philosophy uses adjectives and verbs with the same frequency. In fact, this illustrates the variation that occurs in the use of verbs across registers and disciplines, with verbs being more frequent in philosophy (occurring 95 times per 1,000 words) than in all the other disciplines (in which verbs range in frequency from 27 to 69 times per 1,000 words).

Text excerpt 5.1, which comes from a quantitative research article in biology, illustrates the high reliance on nouns and adjectives in academic journal writing, as well as the relatively low reliance on verbs. This single sentence of 35 words contains 14 nouns (**bolded**) acting as either head nouns or nouns as nominal

premodifiers, 5 adjectives (*italicized*), only two main verbs (underlined), and one non-finite post-nominal clause (double-underlined).

5.1 In particular, while the **species** *sorting* **hypothesis** predicts strong *environmental* **influences**, the *neutral* **theory**, the **mass effect**, and the **patch dynamics frameworks** all predict *differing* **degrees** of *spatial* **structure** resulting from **dispersal** and **competition limitations**. [BIO-QT]

Text excerpt 5.2, on the other hand, illustrates the noteworthy difference that occurs primarily in theoretical philosophy articles: a high reliance on nouns and adjectives, but also a higher use of verbs than other disciplines and journal registers.

5.2 Here I assume, rather than argue, that this **approach** to *moral* **patienthood** and *moral* **considerability** is *correct*. The **non-identity argument** poses a **problem** for *person-affecting* **ethics** and thus requires an **answer** based on **effects** on **individuals**. [PHIL-TH]

Excerpt 5.2, although similar in length to the biology excerpt (36 and 35 words respectively), contains 10 nouns, 5 adjectives, four main verbs and one non-finite post-nominal clause. Thus, the excerpt exemplifies the trends shown in Figure 5.1, with theoretical philosophy relying on nouns to a lesser extent than other disciplines, and exhibiting a more frequent use of verbs, while still maintaining the primarily nominal structure of academic writing.

As expected, however, the variation shown in Figure 5.1 when comparing across disciplines and journal registers is not as great in magnitude as the differences seen in Biber (2006), where a variety of spoken and written registers were compared. However, as the remainder of this section will show, interesting patterns of variation do emerge if we consider the semantic classes of words used. In Sections 5.4.1 and 5.4.2, I look at such semantic classes for two content classes: nouns and verbs.

5.4.1 Nouns

It would seem at first glance from the previous section that research articles in these 6 disciplines do not vary to a great extent in terms of their use of nouns. All disciplines and registers use nouns to a greater extent than any of the other three content classes of words, with philosophy showing a slightly lower reliance on nouns than the other disciplines/registers. However, if we consider the *types* of nouns used in these different registers, a good deal of variation does emerge. Here, I consider semantic sets of nouns, based on the groupings of nouns in Biber (2006; see Appendix C of the book for a listing of these nouns and Section 5.3 for

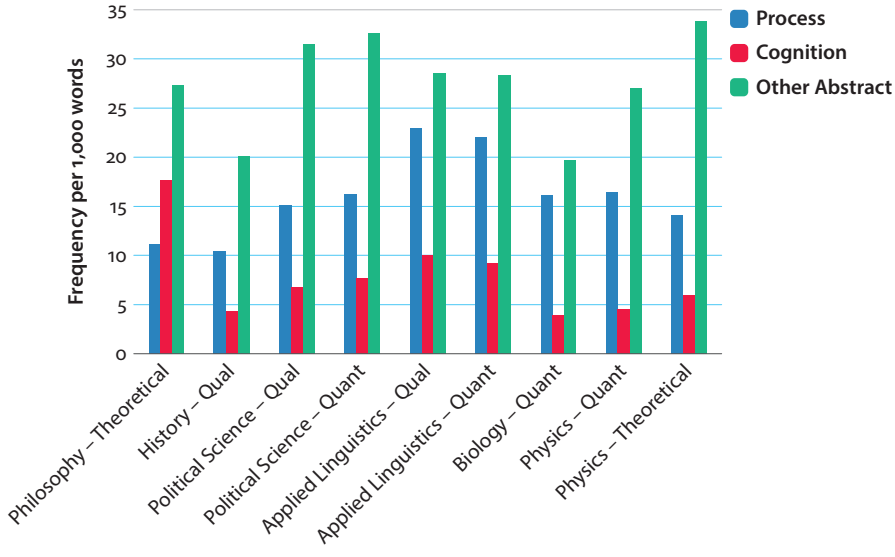


Figure 5.2. Distribution of nouns across registers: Process, cognition and other abstract nouns

a description of how these groupings were created). For ease of presentation, I will discuss the semantic categories of nouns in groups. Figure 5.2 shows the distributions of three types of abstract nouns: process (e.g., *action, application, argument, management, transfer*), cognition (e.g., *ability, belief, hypothesis, idea*), and other abstract nouns (e.g., *alternative, cause, choice, criteria, potential, relationship*).

In the quantitative registers (applied linguistics, biology, physics, and political science), many of these process nouns have to do with the nature of quantitative research, representing qualities and processes inherent in research which has stricter controls on what is being compared, either through experimental research methods or through an attempt to quantitatively compare like with like. As a result, process nouns like *control, test, experiment, result, effect, treatment, question, comparison,* and *procedure* are common in the quantitative registers regardless of discipline (process nouns **bolded**):

5.3 This **experiment** shows the benefits of helping students become competent cross-cultural communicators. [AL-QT]

5.4 Obviously, the total photon energy that is available in our **experiment** is far from enough for the excitation to the 1 u bound state of H [PHYS-QT]

5.5 We have extended the role of a standard **control** variable, marital status, to study the role and **interaction** of polygamy and religion [POLISCI-QT]

5.6 This type of indirect **effect** is often ignored in many fragmentation **studies**. [BIO-QT]

5.7 As noted previously, **treatment** took place over three class sessions [AL-QT]

5.8 Finally, the following **question** has to be addressed: How do the electron and photon irradiation used in the measurements influence the **results?** [PHYS-QT]

Other nouns reflect typical topical content of the disciplines. For example, process nouns dealing with historical and political events and institutions are used in history and political science (e.g., *war, revolution, administration, death, trade, education, marriage, generation, education, discrimination*), while nouns representing processes common in language learning contexts are prevalent in both applied linguistics registers (e.g., *teaching, activity, practice, talk, performance, training, formation, assignment, answer, progress, production, attempt, transfer, strategy*).

Cognition nouns, on the other hand, are highest in theoretical philosophy articles, followed by the social sciences (particularly applied linguistics). In philosophy, the many cognition nouns occur quite frequently, as the content of the discipline is largely concerned with exploring knowledge and reason – both mental processes:

5.9 However, the remaining question is what **reason** there is for thinking the theory true. [PHIL-TH]

5.10 Q-**memory** is a **concept** derived from that of ordinary **memory**. [PHIL-TH]

5.11 …the claim that **knowledge** is more valuable than accidentally true **belief** should serve as an adequacy condition on a **theory** of **knowledge** [PHIL-TH]

Cognition nouns are also relatively common in applied linguistics. Again, this prevalence seems to be related to the nature of what is studied in applied linguistics: nouns related to language ability and mental processes are common, such as *knowledge, experience, assessment, attention, ability, evaluation, examination,* and *understanding.*

5.12 The **ability** for Eduardo to explain the mSSR activity to a newcomer to the class also shows the concomitant developing interactional competence. [AL-QL]

5.13 …requests for confirmation about language form or language choice focus learners' **attention** on a specific language item. [AL-QT]

5.14 These findings provide support for theoretical accounts that associate **knowledge** of these structures with **knowledge** of real words and for instruction oriented toward the development of vocabulary **knowledge**. [AL-QT]

In contrast, cognition nouns are relatively rare in history and the hard sciences of biology and physics. Whereas history is less concerned with mental processes and states of individuals or groups, and more concerned with events and institutional

happenings, the hard sciences have a focus on non-human physical entities that are rarely discussed in terms of mental processes and abilities. Thus, the two disciplines whose focus is on human subjects use cognition nouns to a greater extent than other disciplines.

As shown in Figure 5.2, other abstract nouns are the most commonly-occurring semantic class of nouns in all disciplines and registers. Table 5.2 lists the most commonly-occurring other abstract nouns (occurring more than 50 times per 100,000 words) in each discipline and register. The most frequent nouns in this semantic set appear to primarily relate to the topic or subject matter of the disciplines, with less variation within discipline than seen for other semantic sets of nouns. However, some overlap does exist for disciplines concerned with similar concepts, such as political science and history, which both deal with political concepts:

5.15 The party's **policy** was further tilted again the nationalization of agricultural land [HIST-QL]

5.16 And the differential treatments in criminal **law** that once had favored Roman citizens now were based on the distinction between … [HIST-QL]

5.17 …the USA had secured a government partner willing to implement US neoliberal economic **policy** in El Salvador. [POLI-SCI-QL]

5.18 For **state** and local offices, amateur campaigns are even more common [POLI-SCI-QT]

Figure 5.3 displays the frequencies of concrete, animate, technical, quantity, group, and place nouns across the sub-corpora. Concrete nouns are most frequent in the hard sciences (particularly biology). In biology, physical nouns referring to plants, animals, and some metals are particularly common: *food, soil, muscle, water, body, tissue, chain, plant, copper, seed, heart, tree, solution, leaf, fish, metal, acid, arm, leg, eye, flower, drug,* etc. In contrast, the most common concrete nouns in the two physics registers involve more basic, elemental objects: *crystal, solution, water, mixture, metal,* etc.

Animate nouns are most common in applied linguistics, where most animate nouns refer to various participants in language research: *learner, student, teacher, writer, researcher, participant, reader, speaker, people, person, child, undergraduate, female, child, audience, adult, author, professor,* etc. The remaining humanities and social sciences registers also use animate nouns, but these nouns commonly refer to more general roles than the nouns in applied linguistics. For example, in history and to a lesser extent in political science, nouns such as *people, king, man, family, father, woman, slave, wife, executive, president, secretary, member, son, citizen, author, minister, mother, police, historian,* etc. account for most animate nouns.

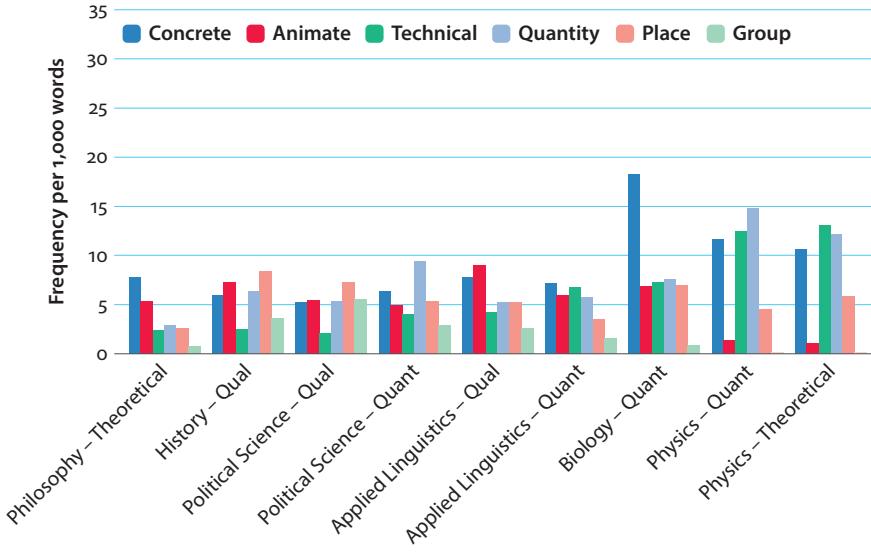


Figure 5.3. Distribution of nouns across registers: Concrete, animate, technical, quantity, place and group nouns

Table 5.2. Most frequent (> 50 times per 100,000 words) 'other abstract' nouns by discipline and register

Philosophy (Theo)	History (Qual)	Poli-Sci (Qual)	Poli-Sci (Quant)	Applied Ling (Qual)	Applied Ling (Quant)	Biology (Quant)	Physics (Quant)	Physics (Theo)
account	end	impact	action	content	content	climate	charge	action
action	law	issue	benefit	context	context	diversity	density	background
content	life	language	capital	culture	information	level	level	charge
context	policy	law	choice	grammar	input	model	model	density
identity	power	level	identity	identity	job	structure	order	direction
law	state	policy	impact	information	language	system	phase	factor
nature	way	power	information	language	level	type	shape	level
principle		role	interest	level	output	value	signal	matrix
respect		security	issue	role	type	variation	source	model
right		state	level	science	way		state	order
sort		support	life	way			structure	phase
state		system	model				system	potential
subject		truth	policy				type	power
truth		way	race				value	state
value			relationship				velocity	structure
way			role					system
			state					value
			support					velocity
			system					

Technical nouns are most common in both physics registers, followed by biology and quantitative applied linguistics. Interestingly, the noun *data* is highly frequent in all disciplines and registers with the exception of history and philosophy, occurring between 30 and 255 times per 100,000 words. In fact,

the noun *data* occurs over 100 times per 100,000 words in quantitative biology, quantitative physics, quantitative political science, and quantitative and qualitative applied linguistics. The noun *data* also illustrates the general trend that for the two disciplines represented by both qualitative and quantitative research, technical nouns are more frequent in the quantitative registers than in the qualitative registers.

In physics, technical nouns are reflective of the content of the discipline: *sample, electron, cloud, wave, ray, nucleus, atom, angle, equation, oxygen, particle, proton, ion, component, light, nuclei, cell,* and *molecule.* While a fair number of technical nouns overlap between physics and biology, the technical nouns in quantitative applied linguistics refer to aspects of language: *word, sentence, list, paragraph, internet, statement.* Reflective of the much more frequent use of technical nouns in physics, there are also many more nouns used quite frequently in physics than in other disciplines.

Like technical nouns, quantity nouns are much more frequent in physics, and to a lesser extent, in biology. Place nouns, on the other hand, are less frequent overall, and are primarily used in history, political science, and biology. In history and both political science registers, place nouns are focused on physical areas defined by political or formal organizational means: *city, land, place, office, country, court, region, organization, area, property, building, county*. In contrast, biology uses a smaller range of place nouns that are more narrowly focused on physical locations defined by natural characteristics of the places: *coast, forest, habitat, field, river, farm, valley, pool, stream.* Group nouns are relatively rare in the corpus. Like nouns, individual lexical verbs can also be grouped according to the most prevalent semantic meaning carried by the verbs. In the next section, I describe the use of four semantic categories of verbs.

5.4.2 Verbs

Figure 5.4 displays the rates of occurrence for verbs in four of the semantic categories from Biber (2006): activity (e.g., *add, bring, divide, produce*), communication (e.g., *address, argue, insist, suggest*), mental (e.g., *assume, believe, determine, interpret*), and existence verbs (e.g., *appear, include, remain, reveal*). Activity verbs are relatively frequent in all disciplines and registers, but are the most frequent in both applied linguistics registers, followed by both physics registers.

Many of the activity verbs that occur most frequently fall into two major uses. The first major use involves verbs (e.g., *use, make, receive, produce, give, obtain*) that convey aspects of methodology, to describe sequences of events, procedures, and methods for conducting the study:

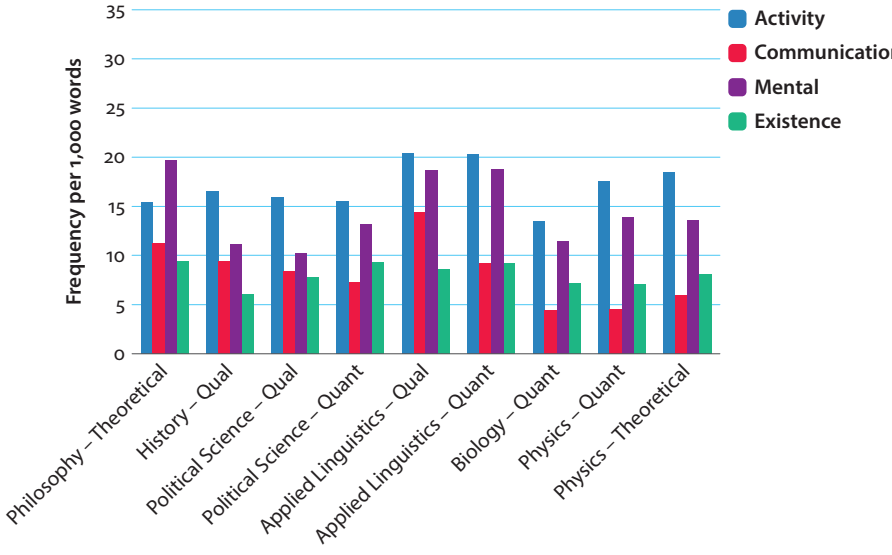


Figure 5.4. Distribution of verbs across registers: Activity, communication, mental and existence verbs

5.19 To measure an incumbent's ties to lobbyists, we **use** the sum of campaign money raised by each member from the D.C. metro area in the 2004 election cycle [POLISCI-QT]

5.20 The teacher **provided** the learners with the URL for a university study skills site and guided them in reading the advice. [AL-QL]

5.21 We can **obtain** these terms by noting that A is defined only up to a gauge transformation. [PHYS-TH]

5.22 One seed from separate treatments was **put** into each pot with a forceps. [BIO-QT]

5.23 Where variances existed, adjustments were **made** after discussion and mutual agreement. [AL-QT]

5.24 Participants were not **made** aware that their knowledge of formulaic sequences played any role in the experiment. [AL-QT]

The second major use of activity verbs (involving verbs such as *show, provide,* and *make*) allows the writer to describe findings, data, or concepts, often linking them to interpretations and claims:

5.25 Figure 7 **shows** the temperature dependences of the squared order parameter of both KDP and DKDP systems along the ferroelectric c axis at atmospheric pressure. [PHYS-TH]

5.26 Indeed, one of our extended examples of the coordinating effects of legal sanctions **makes** clear the paramount importance of these effects in achieving political stability in the midst of (constitutional) regime change. [PHIL-TH]

5.27 Separations **took** place on a 250-mm by 4.6-mm (inner diameter) Supelcogel H column, preceded by a 50-mm by 4.6-mm (inner diameter) Supelguard CH (Supelco) precolumn … [BIO-QT]

5.28 Third, NNSs are inclined to **produce** the politeness marker please frequently and **produce** fewer downtoners and subjective opinions than do NSs of English. [AL-QT]

Communication verbs, on the other hand, are almost twice as common in qualitative applied linguistics (followed by theoretical philosophy) than in the remaining disciplines. Communication verbs are generally more frequent in the humanities and social science disciplines and less frequent in the harder sciences (biology and physics). The most frequently used communication verbs (e.g., *claim, suggest, describe, argue, say, explain, discuss*) are often used to convey information and claims set forth by other researchers or studies (excerpts 5.29–5.31), or to put forward the writers' own claims (excerpts 5.32–5.33).

5.29 Ortiz (2005) **claimed** that if test-takers include those who have not been raised in the culture from which the test samples, which is the situation of the OSSLT for L students, then test validity needs to be discussed with caution. [AL-QT]

5.30 No surviving record **suggests** that this had been done in any previous case. [HIST-QL]

5.31 Hume **explains** the development of the sense of justice and injustice in people via sympathy, which is a psychological mechanism of the human mind. [PHIL-TH]

5.32 We **discuss** these patterns in turn below. [AL-QL]

5.33 In this section, we **argue** that our two-pronged account is superior to the consensus view in addressing questions about armed intervention in the context of such struggles. [PHIL-TH]

Particularly in applied linguistics research articles, communication verbs are also used with the participants of the research as agents of the verbs:

5.34 Other students in Group 2 openly **said**, for example "Je veux parler anglais" (I want to speak English). [AL-QL]

5.35 Veronica, for example, **explained** that she would not **describe** herself as an English Language Learner because of the novice language level she associated with it. [AL-QL]

Like cognition nouns, mental verbs are most frequent in theoretical philosophy, followed by qualitative and quantitative applied linguistics. While the most frequent mental verbs are common across disciplines and registers (e.g., *believe, see, think, need,* and *know*, excerpts 5.36–5.37), philosophy uses a much wider range of mental verbs with greater frequency than the other disciplines. The mental verbs common across disciplines and registers are generally used to convey the writers' thoughts. In philosophy, however, a much wider range of verbs are used for this purpose, and these verbs exhibit more nuanced and personal meanings (excerpts 5.38–5.40): *imagine, conclude, assume, determine, satisfy, understand, feel, justify,* and *expect*.

5.36 We **know** that for ELLs, differentiated instruction, adequate scaffolding, and Teachers skilled enough to work with autonomy are important success factors [AL-QT]

5.37 Further, we **see** evidence that as individuals become more informed they appear to lose reliance upon that social group identity in informing their attitudes and making choices. [POLISCI-QT]

5.38 These three conditions are not unlike the ones that we **expect** a good jury in the courtroom to display. [PHIL-TH]

5.39 Rather we **judge** according to the sentiments that we would have were we to **satisfy** all of the conditions C, C and C [PHIL-TH]

5.40 From this I can **conclude** that the apparent memories that I have are not ordinary memories of my own experiences. [PHIL-TH]

Looking at the general frequencies of the semantic categories of nouns and verbs has provided a broad overview of the types of nouns and verbs that are used across disciplines, and has identified ways in which patterns of noun and verb use vary according to several parameters. For example, this analysis has revealed some disciplinary register differences that seem to correspond to discipline-specific content, while other differences parallel differences across the nature of disciplines that is, across 'hard' and 'soft' disciplines. In fact, this analysis has also shown similarities across social science and hard sciences based on a shared, quantitative methodology.

These analyses have provided the overall distributions of these semantic categories of nouns and verbs, as well as frequency information for specific verbs and nouns used within these categories. Like most linguistic features, it is not the case that absolute differences in the use of nouns and verbs in various semantic groupings exist between disciplines and registers. However, we can see that disciplines and registers rely on the different words to differing extents. An additional way that we can more objectively identify meaningful differences is to

consider the use of each individual semantic category in one discipline/register combination in relation to the overall amount of variation exhibited for the use of these words across these disciplines and registers. In computing a mean frequency of use for each semantic category for each discipline/register combination, we can also identify a standard deviation to describe the degree of dispersion across all sub-corpora.

Thus, Table 5.3 summarizes the degree to which disciplines and registers use the semantic categories of nouns and verbs relative to the overall mean use of that category across the corpus. In this analysis, a frequency of use (for a register/discipline) that fell outside of one standard deviation of the mean was considered a distinctive use of that semantic set. In Table 5.3, each + symbol indicates that the frequency of use of a particular set of nouns or verbs was more than one standard deviation higher in that particular discipline and register than the mean use across all registers and disciplines. Likewise, each – symbol indicates a frequency of use that was lower than the mean by a full standard deviation. In a few instances, ++ is used to indicate that a discipline and register combination exhibited a heavier reliance on a particular semantic set, reflected by the mean frequency of use falling more than two standard deviation from the overall mean frequency of use across the corpus. The absence of a symbol indicates that the mean use of that semantics set fell within one standard deviation of the overall mean.

Table 5.3 confirms the distinctive reliance on cognition nouns and mental and existence verbs in philosophy. Likewise, qualitative applied linguistics shows a reliance on process and animate nouns when compared to all other disciplines and registers, as well as a reliance on activity and mental verbs (a trend which it shares with quantitative linguistics), as well as a high reliance on communication verbs. In contrast, both quantitative and theoretical physics are characterized by their greater reliance on concrete, technical, and quantity nouns and relative lack of use of animate nouns. It is interesting to note that biology and physics, the two disciplines which have so far relied most heavily on nouns are also distinguished from other disciplines by the types of nouns that they use.

The three passages below, selected from theoretical philosophy (5.41), qualitative applied linguistics (5.42), and quantitative physics (5.43) illustrate these trends. In each passage, nouns (**bolded**) and verbs (underlined) which fall into the semantic categories identified as distinctive for each discipline are indicated:

5.41 *Theoretical Philosophy (Collins 2008):*
The **belief** that in this hypothetical case Joe is not Moe is not the negation of the **belief** about Joe and the actual identical twin Moe, the **belief** that Joe is Moe. The hypothetical **thought** experiment 'Moe' is no real or even possible person; it is a mere fiction. Now the mere **fact** that someone systematically confuses one thing for another does not entail that she believes they are

identical. A prospector may never be able to tell the difference between gold and fool's gold (iron pyrites), but she does not believe that they are one and the same. For such a prospector, gold and fool's gold are still discernible in some ways, if not perceptually discernible.

Table 5.3. Summary: Semantic categories across discipline and register based on standard deviations

	Philosophy (Theo)	History (Qual)	Poli-Sci (Qual)	Poli-Sci (Quant)	Applied Ling (Qual)	Applied Ling (Quant)	Biology (Quant)	Physics (Quant)	Physics (Theo)
Nouns									
process	-	-			+	+			
cognition	++								
other abstract		-		+			-		+
concrete							++	+	+
animate					+			-	-
technical								+	+
quantity	-							+	+
place	-	+	+			-			
group			+						-
Verbs									
activity					+	+	-		
communication					+		-	-	
mental	+		-		+	+			
existence	+	-							

Notes:
++ two standard deviations above the mean
+ one standard deviation above the mean
- one standard deviation below the mean

5.42 *Qualitative Applied Linguistics (Frazier 2007):*
A large number of **studies** investigate the **talk** of **students** in writing classrooms; most of these treat the **act** of writing as social in nature. Theories of social **actions** and learning/socialization such as Lev Vygotsky's (1978) are useful in helping **teachers** create practical situations in which writing **students** can learn in social situations. To understand how this learning happens, it is crucial to identify the interactional details of group **work** discourse. Most of the sources that investigate writing **students' talk** (some of which are covered below), however, tend to focus

on purely theoretical concepts, the social power structures inherent in tutor/**peer** relationships, or a priori analyst-imposed categories of group **work talk**. What is missing, and what this paper addresses, therefore, is a close accounting of the structures of **talk** and embodied **action** that occur during group **work** interaction and how group **work participants** themselves orient to their group mates.

5.43 *Quantitative Physics (Ganguly et al. 2007):*
Excited states of 112Sn were populated in the 100Mo(20Ne, an) reaction at a beam **energy** of 136 MeV at the Variable **Energy** Cyclotron Centre, Kolkata. The **target** consisted of isotopically enriched (99.54%) 100Mo, 4.7 mg/cm thick, evaporated on an **aluminium** backing…The raw **data** were sorted into different 4096×4096 matrices after gain matching of all the detectors to a dispersion of 1.0 keV per channel.

The analyses in this section have shown that the text categories in the corpus exhibit preferences for the types of meanings expressed by nouns and verbs, and many of these preferences appear to follow along disciplinary lines. In addition, the excerpts above illustrate the co-occurrence patterns of specific semantic groupings of nouns and verbs that were summarized in Table 5.3. In the philosophy excerpt (5.41), the noun *belief* and the corresponding verbal form *believe* illustrate this disciplines reliance on cognition nouns and mental verbs. Likewise, the applied linguistics excerpt (5.42) demonstrates the use of animate nouns alongside activity, communication, and mental verbs.

5.4.3 The verb phrase: Passive voice

The passive voice has received a good deal of attention in studies of academic writing. Often framed in terms of its role in information flow and the marking of stance, passives have been primarily described in science writing. However, it turns out that in addition to quantitative differences in the use of passives across disciplines and registers, there are noteworthy qualitative differences in the ways that passives are used. Figure 5.5 displays the frequency of use for agentless passives (where the agent of the verb is completely omitted) and by-passives (where the agent is specified in a prepositional phrase beginning with *by* following the passive verb). Agentless passives are much more common than by-passives in all disciplines and registers. Both types of passives occur with increasing frequency as we move from the humanities towards the hard sciences of biology and physics.

In biology and physics, agentless passives are often used with both activity (excerpts 5.44–5.45) and mental (excerpts 5.46–5.47) verbs, and the implied agents of the verbs are often the researchers themselves. This pattern of passive use is in line with previously proposed reasons for passive use, in that the agent of the verb is assumed to be the researcher, and therefore the agent can be safely omitted

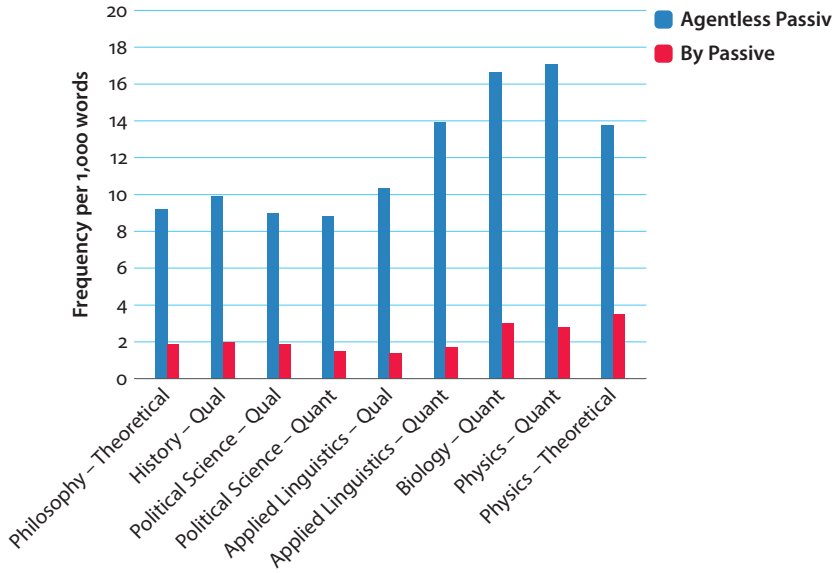


Figure 5.5. Distribution of passive voice verbs across register

to either de-emphasize the role of the agent in the research process or to leave out implied information.

5.44 The deduced B(E) rates **are used** to infer the possible presence of octupole correlations in the 231Ac nucleus. [PHYS-QT]

implied agent: *We use deduced B(E) rates to infer…*

5.45 Results **are expressed** as means (SD) of the values, except in Figs. 5 and 6 where, for clarity, data **are shown** as means SE [BIO-QT]

implied agent: *We express results…, we show data as…*

5.46 The probabilities **are calculated** directly, including the one corresponding to the ionization channel. [PHYS-TH]

implied agent: *We calculate the probabilities directly*

5.47 A clear seasonality **was observed** in the occurrence of some parasite species such as G. proterorhini, H. aduncum, Spiroxys sp., N. rutili, and E. sieboldi, which appeared only in some seasons. [BIO-QT]

implied agent: *We observed a clear seasonality*

In contrast to agentless passives, by-passives in biology and physics (while also occurring with many activity and mental verbs) are frequently used to establish abstract concepts or processes as the agents of the verbs. These agents are often expressed through nominalized processes or non-finite verb phrases.

5.48 This reduction **is caused** by a reduction in velocity correlation time. [PHYS-TH]

5.49 The CSA (in mm) **was determined** by dividing the triceps suræ muscle mass (in mg), by the product of optimal muscle length (in mm) and d, the density of mammalian skeletal muscle (d 1.06 mg mm 3). [BIO-QT]

5.50 Constant A **is obtained** by normalizing the T peak intensity for zero bias condition and it was kept constant throughout the fitting. [PHYS-QT]

Quantitative (and to a lesser extent qualitative) applied linguistics research articles utilize passive voice verbs frequently as well. Like the hard sciences, the researcher can often be assumed to be the agent of the passive verb, as methods and procedures are described:

5.51 Multiple sources of data collection **were used** to investigate this claim [AL-QL]

5.52 Third, data **were analyzed** only for school districts that participated in the voluntary supplementary data collection. [AL-QT]

By-passives are also used in applied linguistics to position processes as agents (excerpt 5.53). However, in contrast to the hard sciences, by-passives are also used to bring human agents into the discourse (excerpts 5.54–55). These human agents are typically third parties (i.e., not the researcher):

5.53 Interrater reliability **was estimated** by examining the correlation coefficients for the raters' scores. [AL-QT]

5.54 Peer group interaction **has also been studied** by L writing researchers. [AL-QL]

5.55 Minfang worked hard to **be accepted** by the learner community at Nanda. [AL-QL]

In history, agentless passives exhibit different patterns than the patterns for long passives discussed to this point. Here, agentless passives often have implied human agents, but that implied agent is not the researcher. That is, agentless passives are not used to describe actions or processes that were carried out by the researcher to conduct the study, but rather by an unnamed historical entity:

5.56 Certain rights **were proscribed** to their indigenous inhabitants [HIST-QL]

5.57 But this does not mean that all barbarians automatically **were considered** citizens. [HIST-QL]

5.58 The coffins **were carried** into the church and placed in the centre of the nave, where they received an absolution from archbishop Toccabelli [HIST-QL]

Qualitative and quantitative political science also exhibit this use of agentless passives, where the focus is on describing events and occurrences (excerpts 5.59–5.60):

5.59 The mayor **was accused** of adding political allies to the city payroll while the city was constantly financially strained. [POLISCI-QL]

5.60 A document summarising policy commitments, the All-Wales Accord, **was produced**. [POLISCI-QT]

However, political science research articles also contain instances of the researcher as the implied agent as the study procedures are described (excerpts 5.61–5.62):

5.61 The regional list vote **was used** because that is the basis of the proportionality calculation. [POLISCI-QT]
implied agent: *We used the regional list vote*

5.62 The weight for our study **was created** within each of the experimental cells [POLISCI-QT]
implied agent: *We* created the weight for our study

Thus, the use of the passive voice varies across discipline and register in terms of (a) the frequency with which the short versus the long passive is used (with agentless passives occurring much more frequently than passives with by-phrases), (b) the overall frequency of use (with both forms of the passive being most frequent in the natural sciences and quantitative applied linguistics), and (c) the various functions the passives is used for. It is interesting to note that, in addition to illustrating these trends, the text excerpts in this section have also exhibited variability in the use of tense marking within these passive verb phrases. In the next section, I examine tense and aspect marking in all verb phrases across the disciplines and registers.

5.4.4 The verb phrase: Tense and aspect

Figure 5.6 displays the distribution of tense and aspect marking across registers and disciplines, showing that in most cases, present tense is used to a greater extent than past tense. This trend is particularly pronounced in philosophy, where present tense verbs are about 9 times as frequent as past tense verbs. In the following excerpt from a theoretical philosophy article, the present tense verbs are **bolded**, while past tense verbs are *italicized*. In this excerpt, the present tense is used to describe the present state or nature of a philosophical construct ('desire').

5.63 *Theoretical Philosophy (Hawkins 2008):*
My thesis **concerns** the narrower pre-philosophical sense of 'desire'. But what exactly **is** this sense? A distinction originally *introduced* by Thomas Nagel may help here. Nagel famously *divided* the broad category of desires into 'motivated desires' and 'unmotivated desires'. The essence of the distinction **has** to do with reasons. Motivated desires **are** states or dispositions that we

have because we **recognize** reasons for having them. Unmotivated desires, by contrast, **are** states or dispositions which **lack** this basis in reasons. They **are** states with which we simply **find** ourselves. Desires, in my preferred sense, **fall** in this second category. The concept of an unmotivated state or disposition **seems** essential to the ordinary pre-philosophical notion of desire. Not only **is** desire not something we **reason** our way to, but once desire **exists** it **is** generally not sensitive to reasons in the way other attitudes **are**.

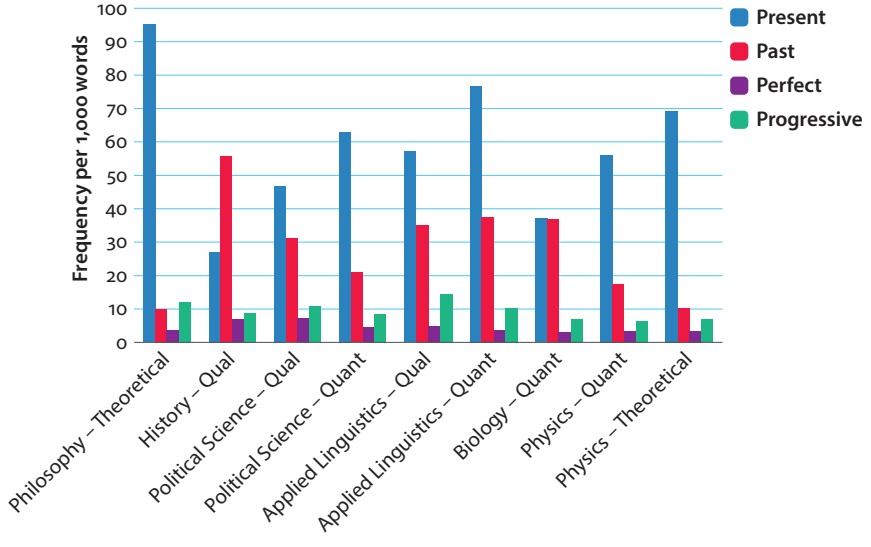


Figure 5.6. Distribution of tense and aspect across register

The two physics registers, particularly theoretical physics, also show a much higher use of present tense verbs. In theoretical articles, present tense is used to establish relationships (often with *be* as a main verb) and describe the effects of a phenomenon:

5.64 *Theoretical Physics (Allen 2008):*
This **is** the discrete version of ⟨formula⟩. If A **is** a function of t, gauge invariance again **holds** but with the scalar potential included. Equation 25 **is** the central result of the present note. This effective Hamiltonian **is** not manifestly Hermitian but it still **conserves** probability and **preserves** the Pauli principle since a straightforward calculation using Eq. 25 in Eq. 14 **gives** ⟨formula⟩ where n **labels** a time-dependent one-electron state. This result also **follows** from the original Schrodinger equation Eq. 1 and the expansion of Eq. 13 since ⟨formula⟩ but it **is** reassuring that the approximation of Eq. 12 **preserves** orthonormality of the time-dependent states.

Although present tense verbs are most common in all disciplines and registers with the exception of history, the magnitude of the differences between the use of past and present tense is much greater in the two theoretical registers. All registers use the past tense to establish previous findings and claims in situating the research study within the canon of knowledge to some extent,[1] as in excerpt 5.65 (finite past tense verbs appear in **bold**):

5.65 *Quantitative Political Science (Wink & Bargen 2008):*
In a recent article, Bullock, Hoffman, and Gaddie (2005) **undertook** a study of Republican gains in U.S. House elections in the South, an area of research that perhaps has been overshadowed by the numerous works devoted to Republican presidential success in the region. The authors **attempted** to answer the questions of exactly when and how Republican U.S. House candidates in the South **were** able to win a majority of the votes of white southerners and maintain that new allegiance to the GOP. They **found** electoral support for Republican House candidates and declines in split-ticket voting by white southerners **began** in the mid to late 1980s, and this trend **continued** through the year 2000.

However, the empirical registers also use the past tense to report methodological steps and procedures in empirical research reports (excerpt 5.66) and to describe the results or outcomes of the research (excerpt 5.67) (finite past tense verbs appear in **bold**):

5.66 *Quantitative Biology (Stewart et al. 2007):*
The South Australian planning region **was divided** into 3119 planning units, each 5 × 5 km. Information on the amount of each biodiversity feature j, in each planning unit i, **formed** the data matrix A = {aij }. Biodiversity features **were identified** from six biophysical data layers that **provided** consistent quality and coverage across the planning region. These **were derived** from government agencies of South Australia and **included** biogeographic regions (mesoscale 1001000s of km); biounits (scale 10-100s of km); marine benthic habitats; coastal saltmarsh and mangrove habitats; species occurrence data (Australian sea lions, Neophoca cinerea; New Zealand fur seals, Arctocephalus forsteri); and bathymetry (depth classes). This **generated** a data matrix of 17,000 records of 102 biodiversity features distributed across 3119 planning units (Stewart et al. 2003). The number of biodiversity features contained within an individual planning unit **ranged** from 2 to 15.

---

1. In fact, this use of the past tense to situate a study in terms of previous research may partially account for the low use of past tense verbs in quantitative physics, which includes much less literature review than other empirical research reports.

5.67 *Qualitative Applied Linguistics (Cheng, Fox & Zheng 2007):*
When we **asked** detailed questions about how students **approached** each of the reading and writing tasks on the OSSLT, we **came** to understand better what **was** in the students' minds when they **tackled** each test task. The following student accounts indicate how the students **said** they **approached** reading on the test. On the whole, L students **seemed** to be more strategic in processing the reading tasks in comparison with their L counterparts, who, in turn, **were** more systematic.

Not surprisingly, history is the only discipline and register to use past tense verbs with a markedly higher frequency than present tense verbs, as the past tense is used to describe and analyze events and happenings (finite past tense verbs appear in **bold**):

5.68 *Qualitative History (Sanford 2006):*
In this they **were strenuously resisted** by Britain and America who **had** an overriding interest in maintaining the Soviet military effort. The fighting on the Eastern Front eventually **broke** the Nazi war machine, thus saving the lives of an enormous number of Western soldiers. After 1943 the Western Allies **sacrificed** not only the objective truth about Katyn but also their Polish wartime ally, although whether they **did** so consciously or otherwise is highly controversial. After Stalingrad in late 1942 Stalin not only **began** to impose his wishes regarding Poland's postwar frontiers but also **did** his utmost to destroy the London Poles, eventually replacing them entirely with an alternative communist leadership which **took** over and **transformed** Poland on his behalf at the end of the second world war.

As can be seen from Figure 5.6 and the text excerpts presented above, perfect and progressive aspect verbs are not highly frequent in any discipline or register.

5.4.5 Personal pronouns

Academic writing is generally characterized by an infrequent use of pronouns relative to the much more frequent use of personal pronouns in spoken language (Biber et al. 1999). However, corpus-based research has investigated personal pronouns (e.g., Harwood 2005a,b; Hyland 2001a; Kuo 1999; Martínez 2005) in published academic writing, and this research has made it clear that personal pronouns fulfill specific discourse functions in academic writing.

In fact, comparing the use of personal pronouns across disciplines and registers reveals that a great deal of variation exists with respect to the use of pronouns, both in terms of the relative frequency of use, and the particular pronouns that are used. Figure 5.7 shows the distribution of first, second, and third person pronouns. Overall, personal pronouns are most frequent in qualitative applied

linguistics and theoretical philosophy, followed by qualitative history. In general, personal pronouns are used to a lesser extent in hard sciences, particularly quantitative biology and physics, than in political science and applied linguistics.

First person pronouns are by far most frequent in theoretical philosophy, followed by theoretical physics and qualitative applied linguistics. In philosophy, the pronouns *I* and *we* are used equally. In particular, *I* is common as the subject of mental verbs (5.69–71) and communication verbs (5.72–74), referring to the writer him/herself:

5.69 As **I understand** it, desire only sometimes gives rise to action. [PHIL-TH]

5.70 **I can infer** that this is what the experience of seeing red is like [PHIL-TH]

5.71 **I conclude** that the most plausible view is that… [PHIL-TH]

5.72 **I have argued** that this strategy can be directed… [PHIL-TH]

5.73 **I will not mention** these requirements… [PHIL-TH]

5.74 What, **I might ask** myself, would be the point of keeping the promise? [PHIL-TH]

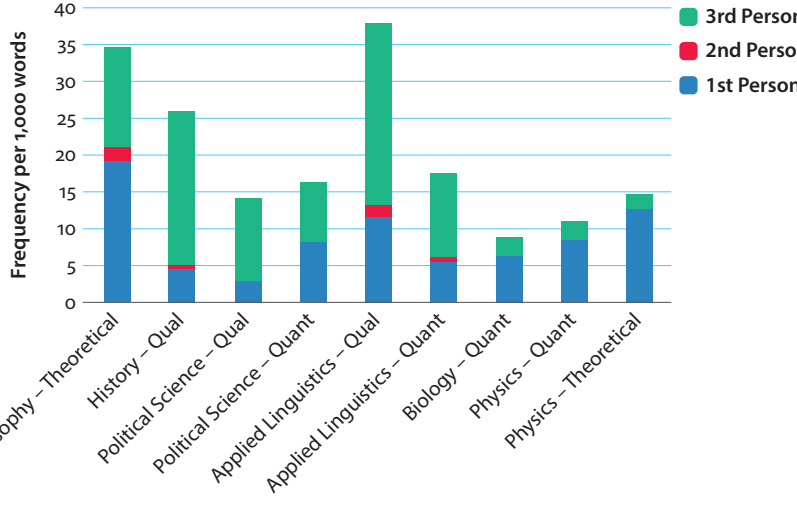


Figure 5.7. Distribution of personal pronouns across register

We, on the other hand, is just as common as *I* in theoretical philosophy, but often refers to the wider human populace rather than the author:

5.75 one quickly realizes that **we** are incredibly lucky to be as **we** are [PHIL-TH]

5.76 the circumstances in which **we** are required to choose between fairness and the all-things-considered well-being [PHIL-TH]

In fact, nearly all philosophy articles in the corpus are written by single authors. At times, *we* also seems to be used to encompass the writer and the reader, as a prompt to follow a line of logic:

5.77 if **we** regard any of the violations of a special relationship as morally wrong, **we** can conclude that TR is false [PHIL-TH]

It is this second use of *we* that is also prevalent in theoretical physics. Although most theoretical physics articles are also written by a single author, the singular first person pronoun *I* is rarely used. Rather, *we* is used to refer to both the writer and the reader, and often expresses actions or specifies steps in an analysis. The effect, then, is that the reader is included in the flow of information and reasoning:

5.78 For this purpose, **we calculate** these trajectories for an initial electronic ensemble. [PHYS-TH]

5.79 Hence, **we can simplify** the counter terms… [PHYS-TH]

5.80 **We define** three holomorphic functions.. [PHYS-TH]

5.81 So far, **we have restricted** the conversion rate [PHYS-TH]

5.82 Similarly, **we see** that **we have** the correct set of gauge fields [PHYS-TH]

The use of *we* to include the reader and the writer appears to be primarily a characteristic of theoretical registers, as this use is not found as readily in other registers. For example, first person pronouns are also common in qualitative applied linguistics, but most of the instances of *we* refer to the authors (articles are often authored by more than one person) and report actions and thoughts, particularly in laying out methodological steps and organizing the text:

5.83 **I observed** and **audiotaped** the classes of three Chilean English teachers and five California ESL instructors, spending 8 hours in each classroom over several weeks. [AL-QL]

5.84 In this exploration, **I define** culture as shared understandings and practices within groups of people [AL-QL]

5.85 In this section, **I review** how these teachers' transnational life experiences helped them to develop intercultural competence [AL-QL]

However, *I* is often used distinctively in qualitative applied linguistics, referring to the participants in the research, as their speech is reproduced in the article as a form of data or evidence for analysis (examples underlined):

5.86 Like Ruby, Paloma laughed when **I asked** about her cultural identity. Still laughing, she replied, "My cultural identity. Um, **I** was born white, Catholic, **I** went to the States, they told me that **I**'m not white, **I**'m Hispanic [AL-QL]

Third person pronouns are most common in qualitative applied linguistics and history, followed by theoretical philosophy, and then political science and quantitative applied linguistics. Third person pronouns are relatively rare in the hard sciences of biology and physics. Third person pronouns are used to refer to the human participants that are the object of study in qualitative applied linguistics (excerpts 5.87–88):

5.87 The same views were shared by Lin, another Chinese student, who told me that **she** felt unprepared to 'memorize' her talk in English [AL-QL]

5.88 The students were aged between 18 and 20, and, as this course was designated 'upper level', **they** had TOEFL scores of 500 or above [AL-QL]

In history texts these pronouns are used to report the actions and thoughts of actors in the events being analyzed and described (that is, the human referents to the pronouns are not typically the object of study), illustrated in excerpts 5.89–5.90:

5.89 The governor-general of Algeria, Jules Cambon, was sympathetic to Ferry's recommendations, however. In 1893, **he** noted to the administrator of the mixed commune of Boghari that the Algerians had been fined [HIST-QL]

5.90 The Poles had stressed that **they** were acting completely independently of the Germans. [HIST-QL]

Second person pronouns are rare in all disciplines and registers. In fact, as seen above in Figure 5.7, when second person pronouns are used, they primarily occur in philosophy and applied linguistics. In philosophy, second person pronouns are used to involve the reader in a vignette or scenario that the author is using in order to create his or her argument, as in excerpts 5.91–5.93. These excerpts also demonstrate that uses of the second person pronoun *you* are often densely clustered near each other and near other personal pronouns as the author sets up and moves through the scenario:

5.91 Suppose I promise to do **you** a favour. If I keep that promise, I do not do it to contribute my share to an activity in which I co-operate with others. [PHIL-TH]

5.92 Having assured me that **you** would be there, **you** must understand that I would feel disappointed if **you** were not; and that **you** would be to blame. Otherwise **you** couldn't have meant what **you** said when **you** said **you** would certainly be there. [PHIL-TH]

5.93 She wants to continue living in your home, but **you** can no longer provide adequate care, and her presence is putting a real strain on your own family. What to do? In the end, **you** decide to put her in a nursing home. **You**'re convinced this is the right thing to do, though **you** know **you**'ll feel guilty for doing it. [PHIL-TH]

In contrast, 2nd person pronouns in applied linguistics articles (and particularly in qualitative articles) most often appear in reported speech which is presented as evidence in the analysis (excerpts 5.94–5.95), and carry meanings typical of second person pronouns in spoken language: as referring to a participant in the conversation, or as a general reference:[2]

5.94 When colleagues called her "near-native," she would respond, "Thank **you** for saying 'near' because if **you** say 'native,' I am not, and I will never be." [AL-QL]

5.95 Li-Ping expresses how she really likes interacting with the news clips: It's much more fun than studying books, I'll say, or just reading an article. Because uh … I mean if **you** just study books, **you** can actually do it everywhere, right?" [AL-QL]

In applied linguistics, *you* is also found in the reporting of statements or questions that were used in the gathering of data from human participants, as in:

5.96 The question ('have **you** ever read Orson Scott Card?') carries a preference structure with a limited set of possible responses [AL-QL]

5.97 The specific sections at the beginning of the grammaticality judgment activity with Robinson (1997) sentences were as follows: All of the following sentences contain invented words. As **you** read each sentence, **you** must decide if the sentence is a possible sentence in English or whether it is an impossible sentence. [AL-QT]

5.5 Summing up: Lexical and grammatical variation

The analyses in this chapter have shown that the use of core grammatical features varies systematically even in a collection of research articles. For example, individual disciplines (regardless of register) have shown shared patterns of use in terms of common meanings expressed by nouns and verbs, with less variation

2. My thanks to an anonymous reviewer who pointed out that most of the instances of *you* in 5.94–95 can be read as general reference.

within two registers in a single discipline (e.g., see Table 5.2). When we look at the use of passive voice, for instance, we see a cline of variation where passives are generally used more frequently in the natural sciences than in other disciplines (e.g., see Figure 5.5).

In addition to variation that occurs across disciplinary lines, however, there are also patterns of variation that correspond to differences in the types of articles published in each discipline. That is, it is not the case that all types of research in a single discipline utilize linguistic features in the same way. For example, going back to passive voice (Figure 5.5) we see that quantitative applied linguistics looks much more like biology and physics in its frequent use of the passive, contrasting with the lower reliance on passive voice in qualitative applied linguistics.

In the next chapter, I move on to a discussion of the use of structural complexity. While the analysis reported on in Chapter 6 also deals with grammatical structures, it differs from the analysis presented in this chapter in two main ways. First, the focus in Chapter 6 is on a set of linguistic features that function similarly to create specific style of discourse – in this case 'compressed' versus 'elaborated' discourse. Second, the analysis in Chapter 6 depends on lexico-grammatical patterns to a much greater extent in order to conduct the investigations.

CHAPTER 6

Structural complexity in journal registers

6.1 Introduction

In Chapter 5, I looked at the general grammatical characteristics of the disciplines and journal registers in the corpus, and summarized the existing body of research related to the grammatical patterns across broad register divisions. Other recent research has focused more specifically on the nominal style of academic writing, and particularly on noun phrase modifiers. This research has shown that pre-modifiers (adjectives and nouns) and post-modifiers (prepositional phrases, relative clauses, appositive noun phrases) are used to a much greater extent in academic writing than in either spoken or other written registers (e.g., see Biber 1988, 1992; Biber & Clark 2002; Biber & Gray 2010; Biber, Gray & Poonpon 2011). This research has challenged the traditional notion that structural complexity refers primarily to the use of clausal structures, and has argued that while the dense use of clausal structures *is* characteristic of structural complexity in spoken language, structural complexity in written registers is phrasal in nature, involving the use of many of the features of the nominal style that has been well-documented for academic writing (e.g., Biber 1988, 1992; Biber et al. 1999; Biber & Gray 2010; Halliday 1989; Wells 1960).

More recently, Biber and Gray (2013) have shown that as this nominal style has developed over the past century, science writing has adopted this style to a much greater extent than non-science writing. This finding suggests the likelihood of substantial synchronic disciplinary variation in the use of various grammatical structures associated with structural complexity. Thus, the analysis in this chapter turns to an investigation of grammatical features associated with structural complexity (Biber & Gray, 2010; Biber, Gray & Poonpon 2011).

The analysis in this chapter differs from the analysis in Chapter 5 in several regards. First, the selection of linguistic features analyzed here is not intended to provide an overview of general grammatical characteristics, but rather an analysis of a targeted, focused set of grammatical features that work together to create a particular functional effect: that of creating informationally dense or compressed, versus elaborated, clausal discourse (see Section 6.2).

Second, the methodology for the present analysis (described in Section 6.3) more closely incorporates lexical information in the analysis. Here, structures such as *to*- and *that*-complement clauses are identified based on controlling words (verbs, nouns, adjectives) previously found to most frequently control these types of clauses (based on information in Biber et al. 1999). Although the focus of analysis is a set of grammatical structures, lexical information is utilized in order to facilitate the automatic processing of the corpus with a greater degree of reliability than otherwise possible.

6.2 Features of elaboration and compression in academic prose

Building upon the substantial research which has established the distinctive structural characteristics of spoken and written English, Biber and Gray (2010) investigate the perception that academic writing is structurally complex and highly elaborated. Motivated by an apparent mismatch between the body of research that had documented the primarily nominal style of academic writing and the verbal style of spoken language, and the persistent stereotype that academic writing is structurally complex and elaborated, Biber and Gray (2010) set out to document the differing nature of structural complexity in spoken versus written language.

While most paradigms of grammatical complexity define complexity based on the use of embedded clauses, Biber and Gray (2010; Biber, Gray & Poonpon 2011) show that such clausal embedding is characteristic of spoken registers, but not written registers (and particularly not academic prose). Rather, they show that structural complexity in academic writing comes from extensive phrasal embedding, often in sentences with quite simple main clause syntax such as those listed in Table 6.1. In addition to multiple prepositional phrases as noun post-modifiers

Table 6.1. Illustrations of extensive phrasal embedding in academic writing

Clausal Structure	Sentence
X is not Y	The **account** [of the **relation** [between ancestry and harm [in **cases** [of *historic* **injustice**]]]] is not categorical. [PHIL-TH]
X has been used in Y	Since then this term has been used in the **study** [of *public* **reactions** [to **circumstances** [of *social* and *political* **change** [at *various historical* **moments**]]]]. [HIST-QL]
X deserves Y	The *distinctive* **effect** [of the **size** [of the *Asian* **population**]] [on *income* **inequality**] certainly deserves *further* **research**. [POLISCI-QT]
X depends on Y	The *reconstruction* **efficiency** depends on the **occupancy** [of the *spectrometer multi wire proportional* **chambers** which is a **function** [of the **luminosity** [of each **exposure** (*target* **length** and *beam* **intensity**)]]]. [PHYS-QT]

(in square brackets), the noun phrases in these sentences also exhibit a dense use of nouns and adjectives as noun pre-modifiers (*italicized*), and at times clauses or appositive noun phrases that modify head nouns (underlined).

Biber and Gray (2010) analyze the use of five types of clausal embedding (finite complement clauses, non-finite complement clauses, finite adverbial clauses, finite relative clauses, and non-finite relative clauses) and four types of phrasal embedding (attributive adjectives, nouns and nominal pre-modifiers, prepositional phrases as nominal post-modifiers, and appositive noun phrases). They link these clausal features to structural elaboration, as the subordinate clauses embed a great deal of information into the main clause. As an example of such elaboration, consider finite relative clauses like in excerpts 6.1–6.2. Here, the relative clauses offer additional information to either describe or specify the referent of the head noun.

6.1 Lower-proficiency learners experienced more difficulty in integrating multiple textual and extra-textual cues (background knowledge) than did higher proficiency **learners**, **who** appeared to know more words in the context. [AL-QT]

6.2 Each neuron has **parameters** **which** are recursively adjusted by learning algorithms. [PHYS-QT]

Adverbial clauses like in 6.3 are likewise elaborating, as they are optional elements that are "added on to the core structure of the main clause to elaborate the meaning of main verbs" (Biber & Gray 2010: 6):

6.3 **Because** the 1996 NATA risk values are calculated for 1990 census tracts, we used geoprocessing in ArcGIS to apply these risk estimates to the 2000 census tract polygons. [POLISCI-QT]

Complement clauses are elaborating structures; these clauses are not optional elements, but rather typically fill the slot of a required clause element. Biber & Gray claim that complement clauses are elaborating because the information from an entire clause is used in a syntactic slot often filled by a noun phrase; The result is more information being packed into the clause, illustrated in excerpt 6.4:

6.4 We **know** **that** this approximation for "undisturbed" propagation is an oversimplification neglecting the effect of the changing temperature gradient (Brunt-Va frequency), the background wind changes and the saturation of waves. [PHYS-QT]

In contrast, Biber and Gray (2010) see embedded phrasal features as indicators of structural compression; information is compressed into noun phrases in optional phrases, many of which can be considered more condensed alternatives to fuller clausal structures. For example, features like prepositional phrases and nouns as

nominal pre-modifiers convey meanings that could be more explicitly conveyed through elaborating clausal structures. The noun phrase "one possible reason for the discrepancy" could be paraphrased as "a possible reason that explains the discrepancy," and the noun phrase "response time" can be paraphrased as "the time that it takes for someone to respond." These brief examples illustrate how phrasal features such as adjectives, nouns, and prepositional phrases can be embedded into the structure of noun phrases to function to elaborate meaning – however, this elaborated meaning is highly compressed in nature.

Historical studies (Biber & Gray 2010, 2013, 2016) have documented an increase over the past century for these types of phrasal modifiers in written registers. Furthermore, academic research articles exhibit particularly dramatic increases in the use of these features, likely related to their informational purpose and highly specialized audience (who use specialized knowledge to comprehend the range of relationships that these structures reflect; see Biber & Gray 2016 for further discussion).

This same line of research has found that these features have increased to a markedly higher degree in the natural sciences when compared to non-science research articles. Biber and Gray (2013) compare science and non-science disciplines, rather than specific disciplines. However, the marked differences between the two general areas of inquiry is indicative of potential variation between specific disciplines, even disciplines that would fall under the 'science' label. Thus, the purpose of this analysis is to investigate the use of elaboration and compression features across the range of disciplines and journal registers represented in the Academic Journal Register Corpus.

6.3 Carrying out a study of structural complexity

Table 6.2 summarizes the major structural categories associated with elaboration and compression that are analyzed in the present study. Four features can be considered 'elaborated' structures: finite complement clauses, which include *that* and *wh*-clauses that function as verb, noun and adjective complements; non-finite complement clauses, which include to-clauses and *ing*-clauses; finite adverbial clauses beginning with adverbial subordinators (e.g., *because, although, if, since, unless, when,* and *while*); and non-finite adverbial clauses (represented by sentence-initial to-clauses and clauses beginning with 'in order to').

This study also analyzes the use of three 'compressed' features: attributive adjectives, nouns as nominal pre-modifiers, and prepositional phrases as nominal post-modifiers. For the purposes of this study, prepositional phrase as nominal post-modifiers are represented by noun + *of* prepositional phrases;

because prepositional phrases can function as noun modifiers or as adverbials, hand coding is necessary to identify prepositional phrases that are functioning specifically as noun modifiers. Prepositional phrases with *of,* however, always function as noun modifiers, and thus could be identified automatically.

Table 6.2. Structural elaboration and compression features (see Biber & Gray 2010; Biber et al. 2011)

'Elaborated' Grammatical Structures	
Finite complement clauses	*These results show that the volumetric body force increases as a function of frequency and applied voltage* *Tuskegee has also been the place where thousands of successful black professionals were educated* *It is not at all clear that such concerns are warranted*
Non-finite complement clauses	*There is a need to fully consider how relationship of power emerge* *Campaign negativity for any office makes people want to stay home from the polls*
Finite adverbial clauses	*This issue of gender is trickier, however, because the archival sources almost always identify X* *If the handwriting of the confession is compared with the complaint, it is evident that X*
Non-finite adverbial clauses	*To avoid this counter-intuitive consequence, we can improve the formulation of a mixed theory* *Religious group is included in the model in order to capture whether members of minority religions feel less satisfied with life than members of the majority…*
Clausal Grammatical Structures Associated with Nominal Style	
Finite relative clauses	*the various ways in which conversational storytellers structure their stories* *every moral theory that gives some consideration to the consequences* *locals who wish to subvert national identity management*
Non-finite relative clauses	*the significant differences shown in model 1* *one piece of evidence supporting this conclusion* *the most effective way to address the participants' concerns*
'Compressed' Grammatical Structures	
Adjectives as nominal pre-modifiers	*common practice, electric field, high rates, federal government, specific instances, sustainable development, complex dynamics*
Nouns as nominal pre-modifiers	*energy transfer, output condition, child support system, ion atom collisions, cash benefit levels, axis ratio distribution details, field strength contribution results*
Prepositional phrases as noun post-modifiers†	*the loss of efficiency, the nature of incidental learning, the observed winter ratio of mean fluctuations, the essence of the brain's representational achievements*

†As represented by noun + *of* sequences for the purposes of the present investigation.

Two features, finite and non-finite relative clauses, can also be considered 'intermediate' structures. Because relative clauses modify a head noun, they can be associated with the nominal style of academic writing. However, at the same time, these structures are clausal in nature. In fact, previous research (Biber & Gray 2010; Biber, Gray & Poonpon 2011) have shown that of all of the clausal structures they examine, finite and non-finite (including *to*-clauses, *ing*-clauses, and *ed*-clauses) relative clauses occur more frequently in academic writing than in conversation (with the exception of *that*-relative clauses). Biber, Gray and Poonpon (2011) recognize these features as 'intermediate' features, yet still group both finite and non-finite relative clauses under 'elaborating' features. In this study, however, I argue that while finite relative clauses are indeed elaborating in the sense taken by this previous research, non-finite relative clauses might better be considered features of compression, a topic which I will return to below in Section 6.4.3.

A specialized computer program was developed in order to analyze the use of elaboration and compression features from Table 6.2. This program, which has also been used in a series of other studies focusing on elaboration and complexity features in a variety of synchronic and diachronic register comparisons (Biber & Gray 2010, 2011, 2016; Biber, Gray & Poonpon 2011), relies on grammatical tags as well as lexico-grammatical patterns. Features such as attributive adjectives, nouns as nominal pre-modifiers, relative clauses, noun + of prepositional phrases can be identified using the grammatical tags assigned to each word in the corpus by the Biber tagger. For features like complement clauses, however, a combination of grammatical tags and lexical information is used, which enables for a more reliable identification of the features of interest. For example, *that*- and *to*-complement clauses were identified based on any occurrence of *that* or *to* tagged as an infinitive marker preceded by one of the common controlling words identified for *that*- and *to*-complement clauses respectively in Biber et al. (1999). In sum, the program relies on both grammatical tags and frequent lexico-grammatical associations to identify the features of interest.

6.4 The use of features of structural elaboration and compression

In this analysis, three types of structures that can be embedded in main clauses and phrases are considered. In Section 6.4.1, I look at the use of embedded clauses as features that elaborate discourse across discipline and registers. In Section 6.4.2, I turn to the use of phrasal modifiers that can function as nominal pre- and post-modifiers to contribute additional information to noun phrases in a highly compressed manner. In Section 6.4.3, I look at clausal post-nominal modifiers, which have some characteristics of both the clausal features of elaboration (they

are clausal in structure) and the phrasal features of compression (they are modifiers within noun phrases).

6.4.1 Clausal elaboration

Figure 6.1 displays the frequency of four types of embedded clauses: finite and non-finite complement clauses, and finite and non-finite adverbials. It is interesting to note here that the relative distributions of these four features within a discipline are generally parallel across the sub-corpora; that is, non-finite complement clause are the most frequent structure in all disciplines, and finite adverbial clauses are the second most frequent, followed by finite complement clauses and then non-finite adverbials (which are relatively rare). The one exception to this is theoretical philosophy, where non-finite complement clauses and finite adverbials are used equally.

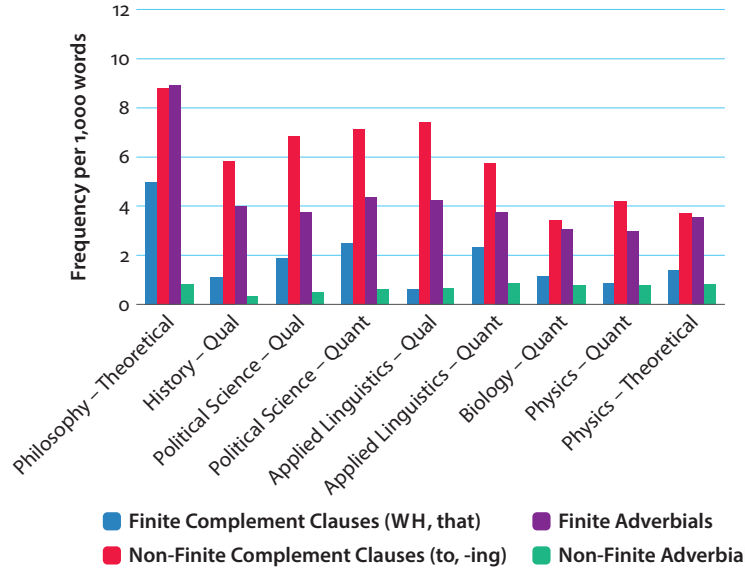


Figure 6.1. Distribution of structures associated with grammatical elaboration: Complement clauses and adverbials

A second trend illustrated in Figure 6.1 is that there is a general pattern of decline in the use of these elaborating features as we move from the pure, soft discipline of philosophy through the social science disciplines (political science and applied linguistics) and to the hard sciences (biology and physics). This trend is particularly observable for finite and non-finite complement clauses, and for finite adverbial clauses to a somewhat lesser extent.

Further trends become apparent when we consider the nature of the non-finite complement clauses, the most frequent elaborating feature. Figure 6.2 shows the use of non-finite complement clauses (ing-clauses and to-clauses combined) controlled by verbs, adjectives, and nouns. For the humanities and social sciences, non-finite verb complement clauses are the most frequent, while non-finite adjective complement clauses are most frequent in the three hard science registers. The higher use of verb complements in the softer disciplines, particularly philosophy, likely correspond to the overall higher use of verbs in these disciplines (see Figure 5.1 in Chapter 5).

However, for this same reason, it is a bit surprising to see the markedly lower use of non-finite noun complement clauses in biology and physics, considering their much higher reliance on nouns in general (see Figure 5.1 in Chapter 5). Rather, physics and biology registers rely more on non-finite adjective complement clauses.

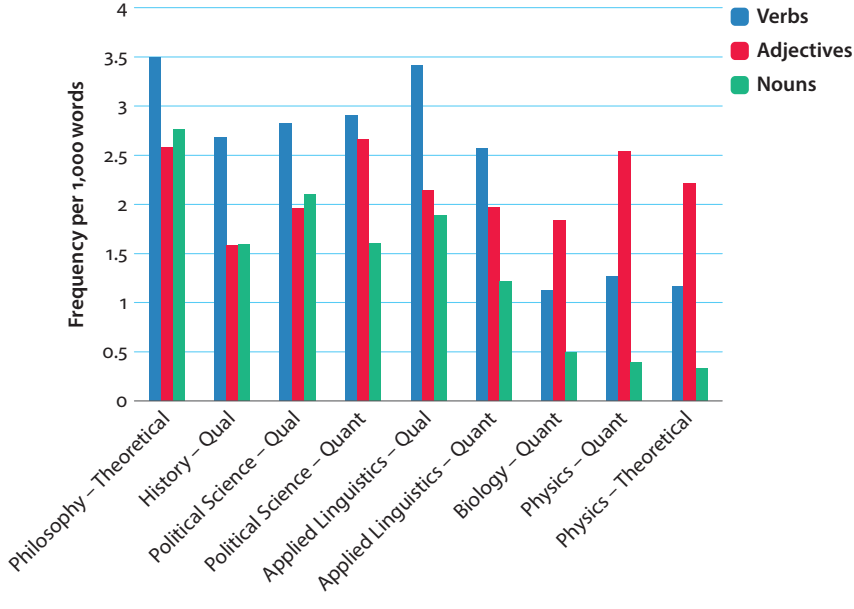


Figure 6.2. Distribution of non-finite complement clauses by controlling word type: Verbs, adjectives and nouns

Despite their overall higher use of nouns, it appears that the nature of the nouns used in physics and biology partially correspond to this low use of non-finite noun complements. That is, many nouns that tend to take non-finite complement clauses are often cognition nouns (excerpts 6.5–6.6), process nouns (6.7–6.8) and other abstract nouns (6.9–6.10), and the previous analysis on the

semantic types of nouns showed that biology and physics do not use these types of nouns frequently, instead exhibiting a high use of concrete, technical, and quantity nouns (see Figures 5.2 and 5.3, and Table 5.5 in Chapter 5).[1]

6.5 The ability **to isolate important causal forces** is important and experiments offer the opportunity. [POLISCI-QT]

6.6 Several actively disliked the thought **of learning more** but expressed the knowledge of benefits derived from acquiring a certain level of linguistic ability. [AL-QT]

6.7 In Apr. 1503, Fabyan was ordered by the court of aldermen to fulfill his agreement **to be alderman** 'upon payne of enprisonemet'. [HIST-QL]

6.8 It would then become interesting to consider the extent to which middle-class parenting practices are as they are because they have the effect **of improving children's chances of future reward** [PHIL-TH]

6.9 the Islamic Republic provides a significant opportunity to Moscow **to expand its influence and interest in both regions**. [POLISCI-QL]

6.10 On his view, intentionality is just a way **of referring to the content of an occurrent mental state**, that in virtue of which it secures its 'aboutness'. [PHIL-TH]

However, like many of the disciplines, physics and biology do use non-finite adjective complement clauses frequently. Across disciplines and registers, these adjective constructions are often used to express personal stance, that is, to mark evaluations and attitudes towards the propositions.

6.11 It is difficult **to know how far we can generalize these results to other L learners of a similar ability**. [AL-QT]

6.12 Also, in nano-robotics, adoption of this type of protective mechanism may be not only helpful **to control the movement,** but also essential **to safeguard the mechanism from overdriving**. [PHYS-TH]

6.13 It is not possible **to derive a principle of rationality or a principle of the right from a theory of the good**. [PHIL-TH]

6.14 First, and most fundamentally, it is necessary **to ask what was Labour's policy on the land question?** [HIST-QL]

1. This is not to say, however, that disciplines like physics and biology do not use non-finite noun complement clauses because of the types of nouns that they use. In fact, it is just as likely that these disciplines do not use the nouns that commonly head non-finite complement clauses because they do not provide elaborating information in this way.

6.15 The residue of discrimination makes some minority members reluctant **to trust these kinds of information sources** [POLISCI-QT]

While finite adverbial clauses participate in the general trend of being more frequent in softer disciplines, the pattern here is much less marked (Figure 6.2, 6.3). In fact, these constructions are about twice as frequent in philosophy as in any other discipline or register. Although finite adverbial clauses tend to be less frequent in the hard sciences than in the social sciences, the difference is much smaller in magnitude. Figure 6.3 shows specific adverbial subordinators that are used to introduce these adverbial clauses by discipline and register. While *if* is one of the most common subordinators in all disciplines and registers, *if* is extremely common in theoretical philosophy; in fact, the use of *if* accounts for a large portion of the difference in adverbial subordination between philosophy and all other disciplines and registers.

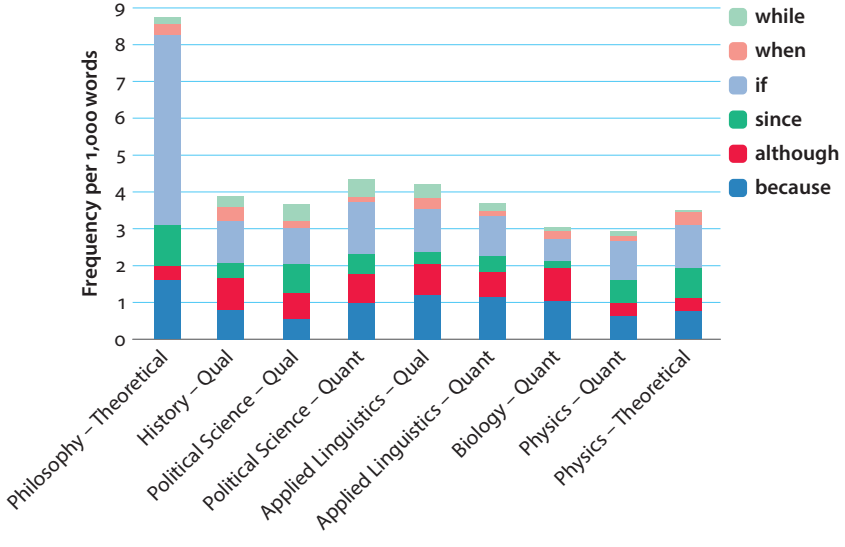


Figure 6.3. Distribution of adverbial subordinators across registers

Excerpt 6.16 from a theoretical philosophy article exemplifies the density of adverbial subordination; adverbial subordination in philosophy is often a way of exploring possibilities and logical relationships.

6.16 *Theoretical Philosophy (Merli 2008):*
Even **if** that belief were settled, there would still be issues of what importance to give it, what to do, and all the rest… Focusing on "what to do" in the sense of all-in endorsement makes the problem of preserving

> disagreement easier, **since** we clash whenever our prescriptions pull in different directions. But judgments of right and wrong, like judgments about what is beautiful or funny, do not by themselves settle what to do, **since** there is conceptual room to make these judgments **while** deciding to do something else. That is, the question of what to do remains open **once** the question of what is morally required is closed. **If** so, the incompatibility between different moral assessments is not exhausted by clashes of all-in prescription, since speakers might differ in their judgments about moral right and wrong **while** agreeing on what to do.

The grammatical features examined in this section all contribute to elaboration in text, and the distributions of use that have been discussed here show that these elaboration features are used much more extensively in philosophy than in any other discipline or registers. However, this analysis has also shown that the use of these features generally decreases as we move from the softer disciplines (philosophy and history), to the social sciences (political science and applied linguistics), to the hard sciences (biology and physics). In the next section, I'll discuss the use of features related to structural compression: phrasal modifiers.

6.4.2 Phrasal compression

Figure 6.4 shows the frequency of use for three types of phrasal modifiers: nouns as nominal pre-modifiers, adjectives as nominal pre-modifiers, and *of*-phrases as nominal post-modifiers. Adjectives as nominal pre-modifiers are the most frequent in all disciplines and registers, occurring 60 to 75 times per 1,000 words. Prepositional phrases with *of* as nominal post-modifiers are about half as frequent as adjectives, occurring between 30 and 40 times per 1,000 words. While these two features are quite common, they do not show a high degree of variation across disciplines and registers when compared to the differences that we've seen with other features. Nouns as nominal pre-modifiers, on the other hand, show a clear, increasing trend as we move from soft disciplines to hard disciplines. In fact, nouns as noun modifiers are almost as common as adjectives as noun modifiers in the hard sciences, particularly the two physics registers. In addition, if we compare registers within disciplines, it is clear that quantitative research uses nouns as noun modifiers more frequently than qualitative research.

Thus, nouns as nominal pre-modifiers are quite characteristic of writing in the hard sciences, and to a lesser degree in the social sciences. These nouns are less characteristic of the humanities disciplines of philosophy and history, and are the least common type of phrasal modifier in these two disciplines. The differing densities of these nouns as noun modifiers are exemplified in the following text excerpts from quantitative physics, quantitative applied linguistics, and qualitative history, where nouns as noun modifiers are **bolded** (head nouns are <u>underlined</u>).

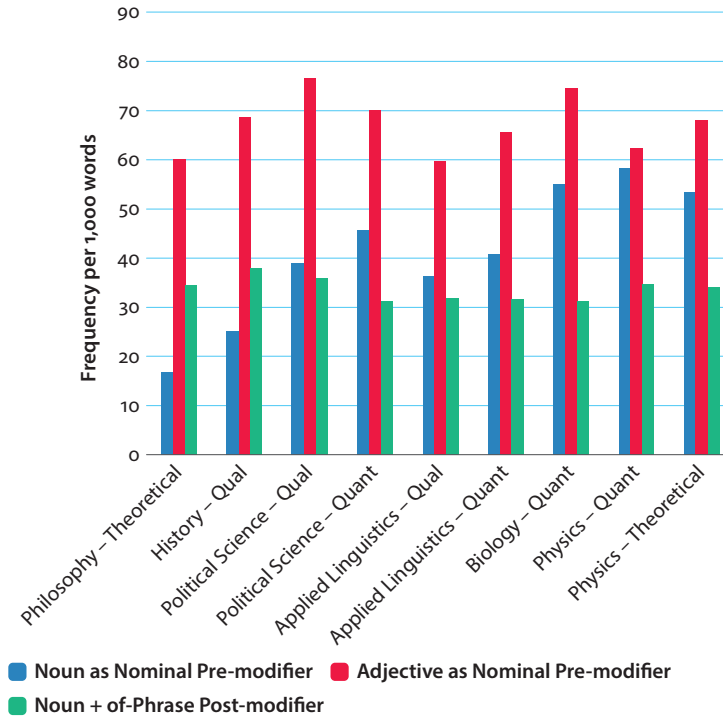


Figure 6.4. Distribution of structures associated with structural compression: Phrasal modifiers

6.17 The **cloud fraction frequency distribution** becomes wider with a shift of the peak toward higher values at coarser resolution. However, at smaller **domain sizes**, the shift of the peak toward larger values due to **pixel effects** is overcompensated by the increase in frequency of smaller **cloud fractions**… An increase in the **cloud size** at coarser resolution shifts the peak of the **size distribution** toward larger values... These results clearly indicate the importance of considering **scale effect** when comparing cloud resolving **model simulations** of **trade wind cumuli** with observations. [PHYS-QT]

6.18 In their study, Brecht et al. (1993) identified salient factors that played a role in second **language gain**. One of the outcomes was that **grammar achievement scores**, as measured by the Qualifying **Grammar Test**, positively related to gains in **speaking**, **reading**, and **listening proficiency**. [AL-QT]

6.19 The relationship between fire and humans has shaped **Plains ecology** and **Plains history** for centuries. The suppression of **prairie fire**, which came with Euro-American settlement in the nineteenth century, was one of the most significant events in **Great Plains** environmental **history**. [HIST-QL]

In this section, it has become clear that all disciplines and registers frequently rely on phrasal modifiers to convey information. One feature in particular, nouns as nominal pre-modifiers, shows substantial differences in terms of frequency of use across registers. Following the opposite trend of the clausal features of embedding found in Section 6.4.1, these nouns as noun modifiers are the most common in hard science disciplines (biology and physics). In the next section, I turn to two features that have characteristics of both clausal elaboration and phrasal compression: relative clauses.

6.4.3 Intermediate features: Clausal modifiers in the noun phrase

Relative clauses are 'elaborating' in the sense that they add supplemental information to noun phrases in order to either specify or describe a head noun. And while relative clauses are clausal in nature (they contain verbs and the clause elements required by the valency patterns of those verbs), they are embedded at the phrasal level, within noun phrases. Figure 6.5 shows the distribution of finite and non-finite relative clauses. Finite relative clauses, which contain subjects and markers of modality, tense, and aspect, are most common in theoretical philosophy. However, unlike some of the clausal features described in the Section 6.4.1 (e.g., finite adverbial clauses), it is not the case that these relative clauses are used to a much

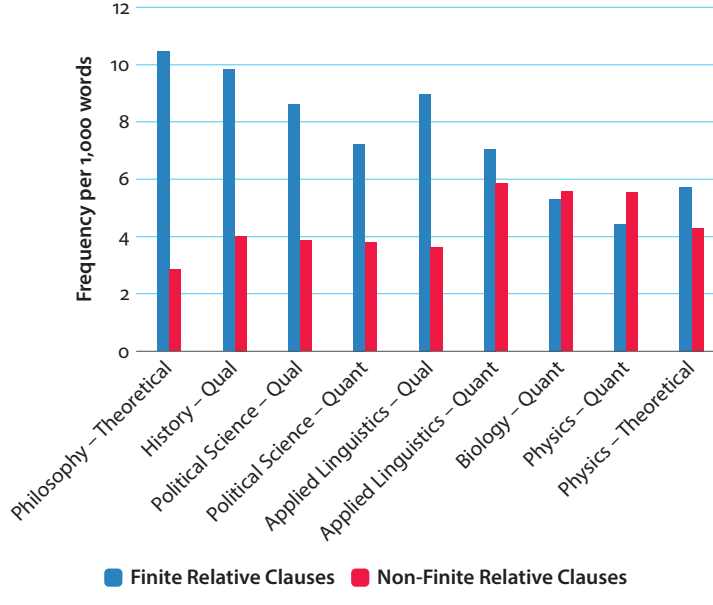


Figure 6.5. Distribution of structures associated with grammatical elaboration and nominal style: Relative clauses

greater extent in philosophy than in most of the other disciplines. Rather, as we move from the softer disciplines towards the harder disciplines (and from qualitative to quantitative research within political science and applied linguistics), there is a general, gradual decrease in frequency for these finite relative clauses (with an interesting spike in use in qualitative applied linguistics). In fact, the trends for these finite relative clauses are more similar to the trends found for other elaborating clausal features in Section 6.4.1 in that they are generally more frequent in the humanities, slightly less frequent in the social sciences, and least frequent in the hard sciences.

In contrast, non-finite relative clauses show the opposite trend: generally *increasing* in frequency as we move from soft to hard disciplines. That is, non-finite relative clauses are least frequent in the philosophy, slightly more frequent in history, political science, and qualitative applied linguistics, and more frequent still in quantitative applied linguistics, biology, and physics. Non-finite relative clauses mimic nouns as nominal pre-modifiers with respect to the patterns of use in these disciplines and registers. In fact, if we compare the functional effect of finite and non-finite relative clauses, these two trends seem logical and lend support to the notion of considering non-finite clauses to be features of compression rather than elaboration.

Finite relative clauses (excerpts 6.20–6.22) contain complete clauses, including subjects and full verb phrases marked for tense, aspect and modality. Therefore, these relative clauses add substantial information to noun phrases in terms of the subjects of verbs as well as details conveyed in the verb phrase such as present tense (6.20), past tense and perfect aspect (6.21), and modality (6.22):

6.20 People **who act unjustly on occasions where they are able to conceal that from others** can still enjoy the benefits resulting from others acting justly. [PHIL-TH]

6.21 The economic model **which had guided El Salvador through the previous 100 years** had been fundamentally altered. [POLISCI-QL]

6.22 Golonka (2000) reported additional evidence **that may contribute to a discussion of metalinguistic awareness and language gain**. [AL-QT]

Non-finite relative clauses, on the other hand, generally do not have subjects, and are not marked for tense, aspect, or modality, resulting in less information being conveyed in non-finite relative clauses. In fact, most non-finite relative clauses could be paraphrased as fuller, finite relative clauses, further supporting the idea that they are compressed – that is, reduced from fuller alternatives that are capable of conveying more specific information.

6.23 Salmonella enterica, the cause of food poisoning and typhoid fever, has evolved sophisticated mechanisms **to manipulate host cell functions**. [BIO-QT]

compare: *mechanisms that manipulate host cell functions*
compare: *mechanisms that salmonella uses to manipulate functions*

6.24 The most probable process **resulting from these reactions** is fission [PHYS-QT]
compare: *process that resulted/could result from these reactions*

6.25 However, orally presented prompt words seem to result in a higher proportion of paradigmatic responses than prompt words **presented visually** [AL-QT]
compare: *prompt words that researchers (had) presented visually*

6.26 The party's 1932 publication The Land and the National Planning of Agriculture argued that nationalization was needed in order to bring farming 'under a proper system of management'.The advantages of nationalization **listed in this pamphlet** were all to do with practicalities and economic efficiency [HIST-QL]
compare: *advantages of nationalization that the Labour Movement party listed in this pamphlet*

In particular, excerpts 6.25 and 6.26 illustrate the compression of information that results from the use of many non-finite relative clauses. These examples both contain non-finite clauses with object gaps (rather than subject gaps), and the subjects are omitted from the non-finite clause. Information about the subject of the verb (as well as about tense, aspect, and modality) are all absent from the clauses, leading to less elaborating information and less explicit statements of aspects of the discourse. This function, along with trends for finite and non-finite relative clauses which mimic the trends for elaborating clausal structures and compressing phrasal modifiers respectively, lead me to argue that non-finite relative clauses are in fact features that function to compress information into noun phrase structures. There is thus both theoretical and empirical support for a framework of compression and elaboration that places finite relative clauses with other elaborating clausal structures, and non-finite relative clauses with other phrasal compression features.

6.5 Summing up: Clausal elaboration and phrasal compression

The analyses presented above have demonstrated that elaborated and compressed grammatical structures vary in use across registers and disciplines, even within the constrained domain of published research articles. These analyses have shown that in general, features of structural elaboration are more common in humanities disciplines (particularly philosophy) than in the social sciences, and even less common in the hard sciences. In contrast, features of structural compression result in information being packed into noun phrases, and are more frequently

used in hard disciplines. Figure 6.6 summarizes these trends, and also illustrates that despite these differences across registers, all disciplines and registers maintain the nominal style of academic writing, relying on phrasal features of compression to much greater extents than clausal embedding. In particular, the placement of both elaboration and compression features in one figure highlights the differences in scales for the rates of occurrence for these features. Regardless of discipline or register, all sub-corpora utilize the compression features to a greater extent than the elaboration features.[2]

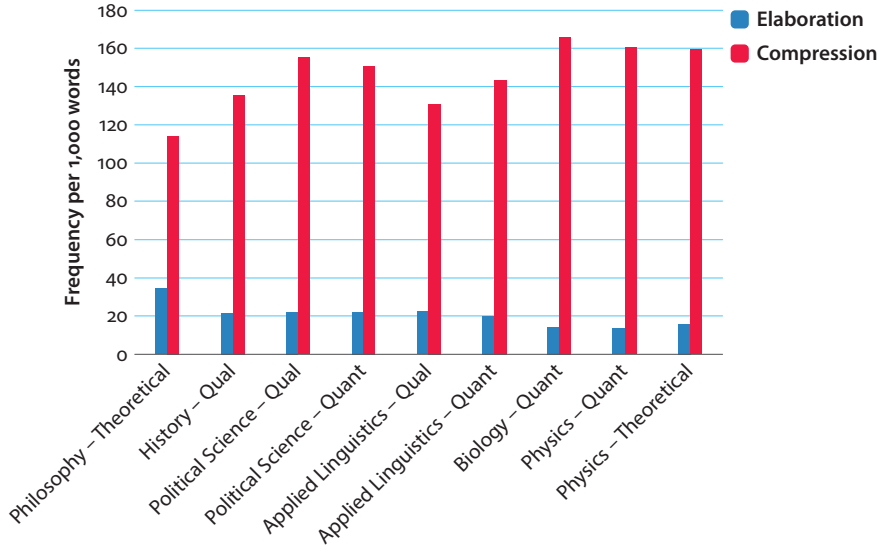


Figure 6.6. Summary of the use of elaboration (including finite relative clauses) and compression (including non-finite relative clauses) features

Excerpts 6.27 and 6.28 illustrate the differing extents to which compression features are used through examples from biology and history. In these excerpts, head nouns of complex noun phrases are **bolded** (for ease of reading, nouns that head phrases with no pre- or post-modification are not marked), nominal pre-modifiers are *italicized*, and nominal post-modifiers are <u>underlined</u> (except for finite relative clauses, an elaborating feature). For reference, the finite main verb

2. In addition, all sub-corpora here use the elaboration features to a lesser extent than spoken language. For example, while finite complement clauses occurred about 5 times per 1,000 words in philosophy articles (the discipline with the highest frequency for this structure), Biber and Gray (2010: Figure 1) report that finite complement clauses occur nearly three times as commonly in conversation – about 14 times per 1,000 words.

phrases in the passage are presented in SMALL CAPS, in both main clauses and embedded clauses.

In 6.27 from a quantitative biology article, there is an extremely dense use of noun modifiers, and the passage illustrates that one head noun is often modified by multiple pre- and/or post-modifiers (e.g., *weak plant growth promoting* **activity,** *protective* **effects** of Bacillus, the **effects** of Bacillus exudates on *fungal* **growth**). There are relatively few main verbs and no finite relative clauses, but there are three non-finite relative clauses (**pathogens** tested, *protective* **effects** observed for certain *Bacillus* **strains**, *In vitro* **experiments** studying the **effects** …).

6.27 *Quantitative Biology (Danielsson et al. 2006):*
The *B. amyloliquefaciens* **strains** SHOWED no or a *weak plant growth promoting* **activity**, whereas the *B. endophyticus* **strain** HAD *negative* **effects** on the plant as revealed by *phenological* **analysis**. On the other hand, two of the *B. amyloliquefaciens* **strains** CONFERRED **protection** of *oilseed* **rape** toward all **pathogens** tested. *In vitro* **experiments** studying the **effects** of Bacillus exudates on *fungal* **growth** SHOWED *clear growth* **inhibition** in several but not all cases. The *protective* **effects** of Bacillus can therefore, at least in part, BE EXPLAINED by **production** of *antibiotic* **substances**, but other mechanisms must also BE INVOLVED probably as a **result** of *intricate plant-bacteria* **interaction**. The *protective* **effects** observed for certain *Bacillus* **strains** MAKE them highly interesting for further studies as *biocontrol* **agents** in *Brassica* **cultivation**.

Excerpt 6.28 comes from a qualitative history article; this excerpt also exhibits the features of compression marked, including adjectives and nouns as nominal pre-modifiers, prepositional phrases and non-finite relative clauses as post-nominal modifiers. Although these features can all be found in this passage, they are not used with the same density as seen in 6.27. In addition, there are more finite main verbs, in addition to several finite relative clauses (which was reconfigured more than once, aberration to which the supposed increase…was primarily ascribed).

6.28 *Qualitative History (Avdela 2008):*
The *political* **dimension** of the **discourses** on 'youth' IS the fourth, in the sense that they WERE EMBEDDED into the *intense political* bipolarism between the Right and the Left, inherited from the Civil War and which WAS RECONFIGURED more than once during the **period** before the *military* **coup** of 1967. In what follows we WILL SEE how, in the **years** after the Civil War, *new* **forms** of entertainment among *young* **people** BECAME the **target** of *multiple* **attacks**, as they WERE CONSIDERED responsible for the *'moral* **aberration'** to which the *supposed* **increase** in *'juvenile* **delinquency'** WAS PRIMARILY ASCRIBED.

The relative densities of elaboration features can be illustrated similarly. Excerpts 6.29 and 6.30 come from theoretical philosophy and quantitative physics,

respectively. Complement clauses (finite and non-finite) are enclosed in [brackets], with the controlling word **bolded.** Finite and non-finite adverbial clauses are italicized, and finite relative clauses are underlined. In 6.29, clausal structures are present in every sentence and include a range of finite and non-finite complement clauses headed by nouns, verbs, and adjectives, as well as several adverbial clauses and finite relative clauses.

6.29 *Theoretical Philosophy (Altman & Wellman 2008):*
As troubling as the risk of abuse, we think, is the **problem** [that even sincere, well-meaning people cannot simply be **trusted** [to make reliable judgments on several essential matters]]. First, *even when a ruler is quite brutal*, his place may simply be taken by someone even more brutal. *If assassinating Saddam had the **consequence** [that his son, Udday, became ruler]*, then the rights of Iraqis might have been violated on even a more massive scale. Second, *even if the successor is not more brutal*, the assassination might have a backlash **effect** in which the public in the state of the now-dead ruler **demands** [that the rights-violating policies of the slain leader be pursued and even intensified]. **Suppose** [that NATO had assassinated Milosevic *in order to stop ethnic cleansing in Kosovo*]. The Serbian public might have become so **inflamed** by the assassination [that it would have been politically **impossible** for any successor [to negotiate a settlement with NATO that would have brought an end to the forced evacuations]

In contrast, excerpt 6.30 represents the much less extensive use of elaborated clausal structures. This quantitative physics excerpt contains only two embedded clauses (although there are several non-finite relative clauses: the specific role played by the 6s states of xenon, mixtures excited by electric discharges, state being populated, processes…correlated to, etc.).

6.30 *Quantitative Physics (Ledru et al. 2007):*
It is well **known** [that, in kryptonxenon mixtures excited by electric discharges, small amounts of xenon lead to the disappearance of the molecular continuum of krypton]. These energy transfers lead, via 5d6p and 6p6s transitions, to the 6s states of xenon being populated. Thus, the specific role played by the 6s states of xenon in several processes leading to the formation of homonuclear or heteronuclear excimers **needs** [to be specified and clarified]. In this paper, we present a spectroscopic and kinetic study of VUV emissions of KrXe mixtures around 150 nm. The aim of this experimental work is the determination of all the processes of formation and decay of heteronuclear excimers correlated to the Xe[6s] states.

While the compression features result in a dense informational style and serve as a means through which authors can pack a great deal of information into complex noun phrases, these elaborating clausal features also carry out particular

functions – often to convey the author's stance (their attitudes towards and evaluations of the propositions that they make).

6.6 Conclusions

While the analyses presented in Chapter 5 and 6 have had a primary focus on grammatical structures, that focus has varied as the studies move from grammatical constructs (Chapter 5) to a unified functional construct represented by a collection of linguistic features (Chapter 6). In the grammatical survey in Chapter 5, the focus was on grammatical categories proposed by linguistic theory, including semantic categories within those grammatical categories. In the study on structural complexity in Chapter 6, the focus was still on the nature of grammar in the Academic Journal Register Corpus. However, these particular linguistic features were chosen because of their utility in characterizing the structural nature of the discourse styles of these registers as 'elaborated' and 'compressed'.

These analyses have further shown that linguistic variation follows a variety of parameters, including differences that correspond to a high degree with simple disciplinary divisions, variation that occurs in patterns linked to the nature of the research (i.e., the type of article), as well as variation that follows the nature of disciplines as classified as 'hard', 'soft', and so on. Both of these analyses have considered the rates of occurrence of individual features. In the final linguistic analysis chapter (Chapter 7), I move to a statistical analysis that accounts for how grammatical, lexical, and lexico-grammatical features co-occur in the texts – a multi-dimensional analysis.

CHAPTER 7

A multi-dimensional analysis of journal registers

7.1 Introduction

In the last two chapters, I focused on describing variation in the use of core grammatical features and a specific collection of features that function to 'elaborate' and 'compress' language. The study reported on in the present chapter takes a different approach, describing the disciplines and registers according to characteristic co-occurrence patterns in the use of a wide range of linguistic features. These linguistic features are not grouped into subsets based on discourse function (as was the case for the analysis of elaboration and compression) prior to the analysis. Rather, the analysis in this chapter uses the statistical method of factor analysis to locate patterns of linguistic features that statistically co-occur, an analytical approach to uncovering register variation cultivated and named multi-dimensional (MD) analysis by Biber (1988, 1995).

Factor analysis is a statistical method for data reduction. Tabachnick and Fidell (2007: 608) describe the goal of factor analysis: "to reduce a large number of observed variables to smaller number of factors". That is, factor analysis serves to identify groups of variables that are correlated with one another (but not correlated with other groups of variables) in order to summarize trends in the data (see Tabachnick & Fidell 2007: 607–609). Biber (1988, 1995) applied statistical factor analysis to detailed linguistic analyses in order to characterize register variation in terms of the use of a much larger contingent of linguistic variables than previously examined in a single study, and to characterize variation that occurs along multiple parameters.

In Biber (1988), exploratory factor analysis was successfully used to identify patterns of variation across a wide range of spoken and written registers. The analysis resulted in the identification of seven factors, which were then interpreted as functional dimensions of variation. That is, each set of co-occurring features (and co-occurring features that were used in complementary distribution) were analyzed according to discourse functions generally common to the set of features included on a factor, with the interpretation aided by a consideration of

how the different registers related to those dimensions of variation (Conrad & Biber 2001: 24). In MD analysis, the term 'dimension' is used to encompass (typically) two groups of linguistic features, where the features within each respective group are highly correlated with the other variables in that group, and the two groups of variables occur in complementary distribution. In the past 20 years, MD analysis has been used in a variety of studies that consider the differences and similarities between highly diverse registers. In addition, MD analysis has been applied to the study of more specialized registers (e.g., see the chapters in Conrad & Biber 2001), and academic registers have been no exception. In the next section, I briefly summarize some of the research that has used MD analysis to study academic writing.

7.2 Background: Multi-dimensional analyses of academic language

As Conrad and Biber (2001) outline, MD analyses can be of two types. In the first type, registers are analyzed in terms of Biber's (1988) dimensions. In the second type, a full MD analysis is undertaken in which new dimensions are formulated based on the registers being investigated. Both approaches offer important information about register variation. When a new set of texts are assigned dimension scores for Biber's (1988) dimensions, they can be compared against the range of registers which have previously been analyzed in this way (as in Biber & Finegan 2001 and Conrad 1996a), adding to our knowledge about the range of registers that people encounter in their lives. The second approach, in which new dimensions are formulated, allows the researcher to discover co-occurring features that are important for a particular set of registers or genres.

In MD analyses of academic language, both types of analyses have been used. Conrad (1996b) uses Biber's (1988) dimensions to analyze textbooks and research articles in ecology and history, comparing them to other registers such as conversation, fiction, and non-fiction.[1] Conrad found that both register (textbooks versus research articles) and discipline corresponded to linguistic differences along the dimensions. For example, along Dimension 1, Conrad found that although research articles and textbooks in both disciplines were highly informational, research articles relied on those informational features a bit more. In contrast, she found that along Dimension 2 (narrative versus non-narrative discourse), variation

1. Conrad (1996a) also uses Biber's (1988) dimensions to analyze student texts; however, for the purpose of this book, I'll discuss only the results for professional academic texts here.

followed more closely along disciplinary lines, with Ecology registers being less narrative than history registers.

Biber and Finegan (2001) apply the MD approach to medical research articles, comparing the 1988 dimension scores for each of the IMRD (Intro-Method-Results-Discussion) sections to each other and to other registers. The analysis locates variation among the IMRD sections, but that variation is limited when compared to the range of variation found across other registers. For example, Biber and Finegan find that all sections of research articles are highly informational along Dimension 1 and generally non-narrative along Dimension 2. On Dimension 5 (impersonal versus non-impersonal), however, methods sections are very highly impersonal. Discussion, results, and introduction sections are also impersonal, but to a lesser extent than the methods sections.

Biber et al. (2004) apply the 1988 dimensions to a wider range of academic language, including both spoken and written registers encountered in university settings. Using the same corpora, Biber (2006: Chapter 7) conducts a new MD analysis to identify co-occurrence patterns that are specific to a more specialized domain (see Biber 2006: 181–182). Although the focus of this new analysis is on describing major register differences (e.g., between classroom teaching, textbooks, institutional writing, service encounters, etc.), Biber also includes a section in which he compares disciplines using classroom teaching and textbooks. If we look at the results specific to textbooks across disciplines, the analysis reveals that along the new Dimension 2 (procedural versus content-focused discourse), natural science textbooks are highly content-focused when compared to other disciplines (although all disciplines fall on the content-focused side of the continuum, see Biber 2006: 204). Biber found a wider range of variation across disciplines along his new Dimension 3 (narrative versus non-narrative orientation), with natural science and engineering textbooks being characterized as highly non-narrative, and education and humanities textbooks having a more narrative orientation.[2]

As Biber (2006: 181–182) notes, the MD approach is useful particularly when looking at many sub-corpora, such as varying registers in multiple disciplines. As summarized above, the previous MD analyses on academic language have been carried out on registers covering a broader range of situational characteristics, such as spoken and written academic registers. In the present study, the focus is on describing a much more specialized, narrow domain of language use. Perhaps more importantly, the focus is only on written registers. Thus, it is likely that the dimensions of variation found to be quite productive in describing

2. It should be noted, however, that classroom teaching in education, humanities, and social science relied on a narrative orientation to a much greater extent than these textbooks.

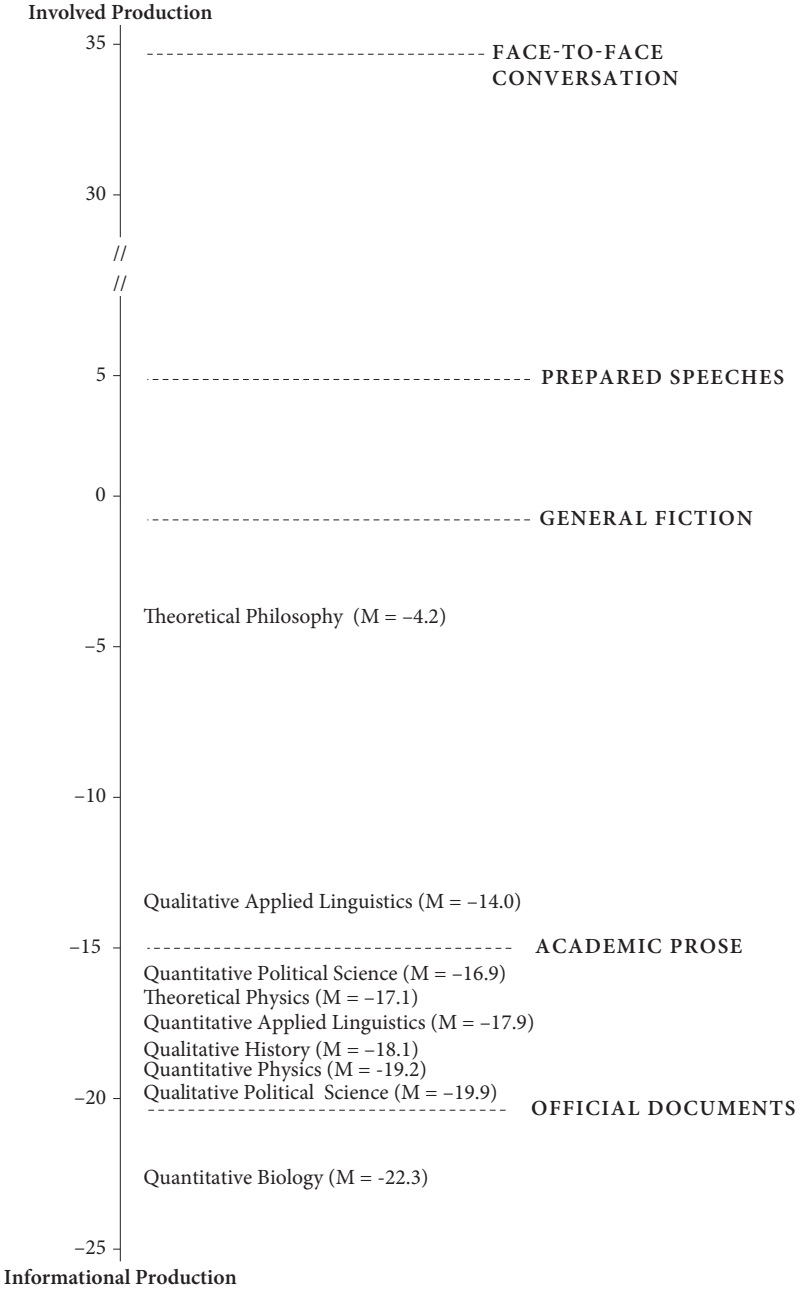


Figure 7.1. Distribution of disciplines and registers along Biber's (1988) Dimension 1 (involved versus informational production), with 5 general registers from Biber (1988) plotted for comparison

register differences in Biber (1988) will not be as useful in describing the range of variation that is important for these particular registers. In fact, we can readily see this by comparing the nine registers being studied in this book along the 1988 dimensions.

The tagcount program, described in Chapter 5, also produces dimension scores for each of the 1988 dimensions (dimension scores will be discussed below in Section 7.3.3). We can then plot these mean dimension scores by register and compare them to the registers studied in Biber (1988). Figure 7.1 shows this comparison along Dimension 1, which has been interpreted as characterizing 'involved' production versus 'informational' production and has represented a clear cline of variation that has consistently distinguished between many spoken and written registers, even across languages (e.g., see Biber 1995). In Figure 7.1, mean dimension scores for the 9 disciplines and registers in this study are plotted, showing that all of these registers fall on the 'informational' side, yet to differing extents. A selection of the registers from Biber (1988) are also plotted, showing that the range of variation among these registers is much broader, while the range of variation among the disciplines and academic journal registers is much more restricted in nature.

Thus, while making comparisons such as this one are interesting in and of themselves, another approach is to conduct a new factor analysis in order to identify the patterns of variation that are the most important in the more specialized domain of types of journal articles across disciplines. In the next section, I describe the methodology for carrying out a new multi-dimensional analysis of academic journal registers across disciplines.

7.3 Carrying out a new multi-dimensional analysis

7.3.1 Initial factor analyses to determine linguistic variables

One of the major strengths of multi-dimensional analyses of language variation is the ability to at once consider the use of very large sets of linguistic features. The data for this MD analysis comes from the 'tagcount' program described briefly in Chapter 3, which provides counts for approximately 130 linguistic features ranging from lexical classes to syntactic structures (see Conrad & Biber 2001: Chapter 2 for a detailed overview).

The first step in a new MD analysis is to select the linguistic features that will be used in the factor analysis. To be suitable for inclusion in a factor analysis, linguistic variables should be conceptually distinct. That is, the linguistic features should be correlated to some degree in terms of use, but care should be

taken that variables largely measuring the same construct or linguistic feature are not all included in the factor analysis. A competing goal is to be as inclusive and specific as possible in the selection of variables for the analysis (Conrad & Biber 2001: 15). For example, the output from the tagcount program includes counts for the overall use of *that*-complement clauses controlled by verbs as well as separate counts for *that*-clauses controlled by stance-carrying verbs grouped into different meaning subsets, such as non-factive, factive, attitudinal, and likelihood verbs. The first count is in essence a composite variable made up of the individual counts for specific subsets of verbs, and thus these variables are not conceptually distinct. Therefore, a choice has to be made whether to include a greater number of more specific variables (one for each type of stance meaning), or a more general overall count. This decision is made by carrying out pilot factor analyses to determine which combinations of variables are able to explain the greatest proportion of variance in the corpus.

In order to determine the set of variables to use in the final factor analysis, a series of initial factor analyses were run using varying groupings of the possible variables. These initial runs served as pilot analyses to identify features which were likely to contribute to the explanation of the linguistic patterns. The various pilot factor analyses were compared in terms of the linguistic features that 'loaded' onto the factors, along with the amount of total variance in the corpus that is accounted for by the various pilot solutions. In general, linguistic features were included in the final analyses only if their communalities exceeded .250 and they loaded on at least one factor with a factor loading[3] of greater than .30.[4] Based on the cumulative patterns found in these pilot factor analyses, 70 features were selected to be included in the final factor analysis, listed in Table 7.1. All words in the semantic sets of nouns, verbs and adjectives are listed in Appendix C.

---

3. Factor loadings range from 0 to 1, and measure how much variance an individual feature has in common with the total shared variance of a factor, thus indicating degree of co-occurrence between feature and the set of features on the factor (Conrad & Biber 2001: 21; Biber 1988).

4. Communalities represent the amount of variance in that variable that is explained by the factor solution (Tabachnick & Fidell 2007: 621). Variables with low communalities are excluded from the factor analysis.

Table 7.1. Summary of linguistic features included in the final factor analysis

Linguistic Feature	Description/Example
A. General	
1. type-token ratio	in the first 400 words of the text
2. word length	average number of letters per word
3. word count	overall number of words per text
B. Nouns and Pronouns	
4. 1st person pronouns	e.g., *I, we*
5. 2nd person pronouns	e.g., *you*
6. 3rd person pronouns	e.g., *he, she, they*
7. pronoun 'it'	all instances of 'it'
8. demonstrative pronouns	*this, these, that, those*
9. nominal pronouns	e.g., *somebody, anyone*
10. all nouns	all words identified as nouns by automatic tagger
11. nominalizations	e.g., *interaction, communication*
12. animate nouns	e.g., *adult, applicant, child, immigrant, patient*
13. process nouns	e.g., *achievement, comparison, effect, formation*
14. cognition nouns	e.g., *ability, decision, concept, idea, knowledge*
15. other abstract nouns	e.g., *advantage, background, culture, model*
16. concrete nouns	e.g., *acid, brain, camera, computer, glacier*
17. technical nouns	e.g., *atom, cell, compound, equation, message*
18. quantity nouns	e.g., *amount, century, frequency, percentage*
19. group nouns	e.g., *church, committee, government, institute*
C. Verbs	
20. possibility, permission and ability modals	*can, could, may, might*
21. prediction modals	*will, would, shall, be going to*
22. necessity and obligation modals	*must, should, had better, have to, got to, ought*
23. verb *BE*	all forms of verb BE
24. verb *HAVE*	all forms of verb HAVE
25. activity verbs	e.g., *bring, combine, encounter, obtain, produce*
26. communication verbs	e.g., *acknowledge, answer, claim, discuss*
27. mental verbs	e.g., *confirm, find, identify, observe, predict, think*
28. causative verbs	e.g., *affect, allow, help, influence, require*
29. existence verbs	e.g., *appear, define, illustrate, indicate, reflect*
30. aspectual verbs	e.g., *begin, complete, continue, keep, start*

(*Continued*)

Table 7.1. (Continued) Summary of linguistic features included in the final factor analysis

Linguistic Feature	Description/Example
D. The Verb Phrase	
31. past tense	e.g., *claimed, concluded, found, reported*
32. perfect aspect	e.g., *had argued, have discussed, has shown*
33. progressive aspect	e.g., *is becoming, is causing, are seeking*
34. agentless passive voice	passive constructions with no specified agent
35. by-phrase passive voice	passive constructions with agent in by-phrase
E. Adjectives	
36. all attributive adjectives	all adjectives occurring as a noun pre-modifier
37. all predicative adjectives	all adjectives occurring in post-predicate position
38. size adjectives (attributive)	e.g., *big, great, large, small*
39. time adjectives (attributive)	e.g., *new, young, old*
40. evaluative adjectives (attributive)	e.g., *best, good, important*
41. relational adjectives (attributive)	e.g., *basic, common, different, major, similar*
42. topical adjectives (attributive)	e.g., *economic, human, international, public*
F. Adverbs	
43. general adverbs	
44. time adverbs	e.g., *again, later, now*
45. stance adverbs	e.g., *obviously, evidently, frankly, surprisingly*
G. Coordination and Subordination	
46. adverbial conjuncts	e.g., *however, therefore, thus*
47. clausal coordinating conjunctions	e.g., *and, or*
48. phrasal coordinating conjunctions	e.g., *but*
49. conditional subordinating conjunctions	e.g., *if, unless*
50. subordinating conjunctions (other)	e.g., *as, except*
H. Clauses Marking Stance	
51. *that*-clause controlled by non-factive (communication) verb	e.g., *argue, claim, show, tell*
52. *that*-clause controlled by factive (certainty) verb	e.g., *demonstrate, conclude*
53. *that*-clause controlled by likelihood verb	e.g., *appear, estimate, seem, suppose, suggest*
54. *that*-clause controlled by factive (certainty) adjective	e.g., *conclude, proves*
55. *that*-clause controlled by likelihood adjective	e.g., *possible, probable*

(*Continued*)

Table 7.1. (Continued)

Linguistic Feature	Description/Example
H. Clauses Marking Stance	
56. *that*-clause controlled by attitudinal adjective	e.g., *afraid, aware, surprised*
57. *that*-clause controlled by factive (certainty) noun	e.g., *conclusion, fact, observation*
58. *that*-clause controlled by likelihood noun	e.g., *assumption, belief, hypothesis*
59. *that*-clause controlled by attitudinal noun	e.g., *hope, fear, view*
60. *to*-clause controlled by speech verb	e.g., *ask, claim, show*
61. *to*-clause controlled by verb of desire	e.g., *agree, hope, intent, prefer*
62. *to*-clause controlled by verb of causality, modality	e.g., *attempt, help, permit, require*
63. *to*-clause controlled by verb of probability	e.g., *appear, seem, tend*
64. all *to*-clauses controlled by stance adjectives	e.g., *certain, worried, appropriate, difficult, easy*
65. all *to*-clauses controlled by stance nouns	e.g., *claim, possibility, assumption, fact*
I. Post-Nominal Modifiers	
66. passive postnominal modifier	non-finite -ed clause postmodifying a noun
67. *that* relative clause	relative clause with *that* as relative pronoun
J. Other	
68. *wh*-questions	all clauses tagged as wh-questions
69. *wh*-clauses	all clauses with wh-complementizer
70. all prepositions	any word tagged as a preposition

7.3.2 Final factor analysis

Two preliminary analyses were used to ensure that the data set of selected variables is appropriate for factor analysis (using SPSS v. 19.0). The Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO = .855, meritorious) meets the minimum requirement for FA (values greater than .6, Tabachnick & Fidell 2007). Likewise, Barlett's Test for Sphericity (Approximate Chi-Square = 11005.23, df = 2415, p = .000) is significant, indicating that the null hypothesis that correlations are 0 can be rejected (Tabachnick & Fidell 2007). Both of these tests indicate that adequate correlations exist in the correlation matrix, and thus, FA is suitable for the data.

The final factor analysis was carried out using Principal Axis Factoring with Promax rotation in IBM SPSS v19.0. The four-factor solution was determined to be the most interpretable when compared with three- and five-factor solutions. Appendix D lists the full (rotated) factorial structure of the four-factor solution, and Appendix E shows the scree plot for the solution. The Initial Eigenvalues (Total Variance Explained) indicate a cumulative percentage of variance explained as 40.26%. All variables with a factor loading of .30 or above were considered important for the analysis for that factor. One variable (*that*-clauses controlled by attitudinal adjectives) did not load on any factor at the specified level.

7.3.3 Calculating and comparing factor scores across disciplines and registers

After identifying the underlying patterns of variation using factor analysis, the next step in the analysis is to characterize each register according to the factors themselves, that is, to quantify the extent to which each register utilizes the patterns of variation uncovered by the factor analysis. First, factor scores were calculated for each dimension for each text in the corpus. To calculate factor scores, z-scores were first computed for each linguistic feature in order to convert the rate of occurrence for each variable into a standardized scale where the mean is equal to 0 and the standard deviation is 1. This serves to equalize the impact of high- and low-frequency variables, so that dimension scores are not disproportionately impacted by high frequency linguistic features (see Biber 1988: 94).

Each factor in this study resulted in two groupings of variables, 'positive' features and 'negative' features, which represent co-occurring sets of features that are in complementary distribution. In other words, texts which rely highly on the positive features rely on the negative features to a lesser extent, and vice versa. To calculate dimension scores, which indicate the degree to which a text can be said to rely on linguistic features on a factor, the standardized z-scores for all of the positive features are added together, and the z-scores for the negative features are subtracted (Biber 1988).

Once a factor score has been computed for each text in the corpus for each dimension, dimension scores for the disciplines/registers can be calculated by taking the mean factor score for each text in sub-corpus. The means and standard deviations for each dimension are included in Appendix F, and Figures 7.2 – 7.5 below plot the registers according to mean dimension scores. One-way ANOVAs were used to test the significance of each dimension (Appendix F, Table F2), and show that all four factors are significant at $\alpha < .05$. The ANOVA results, along with an r^2 value, are presented along with Figures 7.2 – 7.5. The r^2 value is a measure of the proportion of the variance in the dimension scores that can be explained by the register groupings of the texts, and thus indicates how important a dimension

is for explaining the variation in the corpus (see Conrad & Biber 2001: 28). The r^2 values for the four dimensions in this MD analysis range from .81 (Dimension 2) to .43 (Dimension 4).

To test for individual differences between disciplines and registers, the Games-Howell procedure (equal variances not assumed, see Appendix F, Table F4) was used to make post-hoc comparisons. These post-hoc comparisons are listed in Appendix F, Tables F5 – F8. In the next sections, I turn to the interpretation of the four factors, relying cyclically on functional interpretations of the linguistic features that characterize a dimension of variation, as well as a consideration of how the disciplines and registers fall along these parameters.

7.4 Dimensions of variation in academic journal registers in 6 disciplines

Table 7.2 summarizes the four factors, listing the sets of co-occurring linguistic features with factor loadings indicated in parentheses. I have also given each factor a descriptive title that previews the functional analysis of these dimensions:

Dimension 1: Academic Involvement & Elaboration vs. Information Density
Dimension 2: Contextualized Narration vs. Procedural Description
Dimension 3: Human Focus vs. Non-Human Focus
Dimension 4: 'Academese'

In the sections that follow, I explore each of dimensions of variation in detail. More specifically, I discuss the functional underpinnings of these groups of features in relation to the ways in which the disciplines and registers are distributed along these dimensions of variation, and illustrate this analysis with text excerpts throughout.

7.4.1 Dimension 1: Academic involvement and elaboration vs. informational density

Dimension 1, labeled academic involvement and elaboration versus informational density, is made up of 26 features on the positive end of the factor, and 8 features on the negative end of the factor. Despite the fact that this dimension has been extracted based on only written registers in a fairly specialized domain, there is a good deal of overlap in the features that comprise this factor, Biber's (1988) Dimension 1 (involved versus informational discourse) and Biber's (2006) Dimension 1 (oral versus literate texts). In fact, this overlap is particularly apparent for the negative features on this dimension, where all but two of the features (process nouns, past tense) were also negative features in Biber's (1988) and/or (2006) dimensions:

Table 7.2. Structure of four-factor solution

Dimension 1: Academic Involvement & Elaboration vs. Information Density

Positive features:

Pronouns: nominal pronouns (.69), pronoun *it* (.62), 1st person pronouns (.58), demonstrative pronouns (.52)

Nouns: nouns of cognition (.57)

Adjectives: predicative adjectives (.70), evaluative attributive adjectives (.33)

Verbs: verb b*e* (.79), verb *have* (.67), causative verbs (.34)

Modal Verbs: modals of prediction (.69), modals of possibility (.66), modals of necessity (.65)

Adverbs: general adverbs (.54), stance adverbials (.47), adverbials of time (.34)

Conjunctions: Subordinating conjunction – conditional (.83), adverbial conjuncts (.48), subordinating conjunctions (.39)

Finite Clauses: *that*-clauses controlled by nouns of likelihood (.65), *that*-clauses controlled by verbs of likelihood (.59), *that*-clauses controlled by factive adjectives (.48), *that*-clauses controlled by attitudinal nouns (.47), *that*-clauses controlled by factive nouns (.44), *wh*-clauses (.34)

Non-Finite Clauses: *to*-clauses controlled by stance adjectives (.37), *to*-clauses controlled by verbs of probability

Negative features:

Nouns: nouns (–.75), process nouns (–.40)

Verbs: past tense verbs (–.67)

Passives: passive postnominal modifiers (–.53), agentless passive voice verbs (–.32)

Other: prepositions (–.39), type-token ratio (–.35), word length (–.31)

Dimension 2: Contextualized Narration vs. Procedural Description

Positive features:

Pronouns: 3rd person pronouns (.65)

Nouns: group nouns (.49), nominalizations (.32), animate nouns (.43)

Adjectives: topical attributive adjectives (.53), attributive adjectives indicating time (.47)
Verbs: past tense verbs (.55), aspectual verbs (.52), perfect aspect verbs (.48), communication verbs (.47), present progressive verbs (.42)

Conjunctions: phrasal coordinating conjunctions (.51), clausal coordinating conjunctions (.35)

Finite Clauses: *that*-relative clauses (.46), *that*-clauses controlled by non-factive verbs (.45), *wh*-questions (.32)

Non-Finite Clauses: *to*-clauses controlled by verbs of modality, causation and effort (.57), *to*-clauses controlled by verbs of desire (.41), *to*-clauses controlled by stance nouns (.35)

Other: word length (.52), word count (.36), type-token ratio (.31)

(*Continued*)

Table 7.2. (Continued)

Negative features:

Nouns: technical nouns (–.61), quantity nouns (–.46), concrete nouns (–.37)

Adjectives: attributive adjectives indicating size (–.37)

Passives: agentless passive voice verbs (–.52), passive voice verbs with *by*-phrases (–.47)

Dimension 3: Human vs. Non-human Focus

Positive features:

Pronouns: 2nd person pronouns (.40), 3rd person pronouns (.35)

Noun: process nouns (.50)

Verbs: mental verbs (.65), activity verbs (.60), communication verbs (.51), present progressive verbs (.49)

Finite Clauses: *that*-clauses controlled by factive verbs (.42), *wh*-clauses (.33)

Non-Finite Clauses: *to*-clauses controlled by verbs of desire (.46), *to*-clauses controlled by speech verbs (.45)

Negative features:

Adjectives: attributive adjectives (–.54), attributive adjectives indicating topic (–.42)

Adverbs: general adverbs (–.31)

Other: prepositions (–.35)

Dimension 4: 'Academese'

Positive features:

Nouns: nominalizations (.43), process nouns (.38), other abstract nouns (.32)

Adjectives: relational attributive adjectives (.45)

Verbs: existence verbs (.37)

Finite Clauses: *that*-clauses controlled by likelihood adjectives (.38), *to*-clauses controlled by stance adjectives (.33)

Other: word length (.58)

Negative features:

Adverbs: time adverbials (–.47)

all nouns, passive post-nominal modifiers, agentless passive verbs, prepositions, type-token ratios, and word length. In previous multi-dimensional research on a range of registers, these features have been associated with informational purposes, particularly in written registers. It is interesting that this combination of features still emerges as a dimension of variation that distinguishes amongst written texts which all have a primary informational purpose. This finding means that we must take a deeper look at how these features are functioning in the particular

registers in the study. When we do so (illustrated below), we see that these features can also be associated more specifically with *informational density* – that is, with highly compressed styles of discourse in which a great deal of information is presented in a dense manner.

On the positive end of the factor, features include various types of pronouns (nominal pronouns, 'it', 1st person pronouns, and demonstrative pronouns), modifiers (predicative adjectives, evaluative attributive adjectives, general adverbs, adverbs of time), explicit markers of logical and grammatical relationships (conditional and other subordinating conjunctions, adverbial conjuncts), and structures that convey personal stance meanings (possibility/permission/ability modals, *that*- clauses controlled by likelihood nouns, verbs and adjectives, factive nouns and adjectives, and attitudinal nouns; *to*-clauses controlled by stance adjectives and verbs of probability).

Although the degree of correspondence between the positive features on this dimension and Biber's (1988, 2006) dimensions is not as strong as for the negative features, some functional overlap does exist. For example, Biber's (1988) dimensions also contained many stance markers (including emphatics, hedges, and amplifiers),[5] types of pronouns, and explicit grammatical links (e.g., causative subordination and non-phrasal coordination, see Biber 1988). Further overlap exists between Biber's (2006) Dimension 1, including some of these same features, as well as conditional and wh-clauses.

In contrast, many of the features which were important on Biber's earlier dimensions, but which are not important features along Dimension 1 in the present study, are those that are highly correlated with interactional spoken language: contractions, indefinite pronouns, *wh*-questions, *that*-deletion, and discourse particles (see Biber 1988, 2006; Friginal 2009). In fact, many of these features were simply not included in the factor analysis conducted in this study because they did not exhibit high frequencies or substantial variation across these registers during the initial pilot factor analyses. That is, they did not contribute to understanding linguistic variation in these registers (as they are primarily characteristics of spoken language) and were excluded from further analysis.

Taking into account these two sets of features, it appears that Dimension 1 in this study reflects a common distinction that has been consistently identified in several different multi-dimensional analyses: that of high-density informational language and more involved language production. It should be noted, however,

---

5. In fact, these linguistic features from the (1988) dimensions were not considered in the present factor analysis, as these categories were largely included in various other stance categories from the newer stance framework developed based on the LGSWE (Biber et al. 1999), which has been used in more recent studies.

that Dimension 1 here reveals a narrower scope of involvement, more focused on those involvement features that are key to academic writing. Support for this interpretation is apparent if we look at the distribution of the disciplines and registers in the present study, plotted by mean dimension score in Figure 7.2. While not an exact match, the general pattern of how registers fall along Dimension 1 in the present study is quite similar to how they fall along Biber's (1988) Dimension 1, displayed above in Figure 7.1. Most noticeably, Figure 7.2 shows that Dimension 1 reflects a dichotomy between philosophy and all other disciplines. Philosophy has an *extremely* high positive dimension score (41.7) while the other disciplines range from a low of -15.9 (biology) to 2.1 (quantitative political science). This is the same overall pattern as seen above in Figure 7.1. The new Dimension 1 thus appears to be reflecting the same general underlying pattern, yet results in more precise information as to the involvement structures that are playing important roles in academic writing specifically (or at least in theoretical philosophy).

The issue of compression in academic language has been addressed to a certain extent in the last chapter; however, the statistical co-occurrence of some of these same compression features confirms the importance of these densification structures for academic writing. Excerpt 7.1 illustrates the high information density, where various structures are embedded into complex noun phrases. The last half of this excerpt is particularly reflective of the dense information structure that results from the use of the negative features on this dimension, where appositive noun phrases are used in abundance to pack concepts into a sentence structure with relatively few verbs. In fact, in the passage of 161 words, only five main clause verbs phrases (*were used, were delimited, were not included, were collected,* and *were collected*) are used, all of them passive. In the excerpt, the negative features on Dimension 1 are marked: nouns are **bolded**, prepositions are underlined, past tense verbs are *italicized,* passive structures are double-underlined, and passive postnominal modifiers are additionally in SMALL CAPS.

7.1 *Quantitative Biology (Hoeinghaus, Winemiller, & Agostinho 2008):*
Stable **isotopes** of **carbon** and **nitrogen** *were* used to estimate **food-chain length** and identify **patterns** of **material flow** through dominant trophic **pathways** for each **food web** (Hoeinghaus et al. 2007a). The aquatic **food webs** ANALYZED in this **study** were delimited by **fishes** as **consumers** plus their aquatic and riparian **prey** and **production sources** CONSUMED throughout the **web** leading to those **consumers**. **Parasites** and non-aquatic **organisms** that feed on **fish**, such as **birds** and **humans**, *were* not included. **Samples** for isotopic **analysis** *were* collected between **September** and early **December** 2003 (late dry **season**), prior to seasonally rising **water levels** and **fish migrations**. At each **location**, representative riparian and aquatic **carbon sources** (C **plants** and C **grasses**, fine **particulate** organic **material**, coarse **detritus**, **periphyton**, and **phytoplankton**), primary **consumers**

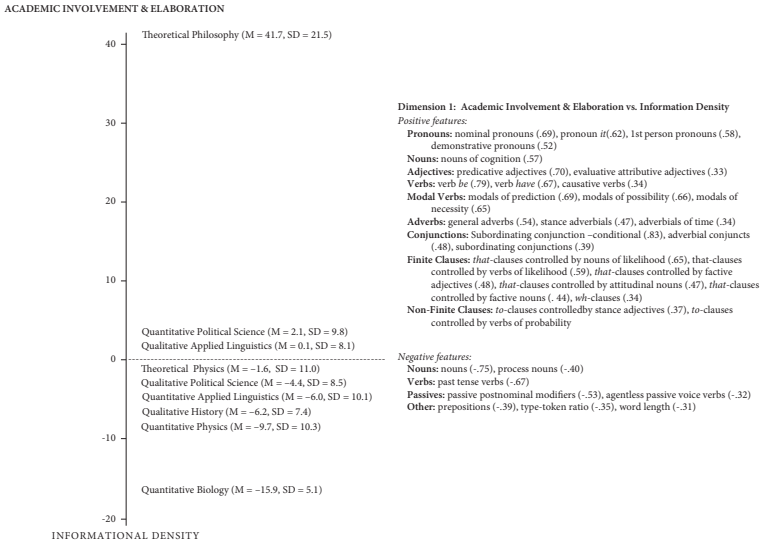


Figure 7.2. Distribution of disciplines and registers along Dimension 1: Academic involvement and elaboration versus informational density. One-way ANOVA results: F = 66.62, p = .000, r^2 = .67. Post-hoc comparisons are listed in Appendix F

(**snails**, **bivalves**, **zooplankton**, and herbivorous and detritivorous **fishes**) and secondary **consumers** (omnivorous and carnivorous **fishes**) *were* collected at multiple **points** along a 2A5 **km sample reach** to characterize trophic **pathways** from **source** to top **consumer**.

While the excerpt in 7.1 shows almost no use of the positive features on Dimension 1, it is not the case that registers with highly negative dimension scores on Dimension 1 lack all positive features. Rather, if we mark another passage from quantitative biology for both the negative and positive features on Dimension 1, we can see that the use of the positive features is simply overshadowed by the density of the negative features. In the first mark-up of this excerpt, the negative features of Dimension 1 are marked: nouns are **bolded**, prepositions are underlined, past tense verbs are *italicized,* passive structures are double-underlined, and passive postnominal modifiers are additionally in SMALL CAPS.

7.2 *Quantitative Biology (Kelly, Macisaac, & Heath 2006):*
The **beam waist** at 7 mm along the **beam** from the minimum **waist** is increased only by 10%. This **length** of the **beam** is within the main **viewing angle** of the **detector**. The **signal count** PRODUCED BY THE LASER is then proportional to k, with j as **photon density** and k as the **number** of **photons** USED TO PRODUCE THE **SIGNAL** in the **fragmentation process**. Due to the large **number** of **results** DESCRIBED HERE, we make a stepwise and systematic **approach**, proving each **step** carefully even if they might have been proved previously in other **publications** (ALWAYS REFERENCED). This means also systematically, that we do not use the disproved **models** to attempt to interpret later **steps** in the **development**, to avoid **confusion**.

In the second mark-up of this same excerpt, the positive features are marked: (main clause verbs are in bolded **SMALL CAPS**, pronouns are underlined, general adverbs and predicative adjectives are underlined, prediction and possibility modal verbs are *italicized*, and clausal structure are indicated with head words **bolded** and corresponding [square brackets]).

7.3 *Quantitative Biology (Kelly, Macisaac, & Heath 2006):*
The beam waist at 7 mm along the beam from the minimum waist **IS INCREASED** only by 10%. This length of the beam **IS** within the main viewing angle of the detector. The signal count produced by the laser **IS** then proportional to k, with j as photon density and k as the number of photons used to produce the signal in the fragmentation process. Due to the large number of results described here, we make a stepwise and systematic approach, proving each step carefully even **if**$_1$ [they *might* have been proved previously in other publications (always referenced).]$_1$ This **MEANS**$_2$ also systematically, [that we do not use the disproved models to attempt to interpret later steps in the development, to avoid confusion.]$_2$

The first excerpt from quantitative biology (excerpt 7.1), which focused on describing research procedures, used very few positive features along Dimension 1. In contrast, the passage in 7.2 and 7.3 exhibits some instances of positive features along Dimension 1. However, in addition to illustrating the (still) much less frequent use of positive features, this passage also demonstrates that in biology the positive features on Dimension 1 are used at points in the discourse where the writer is commenting on the research, making connections outside of the actual events of the research, and discussing the implications of the research. This contrasts to the use of the positive features of Dimension 1 in philosophy (excerpt 7.4 below), where the features are relied upon more pervasively throughout the discourse.

In the philosophy excerpt below, the positive features on Dimension 1 seem to be used for two main purposes. First, as in Biber's earlier dimensions, the use of personal pronouns creates a sense of interaction in the discourse, with authors explicitly talking about themselves (as in the biology example above), and bringing others into the study in order to illustrate points or provide evidence. Second, the features on Dimension 1 are used to elaborate discourse and show the involvement of the writers by providing personal stance and evaluation. These explicit evaluations, along with other types of subordinating conjunctions, serve to make explicit connections between meaning relationships, interpretations, and the authors' own evlautions, as illustrated in excerpt 7.4 (The positive features are again marked as above, and additional positive features such as cognition nouns and wh-clauses are underlined):

7.4 *Theoretical Philosophy (Roache 2006):*
Wollheim's **point**$_1$ [that in order to have q-memories of everything that his father experienced on his childhood walks, **he** *must* inherit not only his father's memories of these walks, but also **their** psychological context?]$_1$ Well, it *may* prove a problem for Wollheim's **claim**$_2$ [that **it** **IS** **impossible**$_3$ [to isolate memories from their psychological context,]$_3$]$_2$ but **it** does not undermine our weaker **claim**$_4$ [that memories are weakened when **they** are isolated in this way]$_4$. Martin does not tell **us** how much of the Spanish conversations **he** now remembers, but **it** seems **plausible**$_5$ [to **suppose**$_6$ [that the **fact**$_7$ [that **he** no longer understands Spanish]$_7$ *will* have resulted in his now not being **able**$_7$ [to recall some of the details of the conversations.]$_5$]$_6$]$_7$ **It** seems generally **true**$_8$ [that a loss of psychological context results in an impoverishment in the content of a memory.]$_8$ [PHIL-TH]

In contrast, we can look at this same passage with the negative features highlighted in order to illustrate the less frequent use of the negative features of Dimension 1: nouns are **bolded**, prepositions are underlined, passive structures are double-underlined, and passive postnominal modifiers are additionally in SMALL CAPS.

7.5 *Theoretical Philosophy (Roache 2006):*
Wollheim's point that in order to have q-**memories** of everything that his **father** experienced on his **childhood walks**, he must inherit not only his **father's memories** of these **walks**, but also their psychological **context**? Well, it may prove a **problem** for **Wollheim's claim** that it is impossible to isolate **memories** from their psychological **context**, but it does not undermine our weaker **claim** that **memories** are weakened when they are isolated in this **way**. **Martin** does not tell us how much of the Spanish **conversations** he now remembers, but it seems plausible to suppose that the **fact** that he no longer understands **Spanish** will have resulted in his now not being able to recall some of the **details** of the **conversations**. It seems generally true that a **loss** of psychological **context** results in an **impoverishment** in the **content** of a **memory**. [PHIL-TH]

Biology and philosophy illustrate the two extremes of Dimension 1, but Figure 7.2 shows less variation for the remaining disciplines and registers. While quantitative physics is also highly negative, the remaining disciplines (applied linguistics, political science, history, and theoretical physics) all have mean dimension scores that fall within an approximately 6 point range near 0. This clustering of disciplines and registers around 0 on the Dimension 1 scale shows that these disciplines and registers rely on the elaborating features and the information density features in more balanced ways than philosophy and biology. We can see this more balanced use of elaborated and densification features in the following excerpts. Here, for the sake of clarity, all positive features along Dimension 1 are **bolded**, while all negative features are underlined. In excerpt 7.6, we can see the overall frequent use of nouns and prepositions that help maintain a nominal style of writing, while we can also see cognition nouns (e.g., *theories*, *concepts*), *be* verbs, predicative adjectives (e.g., *are useful, is crucial*), modal verbs, general adverbs, adverbial conjuncts (e.g., *however*) and stance to-clauses (e.g., *crucial to identify X*).

7.6 *Qualitative Applied Linguistics (Frazier 2007):*
A large number of studies investigate the talk of students in writing classrooms; most of these treat the act of writing as social in nature. **Theories** of social actions and learning/socialization such as Lev Vygotsky's (1978) are **useful** in helping teachers create practical situations in which writing students **can** learn in social situations. To understand **how this learning happens**, it **is crucial to identify the interactional details of group work discourse**. Most of the sources that investigate writing students' talk (some of which are covered below), **however**, tend to focus on **purely** theoretical **concepts**, the social power structures inherent in tutor/peer relationships, or a priori analyst-imposed categories of group work talk.

In theoretical physics, we see the same dense use of nouns and prepositions, past tense verbs (e.g., *used, obtained*). However, we also see positive features such as adverbial conjuncts (e.g., *however, in addition*), personal and demonstrative pronouns, *be* as main verb, and several *that*-complement clauses. In this excerpt, previous research is being discussed and connected to the issues relevant for the study to be reported on.

7.7 *Theoretical Physics (Thomas, Christakis, & Jorgensen 2006):*
However, Menger and D'Angelo used 13C NMR to observe the conformational equilibria of undecane-2,5-di-13C in solvents ranging from chloroform to aqueous ethanol; by measuring 3JCC, **they** obtained the fraction of trans and gauche for the C–C bond. The result **was** 76% trans in all solvents. **In addition**, small-angle neutron scattering (SANS) studies by Dettenmaier and GoodsaidZalduondo and Engelman and Raman studies by Fischer **agree** [**that** the influence of intermolecular interactions on individual monomer conformation in the liquid state] is **negligible** and [**that** the conformations of n-alkanes in the condensed phase **are similar** to **those** populated in isolation.]

As mentioned above, the distribution of the registers along Dimension 1 show a primarily dichotomous relationship. This dichotomy is further supported by the post-hoc comparisons, listed in Appendix F. These comparisons show that philosopy and quantitative biology are significantly different from nearly every other register (with one exception: biology and quantitative physics are not significantly different). The remaining post-hoc comparisons reveal only a few other significant differences (e.g., quantitative physics is significantly different from quantitative political science and qualitative applied linguistics).

However, the placement of two registers is somewhat surprising: theoretical physics and qualitative history. Looking past the dichotonomy between theoretical philosophy and the other disciplines, at first glance it appears that there is also a distinction between the natural sciences and the remaining non-science disciplines, with quantitative biology and quantitative physics having the largest negative scores along Dimension 1. However, we then notice that theoretical physics has a mean dimension score near zero, while three non-science registers have more highly negative scores. Particularly surprising here is qualitative history, with the lowest negative score of the non-science disciplines (-6.2). I'll return to history in a moment, but first let's take a look at theoretical physics.

Excerpt 7.7 illustrated how the positive and negative features were used in combination during a discussion of previous research and its relation to the upcoming study. However, like theoretical philosophy, theoretical physics relies on the logical progressions of evidence in order to support claims and present

research, and Excerpt 7.8 employs positive features (all **bolded**) such as 1st person pronouns, modal verbs, time adverbs, and demonstrative pronouns in order to involve and guide the reader through the steps in the analyses. However, the use of nouns, prepositions, and passive verbs are also still prevalent in the excerpt.

7.8 *Theoretical Physics (Cacciatori et al. 2008):*
The second constraint, in particular Eq. (3), **can** be used to restrict the search for the [formula] functions [formula] to **that** of a single one, for which **we** choose [formual] with [formula]. **We** work out the details of this reduction. In particular, **we** give the constraints which the function [formual] **should** satisfy and given this function **we** define functions [formula], for all even characteristics [symbol], which satisfy the constraints from Section 2.3. Let [symbol] **be** the subgroup of Sp(2g, Z) which fixes the characteristic [formula]:

[formula]

For [symbol] **we** required [that [formula]], that is, [formula] a modular form on [symbol] of weight 8. Given such a modular form [formula] **we now** define, for each even characteristic a function [formula] in such a way that Eq. (3) holds.

While it was a bit unexpected to see the use of positive features along Dimension 1 in physics, it was also surprising to see the frequent use of negative features in qualitative history. In the following excerpt, negative features along Dimension 1 are marked: nouns are **bolded**, prepositions are underlined, past tense verbs are *italicized,* passive structures are double-underlined, and passive postnominal modifiers are additionally in SMALL CAPS.

7.9 *Qualitative History (Lopes Don 2006):*
By the early 1540s, a **consensus** *developed* in **councils** of the Spanish **government** that the **use** of the **Inquisition** to induce religious **orthodoxy** among the new **converts** *was* inappropriate and possibly dangerous for the **security** of the **colony**. The **Indian Inquisition** *ended* when the **Council** of the **Indies** *revoked* the **bishop's** inquisitional **powers** in 1543. In 1571, when **Philip** II formally *established* a Holy **Office** in **New Spain**, he specifically *prohibited* **trials** against indigenous **colonists**. Most **historians** have attributed this **decision** to the **failure** of the earlier **Indian Inquisition**.

This excerpt is typical of qualitative history reports, showing a dense use of nouns and prepositions, many of which function as noun modifiers and result in compressed noun phrase structures. However, unlike the natural sciences, passive constructions are not notably frequent (also documented in Chapter 5). Rather, the use of the past tense is extremely frequent as past events are described and reported. Combining with this frequent use of nouns, prepositions, and past tense

verbs is the relative lack of many positive features along Dimension 1, such as modal verbs and stance-conveying constructions – linguistic features not used frequently in the reporting of the events and states of the past.

In sum, Dimension 1 illustrates (A) a dichotomous division between philosophy and other registers, (B) a hint of a pattern in which empirical, natural science research relies on features related to informational density to a greater extent than social science research, and (C) that individual registers and disciplines exhibit somewhat idiosyncratic patterns of use based on situational characteristics unique to that disipline or discipline/register combination. In the following section, however, we see a different type of pattern of variation.

7.4.2 Dimension 2: Contextualized narration vs. procedural discourse

A total of 28 features have factor loadings greater than or equal to .30 on Dimension 2, with 22 positive features and 6 negative features. Positive features include word classes that refer to human constituents (3rd person pronouns, group nouns, animate nouns), adjectives indicating time and topic, tense and aspect markers (past tense, perfect and progressive aspect verbs), *that*-relative clauses, and communication verbs (including *that*-clauses controlled by non-factive/communication verbs). In addition, phrasal and clausal coordinating conjunctions loaded with the positive features on Dimension 2. Negative features on Dimension 2 include three specific types of nouns (technical nouns, quantity nouns, and concrete nouns), attributive adjectives indicating size, and agentless and by-phrase passive voice verbs.

Figure 7.3 plots the mean dimension scores by register along Dimension 2. In contrast to Dimension 1, where the nature of the discipline reflected the two poles of the variation, a different organization of registers emerges for Dimension 2. The three qualitative registers (history, political science, and applied linguistics) have the highest positive dimension scores, followed by theoretical philosophy. The hard sciences, quantitative biology and quantitative and theoretical physics, have highly negative dimension scores, while the quantitative registers in the social sciences (political science and applied linguistics) fall in the middle of this dimension. As I will explore in the rest of this section, it appears that Dimension 2 is highly correlated with the primary way in which the disciplines and registers present evidence. That is, the three qualitative registers and theoretical philosophy all rely upon extensive prose to present evidence, and this prose often provides a narration of what happened during a research study with participants (i.e., qualitative applied linguistics) or during a particular time period or political situation (i.e., history and political science).

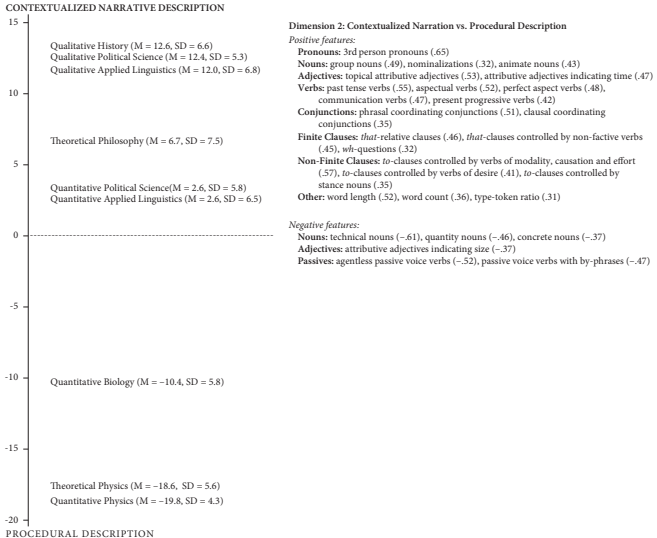


Figure 7.3. Distribution of disciplines and registers along Dimension 2: Contextualized narration versus procedural description. One-way ANOVA results: $F = 138.99$, $p = .000$, $r^2 = .81$. Post-hoc comparisons are listed in Appendix F

In fact, many of the positive features have been linked in the past to narrative discourse (e.g., see Biber 1988: Dimension 2; Biber 2006: Dimension 3), particularly 3rd person pronouns, human/animate nouns, past tense and perfect aspect verbs, and communication verbs. This interpretation also fits with the way in which these features are used in the three qualitative sub-corpora. In the following excerpts, 3rd person pronouns, group nouns, and animate nouns are underlined, conjunctions are *italicized*. Past tense, perfect aspect, and progressive verbs are **bolded,** relative clauses with *that* are in [square brackets], *that*-clauses controlled by non-factive/communication verbs and non-finite *to*-clauses controlled by verbs of modality/ causation/effort and desire are underlined:

7.10 The Charkaoui v. Canada case involves two permanent residents *and* one refugee, all of Arab origin. All **were detained** indefinitely pending a deportation$_1$ [that may never happen,]$_1$ on the authority of ministerial "security certificates" based on secret evidence$_2$ [that **led** government to believe that they might be dangerous.]$_2$ None of them **were even suspected** of having committed a crime in Canada *or* elsewhere. [POLISCI-QL]

7.11 At the beginning of this interaction, Student 1 **made** a mistake with the adjective ending to groB, a grammatical feature$_1$ [that **had not yet been covered** in the class.]$_1$ Furthermore, Student 1 **did not use** the preferred German adjective for long. Florian **provided** an implicit lexical recast, using the preferred German adjective for this meaning, *and* simultaneously **supplied** the correct adjective ending. He later **noted**, however, that this **was** not an intentional corrective recast. [AL-QL]

In these two excerpts, the focus is on describing, in detail, an account of events, and these events serve as the evidence on which an analysis is built. In the qualitative research paradigm, the purpose is to describe the natural course of events or actions, and then build interpretations upon those observations. Thus, there is a focus on establishing a narrative that sets up a reconstructed event to serve as evidence for the writers' claims and interpretations. Past tense, perfect aspect, aspectual verbs, and time adjectives are particularly common in establishing the sequence and timing of events and happenings, while communication verbs are used to report speech from the often human actors that are represented by 3rd person pronouns, group nouns, and animate nouns.

The negative pole of this dimension, on the other hand, is characterized by the use of technical, quantity, and concrete nouns, as well as passive voice verbs. Theoretical and quantitative physics have the largest negative dimension scores for Dimension 3, and excerpts 7.12 and 7.13 show how these features are employed

in theoretical physics. Technical, quantity, and concrete nouns are **bolded,** passive main verbs are underlined, and attributive size adjectives are *italicized*:

7.12 Therefore, Sz is a conserved **quantity**. As a result, in terms of the standard basis vectors [symbols] and [symbol], the reduced two-spin density matrix, which is constructed by tracing out the rest spin degrees of freedom, can be written as [formula]. In terms of the spin correlation functions, the elements of i,i+1 in **equation** (2) are given by [formula]. Obviously, by its definition, [formula] is a semi-positive definite matrix. Let [formula] be its eigenvalues. We define the two- site local entanglement Ev of the system to be the 4 von Neumann entropy of [formula], i.e., we have [formula]. As is well known, on a d -dimensional simple cubic lattice, the phase **diagram** of the antiferromagnetic XXZ model is divided into three parts by two phase transition points 1 = -1 and 2 = 1. [PHYS-TH]

7.13 In writing [*equation*], the second term, F is of *lower* order in powers of n compared to the first one, since D is of order n. (Recall that D 2 = -Cn is of order n.) Thus N is the leading term we need at *large* n, the corrections to (57) are subdominant. They will be needed for *higher* dimensional spaces, as we shall see later. [PHYS-TH]

Excerpt 7.12 illustrates one of the key characteristics of theoretical physics with respect to Dimension 2: the presentation of evidence using mathematical formulas. This focus on mathematics clearly relates to the the nouns and adjectives that seek to quantify or describe the size of objects (which also load on this dimension). In addition, the passive voice is used frequently in the prose introducing and explaining the formulas, and functions to establish the procedural steps in the analysis related to the computations.

Quantitative physics has a Dimension 2 score as low as theoretical physics, and while some evidence is presented as mathematical formulas, this use is not as prevalent in quantitative physics. Excerpt 7.14 illustrates that in addition to passive voice being used to establish steps in the analysis or procedural methodology, the nature of the discipline of physics is also explanatory, as the discipline is focused on presenting quantitative displays of evidence in figures, tables, and some formulas.

7.14 Four **data samples** each are selected from the core and segment **data** sets. The **data samples** are defined by the **energy** measured in the core and are labeled: DEP: The **sample** contains events with a core **energy** in the region of (1593 ± 5) keV. These events are associated with the double escape **peak** of the 2615 keV 208 Tl photon. The photon produces electronpositron pairs of which the positron subsequently annihilates. [PHYS-QT]

Quantitative biology also has a negative dimension score on Dimension 2, yet there is an almost 10 point difference between the physics registers. In biology, the primary focus within the context of this dimension is in the description of the methodological steps carried out by the researchers. Biology typically relies on quantitative displays of evidence, but does not use explications of mathematical formula in its arguments, one area in which physics relied on these negative features:

7.15 After selecting relevant variables using a stepwise procedure, a partial regression analysis was carried out only with those variables that were consistently selected by most richness and diversity indices (Table 2). The interpretation of NMDS and PCA axes used in this analysis is presented in Table 1, and details of their loadings can be found in Appendices B and C. [BIO-QT]

Quantitative political science and applied linguistics have low positive dimension scores, and an examination of research articles in these two categories reveals that the quantitative social sciences rely on both the positive and negative features of Dimension 2. This may be the result of two competing factors. First, there could be some influence of discipline, since the qualitative registers in these two disciplines have the highest dimension scores on Dimension 2. Second, as quantitative research, these two registers also share characteristics related to research paradigm with the quantitative sciences. That is, these quantitative political science and applied linguistics still have an underlying focus on human subjects or events (which inherently involve people), and they exhibit tendencies to provide elaborated introductions and literature reviews in which they establish the scene for research (often through narration of a situation) and place themselves in the existing body of literature. This is illustrated in excerpt 7.16, where 3rd person pronouns, group nouns, and animate nouns are underlined, past tense, perfect aspect, and progressive verbs are **bolded.** Relative clauses with *that* are in [square brackets], *that*-clauses controlled by non-factive/communication verbs and non-finite *to*-clauses controlled by verbs of modality/ causation/effort and desire are underlined. Passive voice main verbs are **underlined and bolded**:

7.16 *Quantitative Political Science (Ramírez 2007):*
The transformations of the composition of the workforce, of the student population, and of the electorate are among the most significant changes$_1$ [that **have accompanied** the population change of the past 20 years.]$_1$ It is particularly important to recognize the role$_2$ [that Latinos and specifically Latino immigrants play in these transformations.]$_2$ Both scholars and policy makers recognize the significant role$_3$ [that Latinos **played** in the transformation of the workforce and the education]$_3$ and **have considered** the consequences of this change (Fullerton, 1997; Fullerton & Toossi, 2001; Passel & Suro, 2003; Vernez, 1998, 1999; Vernez, Krop, & Rydell, 1999). Less **has**

been done to understand the contemporary and future ramification of the influx of both native-born and naturalized Latinos in the electorate or their respective patterns of participation. In 1980, there **were** 2.5 million Latinos voters in the United States. By 2000, this figure **had more than doubled** to 5.5 million voters. Although Latinos **composed** less than 6% of the electorate in 2000,1 the growth rate of that group **has been** impressive.

However, quantitative political science and applied linguistics also carry the methodological values of quantitative research to exercise control or research contexts and to produce concise, informative descriptions of the procedures followed to collect and analyze data. There are many instances of passive voice verbs (negative Dimension 2 features) to report research procedures, as well as past tense and perfect aspect verbs (positive Dimension 2 features) to report the results of studies:

7.17 *Quantitative Applied Linguistics (Siyanova & Schmitt 2008):*
The 31 frequent and 31 infrequent collocations **were combined**, in random order, and **attached** to the collocation instrument. Participants **were asked** to rate all 62 collocations on the basis of their commonness in the English language. Although we **were** interested in judgements about the acceptability of the collocations, collocations$_1$ [that **are used** frequently by natives]$_1$ are clearly acceptable, while collocations$_2$ [that do not occur in 100 million words]$_2$ are much less likely to be so. We **felt** that a judgement task relating to frequency would be more transparent to our participants than a task asking them to rate acceptability. Therefore, the instructions **required** the participants to rate the collocations according to frequency on a six-point scale.

In the next section, we see yet another pattern with respect to the organization of disciplines and registers along a dimension of variation. Dimension 3 cuts across disciplines which have at their foundation an inquiry into the workings of the human mind and disciplines with other areas of inquiry.

7.4.3 Dimension 3: Human vs. non-human focus

Dimension 3 consists of 11 positive features and 4 negative features. The positive features include 2nd and 3rd person pronouns, process nouns, mental, activity, and communication verbs, and progressive aspect. In addition to general wh-clauses, several stance features also characterize the positive end of this dimension: *that*-clauses controlled by factive verbs, and *to*-clauses controlled by verbs of desire and speech verbs. There are only four negative features on this dimension: all attributive adjectives, attributive adjectives indicating topic, general adverbs, and prepositions.

Figure 7.4 shows the distribution of the disciplines and registers along Dimension 3. Dimension 3 appears indicate a dichotomy between two types of disciplines: (1) disciplines which have human beings and their mental/cognitive activities at the

heart of their subjet domain (applied linguistics and philosophy), and (2) disciplines whose topic domain is not focused on the cognitive activities of human beings. It should be noted that because there are few variables on the negative set of features for Dimension 3, interpretations of this part of the dimension is limited and preliminary.

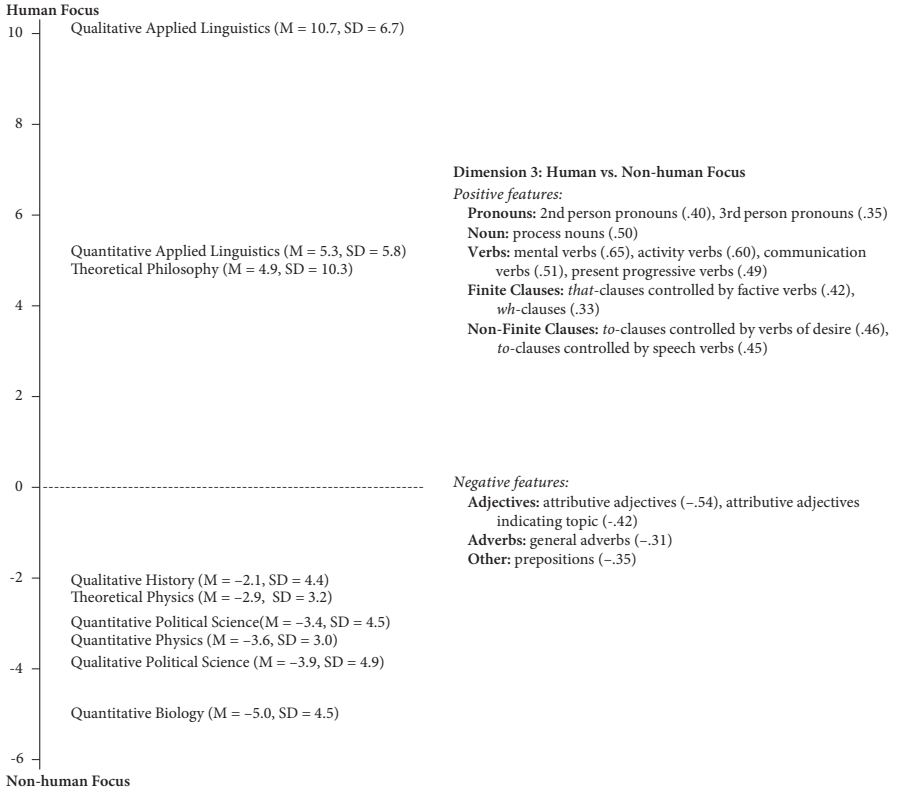


Figure 7.4. Distribution of disciplines and registers along Dimension 3: Human versus non-human focus. One-way ANOVA results: F = 28.42, p = .000, r^2 = .47. Post-hoc comparisons are listed in Appendix F

As the situational analysis in Chapter 4 summarized, the discipline of applied linguistics is concerned with humans and language, with the ways in which we use language, how we learn and acquire language, and the ways of teaching langauge to promote language acquisition (to name a few). Philosophy, on the other hand, is not focused on language specifically, but on the broader human mind and how we can use logic to understand human nature, human problems, and human cognition.

Reflecting the nature of the object of inquiry in these two disciplines, Dimension 3 shows parallels between the use of a variety of verbal structures that quite

often take human agents as the subjects of the verbs. These verbal structures include semantic categories of verbs that we use to describe mental processes, activities, and communicative acts. They also include stance clauses, and the human agents can be explicitly used as the subject (verbs marked 2, 4, 5, 6), or implied based context (verbs marked 1, 3, 7):

7.18 Teachers built background knowledge by **asking**[1 – implied 'teachers'] learners what they **thought**[2 – explicit 'learners'] the site would be about [AL-QL]

7.19 Indeed, immunology **is often decribed**[3 – implied 'people'] as the science of discrimination between self and non-self [PHIL-TH]

7.20 We **discuss**[4 – explicit authors as 'we'] these patterns in turn below [AL-QL]

7.21 Egalitarians can be pluralists about value. They **think**[5 – explicit egalitarians as 'they'] that there is a reason to level down–where doing so will make a distribution fairer–but **recognize**[6 – explicit egalitarians as 'they'] that there are also reasons not to… The (non-person-affecting) reason to level down is unlikely to outweigh the reasons there are to **prefer**[7 – implied 'we' as humankind] people [to be better rather than worse off.] [PHIL-TH]

In addition, the examples illustrate the variety of roles that these explicit and implied human agents can have. For example, mental, activity, and communication verbs, as well as *to*-clauses controlled by desire and speech verbs, are used to portray the thoughts and ideas of participants in the research (excerpts 7.22–7.24). In philosophy, this is often in the form of unreal characters (as in excerpt 7.25), or indefinite pronouns (excerpt 7.26) to refer generally to human beings that are used in the article to explore and illustrate logical processes.

7.22 Regarding the kind of difficulties they **experienced**, some **commented** on the inferential reading questions and writing conventions. Many L students **mentioned** vocabulary as one of the key difficulties in **taking** the test. [AL-QL]

7.23 After **reading** these instructions, each participant completed a series of practice sentences (see Appendix D) and **made** practice grammaticality judgements for these sentences. [AL-QT]

7.24 Henry **discussed** how good writers from another elite university in the same city mediated his writing although he **knew** none of them [AL-QL]

7.25 Mary **knows** all there is to **know** about physics, chemistry and neurophysiology, yet has never **experienced** colour. Most philosophers **think** that if Mary **learns** something genuinely new upon seeing colour for the first time, then physicalism is false. [PHIL-TH]

7.26 Memory is ordinarily taken to be factive. One cannot **remember** that which did not happen. [PHIL-TH]

While these examples have subjects referring to the entities being studied, these same verbs and structures are also often used to make connections to previously established theories and findings, as in excerpts 7.27 and 7.28. This second use is not unique to philosophy and applied linguistics, although a preliminary analysis suggests that this use is highly prevalent in these disciplines as the writers provide extensive theoretical grounding for both introducing studies and concepts, as well as interpreting results.

7.27 Strawson **has argued** that our ordinary conception of moral responsibility requires a kind of ultimate self-creation that is incoherent. Strawson **gives** various different formulations of the argument, but I **find** the versions presented in his article, "The Bounds of Freedom," particularly lucid. [PHIL-TH]

7.28 Goss, Ying-Hua and Lantolf (1994), who **compared** grammatical judgement tasks **completed** individually and in pairs by learners of Spanish, **found** modest differences in favour of pairs and only on some grammatical features. [AL-QT]

Finally, these same structures are used with the authors/researchers as the grammatical subjects of the verbs. Again, this use is not limited to philosophy and applied linguistics by any means. However, the situational analysis presented in Chapter 4 indicated that these two disciplines are among those that most explicitly and extensively state the purpose of the research, as well as discuss the nature of data and procedures (particularly applied linguistics).

7.29 In order to **see** whether applying the cognitive typology can be of assistance in resolving some problems in a particular context of translation, I **analysed** original transcripts of police interviews with Spanish-speaking witnesses and suspects, with translation into English **provided** by certified court interpreters. [AL-QL]

7.30 I **argue** for an alternative justification for conservation in the capacity of foresight, which **requires** us [to act not only upon duties that we have now, but also upon those that we will predictably have in the future.] [PHIL-TH]

7.31 We **noticed** two aspects that were not **addressed** in the literature. First, none of the prereading methods **studied explored** the possibilities of content-area materials available from authentic texts within the discipline. While EAP reading at the higher level may be more of a reading problem than a language problem, we **believe** it is worthwhile to **explore** the utility of content materials for EAP reading intervention. [AL-QT]

As noted in the discussion above, not all of these uses of mental, activity, and communication verbs are limited to use in philosophy and applied linguistics. In addition, we could also argue that history and political science, in studying historical

events and social organization are also inherently human-based disciplines. We might assume that because the events, processes and situations that are described in history and political science are carried out by human beings, these disciplines would look more similar to applied linguistics and philosophy than is demonstrated by Figure 7.5. Yet, history and political science are generally *not* significantly different (see Appendix F for post-hoc analyses) from the hard sciences along this Dimension. Therefore, it seems that there is a fundamental difference between the two disciplines that explore mental processes or phenomena that are connected to the cognitive abilities of humans and other disciplines. History and political science fall in with physics and biology along this dimension because object of study is typically not on humans and their cognitive processes or communicative roles, but rather on the *events* and *situations* that make up human history and society.

To illustrate, we can look at the following excerpts, one from a history article and one from a quantitative political science article. These two excerpts illustrate that although the phenomena under investigation are events and trends carried out by human beings, the focus of these analyses is not on understanding the cognitive characteristics of human beings, but rather practices, actions/events, and so on:

7.32 *Qualitative History (Fette 2007):*
Women's battles and breakthroughs in the liberal professions during the Third Republic–their struggle for entry in the late nineteenth century, their growth and progress, and the resistance they encountered anew in the 1930s–are one chapter in a bigger story of professional exclusion. Women were in fact only one unwanted social category among many in the French professions; foreigners, naturalized citizens, and the lower social classes also served as scapegoats for supposed overcrowding and loss of tradition. Of course not all women in pursuit of medical or legal careers were bourgeois and French, and thus faced multiple prejudices.

7.33 *Quantitative Political Science (Pacek & Radcliff 2008):*
Critically, decommodification reflects the quality as well as quantity of social rights and entitlements; the mere presence of social assistance or insurance may not necessarily bring about significant decommodification if they do not substantially emancipate citizens from market dependence. Citizens are "emancipated" from the market in the sense that they can freely opt out of work, when necessary, without risking their jobs, incomes, or general welfare. [POLISCI-QT]

In sum, Dimension 3 distinguishes between disciplines which have humans, their thoughts, knowledge, and communication practices at the heart of their evidenciary practices, and disciplines which focus on either historical events and

conditions (history and political science) or the non-human natural world (biology and physics).

7.4.4 Dimension 4: 'Academese'

Dimension 4 is characterized by the use of the fewest features, with 8 positive features and only 1 negative feature. The positive features include three sets of abstract nouns: process nouns, other abstract nouns, and nominalizations. Along with these three types of abstract nouns, existence verbs and adjectival structures (relational adjectives, *that*-clauses controlled by likelihood adjectives, and *to*-clauses controlled by all stance adjectives) also load positively on Dimension 4. Because there are so few features on this dimension, the interpretations offered here will be brief and are preliminary in nature. In fact, it is even more important to characterize the functional interpretation of this dimension while also considering the distribution of registers along the dimension. Figure 7.5 displays the mean scores for each discipline and register for Dimension 3.

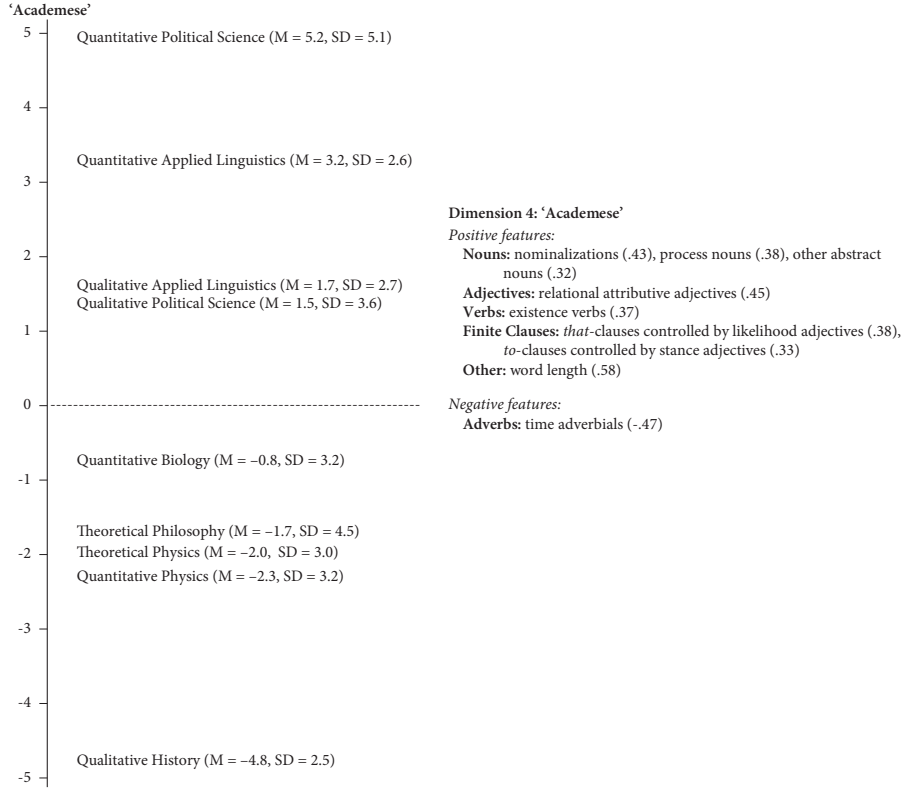


Figure 7.5. Distribution of disciplines and registers along Dimension 4: 'Academese'. One-way ANOVA results: F = 24.33, p = .000, r^2 = .43. Post-hoc comparisons are listed in Appendix F

Along this dimension, only four registers have positive scores: quantitative political science and applied linguistics have the highest positive dimension scores, followed by qualitative political science and applied linguistics. Qualitative history has the highest negative dimension score, followed by quantitative and theoretical physics and theoretical philosophy. When we compare the situational characteristics of these two groups of disciplines (those with positive scores and those with negative scores), we see that the major situational difference between the two groups is the extent to which the disciplines are concerned with overtly representing themselves as empirical research through the inclusion of situational features like explicitly stated research designs, research questions, integrated citations, and labeling data and research processes.

To take a step back, the two theoretical disciplines (philosophy and physics) are clearly not empirical research paradigms, as evidenced by the very definition of 'theoretical' texts. However, during the interview processes described in Chapter 2, it came out that academics within the disciplines of history and quantitative physics, although technically empirical in nature (as they observe data which they use to draw conclusions from), would not necessarily overtly use the term 'empirical' to describe research practices in the disciplines (T. Porter, personal communication, May 6, 2009; G. Lubick, personal communication, January 26, 2010).

Thus, since the negative pole of this dimension is populated by the two theoretical registers, along with qualitative history and quantitative physics, it is possible that this dimension is distinguishing between disciplines and registers which are explicitly promoted as 'empirical' in nature and those which are less concerned with being labeled empirical (regardless of whether or not they are empirical in nature, as is the case for qualitative history and quantitative physics). It appears that registers which are expressly concerned with packaging themselves as 'scientific' inquiry (such as the social sciences) employ features such as absract nouns, process nouns, and nominalizations, mimicing the process of grammatical metaphor that Halliday's work has documented so extensively in scientific discourse.

If this is the case, it is interesting to note that the two quantitative social sciences (applied linguistics and political science) have the highest positive scores on this dimension, followed by the two qualitative registers in these same disciplines. When we look at the features that characterize the positive end of this dimension of variation, we see several adjectival structures, such as relational adjectives (e.g., *basic, common, different, general, individual, main, particular, same, similar, various*), existence verbs (e.g., *appear, contain, defined, exist, illustrate, include, indicate, reflect, represent, stay, tend, vary*). In addition, *to* and *that*-clauses headed by certain stance adjectives are also important for this dimension, as well as nominalizations and process and other abstract nouns. These patterns

are illustrated in the two excerpts below (positive features bolded), and we can see that these existence verbs and adjectival structures are often used to describe and interpret the results of the data analysis, a discourse function highly associated with empirical research.

7.30 All round then, given the **various** terms and working **definitions** used in **studies** of recasts, it may be **unhelpful** to lump together all the **different** types under the single label of 'recasts' and also to assume that all recasts are categorically implicit. [AL-QT]

7.31 On the issue of **same** sex **marriage**, uninformed women were approximately 20 per cent more **likely** to support **same** sex **marriage** than were men, and **information** had the effect of widening that gap, bringing the probability of **support** up to nearly 29 per cent **higher** than men. **Information** had a **similar** (and **larger**) **effect** on the **support** for easier access to **abortion**… the informed nonreligious are 28 per cent more **likely** to support same sex marriage than are the informed religious and 34 per cent more **likely** to support abortion. [POLISCI-QT]

Thus, although this analysis is preliminary, it seems that dimension 4 distinguishes between disciplines and registers that are overtly empirical in nature, and the linguistic features on this dimension help researchers characterize and make sense of results in order to offer interpretations of that data.

7.5 Conclusions

The new multi-dimensional analysis carried out in this chapter has revealed that linguistic variation in published academic research articles varies according to multiple parameters. Discipline is only one of the important characteristics of a research article that corresponds to linguistic variation. Rather, variation occurs along multiple parameters that have generally been unrecognized or disregarded in previous studies of disciplinary variation. Parameters such as the nature of evidence, the presence or absence of data, research methods (qualitative vs. quantitative vs. theoretical research), and the object of study. That is, while previous research has largely disregarded the influence of the nature of research in studies of linguistic variation in research articles, this study has highlighted that such distinctions do matter. Each of these characteristics corresponds to specific linguistic patterns, and linguistic resources work together to construct meaningful discourse that is reflective of the nature of the discipline as well as the research paradigm within which the scholarly work fits.

In addition, the distinct ways in which the registers and disciplines distributed along the four dimensions highlights the complex nature of academic writing.

Writers utilize a multitude of linguistic resources that correspond to an equally complex set of situational, or non-linguistic, characteristics. For example, articles in theoretical physics both follow linguistic traditions of the natural sciences (procedural discourse and a non-human focus in Dimensions 2 and 3), while also creating involved forms of argument in which the reader is engaged with the procedural steps that are presented as evidence of the author's findings, as illustrated in Dimension 1. Likewise, quantitative applied linguistics articles utilize both moderately dense expression of information like scientific writing more generally (Dimension 1), a certain degree of narrative orientation (Dimension 2), along with a major focus on human actors (Dimension 3).

Somewhat surprisingly, it turns out that one of those dimensions of variation continues to highlight the differing degree to which texts balance between the processes of elaboration/involvement and informational density. This finding comes despite the much narrower domain of inquiry into sub-registers that all belong to a larger, informational, written register. Despite a few differences in the specific features that loaded on Dimension 1 in the present analysis, the degree of correspondence also supports the idea that this dimension of variation is a universal construct in describing register variation in English. More specifically, this new dimension reflects those grammatical complexity features which are most important for professional academic writing specifically.

The MD approach is unique because the underlying patterns of variation are derived *inductively* and *quantitatively* from the analysis of the corpus, and then interpreted functionally to explain the observed patterns. In addition, the dimensions of variation that emerge from the analysis are typically considered previously unrecognized constructs (see Biber 2010). In this approach, no organization of the texts is overtly placed on the corpus during the statistical process of identifying co-occurrence patterns. Such organization of texts into sub-corpora to calculate per-register/discipline dimension scores is carried out only after the co-occurrence patterns have been established based on the inductive analysis of the corpus.

In the next chapter, I consider the results of the three analyses presented in Chapters 5, 6, and 7, with the goal of synthesizing the results into statements about what we know about these disciplines, and linking these findings more explicitly to the situational characteristics of the disciplines and registers that were described in Chapter 4.

CHAPTER 8

A Synthesis

What do we know?

8.1 Introduction

The linguistic analyses presented in the last three chapters have uncovered complex patterns of variation. On one hand, some linguistic features have been shown to vary along disciplinary lines, and more generally along the parameters of traditional discipline groupings (such as the humanities, social sciences, and hard sciences). On the other hand, these studies have also highlighted variation that occurs irrespective of discipline, instead seeming to follow along parameters related to other situational characteristics such as the purpose of the research, the nature of the evidence used in the research, and so on.

While I have made some connections between the linguistic findings in Chapters 5 through 7 and the non-linguistic characteristics of texts described in Chapter 4, in this chapter I will summarize several of the central patterns of variation that appear to be associated with specific situational characteristics of the texts. In other words, the first goal of this chapter (Section 8.2) is to synthesize the linguistic (Chapters 5–7) and non-linguistic analyses (Chapters 2, 4) undertaken in the book in order to concisely address the primary goal of the study: *to investigate linguistic variation across disciplinary and register boundaries while acknowledging the varied nature of research articles within and across disciplines*.

In Section 8.3, I focus on the relationships between the three analyses that I carried out in the study, and also propose ways in which additional corpus analytical techniques could be utilized to provide further comprehensive descriptions of language variation in disciplinary registers. Finally, I conclude the book with a discussion of the results of this research in terms of the methodological implications for future corpus studies of academic writing.

8.2 Summing up: Linguistic variation in the Academic Journal Register Corpus

As mentioned above, the three linguistic analyses reported on in this book have revealed variation that occurs in complex ways and reflects a variety of factors. In this section, I am concerned with three major patterns of variation. First, I focus on variation that occurs along disciplinary lines. This variation constitutes two general types: (1) variation that occurs along individual disciplinary lines (e.g., philosophy versus applied linguistics), and (2) variation that follows the parameters of traditional discipline groupings (such as the humanities, social sciences, and hard sciences). This second type of variation often exhibits a directional cline of variation in which 'soft' disciplines such as philosophy rely on features to a certain extent, the 'hard' or natural science disciplines rely on those same feature in a much more frequent (or much less frequent) manner, and the social science disciplines exhibit patterns of use somewhere between the 'soft' and 'hard' disciplines. In Section 8.2.1, I focus on these two types of disciplinary variation. In Section 8.2.2, I turn to linguistic variation that appears to cut across disciplinary lines, instead occurring alongside register differences in academic journal articles.

8.2.1 How does language use vary across discipline?

Throughout these studies, a complex picture of disciplinary variation has been uncovered, and most variation appears to be influenced by multiple factors. However, there have been some characteristics that are distinctive of a particular discipline. For example, Table 8.1 below summarizes the features that characterized a discipline. For disciplines represented by a single register, this means that the use of the feature was much more common in that discipline than in most others. In disciplines represented by two registers, this means that the use of that feature was distinctive for that discipline, and both registers representing the discipline showed a similar pattern of use.

Table 8.1 shows the frequency of each feature relative to other disciplines and is based on an examination of the mean rates of occurrence for each feature by sub-corpus. A single + symbol indicates that the feature occurred more frequently in that discipline than the average rate of occurrence across all disciplines. Two ++ symbols indicate that the feature was markedly more frequent in that discipline, often occurring with a frequency of one standard deviation above the mean. Many of these features were identified as characteristic of the discipline through the grammatical analyses in Chapters 5 and 6, and others came out as features which loaded on factors in the multi-dimensional analysis (Chapter 7) that highly

characterized the discipline. In fact, many of the findings from the three analyses complement and validate each other.

For example, the analysis in Chapter 5 revealed that past tense verbs were more frequent than present tense verbs in only one discipline: history. In the multi-dimensional analysis in Chapter 7, past tense was one of the most important features to load on Dimension 2 (with a factor loading of .55). As shown by Figure 7.3, history had the highest mean dimension score on Dimension 2, and the analysis of the text excerpts in history confirmed the prevalent use of the past tense as history writers present evidence regarding historical events and happenings and describe the historical contexts under investigation. Although this is a simplified example for the purposes of illustration (i.e., there are many other important factors along Dimension 2 that also lead history to have a high positive dimension score), the importance of past tense as a characterizing feature of qualitative history articles is confirmed by multiple analyses. In fact, overlap in the findings from the three analyses illustrates the confirmatory power of taking multiple analytical approaches to the description of language variation across registers.

The third column of Table 8.1 summarizes situational characteristics from the analysis in Chapter 4 that can be associated with the use of features in these particular disciplines. As can be seen by this juxtaposition of linguistic features and situational characteristics, many of the features that appear to be highly associated with particular disciplines are inherently connected to the subject matter of the discipline, and by extension, the nature of the object under study. Purpose also plays a role here. Philosophy is a good example of the influence of all three of these factors. As the subject matter of philosophy typically encompasses aspects of human cognition, and the purpose is to explore and interpret the current and ongoing or universal state of these cognitive phenomena, it is logical that philosophy would rely on mental verbs, features associated with a human focus (Dimension 3), and present tense verbs in order to discuss and argue about the nature of the human mind.

Table 8.1. Summary: Distinctive characteristics by discipline

Discipline	Characteristic Linguistic Feature	Corresponding Situational Characteristics
Philosophy	++ present tense ++ adverbial subordinators + human focus (Dimension 3) + mental verbs + to-clauses (verbs of desire)	subject matter of human cognitive states and processes, with focus on analyses of the present, continuous nature of those phenomena
History	+ past tense ++ third person pronouns	narrative purpose focused on human events and actions

(*Continued*)

Table 8.1. (Continued) Summary: Distinctive characteristics by discipline

Discipline	Characteristic Linguistic Feature	Corresponding Situational Characteristics
Political Science	+ attributive adjectives + attributive adjectives (topic) ++ prepositions	–
Applied Linguistics	++ process nouns + activity verbs + mental verbs + balanced use of present & past tense ++ human focus (Dimension 3) + *that*-clause (factive verbs) + to-clauses (verbs of desire)	subject matter of applied linguistics as processes associated with learning, teaching and using language, the basis of which are human participants
Biology	+ attributive adjectives + nouns (all) ++ concrete nouns	subject matter of inanimate objects, with goal of describing items in detail
Physics	+ concrete nouns + technical nouns + quantity nouns + present tense ++ prepositions + size adjectives + passives with by-phrase	subject matter/object of study as concrete, measurable aspects of the physical world; explicit, precise presentation of research procedures

Perhaps an even more prevalent trend that has come out of the analyses in this book is that variation in the use of these linguistic features often follows along a fairly linear cline of variation across academic disciplines as we move from soft disciplines such as philosophy to social science disciplines to hard, natural science disciplines. At the same time, this results in natural groupings of disciplines that use features in similar ways. What is noteworthy here is that these groupings often reflect the tradition of grouping disciplines into major categories like humanities, social sciences, and hard sciences. Table 8.2 is laid out similarly to Table 8.1, this time with the three discipline groupings as the organizing principle. For the purposes of this summary, humanities are represented by philosophy and history, social sciences are represented by applied linguistics and political science, and hard sciences by biology and physics.

Table 8.2 shows that many of the features that vary along this cline are those associated with structural complexity, with elaborating features occurring much more commonly in the humanities, less so in the social sciences, and much less frequently in the hard sciences. The compression features, on the other hand, show the opposite trend, occurring with much higher frequencies in the hard sciences than in the humanities, and with medium (comparative) frequency in the social

sciences. Thus, the features summarized in Table 8.2 primarily come from the studies of structural complexity and those features related to Dimension 1: Academic Involvement and Elaboration vs. Informational Density.

The differing 'levels' of use are represented by the following symbols to indicate relative reliance on those features based on visual examinations of the mean rates of occurrence for these features and of the figures provided in Chapters 5, 6, and 7: the symbol – indicates that the feature was less common than in other disciplines, the symbol + indicates that the feature is relatively common, and two ++ symbols indicate that the feature is much more common than in other disciplines.

In contrast to the features that varied along individual disciplinary divisions (where differences appeared to reflect subject matter and the object of study), Table 8.2 illustrates that the variation in the use of many of these features is likely related to different situational characteristics. For example, one of the key differences that has been revealed through the complexity study and the multi-dimensional analysis study (Dimension 1) is the relatively higher use of clausal features such as finite complement clauses (particularly verb complement clauses) and stance complement clauses in the humanities, followed by the social sciences, and showing an infrequent use in the hard sciences. This may reflect, in part, the use of in-text citations and the explicitness (and nature) of stated purposes. That is, these clausal features are often used to relate to previous research with in-text citations, position others' claims into the discourse so that they can be analyzed or interpreted, and introduce authors' purposes. In the humanities, these things are nearly always done, and in the social sciences these things are usually done. This trend contrasts with the hard sciences, which often have fewer, less explicit statements of purpose, and typically reference other research through endnotes or footnotes in which the cited authors' names do not appear in the prose of the article (and thus cannot lead to the use of reporting verbs and clauses).

Another possible connection between the cline of variation that occurs for compression features and the situational characteristics of the disciplines concerns the factor 'explicitness of research design'. That is, the hard sciences use compression features such as noun premodifiers, non-finite relative clauses, and passive voice verbs to a much greater extent than humanities. It also happens that these features are often used to describe and outline research procedures (as evidenced by the discussions in Chapters 5, 6, and 7). While the hard sciences always contain explicit descriptions of data and methods (with the exception of theoretical physics), this is almost never done in the humanities, and is variable within the social science disciplines and registers.

Thus, as this synthesis has shown, there is a complex interplay between various situational characteristics, and these interactions have clear links to the

patterns of linguistic variation that we see across disciplines. In the next section, I turn to variation across registers within these disciplines.

Table 8.2. Summary: Distinctive characteristics by discipline type

Discipline Grouping	Characteristic Linguistic Features	Corresponding Situational Characteristics
Humanities (*philosophy, history*)	– passive voice (agentless) + pronouns (*it,* demonstratives, nominal) ++ finite complement clauses (*philosophy) (++ finite verb complement clauses) ++ non-finite complement clauses ++ finite adverbials ++ finite relative clauses – non-finite relative clauses – nouns as nominal premodifiers ++ academic involvement and elaboration features (Dim 1, *philosophy) – procedural discourse features (Dim 2)	longer texts; purposes to use logic to explore issues (philosophy) and describe and interpret historical events or trends (history); typical purpose is to explore or discuss and then offer arguments
Social Sciences (*political science, applied linguistics*)	+ passive voice (agentless) + finite verb complement clauses + non-finite complement clauses + finite adverbials + finite relative clauses + finite relative clauses + nouns as nominal premodifiers + academic involvement and elaboration features (Dim 1) + informational density features (Dim 1) ++ overt empiricism features (Dim 3)	equal focus on establishing connections with previous research, describing methods including data and procedures; relatively equal use of references within the text and outside the text; purpose is focused on both what the research did and the implications of that research
Hard Sciences (*biology, physics*)	++ passive voice (agentless, by-phrase) – finite complement clauses (especially low in verb clauses) – non-finite complement clauses – finite adverbials – finite relative clauses ++ non-finite relative clauses ++ nouns as nominal premodifiers ++ informational density features(Dim 1) ++ procedural discourse features (Dim 2) – overt empiricism features (Dim 3)	shorter texts; rarely uses in-text citations; purposes less explicitly stated; focus on describing in detail methodological procedures; purpose is focused on (and stated in terms of) what the research did

8.2.2 How does language use vary across academic journal registers?

One key difference between this study and previous studies on disciplinary writing has been the inclusion of the type of research article, or register, as a component of the corpus design and thus, as a factor that can be systematically linked to patterns of language use. It turns out that the more specific journal register of journal

articles (i.e., different types of journal articles) can indeed be linked to specific patterns of linguistic variation. In this section, I continue with the approach of providing a summary of features which have shown to be characteristic, or distinctive, of a group of texts, this time focusing on texts grouped into the three registers included in this study: theoretical, qualitative, and quantitative research. This summary is presented in Table 8.3 below.

Table 8.3. Summary: Distinctive characteristics by academic journal register

Register	Characteristic Linguistic Features	Corresponding Situational Characteristics
Theoretical (*as compared to empirical articles*)	++ 1st person pronouns – overt empiricism (Dim 4) + modals of possibility, permission, ability + adverbs of time + adverbial conjuncts	evidence is a logical (ordered) progression of ideas/formulas; no observed data; purpose is to work through an issue/topic
Quantitative (*as compared to qualitative articles*)	+ passive voice – finite relative clauses + non-finite relative clauses + quantity nouns + predicative adjectives	typically follows IMRD organization with explicit statements of data and methods; frequent use of visual elements to display analysis; quantitative data
Qualitative (*as compared to quantitative articles*)	+++ contextualized narrative description features (Dim 2) (++ 3rd person pronouns) (+ group nouns) (+ perfect aspect) (+ aspectual verbs) + adverbs of time + stance adverbs + communication verbs ++ pronouns + finite complement clauses + finite relative clauses – non-finite relative clauses	purpose of qualitative research is to comprehensively describe a context and make interpretations and arguments based on those observations; less explicit (and extensive) statements of data and procedures

For this analysis, characteristic features were determined through two main processes. For the disciplines represented by two registers, a comparison was made (using mean frequencies of occurrence) to identify features that varied in use *within* the discipline. Most of the results for this analysis came from the multi-dimensional analysis, and the relevant dimensions are listed in Table 8.3, with features from those dimensions enclosed in (parentheses). I have only listed features associated with the overall dimension if the individual rates of occurrence for that feature also showed a correspondence specifically to the register of interest (and not other registers). I also compared the mean frequencies for

all variables for this analysis. In general, relative frequencies are again marked with + and –. However, in this analysis these symbols represent frequencies *relative to* contrasting register(s). That is, quantitative and qualitative research articles are compared against each other, and theoretical articles are compared against all empirical articles.

Despite the dramatic differences in the subject matter of the two disciplines representing theoretical articles, a few unifying features have emerged from this analysis. As explored in Chapter 5, first person pronouns are particularly distinctive of this register. In my analysis in Chapter 5, I commented that the use of first person pronouns helped the author involve the reader in the logical progression of evidence (albeit in different formats, with physics relying on mathematical formulas and simulations, and philosophy relying on unreal situations and vignettes). In fact, the other features listed above also contribute to this function in theoretical articles. That is, adverbial conjuncts, adverbs of time, and modals of possibility help the writer explicitly label logical relationships within (and across) clauses, thus overtly marking thought processes for the reader and helping readers to follow along with the logical progression of evidence.

Despite these similarities, however, theoretical physics and philosophy demonstrate that register is not the only factor that influences the use of linguistic features. Rather, there is a complex interplay between register and discipline, in which the same linguistic features that are associated with a particular register are in reality employed in distinct ways in different disciplines.

In theoretical philosophy, evidence is presented through extensive prose discussions, and the markers of overt relationships and meanings are embedded within that prose, organizing it for the reader. In excerpt 8.1, linguistic items that explicitly mark relationships between clauses or ideas and help involve the reader in the progression of the argument are **bolded**, occurring throughout the prose example the author is using to present his or her argument:

8.1 *Theoretical Philosophy (Hawkins 2008):*
In addition, she tries to make it turn out that eating each item requires the same number of bites. **Thus if** it takes her twelve bites to eat her sandwich, she must adjust her bites of pickle **so that** it takes exactly twelve bites to eat the pickle (and **similarly** exactly twelve sips to drink the milk and twelve bites to eat the Chompo bar). **Then** the last bite of sandwich is followed by the last bite of pickle and so on, and her lunch has 'come out even'. **Why does Frances do this?** The best answer seems to be that the ritual just appeals to her. She enjoys it in her little way. It strikes her as a good thing to do at the time, **even though** she could offer no reasons for believing it has worth. To say that she values the ritual, or has some belief about its worth,

> as we seem compelled to do **if** desire is recast as evaluative belief, **would** seriously distort the case. Frances' ritual seems to be a good example of an action which Velleman **would** forbid treating as a case of earnest evaluative pursuit; to that extent, I agree. **However**, it also seems natural to say of Frances that in desiring to make her lunch come out even, she sees something good in her ritual. While we should not collapse desire into evaluative belief,

In contrast, theoretical physics also presents evidence in a series that the reader must follow along with in order to make sense of the text. However, theoretical physics does *not* use extensive prose analyses to present evidence. Rather, series of mathematical formulas are used, and the authors include many of these same 'overt marking' features such as personal pronouns, adverbs, and adverbial conjuncts as frames for introducing and moving between steps in the mathematical analysis. Excerpt 8.2 below exemplifies this:

8.2 *Theoretical Physics (Lee & Yang 2007):*
Four-tachyon scatterings with NR = NL = 0
We are **now** ready to calculate the string scattering amplitudes. Let **us first** calculate the case with [formula].
[formula]
where we have used [formula] for closed string propagators [formula].
Note that for this **simple** case, Eq. (2.7) implies either m = 0 or n = 0. **However**, **we** will keep track of the general values of (m, n) **here** for the reference of future calculations. By using the formula [formula] **we** obtain [formula] where **we** have used [formula]. **In the above calculation**, **we** have used the following **well-known** formula for gamma function [formula].
High energy massive scatterings for general NR + NL. **We now** proceed to calculate the high energy scattering amplitudes for general higher mass levels with fixed NR + NL. **We now** have more mass parameters to define the "high energy limit". **So** let **us first** clear and redefine the concept of "high energy limit" in **our** following calculations.

Thus, despite a similarity in the need to guide readers through the evidence provided in theoretical articles, that evidence takes very different forms depending on discipline. This leads philosophy and physics to appear much less similar to each other if we look at the linguistic features that seem to characterize theoretical registers within the context of the individual disciplines.

When we compare quantitative and qualitative research articles, despite both registers being reports of empirical research, we also see substantial variation. Based on the summary presented in Table 8.3, we see that Dimension 2 (Contextualized Narration vs. Procedural Discourse) accounts for much of the variation between these two registers. Going back to Figure 7.4, it becomes clear that

Dimension 2 is actually a very strong descriptor of qualitative research, while quantitative research (particularly the hard sciences) typically falls on the more procedural cline of the dimension. Quantitative research in the social sciences, however, falls in the middle of this dimension. This major division likely reflects the explicit attention paid to describing data and methods in the hard sciences. Quantitative research in the social sciences also generally provides these descriptions, but qualitative research often does not include explicit marking of data and methods, a fact which has been previously acknowledged about qualitative research (see, e.g., Moilanen 2000).

However, more important than this relative lack of explicit procedural discourse, is the reliance on contextualized narration, which highly reflects the nature of qualitative inquiry as "a situated activity that locates the observer in the world" (Denzin & Lincoln 2000: 3), a characteristic that has been described to me as the desire to "jump into the story" (S. Wright, personal communication, September 10, 2009). Features such as 3rd person pronouns, perfect aspect, aspectual verbs (e.g., *begin, cease, complete, continue, end, finish, keep, start, stop*), and adverbs of time all help to create that narrative inquiry. In qualitative registers, evidence is presented through rich, prose descriptions that employ many of these narrative features.

In contrast, the analyses in this book have characterized quantitative research as more procedural. To illustrate this complex interplay between register and discipline, let's consider two excerpts from political science. More specifically, these two excerpts come from parts of the research article in which evidence is being presented and analyzed. In the first example from qualitative political science, the authors position the research as an analysis of the effects of a judicial decisions and social movements, and to do so, present evidence: a description of a specific court case. This excerpt relies on many of the features of contextualized narration, as summarized above.

8.3 *Qualitative Political Science (Meyer & Boutcher 2007):*
The unanimous decision in Brown v. Board of Education looms large in virtually every narrative of the civil rights movement. Although African-Americans had been organizing for civil rights for decades beforehand, the decision marked a national political breakthrough. To be sure, civil rights had appeared episodically in presidential politics in the years prior: Harry Truman desegregated the armed forces by executive order in 1947, as part of a larger effort to ramp up American foreign and military policies; Truman's tumultuous 1948 reelection campaign was marked by pressure for government action by reformers within the Democratic party, most notably Minnesota Senator Hubert Humphrey, and by the first exit of Southern Democrats from the party over the issue, led by Senator Strom Thurmond

> and his "Dixiecrats." Still, Brown promised-or threatened-to step into the day-to-day lives of black and white Americans on a scale previously unimaginable, with children on the front edge of massive social change.

In contrast, quantitative political science relies on representations of quantitative data, and the findings that result from the analysis of that data. Thus, the following excerpt is more procedural in nature, referring the reader to the graphical representation of data and then describing it in language that relies on features such as passive voice, relative clauses, predicative adjectives, and so on (see Table 8.3).

> 8.4 *Quantitative Political Science (Inglehart, Moaddel & Tessler 2006):*
> Figure 1 shows the percentage of the public who indicated that they would not want to have foreigners as neighbors, in countries on all six inhabited continents. A more specific version of this question was asked in Iraq: the public was asked about various specific groups of foreigners, ranging from Westerners (the Americans, British and French) to neighboring Islamic publics (Iranians, Turks, Kuwaitis and Jordanians) and also including various groups within Iraqi society.
> Perhaps not surprisingly under current conditions, the nationalities of the two main occupying powers were highly unpopular: Americans and British were both rejected as neighbors by overwhelming majorities of 87 percent among the Iraqi public as a whole.

Through this synthesis, some connections between the information that was gained through the three analyses in the book have been illustrated. In the next section, however, I turn to this issue more specifically.

8.3 Three grammatical analyses and future directions

In this section, I discuss how the three studies of grammatical variation in the Academic Journal Register Corpus complement one another and lead to greater understanding of linguistic variation across registers and disciplines. In the first section, I focus on these points specifically for the three approaches that I have taken in the analyses reported on here. But first, let's review the approaches that I've taken in the linguistic analyses in this book.

In the first analysis, I relied on previously established semantic sets of nouns and verbs, as well as linguistic features identified based on automatic tagging of texts for parts of speech. This approach relies upon previously established linguistic concepts, investigates their use in each text in the corpus, and then imposes order on that data by calculating means for each sub-corpus. The focal linguistic variables in this analysis are general lexical and grammatical features.

The second analysis investigates a functional construct. That is, it investigates a group of features that have been compiled based on the premise that there are underlying communicative functions being carried out by the features in tandem. The study relies upon sets of pre-selected words and combines that lexical information with grammatical tagging information in order to more reliably identify the instances of language which should be counted for the analysis. These pre-selected words were derived empirically by Biber and colleagues for *The Longman Grammar of Spoken and Written English* – words occurring in a particular syntactic slot were analyzed, and then the most frequently occurring ones are grouped together to form the sampling domain. While these analyses also resulted in rates of occurrence reported as means for the sub-corpora, the fact that the features can be group based on theoretical relationships allows for a comprehensive analysis of functional constructs – elaboration and compression.

The third analysis uses the statistical technique of factor analysis to inductively identify co-occurring sets of linguistic features. The analysis relies on the entire data set (with no register distinctions overtly marked) and analyzes the corpus from the bottom-up in order to derive sets of linguistic features on the basis of co-occurrence patterns alone. Thus, the features are grouped based on their observed patterns, rather than on any pre-existing theories of discourse functions or styles. The question is, however, what have we learned from these three approaches?

8.3.1 What have we learned from three complementary approaches?

The picture of academic writing that has been painted by the analyses in this book is complex and multi-layered. The individual descriptions of the use of linguistic features have each contributed on their own to our knowledge about how and why professional academic writers use language in the way that they do. Each individual analysis, along with the analysis of the non-linguistic characteristics of the texts, allows us to comment on a particular component of academic writing. To use the example of first person pronouns, the analyses have revealed that theoretical articles in physics and philosophy both rely heavily upon first person pronouns in order to lead the reader along the argument they are making. Because the scope of the investigation allowed for a comprehensive look at the forms and uses of first person pronouns, a great deal of detailed information on how the feature is used is available. When we turn to, for example, the comprehensive register study using multi-dimensional analysis and see first person pronouns on Dimension 1, we can use the more detailed functional analysis to more fully interpret the relationship of the feature to the overall factor. We can even use this information in interpreting the distribution of the mean dimensions scores for each discipline and register. This technique

was applied in Section 8.2.2 above in the interpretation of the characteristic features of theoretical articles.

In the second analysis of the functional grouping of features, we learned less about the variation of each individual feature (an issue discussed below), but we do see how language users rely on a variety of linguistic choices to create discourse of a particular style. Specifically in reference to the study of structural complexity, we saw a clear, consistent, and almost linear cline of variation as we moved along the continuum of soft to hard academic disciplines. This was in contrast to the more varied nature of the results in the general lexical and grammatical survey.

In contrast, the multi-dimensional analysis revealed much more complex patterns of variation across a wider range of linguistic features. While the resulting dimension 1 in part mimicked the cline of variation seen in the structural complexity study, it gave additional information about other co-occurring features, and the knowledge that was gained from the first two analyses aided in the interpretation of the underlying function of dimension 1. Then, as the other three dimensions were explored, further parameters underlying variation emerged. These dimensions helped make sense of some of the results seen during the initial two analyses, and here I am particularly reminded of the analysis of past tense and pronoun uses in the grammatical survey. In fact, the way in which these multiple analytical approaches complement each other is highly cyclical in nature, as we can revisit analyses to make further sense of findings and offer more comprehensive functional descriptions of the patterns that we find at various levels. For example, knowledge of how the different disciplines relied on semantic sets of verbs to differing degrees provided a building block for the analysis of passive voice. As an initial investigation utilizing complementary approaches, this study has also provided the basis for recommendations for further research along these same lines.

8.3.2 Future research using corpus analytical techniques to investigate variation in academic journal registers

A great deal has been learned about academic journal registers in these six disciplines through these analyses, and this knowledge has helped us to understand some of the fundamental differences in research in these disciplines. Still, this research has also indicated potential areas for future research. These future areas involve the application of a variety of analytical techniques, but also a variety of additional linguistic features that may be important markers of variation across academic disciplines and registers.

First, many of the analyses presented here could be developed to take on a more corpus-driven perspective, with the goal of identifying features that match the

specialized domain under investigation. For example, the words included in the semantic sets of nouns, verbs and adjectives have been arrived at from a corpus-based approach, where the most frequent nouns occurring in (typically) a broad range of registers are analyzed. However, an alternative approach would be to begin with the frequently occurring words in the specific registers under investigation and derive semantic sets from the actual words that are frequent in this corpus. This could also involve, for example, a re-thinking of some of the words which are included in the semantic categories. Biber (2006: Appendix A) notes that these words have been assigned to semantic categories according to their most common, basic use, which was determined by looking at a range of spoken and written texts. Since the domain of academic journal article registers in 6 disciplines is more restricted/specialized, it's possible that words would carry different primary meanings based on their use in a particular register. Having semantic sets of words that are individualized to the specific domain of language use may reveal previously un-documented variation.

A second study that would move the current analyses into a more inductive approach would be to utilize a technique like cluster analysis. In most corpus-based studies in which registers are compared, register or text is approached from a top-down perspective. That is, the texts in the corpus or corpora have been assigned a register label and belong to a corpus or sub-corpus with other texts of the same register. However, it is also possible to begin with a set of texts and then perform linguistic analyses that indicate how texts group together because of linguistic similarities. Cluster analysis has been used for this purpose (Biber & Finegan 1988, 1989) to inductively identify text types.

Such an inductive text type approach would be useful in the study of disciplinary writing, as it allows texts to be grouped because of their linguistic similarities rather than pre-determined text categories. Not only can this be indicative of within-register variation, but it also allows the researcher to discover underlying similarities across registers and locate less readily-apparent functional and situational characteristics of texts that are grouped together. For example, such an approach may identify groups of articles within and across disciplines and registers that are unified based on linguistic features associated with, for example, a specific methodology (e.g., ethnographic research in applied linguistics and political science), a specific topic/subject matter (e.g., studies about the characteristics of language versus the processes of acquiring language within applied linguistics), specific data types (e.g., historical records in history and qualitative political science) and so on.

8.3.3 Implications: Future linguistic features of interest

The general picture of disciplinary variation that has emerged from this research is that there is a fundamental cline of variation that involves the degree to which

discourse in the disciplines is elaborated and compressed (as illustrated by all three analyses, but particularly dimension 1 in the MD analysis). This cline of variation exists both across disciplines, and within disciplines, and is highly related to the situational characteristics of these texts. In particular, one aspect that was not coded in the present study, but which varied markedly across the texts in my corpus, is the degree and manner in which literature reviews are used across the disciplines, and in different registers. Thus, it seems that the following features may be indicated as potential markers of variation along these same parameters.

First, the analyses could be expanded to include additional linguistic features. For example, previously studied compression features include appositive noun phrases and all prepositional phrases as post-nominal modifiers (the present study looked only at noun +*of*-phrase sequences). In addition, the complexity study focused on the broader categories of grammatical structures (e.g., all finite complement clauses), whereas it may be the case that different disciplines and registers use these structures in more nuanced ways not captured by the approach taken here. For example, the brief look at the controlling words for *that*-complement clauses presented in Chapter 6 illustrates that the disciplines and registers relied on different controlling words. More detailed studies of these features would likely add to our knowledge about the nature of compression and elaboration in written academic texts.

Second, a related theme in the analyses presented here has been the degree to which logical and grammatical relationships are explicitly or not explicitly marked. Studies which investigate other features that make these explicit connections would likely reveal more interesting patterns across registers and disciplines. Possible features include information structuring (e.g., fronting, extraposition), cohesive devices (e.g., shell nouns in their various syntactic environments), lexical bundles (particularly those with referential and discourse organizing structures), linking adverbials, and markers of stance (both lexical and grammatical in nature).

Third, as the situational analysis in Chapter 4 illustrated, a great deal of variation exists as to the various sections research articles have, how they label them, and the degree to which they use these sections to explicitly state information about data and methods. While most research has focused on smaller units (i.e., moves and steps, Swales 1990), or on disciplines which do contain those basic sections in a fairly standardized manner, little research has addressed the correspondence between the rhetorical structures of research across disciplinary lines.

Finally, while this study has focused on the results of research articles and the sub-registers within that larger domain, the analysis presented in Chapter 2 has also shown that a variety of other texts also form the canon of knowledge in

academic journals, such as reviews, synthesis papers, editorials, and forums. These registers also warrant description.

8.3.4 Implications: Corpus design for studies of disciplinary writing

The final topic that I would like to address is that of registers within academic research articles. At the start of this study, very little linguistic research had acknowledged, let alone investigated, the possibility that we should consider different types of research articles in corpora that are designed to represent journal article writing. While some research may comment on the differences in the situational characteristics of research articles across disciplines in minimal ways, the lack of consideration of variation within disciplines is part of the motivation for this research.

In discussing this issue, I assume that several premises are relevant in corpus design, namely that the goal of corpus-based research on research articles, unless otherwise stated, is to analyze the overall register of 'research articles'. It seems to me that this is the approach that has been overwhelmingly adopted, in my own research and in the field in general. However, the results of these analyses show that pure convenience samples of research articles with no consideration of the nature of research being reported may prove problematic for the generalizability of corpus-based studies, limiting the claims that we can make about our results reflecting the broader register of research articles.

When I began the research reported on in this book, I felt confident that I would identify significant differences across disciplines. I also felt confident that I would locate interesting patterns that I could associate, at least descriptively, with the different registers that I identified. However, the results of this research have gone beyond these expectations, and statistically significant differences have been found at the most detailed level of comparisons. That is, looking at the significance testing results for the multi-dimensional analysis, and more specifically at the post-hoc analyses, it turns out that the patterns of variation uncovered here statistically differentiate between registers within the same discipline (in particular, see the results for Dimensions 2, 3 and 4 for applied linguistics and political science). So, what does this mean for the design of corpora for linguistic descriptions of research articles?

In building a corpus, it is not always feasible (or necessary for the research goals) to apply the degree of analysis into the specific texts included in the corpus that I have applied in constructing this corpus. However, we can follow several principles to minimize the effect of register differences within a discipline, particularly for research which seeks to analyze disciplinary writing but not necessarily the sub-registers that occur within those disciplines (e.g., as is done in most

studies comparing multiple disciplines). In the first approach, we can limit the corpus design to one research article type that is prevalent in the discipline, and that is straightforward to identify. This would mimic a corpus design approach, for example, that focuses on a single sub-discipline within a discipline, rather than trying to represent many sub-disciplines or disregarding sub-discipline altogether. In the second approach, care could be taken to ensure representation of the different registers while not requiring an exact balanced representation of those articles. As research continues to explore the characteristics of academic writing, it is my hope that by considering aspects such as the type of research being reported on, we will be able to more fully understand writers' motivations for producing the language patterns in the texts that the academic community values so highly.

References

Afros, Elena & Schryer, Catherine F. 2009. Promotional (meta)discourse in research articles in language and literary studies. *English for Specific Purposes* 28: 58–68. DOI: 10.1016/j.esp.2008.09.001

Aktas, Rahime N. & Cortes, Viviana. 2008. Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes* 7: 3–14. DOI: 10.1016/j.jeap.2008.02.002

Allen, Roland. 2008. Coupling of electrons to the electromagnetic field in a localized basis. *Physical Review B* 78: 1–6.

Altenberg, Bengt & Granger, Sylviane. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22(2): 173–195. DOI: 10.1093/applin/22.2.173

Altman, Andrew & Wellman, Christopher Heath. 2008. From humanitarian intervention to assassination: Human rights and political violence. *Ethics* 118(2): 228–257. DOI: 10.1086/526543

Archer, Arlene. 2008. 'The place is suffering': Enabling dialogue between students' discourses and academic literacy conventions in engineering. *English for Specific Purposes* 27: 255–266. DOI: 10.1016/j.esp.2007.10.002

Avdela, Efi. 2008. 'Corrupting and uncontrollable activities': Moral panic about youth in the post-civil-war Greece. *Journal of Contemporary History* 43(1): 25–44. DOI: 10.1177/0022009407084556

Baker, Gordan. 2002. Wittgenstein on metaphysical/everyday use. *Philosophical Quarterly* 52(208): 289–302. DOI: 10.1111/1467-9213.00269

Banks, David. 2005. On the historical origins of nominalized process in scientific text. *English for Specific Purposes* 24: 347–357. DOI: 10.1016/j.esp.2004.08.002

Banks, David. 2008. *The Development of Scientific Writing: Linguistic Features and Historical Context*. London: Equinox.

Basso, Keith. 1974. The ethnography of writing. In *Explorations in the Ethnography of Speaking*, Richard Bauman & Joel Sherzer (eds), 425–432. Cambridge: CUP.

Basturkmen, Helen. 2009. Commenting on results in published research articles and masters dissertations in Language Teaching. *Journal of English for Academic Purposes* 8: 241–251. DOI: 10.1016/j.jeap.2009.07.001

Basturkmen, Helen. 2012. A genre-based investigation of discussion sections of research articles in dentistry and disciplinary variation. *Journal of English for Academic Purposes* 11: 134–144. DOI: 10.1016/j.jeap.2011.10.004

Bazerman, Charles. 1994. *Constructing Experience*. Carbondale IL: Southern Illinois University Press.

Becher, Tony. 1981. Towards a definition of disciplinary cultures. *Studies in Higher Education* 6(2): 109–122. DOI: 10.1080/03075078112331379362

Becher, Tony. 1994. The significance of disciplinary differences. *Studies in Higher Education* 19: 151–161. DOI: 10.1080/03075079412331382007

Belcher, Diane & Hirvela, Alan. 2005. Writing the qualitative dissertation: what motivates and sustains commitment to a fuzzy genre? *Journal of English for Academic Purposes* 4: 187–205. DOI: 10.1016/j.jeap.2004.07.010

Bhatia, Vijay. 1997. Genre-mixing in academic introductions. *English for Specific Purposes* 16(3): 181–195. DOI: 10.1016/S0889-4906(96)00039-7

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP.

Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15: 133–163. DOI: 10.1080/01638539209544806

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4): 243–257. DOI: 10.1093/llc/8.4.243

Biber, Douglas. 1994. An analytical framework for register studies. In *Sociolinguistic Perspectives on Register,* Douglas Biber & Edward Finegan (eds), 31–56. Oxford: OUP.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: CUP.

Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins.

Biber, Douglas. 2010. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford Handbook of Linguistic Analysis*, Bernd Heine & Heiko Narrog (eds), 160–191. Oxford: OUP.

Biber, Douglas & Clark, Victoria. 2002. Historical shifts in modification patterns with complex noun phrase structures: How long can you go without a verb? In *English Historical Syntax and Morphology* [Current Issues in Lingistic Theory 223], Teresa Fanego, Javier Pérez-Guerra & José López-Couso (eds), 43–66. Amsterdam: John Benjamins.

Biber, Douglas, Connor, Ulla & Upton, Thomas A. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure* [Studies in Corpus Linguistics 28]. Amsterdam: John Benjamins.

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre and Style*. Cambridge: CUP.

Biber, Douglas, Conrad, Susan & Cortes, Viviana. 2004. 'If you look at...': Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25: 371–405. DOI: 10.1093/applin/25.3.371

Biber, Douglas, Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Pat & Helt, Marie. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36: 9–48. DOI: 10.2307/3588359

Biber, Douglas, Csomay, Eniko, Jones, James & Keck, Casey. 2004. A corpus linguistic investigation of vocabulary-based discourse units in university registers. In *Applied Corpus Linguistics: A Multi-Dimensional Perspective,* Ulla Connor & Thomas Upton (eds), 53–72. Amsterdam: Rodopi.

Biber, Douglas & Finegan, Edward. 1988. Adverbial stance types in English. *Discourse Processes* 11: 1–34. DOI: 10.1080/01638538809544689

Biber, Douglas & Finegan, Edward. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9: 93–124. DOI: 10.1515/text.1.1989.9.1.93

Biber, Douglas & Finegan, Edward. 1994. *Sociolinguistic Perspectives on Register*. New York: OUP.

Biber, Douglas & Finegan, Edward. 2001. Intra-textual variation within medical research articles. In *Variation in English: Multi-dimensional Studies,* Susan Conrad & Douglas Biber (eds), 108–137. London: Longman.

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9: 2–20. DOI: 10.1016/j.jeap.2010.01.001

Biber, Douglas & Gray, Bethany. 2011. Grammar change in the noun phrase: The influence of written language use. *English Language & Linguistics* 15(2): 223–250. DOI: 10.1017/S1360674311000025

Biber, Douglas & Gray, Bethany. 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2): 104–134. DOI: 10.1177/0075424212472509

Biber, Douglas & Gray, Bethany. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge: CUP.

Biber, Douglas, Gray, Bethany, Honkapohja, Alpo & Pahta, Päivi. 2011. Prepositional modifiers in early English medical prose: A study ON their historical development IN noun phrases. In *Communicating Early English Manuscripts*, Päivi Pahta & Andreas H. Jucker (eds), 197–211. Cambridge: CUP.

Biber, Douglas, Gray, Bethany & Poonpon, Kornwipa. 2011. Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly* 45(1): 5–35. DOI: 10.5054/tq.2011.244483

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas & Jones, James. 2009. Quantitative methods in corpus linguistics. In *Corpus Linguistics: An International Handbook,* Anke Lüdeling & Merja Kytö (eds), 1286–1304. Berlin: Walter de Gruyter.

Brett, Paul. 1994. A genre analysis of the results sections of sociology articles. *English for Specific Purposes* 13(1): 47–59. DOI: 10.1016/0889-4906(94)90024-8

Bruce, Ian. 2008. Cognitive genre structures in Methods sections of research articles: A corpus study. *Journal of English for Academic Purposes* 7: 38–54. DOI: 10.1016/j.jeap.2007.12.001

Bruce, Ian. 2009. Results sections in sociology and organic chemistry articles: A genre analysis. *English for Specific Purposes* 28: 105–124. DOI: 10.1016/j.esp.2008.12.005

Bunton, David. 2005. The structure of PhD conclusion chapters. *Journal of English for Academic Purposes* 4: 207–224. DOI: 10.1016/j.jeap.2005.03.004

Cacciatori, Sergio, Dalla Piazza, Francesco & van Geemen, Bert. 2008. Modular forms and three-loop superstring amplitudes. *Nuclear Physics B* 800(3): 565-590. DOI: 10.1016/j.nuclphysb.2008.03.007

Callies, Marcus. 2013. Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. *International Journal of Corpus Linguistics* 18(3): 357–390. DOI: 10.1075/ijcl.18.3.05cal

Cao, Feng & Hu, Guangwei. 2014. Interactive metadiscourse in research articles: A comparative study of paradigmatic and disciplinary influences. *Journal of Pragmatics* 66: 15–31. DOI: 10.1016/j.pragma.2014.02.007

Charles, Maggie. 2006a. Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes* 25: 310–331. DOI: 10.1016/j.esp.2005.05.003

Charles, Maggie. 2006b. The construction of stance in reporting clauses: A cross-disciplinary study of theses. *Applied Linguistics* 27: 492–518. DOI: 10.1093/applin/aml021

Charles, Maggie. 2007. Argument or evidence? Disciplinary variation in the use of the Noun that pattern in stance construction. *English for Specific Purposes* 26: 203–218. DOI: 10.1016/j.esp.2006.08.004

Chen, Qi & Ge, Guang-chun. 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes* 26: 502–524. DOI: 10.1016/j.esp.2007.04.003

Cheng, Liying, Fox, Janna & Zhen, Ying. 2007. Student accounts of the Ontario Secondary School Literacy Test: A case for validation. *Canadian Modern Language Review* 64(1): 69–98. DOI: 10.3138/cmlr.64.1.069

Collier, David. 1993. The comparative method. In *Political Science: The State of the Discipline II,* Ada W. Finifter (ed), 105–120. Washington DC: The American Political Science Association.

Collins, John. 2008. Content externalism and brute logical errors. *Canadian Journal of Philosophy* 38(4): 549–574. DOI: 10.1353/cjp.0.0031

Conrad, Susan. 1996a. Academic Discourse in Two Disciplines: Professional Writing and Student Development in Biology and History. PhD dissertation, Northern Arizona University.

Conrad, Susan. 1996b. Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education* 8: 299–326. DOI: 10.1016/S0898-5898(96)90025-X

Conrad, Susan & Biber, Douglas. 2001. *Variation in English: Multi-dimensional Studies*. London: Longman.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23: 397–423. DOI: 10.1016/j.esp.2003.12.001

Cortes, Viviana. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes* 12: 33–43. DOI: 10.1016/j.jeap.2012.11.002

Crystal, David & Davy, Derek. 1969. *Investigating English Style*. Harlow: Longman.

Dahl, Trine. 2008. Contributing to the academic conversation: A study of new knowledge claims in economics and linguistics. *Journal of Pragmatics* 40: 1184–1201. DOI: 10.1016/j.pragma.2007.11.006

Danielsson, Jesper, Reva, Oleg & Meijer, Johan. 2006. Protection of oilseed rape (Brassica napus) toward fungal pathogens by strains of plant-associated Bacillus amyloliquefaciens. *Microbial Ecology* 54(1): 134–140. DOI: 10.1007/s00248-006-9181-2

Davis, Scott & Tanner, Julian. 2003. The long arm of the law: Effects of labeling on employment. *The Sociological Quarterly* 44(3): 385–404. DOI: 10.1111/j.1533-8525.2003.tb00538.x

Denzin, Norman & Lincoln, Yvonna. 2000. Introduction: The discipline and practice of qualitative research. In *Handbook of Qualitative Research,* 2nd edn, Norman Denzin & Yvonna Lincoln (eds), 1–29. London: Sage.

Diani, Giuliana. 2008. Emphasizers in spoken and written academic discourse: The case of really. *International Journal of Corpus Linguistics* 13(3): 296–321. DOI: 10.1075/ijcl.13.3.04dia

Dueñas, Pilar M. 2007. 'I/we focus on...': A cross-cultural analysis of self-mentions in business management research articles. *Journal of English for Academic Purposes* 6: 143–162. DOI: 10.1016/j.jeap.2007.05.002

Fang, Zhihui, Schleppegrell, Mary J. & Cox, Beverly E. 2006. Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research* 38(3): 247–273. DOI: 10.1207/s15548430jlr3803_1

Fette, J. 2007. Pride and prejudice in the professions: Women doctors and lawyers in Third Republic France. *Journal of Women's History* 19(3): 60–86.

Flowerdew, John. 2006. Use of signaling nouns in a learner corpus. *International Journal of Corpus Linguistics* 11(3): 345–362. DOI: 10.1075/bct.17.06flo

Flowerdew, John & Dudley-Evans, Tony. 2002. Genre analysis of editorial letters to international journal contributors. *Applied Linguistics* 23(4): 263–489. DOI: 10.1093/applin/23.4.463

Fortanet, Inmaculada. 2008. Evaluative language in peer review referee reports. *Journal of English for Academic Purposes* 7: 27–37. DOI: 10.1016/j.jeap.2008.02.004

Frazier, Stefan. 2007. Tellings of remembrances 'touched off' by student reports in group work in undergraduate writing classes. *Applied Linguistics* 28(2): 189–219. DOI: 10.1093/applin/amm002

Freddi, Maria. 2005. Arguing linguistics: Corpus investigation of one functional variety of academic discourse. *Journal of English for Academic Purposes* 4: 5–26.

Friginal, Eric. 2009. *The Language of Outsourced Call Centers: A Corpus-Based Study of Cross-Cultural Interaction* [Studies in Corpus Linguistics 34]. Amsterdam: John Benjamins.

Fuertes-Olivera, Pedro A. 2007. A corpus-based view of lexical gender in written Business English. *English for Specific Purposes* 26: 219–234. DOI: 10.1016/j.esp.2006.07.001

Ganguly, S., Banerjee, P., Ray, I., Kshetri, R., Raut, R., Bhattacharya, S., Saha-Sarkar, M., Goswami, A., Mukhopadhyay, S., Muhkerjee, A., Mukherjee, G. & Basu, S.K. 2007. Study of intruder band in ^{112}Sn. *Nuclear Physics A* 789: 1–14. DOI:10.1016/j.nuclphysa.2007.01.092

Gardner, Sheena & Nesi, Hilary. 2012. A classification of genre families in university student writing. *Applied Linguistics* 34(1): 1–29. DOI: 10.1093/applin/ams024

Giannoni, Davide S. 2008. Popularizing features in English journal editorials. *English for Specific Purposes* 27: 212–232.

Gillaerts, Paul & Van de Velde, Freek. 2010. Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes* 9: 128–139. DOI: 10.1016/j.jeap.2010.02.004

Goodin, Robert E. (ed.) 2009. *The Oxford Handbook of Political Science*. Oxford: OUP.

Goodin, Robert E. & Klingermann, Hans-Dieter. 1998. *A New Handbook of Political Science*. Oxford: OUP.

Golonka, Ewa. 2006. Predictors revised: Linguistic knowledge and metalinguistic awareness in second language gain in Russian. *Modern Language Journal* 90(4): 496-505. DOI: 10.1111/j.1540-4781.2006.00428.x

Gosden, Hugh. 1992. Discourse functions of marked theme in scientific research articles. *English for Specific Purposes* 11: 207–224. DOI: 10.1016/S0889-4906(05)80010-9

Grant, Leslie & Ginther, April. 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing* 9(2): 123–145. DOI: 10.1016/S1060-3743(00)00019-9

Gray, Bethany. 2010. On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose. *Journal of English for Academic Purposes* 9: 167–183. DOI: 10.1016/j.jeap.2009.11.003

Green, Christopher F., Christopher, Elsie R. & Mei, Jaquelin L.K. 2000. The incidence and effects on coherence of marked themes in interlanguage texts: A corpus-based enquiry. *English for Specific Purposes* 19: 99–113. DOI: 10.1016/S0889-4906(98)00014-3

Groom, Nicholas. 2005. Pattern and meaning across genres and disciplines: An exploratory study. *Journal of English for Academic Purposes* 4: 257–277. DOI: 10.1016/j.jeap.2005.03.002

Halliday, M.A.K. 1978. *Language as a Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: OUP.

Halliday, M.A.K. 2004. *The Language of Science*. London: Continuum.

Hardy, Jack & Römer, Ute. 2013. Revealing disciplinary variation in student writing: A multi-dimensional analysis of the Michigan Corpus of Upper-Level Student Papers (MICUSP). *Corpora* 8(2): 183–207. DOI: 10.3366/cor.2013.0040

Harwood, Nigel. 2005a. 'I hoped to counteract the memory problem, but I made no impact whatsoever': Discussing methods in computing science using I. *English for Specific Purposes* 24: 243–267. DOI: 10.1016/j.esp.2004.10.002

Harwood, Nigel. 2005b. '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. DOI: 10.1093/applin/ami012

Hawkins, Jennifer. 2008. Desiring the bad under the guise of the good. *Philosophical Quarterly* 58(231): 244–264. DOI: 10.1111/j.1467-9213.2007.520.x

Heineman, Robert. 1995. *Political Science*. New York NY: McGraw Hill.

Hemais, Barbara. 2001. The discourse of research and practice in marketing journals. *English for Specific Purposes* 20: 39–59. DOI: 10.1016/S0889-4906(99)00021-6

Hewings, Martin & Hewings, Ann. 2002. "It is interesting to note that...": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes* 21: 367–383. DOI: 10.1016/S0889-4906(01)00016-3

Hewings, Ann, Lillis, Theresa & Vladimirou, Dimitra. 2010. Who's citing whose writings? A corpus based study of citations as interpersonal resource in English medium national and English medium international journals. *Journal of English for Academic Purposes* 9: 83–150. DOI: 10.1016/j.jeap.2010.02.005

Hinkel, Eli. 2003. Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly* 37(2): 275–301. DOI: 10.2307/3588505

Hirano, Eliana. 2009. Research article introductions in English for specific purposes: A comparison between Brazilian Portuguese and English. *English for Specific Purposes* 28: 240–250. DOI: 10.1016/j.esp.2009.02.001

Hoeinghaus, David, Winemiller, Kirk & Agostinho, Angelo. 2008. Hyrdogeomorphology and river impoundment affect food-chain length of diverse Neotropical food webs. *Oikos* 117: 984–995. DOI: 10.1111/j.0030-1299.2008.16459.x

Holmes, Richard. 1997. Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes* 16: 321–337. DOI: 10.1016/S0889-4906(96)00038-5

Hyland, Ken. 1996. Writing without conviction? Hedging in science research articles. *Applied Linguistics* 17(4): 433–454. DOI: 10.1093/applin/17.4.433

Hyland, Ken. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text* 18: 349–382. DOI: 10.1515/text.1.1998.18.3.349

Hyland, Ken. 1999a. Talking to students: Metadiscourse in Introductory coursebooks. *English for Specific Purposes* 18(1): 3–26.

Hyland, Ken. 1999b. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20(3): 341–367. DOI: 10.1093/applin/20.3.341

Hyland, Ken. 2001a. Bringing in the reader: Addressee features in academic articles. *Written Communication* 18(4): 549–574. DOI: 10.1177/0741088301018004005

Hyland, Ken. 2001b. Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes* 20: 207–226. DOI: 10.1016/S0889-4906(00)00012-0

Hyland, Ken. 2002a. Directives: Argument and engagement in academic writing. *Applied Linguistics* 23: 215–239. DOI: 10.1093/applin/23.2.215

Hyland, Ken. 2002b. Authority and invisibility: authorial identity in academic writing. *Journal of Pragmatics* 34: 1091–1112. DOI: 10.1016/S0378-2166(02)00035-8

Hyland, Ken. 2007. Applying a gloss: Exemplifying and reformulating in academic discourse. *Applied Linguistics* 28(2): 266–285. DOI: 10.1093/applin/amm011

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27: 4–21. DOI: 10.1016/j.esp.2007.06.001

Hyland, Ken. 2010. Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes* 9: 116–127. DOI: 10.1016/j.jeap.2010.02.003

Hyland, Ken & Tse, Polly. 2005. Hooking the reader: A corpus study of evaluative that in abstracts. English for Specific Purposes, 24, 123–139. DOI: 10.1016/j.esp.2004.02.002

Hyland, Ken & Tse, Polly. 2007. Is there an "academic vocabulary"? *TESOL Quarterly* 4(2): 235–253. DOI: 10.1002/j.1545-7249.2007.tb00058.x

Hymes, Dell. 1974. *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia PA: University of Pennsylvania Press.

Inglehart, Ronald, Moaddel, Mansoor & Tessler, Mark. 2006. Xenophobia and in-group solidarity in Iraq: A natural experiment on the impact of insecurity. *Perspectives on Politics* 4(3): 495–505. DOI: 10.1017/S1537592706060324

Jarvis, Scott, Grant, Leslie, Bikowski, Dawn & Ferris, Dana. 2003. Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing* 12: 377–403. DOI: 10.1016/j.jslw.2003.09.001

Jordanova, Ludmilla. 2000. *History in Practice*. London: Arnold.

Kanoksilapatham, Budsaba. 2005a. Rhetorical structure of biochemistry research articles. *English for Specific Purposes* 24: 269–292. DOI: 10.1016/j.esp.2004.08.003

Kanoksilapatham, Budsaba. 2005b. A Corpus-Based Investigation of Scientific Research Articles: Linking Move Analysis with Multidimensional Analysis. PhD dissertation, Georgetown University.

Kaplan, R. 2010. *The Oxford Handbook of Applied Linguistics,* 2nd edn. Oxford: OUP.

Kelly, David W., Macisaac, Hugh J. & Heath, Daniel D. 2006. Vicariance and dispersal effects on phylogeographic structure and speciation in a widespread estuarine invertebrate. *Evolution* 60(2): 257–267.

Khedri, Mohsen, Heng, Chan Swee & Ebrahimi, Seyed Foad. 2013. An exploration of interactive discourse markers in academic research article abstracts in two disciplines. *Discourse Studies* 15(3): 319–331. DOI: 10.1177/1461445613480588

Kim, Seung-Hwan, Holway, Antonia H., Wolff, Suzanne, Dillin, Andrew & Michael, W. Matthew. 2007. SMK-1/PPH-4.1mediated silencing of the CHK-1 response to DNA damage in early C. elegans embryos. *The Journal of Cell Biology* 179(1): 41–52. DOI: 10.1083/jcb.200705182

Koutsantoni, Dimitra. 2004. Attitude, certainty and allusions to common knowledge in scientific research articles. *Journal of English for Academic Purposes* 3: 163–182. DOI: 10.1016/j.jeap.2003.08.001

Koutsantoni, Dimitra. 2006. Rhetorical strategies in engineering research articles and research theses: Advanced academic literacy and relations of power. *Journal of English for Academic Purposes* 5: 19–36. DOI: 10.1016/j.jeap.2005.11.002

Kuo, Chih-Hua. 1999. The use of personal pronouns: Role relationships in scientific journal articles. *English for Specific Purposes* 18: 121–138. DOI: 10.1016/S0889-4906(97)00058-6

Kwan, Becky S.C., Chan, Hang & Lam, Colin. 2012. Evaluating prior scholarship in literature reviews of research articles: A comparative study of practices in two research paradigms. *English for Specific Purposes* 31: 188–201. DOI: 10.1016/j.esp.2012.02.003

Ledru, Gerald, Marchal, Frédéric, Merbahi, Nofel, Gardou, J.P. & Sewraj, N. 2007. Study of the formation and decay of KrXe* excimers at room temperature following selective excitation of the xenon 6s states. *Journal of Physics B* 40(10): 1651.

Lee, David Y.W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3): 37–72.

Lee, David Y.W. & Chen, Sylvia X. 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18: 149–165. DOI: 10.1016/j.jslw.2009.07.003

Lee, Jen-Chi & Yang, Yi. 2007. Linear relations and their breakdown in high energy massive string scatterings in compact spaces. *Nuclear Physics B* 784: 22–35. DOI:10.1016/j.nuclphysb.2007.06.005

Lewin, Beverly. 2005. Contentiousness in science: The discourse of critique in two sociology journals. *Text* 25: 723–744. DOI: 10.1515/text.2005.25.6.723

Li, Li-Juan & Ge, Guang-Chun. 2009. Genre analysis: Structural and linguistic evolution of the English-medium medical research article (1985-2004). *English for Specific Purposes* 28: 93–104. DOI: 10.1016/j.esp.2008.12.004

Lim, Jason. 2006. Method sections of management research articles: A pedagogically motivated qualitative study. *English for Specific Purposes* 25: 282–309. DOI: 10.1016/j.esp.2005.07.001

Lin, Ling & Evans, Stephen. 2012. Structural patterns in empirical research article: A cross-disciplinary study. *English for Specific Purposes* 31: 150–160. DOI: 10.1016/j.esp.2011.10.002

Lopes Don, Patricia. 2006. Franciscans, Indian sorcerers, and the Inquisition in New Spain, 1536–1543. *Journal of World History* 17(1): 27–49.

Loudermilk, Brandon. 2007. Occluded academic genres: An analysis of the MBA thought essay. *Journal of English for Academic Purposes* 6: 190–205. DOI: 10.1016/j.jeap.2007.07.001

Magnet, Anne & Carnet, Didier. 2006. Letters to the editor: Still vigorous after all these years? A presentation of the discursive and linguistic features of the genre. *English for Specific Purposes* 25: 173–199. DOI: 10.1016/j.esp.2005.03.004

Marco, Maria J. L. 2000. Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes* 19: 63–86. DOI: 10.1016/S0889-4906(98)00013-1

Martínez, Iliana. 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. DOI: 10.1016/S1475-1585(03)00003-1

Martínez, Iliana. 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. DOI: 10.1016/j.jslw.2005.06.001

Martínez, Iliana, Beck, Silvia, & Panza, Carolina. 2009. Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes* 28: 183–198. DOI: 10.1016/j.esp.2009.04.003

McEnery, Tony, Xiao, Richard & Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York NY: Routledge.

Merli, David. 2008. Expressivism and the limits of moral disagreement. *Journal of Ethics* 12: 25–55. DOI: 10.1007/s10892-007-9022-7

Meyer, Davis & Boutcher, Steven. 2007. Signals and spillover: Brown v. Board of Education and other social movements. *Perspectives on Politics* 5(1): 81-93. DOI: 10.1017/S1537592707070077

Moilanen, Pentti. 2000. Interpretation, truth and correspondence. *Journal of the Theory of Social Behavior* 30(4): 377–390. DOI: 10.1111/1468-5914.00136

Moore, Tim. 2002. Knowledge and agency: A study of 'metaphenomenal discourse' in textbooks from three disciplines. *English for Specific Purposes* 21: 347–366. DOI: 10.1016/S0889-4906(01)00030-8

Munslow, Alun. 1997. *Deconstructing History*. London: Routledge.

Nesi, Hilary & Gardner, Sheena. 2012. *Genres across the Disciplines: Student Writing in Higher Education*. Cambridge: CUP.

Norman, Guy. 2003. Consistent naming in scientific writing: Sound advice or shibboleth? *English for Specific Purposes* 22: 113–130. DOI: 10.1016/S0889-4906(02)00013-3

Nwogu, Kevin. 1997. The medical research paper: Structure and functions. *English for Specific Purposes* 16: 119–138. DOI: 10.1016/S0889-4906(97)85388-4

Ozturk, Ismet. 2007. The textual organization of research article introductions in applied linguistics: Variability within a single discipline. *English for Specific Purposes* 26: 25–38. DOI: 10.1016/j.esp.2005.12.003

Parkinson, Jean. 2013. Representing own and other voices in social science research articles. *International Journal of Corpus Linguistics* 18(2): 199–228. DOI: 10.1075/ijcl.18.2.02par

Parkinson, Jean & Musgrave, Jill. 2014. Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes* 14: 48–59. DOI: 10.1016/j.jeap.2013.12.001

Peacock, Matthew. 2006. A cross-disciplinary comparison of boosting in research articles. *Corpora* 1(1): 61–84. DOI: 10.3366/cor.2006.1.1.61

Posteguillo, Santiago. 1999. The schematic structure of computer science research articles. *English for Specific Purposes* 18(2): 139–160. DOI: 10.1016/S0889-4906(98)00001-5

Pacek, Alexander & Radcliff, Benjamin. 2008. Assessing the welfare state: The politics of happiness. *Perspectives on Politics* 6(2): 267–277.

Ramírez, Ricardo. 2007. Segmented mobilization: Latino nonpartisan get-out-the-vote efforts in the 2000 general election. *American Politics Research* 35(2): 155–175. DOI: 10.1177/1532673X06296578

Reece, Jane B., Urry, Lisa A., Cain, Michael L., Wasserman, Steven A., Minorsky, Peter V. & Jackson, Robert B. 2010. *Campbell Biology,* 9th edn. Boston MA: Benjamin Cummings.

Roache, Rebecca. 2006. A defence of quasi-memory. *Philosophy* 81(2): 323–355. DOI: 10.1017/S0031819106316075

Robinson, Marin, Stoller, Fredricka, Costanza-Robinson, Molly & Jones, James K. 2008. *Write like a Chemist: A Guide and Resource*. Oxford: OUP.

Römer, Ute & Swales, John. 2010. The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes* 9: 249.

Ruiying, Yang & Allison, Desmond. 2004. Research articles in applied linguistics: Structures from a functional perspective. *English for Specific Purposes* 23: 264–279. DOI: 10.1016/S0889-4906(03)00005-X

Russell, David. 1991. *Writing in the Academic Disciplines, 1870-1990: A Curricular History*. Carbondale IL: Southern Illinois University Press.

Salager-Meyer, Françoise. 1994. Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes* 13: 149–170. DOI: 10.1016/0889-4906(94)90013-2

Samraj, Betty. 2002. Introductions in research articles: Variations across disciplines. *English for Specific Purposes* 21: 1–17. DOI: 10.1016/S0889-4906(00)00023-5

Samraj, Betty. 2004. Discourse features of the student-produced academic research paper: variations across disciplinary courses. *Journal of English for Academic Purposes* 3: 5–22. DOI: 10.1016/S1475-1585(03)00053-5

Samraj, Betty. 2005. An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes* 24: 141–156. DOI: 10.1016/j.esp.2002.10.001

Samraj, Betty. 2008. A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes* 7: 55–67. DOI: 10.1016/j.jeap.2008.02.005

Sanford, George. 2006. The Katyn Massacre and PolishSoviet relations, 1941–43. *Journal of Contemporary History* 41(1): 95-111. DOI: 10.1177/0022009406058676

Scott, Virginia & De La Fuente, María José. 2008. What's the problem? L2 learners' use of the L1 during consciousness-raising, form-focused tasks. *Modern Language Journal* 92(1): 100–113. DOI: 10.1111/j.1540-4781.2008.00689.x

Schleppegrell, Mary. 1996. Conjunction in spoken English and ESL writing. *Applied Linguistics* 17(3): 271–285.

Schmitt, Norbert. 2010. *An Introduction to Applied Linguistics,* 2nd edn. Abingdon: Hodder Education.

Siyanova, Anna & Schmitt, Norbert. 2008. L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review* 64(3): 429–458. DOI: 10.3138/cmlr.64.3.429

Spycher, Pamela. 2007. Academic writing of adolescent English learners: Learning to use "although". *Journal of Second Language Writing* 16: 238–254. DOI: 10.1016/j.jslw.2007.07.001

Stewart, Romola, Ball, Ian & Possingham, Hugh. 2007. The effect of incremental reserve sesign and changing reservation goals on the long-rerm efficiency of reserve systems. *Conservation Biology* 21(2): 346-354. DOI: 10.1111/j.1523-1739.2006.00618.x

Stoller, Fredricka & Robinson, Marin. 2013. Chemistry journal articles: An interdisciplinary approach to move analysis with pedagogical aims. *English for Specific Purposes* 32: 45–57. DOI: 10.1016/j.esp.2012.09.001

Stotesbury, Hilkka. 2003. Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes* 2: 327–341. DOI: 10.1016/S1475-1585(03)00049-3

Swales, John. 1990. *Genre Analysis: English for Academic and Research Settings*. Cambridge: CUP.

Swales, John, Ahmad, Ummul, Chang, Yu-Ying, Chavez, Daniel, Dressen, Dacia & Seymour, Ruth. 1998. Consider this: The role of imperatives in scholarly writing. *Applied Linguistics* 19: 97–121. DOI: 10.1093/applin/19.1.97

Tabachnick, Barbara & Fidell, Linda. 2007. *Using Multivariate Statistics,* 5th edn. Boston MA: Pearson.

Tarone, Elaine, Dwyer, Sharon, Gillette, Susan & Icke, Vincent. 1998. On the use of the passive and active voice in astrophysics journal papers: With extensions to other language and other fields. *English for Specific Purposes* 17: 113–132. DOI: 10.1016/S0889-4906(97)00032-X

Thomas, Sarah & Hawes, Thomas. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes* 13: 129–148. DOI:10.1016/0889-4906(94)90012-4

Thomas, Laura, Christakis, Theodore & Jorgensen, William. 2006. Conformation of alkanes in the gas phase and pure liquids. Journal of Physical Chemistry B 110(42): 21198-21204. DOI: 10.1021/jp064811

Tosh, John. 2000. *The Pursuit of History,* 3rd edn. Harlow: Longman.

Tucker, Paul. 2003. Evaluation in the art-historical research article. *Journal of English for Academic Purposes* 2: 291–312. DOI: 10.1016/S1475-1585(03)00047-X

Turner, Bryan. 2006. Discipline. *Theory, Culture & Society* 23: 183–197. DOI: 10.1177/0263276406062698

Vongpumivitch, Viphavee, Huang, Ju-yu & Chang, Yu-Chia. 2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes* 28: 33–41. DOI: 10.1016/j.esp.2008.08.003

Warchal, Krystyna. 2010. Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. *Journal of English for Academic Purposes* 9: 140–150. DOI: 10.1016/j.jeap.2010.02.002

Webber, Pauline. 1994. The function of questions in different medical journal genres. *English for Specific Purposes* 13: 257–268. DOI: 10.1016/0889-4906(94)90005-1

Wells, Rulon. 1960. Nominal and verbal style. In *Style in Language,* Thomas A. Sebeok (ed.), 213–220. Cambridge: CUP.

Williams, Ian. 1996. A contextual study of lexical verbs in two types of medical research report: Clinical and experimental. *English for Specific Purposes* 15: 175–197. DOI: 10.1016/0889-4906(96)00010-5

Wilson, Norman. 1999. *History in Crisis? Recent Directions in Historiography*. Upper Saddle River NJ: Prentice Hall.

Wink, Kenneth & Bargen, Andrew. 2008. The consolidation of the white southern congressional vote: The roles of ideology and party identification. *Politics and Policy* 36(3): 376–399. DOI: 10.1111/j.1747-1346.2007.00113.x

Vande Kopple, William. 1994. Some characteristics and functions of grammatical subjects in scientific discourse. *Written Communication* 11(4): 534–564. DOI: 10.1177/0741088394011004004

Yeung, Lorrita. 2007. In search of commonalities: Some linguistic and rhetorical features of business reports as a genre. *English for Specific Purposes* 26: 156–179. DOI: 10.1016/j.esp.2006.06.004

Zeiger, Mimi. 1999. *Essentials of Writing Biomedical Research Papers*. New York NY: McGraw Hill.

APPENDIX A

Journals examined during taxonomy development

Discipline	Journal Title	Year
Applied Linguistics	*Text and Talk*	2009
	Applied Linguistics	2006
	Studies in Second Language Acquisition	2006
	TESOL Quarterly	2008
Chemistry	*Inorganic Chemistry*	2007
	Chemistry of Materials	2007
	Journal of the American Chemical Society	2008
	Journal of Chemical Ecology	2007
	Journal of Organic Chemistry	2007
	Macromolecules	2008
Economics	*Quarterly Journal of Economics*	2003
	Economic Inquiry	2003
	Journal of Economic Issues	2003
	Population and Development Review	2005
	Computational Economics	2004
	Econometrica	2004
Engineering (General & Civil)	*Journal of Mechanical Design*	2001
	Journal of Heat Transfer	2007
	Journal of Engineering Materials and Technology	2005
	Journal of Structural Engineering	2003
Geology	*Journal of Sedimentary Research*	2005
	Environmental Geology	2005
Medicine (Pediatrics)	*Developmental Medicine and Child Neurology*	2002
	Clinical Pediatrics	2003

(*Continued*)

Appendix A. (Continued)

Discipline	Journal Title	Year
	Children's Health Care	2003
	Developmental and Behavioral Pediatrics	2004
Philosophy	*American Philosophical Quarterly*	2008
	Mind: A Quarterly Review of Philosophy	2007
	Philosophy and Phenomenological Research	2007
	Journal of Philosophy	2008
	Philosophical Quarterly	2002
Physics	*American Journal of Physics*	2009
	Foundations of Physics	2008
	Applied Physics	2008
Political Science	*Political Theory*	2006
	Journal of Theoretical Politics	2006
	Political Studies	2003
	The Political Quarterly	2003
	Policy and Politics	2009
Psychology	*Journal of Experimental Psychology*	2003
	Journal of Applied Psychology	2003
	Journal of General Psychology	2003
	Developmental Psychology	2003
Sociology	*Journal of Social Policy*	2003
	Community Development Journal	2007
	Theory and Society	2004
	Sociological Perspectives	2003
	Sociological Quarterly	2003
	Qualitative Sociology	2001
	Journal of Sociology	2007

APPENDIX B

Reliability of automatic tags

Linguistic Feature/Tag	Initial Reliability Rates			Final Reliability Rates (after scripts)		
	Precision	Recall	Overall	Precision	Recall	Overall
Prepositions (all)	99%	99%	100%	100%	99%	100%
Pronouns (all)	100%	77%	77%	100%	96%	96%
Adjectives (all)	95%	97%	98%	97%	98%	99%
Adjectives (attributive)	99%	97%	99%	99%	98%	98%
Adjectives (predicative)	98%	94%	96%	99%	95%	96%
Nouns (all)	98%	98%	99%	98%	99%	100%
Adverbs (general)	100%	99%	99%	100%	99%	99%
That-clauses (adjectives)	100%	80%	80%	100%	90%	90%
That-clause (nouns)	83%	71%	86%	90%	85%	95%
That-clauses (verbs)	95%	84%	89%	97%	92%	95%
That relative clauses	83%	73%	88%	91%	89%	98%
To-infinitive clauses	96%	95%	100%	96%	97%	99%
Modal verbs (all)	100%	99%	99%	100%	99%	99%
Base form of *be* as aux verb	98%	98%	100%	99%	99%	100%
Base form of *be* as main verb	98%	98%	100%	99%	99%	100%
Base form of *do* as aux verb	100%	90%	90%	100%	90%	90%

(*Continued*)

Appendix B. (Continued)

Linguistic Feature/Tag	Initial Reliability Rates			Final Reliability Rates (after scripts)		
	Precision	Recall	Overall	Precision	Recall	Overall
Base form of *do* as main verb	79%	100%	73%	72%	100%	62%
Base form of *have* as aux verb	100%	98%	98%	100%	98%	98%
Base form of *have* as main verb	94%	100%	94%	94%	100%	94%
Base form (all verbs)	92%	94%	98%	95%	97%	99%
Past form of *be* as aux verb	99%	97%	98%	99%	98%	98%
Past form of *be* as main verb	96%	99%	97%	97%	99%	98%
Past form of *do* as aux verb	100%	80%	80%	100%	88%	88%
Past form of *do* as main verb	62%	100%	38%	73%	100%	63%
Past form of *have* as aux verb	100%	100%	100%	100%	100%	100%
Past form of *have* as main verb	100%	100%	100%	100%	100%	100%
Past tense verbs	89%	87%	97%	92%	95%	97%
3rd person singular *be* as aux verb	96%	99%	97%	99%	100%	99%
3rd person singular *be* as main verb	99%	99%	100%	100%	100%	100%
3rd person singular *do* as aux verb	100%	74%	74%	100%	90%	90%
3rd person singular *do* as main verb	44%	100%	-25%	60%	100%	33%

(*Continued*)

Appendix B. (Continued)

Linguistic Feature/Tag	Initial Reliability Rates			Final Reliability Rates (after scripts)		
	Precision	Recall	Overall	Precision	Recall	Overall
3rd person singular *have* as aux verb	100%	98%	98%	100%	98%	98%
3rd person singular *have* as main verb	96%	100%	95%	96%	100%	95%
3rd person singular verb	96%	91%	95%	97%	95%	99%
Infinitive verb	97%	90%	93%	98%	94%	95%
wh-relative clauses	99%	97%	98%	99%	97%	98%
wh-questions	100%	100%	100%	100%	100%	100%
Passive voice verb	97%	92%	95%	97%	98%	100%
Perfect aspect verb	100%	98%	98%	100%	98%	98%
Progressive aspect verb	81%	63%	77%	95%	83%	87%
–ing postnominal modifier	47%	87%	14%	53%	91%	27%
–ed postnominal modifier	52%	44%	83%	69%	62%	91%
–ing form of noun	78%	49%	64%	85%	71%	83%
–ing nouns and adjectives (non-verbs)	83%	72%	87%	87%	82%	94%
Articles	100%	100%	100%	100%	100%	100%
Coordinating conjunctions	100%	100%	100%	100%	100%	100%
Ordinal numbers	98%	100%	98%	100%	100%	100%
Cardinal numbers	98%	100%	98%	100%	100%	100%

(*Continued*)

Appendix B. (Continued)

Linguistic Feature/Tag	Initial Reliability Rates			Final Reliability Rates (after scripts)		
	Precision	Recall	Overall	Precision	Recall	Overall
Existential there	100%	100%	100%	100%	100%	100%
Qualifiers	100%	99%	99%	100%	99%	99%
Conditional subordinating conjunctions	100%	100%	100%	100%	100%	100%
Concessive subordinating conjunctions	100%	100%	100%	100%	100%	100%
Causative subordinating conjunctions	100%	100%	100%	100%	100%	100%
Demonstrative pronouns	80%	64%	80%	94%	94%	100%
Demonstrative determiners	74%	98%	67%	90%	98%	91%

APPENDIX C

Semantic classes of nouns, verbs, and adjectives

Table C1. Semantic classes of nouns (see Biber 2006)

Cognition Nouns

ability, analysis, assessment, assumption, attention, attitude, belief, calculation, concentration, concept, concern, conclusion, consciousness, consequence, consideration, decision, desire, emotion, evaluation, examination, expectation, experience, fact, feeling, hypothesis, idea, judgment, knowledge, look, memory, need, notion, observation, opinion, perception, perspective, possibility, probability, reason, recognition, relation, responsibility, sense, theory, thought, understanding, view

Group Nouns

airline, bank, church, college, colony, committee, community, company, congress, firm, flight, government, home, hospital, hotel, house, household, institute, institution, lab, laboratory, school, university

Animate Nouns

American, Indian, accountant, adult, adviser, adviser, agent, aide, ancestor, animal, anthropologist, applicant, archaeologist, artist, artiste, assistant, associate, attorney, audience, auditor, author, baby, bachelor, bird, boss, boy, brother, Buddha, buyer, candidate, cat, child, citizen, client, colleague, collector, competitor, consumer, counselor, couple, critic, customer, daughter, dean, deer, defendant, designer, developer, director, doctor, dog, dr., driver, economist, employee, employer, engineer, engineer, executive, expert, faculty, family, farmer, father, female, feminist, freshman, friend, geologist, girl, god, graduate, guy, hero, historian, host, hunter, husband, immigrant, individual, infant, instructor, investor, Jew, judge, kid, king, lady, lawyer, leader, learner, listener, maker, male, man, manager, manufacturer, member, miller, minister, mom, monitor, monkey, mother, Mr., neighbor, observer, officer, official, owner, parent, participant, partner, patient, peer, people, person, personnel, physician, plaintiff, player, poet, police, president, processor, professional, professor, provider, psychologist, reader, researcher, resident, respondent, schizophrenic, scholar, scientist, secretary, server, shareholder, Sikh, sister, slave, son, speaker, species, spouse, student, supervisor, supplier, teacher, theorist, tourist, undergraduate, user, victim, wife, woman, worker, writer

Technical Nouns

angle, atom, bacteria, bill, carbon, cell, center, chapter, chromosome, circle, cloud, component, compound, data, diagram, DNA, electron, element, equation, exam, fire, formula, gene, graph, hydrogen, internet, ion, iron, isotope, jury, layer, lead, letter, light, list, margin, mark, matter, message, mineral, mineral, molecule, neuron, nuclei, nucleus, organism, oxygen, page, paragraph, particle, play, poem, proton, ray, sample, schedule, sentence, software, solution, square, star, statement, thesis, unit, unit, virus, wave, web, word

(*Continued*)

Table C1. (Continued) Semantic classes of nouns (see Biber 2006)

Other Abstract Nouns

absence, account, action, address, advantage, aid, alternative, aspect, authority, axis, background, balance, base, beginning, benefit, bias, bond, capital, care, career, cause, characteristic, charge, check, choice, circuit, circumstance, climate, code, color, column, combination, complex, condition, connection, constant, constraint, contact, content, context, contract, contrast, crime, criteria, cross, culture, current, curriculum, curve, debt, density, design, detail, dimension, direction, disorder, diversity, economy, emergency, emphasis, employment, end, equilibrium, equity, error, expense, facility, factor, failure, fallacy, feature, format, freedom, fun, gender, goal, grade, grammar, health, heat, help, identity, image, impact, importance, influence, information, input, interest, issue, job, kind, labor, language, law, leadership, level, life, link, manner, math, matrix, meaning, model, music, name, nature, network, objective, opportunity, option, order, origin, output, past, pattern, phase, philosophy, plan, policy, position, potential, power, prerequisite, presence, pressure, principle, profile, profit, proposal, psychology, quality, quiz, race, reality, relationship, religion, requirement, resource, respect, rest, return, right, risk, role, rule, scene, science, security, series, set, setting, sex, shape, share, show, side, sign, signal, situation, skill, sort, sound, source, spring, stage, standard, start, state, stimulus, strength, stress, structure, style, subject, substance, success, support, survey, symbol, system, topic, track, trait, trouble, truth, type, value, variation, variety, velocity, version, way, whole

Place Nouns

apartment, area, bathroom, bay, bench, bookstore, border, bottom, boundary, building, campus, canyon, cave, city, class, classroom, coast, continent, country, county, court, delta, desert, district, earth, environment, estuary, factory, farm, field, floor, forest, front, ground, habitat, hall, hell, hemisphere, hill, hole, horizon, interior, lake, land, lecture, left, library, location, market, middle, moon, mountain, museum, neighborhood, north, ocean, office, opposite, orbit, orbital, organization, outside, parallel, park, passage, place, planet, pool, prison, property, region, residence, restaurant, river, road, room, sector, shaft, shop, southwest, station, store, stream, territory, top, valley, village

Process Nouns

accounting, achievement, act, action, activity, addition, administration, admission, agreement, answer, application, approach, argument, arrangement, assignment, attempt, attendance, birth, break, change, claim, comment, comparison, competition, conflict, construction, consumption, contribution, control, counseling, criticism, deal, death, debate, definition, demand, description, development, discrimination, discussion, distribution, division, education, effect, eruption, evolution, exchange, exercise, experiment, explanation, expression, flow, formation, function, generation, graduation, management, marketing, marriage, mechanism, meeting, method, operation, orientation, performance, practice, presentation, procedure, process, production, progress, question, reaction, registration, regulation, research, result, revolution, selection, service, session, strategy, study, talk, task, teaching, technique, test, trade, tradition, training, transfer, transition, treatment, trial, use, war, work

Quantity Nouns

age, amount, century, cycle, date, day, energy, frequency, future, half, heat, height, hour, length, lot, measure, meter, mile, minute, moment, month, morning, part, per, percent, percentage, period, portion, quantity, quarter, rate, ratio, second, section, semester, summer, temperature, term, time, today, volt, voltage, volume, week, weekend, weight

(*Continued*)

Table C1. (Continued) Semantic classes of nouns (see Biber 2006)

Concrete Nouns

acid, alcohol, aluminum, arm, artifact, asteroid, automobile, award, bag, ball, banana, band, bar, basin, bed, bell, belt, block, board, boat, body, bone, book, box, brain, branch, bubble, bud, bulb, bulletin, button, cake, camera, cap, car, card, case, cent, chain, chair, chart, clay, clock, clothing, club, comet, computer, copper, copy, counter, cover, crop, crystal, cylinder, deposit, desk, device, dinner, disk, document, dollar, door, dot, drain, drawing, drink, drop, drug, dust, edge, engine, envelope, equipment, eye, face, fiber, fig, file, film, filter, finger, fish, flower, food, foot, frame, fruit, furniture, game, gap, gate, gel, gift, glacier, grain, gun, hair, hand, handbook, handout, head, heart, ice, instrument, item, journal, key, knot, lava, leaf, leg, lemon, liquid, load, machine, magazine, magnet, mail, manual, map, marker, match, metal, mixture, modem, mole, motor, mound, mouth, movie, mud, muscle, mushroom, nail, newspaper, node, note, notice, novel, oak, object, package, page, paper, peak, pen, pencil, phone, picture, pie, piece, pipe, plant, plate, pole, portrait, post, pot, pottery, radio, rain, reactor, resistor, retina, ridge, ring, ripple, rock, root, salt, sand, score, screen, sculpture, seat, seawater, sediment, seed, sheet, shell, ship, silica, slide, slope, snow, sodium, soil, solid, solution, space, sphere, spot, statue, steam, steel, stem, step, stick, stone, strata, string, sugar, syllabus, table, tank, tape, target, telephone, telescope, textbook, ticket, tip, tissue, tool, tooth, train, transcript, transistor, tree, truck, tube, vehicle, vessel, video, visa, wall, water, water, wheel, window, wire, wood

Table C2. Semantic classes of verbs (see Biber 2006)

Activity Verbs

accompany, acquire, add, advance, apply, arrange, beat, behave, borrow, bring, burn, buy, carry, catch, check, clear, climb, combine, come, control, cover, defend, deliver, dig, divide, earn, eat, encounter, engage, exercise, expand, explore, fix, form, get, give, go, hang, hold, left, lie, lose, made, meet, move, obtain, obtain, open, pay, pick, play, produce, provide, pull, put, react, receive, reduce, repeat, run, save, sell, send, shake, share, show, sit, smile, smile, spend, stare, take, throw, try, turn, use, visit, wait, walk, watch, wear, win, work

Aspectual Verbs

begin, cease, complete, continue, end, finish, keep, start, stop

Causative Verbs

affect, allow, assist, cause, enable, ensure, force, guarantee, help, influence, let, permit, prevent, require

Communication Verbs

accuse, acknowledge, address, advise, announce, answer, appeal, argue, ask, assure, challenge, claim, complain, consult, convince, declare, demand, deny, describe, discuss, emphasize, encourage, excuse, explain, express, inform, insist, invite, mention, offer, offer, persuade, phone, pray, promise, propose, question, quote, recommend, remark, reply, response, say, shout, sign, sing, speak, specify, state, suggest, swear, teach, tell, thank, threaten, urge, warn, welcome, whisper, write

Existence Verbs

appear, concern, constitute, contain, define, derive, deserve, exist, fit, illustrate, imply, include, indicate, involve, lack, live, look, matter, owe, possess, reflect, relate, remain, represent, reveal, see, sound, stand, stay, suit, tend, vary

(*Continued*)

Table C2. (Continued) Semantic classes of verbs (see Biber 2006)

Mental Verbs

accept, afford, agree, appreciate, approve, assess, assume, bear, believe, blame, bother, calculate conclude, care, celebrate, compare, confirm, consider, count, dare, decide, deserve, detect, determine, discover, dismiss, distinguish, doubt, enjoy, examine, expect, experience, face, fear, feel, find, forget, forgive, guess, hate, hear, hope, identify, ignore, imagine, impress, intend, interpret, judge, justify, know, learn, like, listen, love, mean, mind, miss, need, notice, observe, perceive, plan, predict, pretend, prove, read, realize, recall, reckon, recognize, regard, remember, remind, satisfy, see, solve, study, suffer, suppose, suspect, think, trust, understand, want, wonder, worry

Occurrence Verbs

arise, become, change, develop, die, disappear, emerge, fall, flow, grow, happen, increase, last, occur, rise, shine, sink, slip

Table C3. Semantic classes of adjectives (see Biber 2006)

Evaluative Adjectives

best, good, important, nice, right, simple, special

Relational Adjectives

basic, common, different, final, following, full, general, higher, individual, lower, main, major, particular, same, similar, specific, total, various, whole

Size Adjectives

big, great, high, large, little, long, low, small

Topical Adjectives

economic, human, international, local, national, natural, normal, oral, physical, political, public, public, sexual, social

APPENDIX D

Full factorial structure matrix for the four-factor solution

Pattern Matrix[a]

	Factor			
	1	2	3	4
type-token	–.351	.308	–.205	–.091
wrd_len	–.313	.520	–.264	.580
wrd_cnt	.203	.361	–.070	–.167
pro_2p	.234	.215	.401	–.154
pro_dem	.517	–.176	.068	–.036
pro_1p	.578	–.183	.148	–.138
pro_it	.615	.284	–.104	–.198
v_be	.787	–.037	.068	.098
pro_nom	.690	.063	.068	–.103
wh_q	.241	.318	.273	–.084
mod_poss	.655	–.151	.105	.090
wh_clause	.335	.233	.328	.049
NOUN	–.748	–.073	–.084	.253
PREP	–.394	.055	–.354	–.188
adj_attrb	–.122	.152	–.539	.268
v_past	–.672	.552	.224	–.245
pro_3p	–.127	.645	.346	–.203
v_perf	–.109	.483	–.085	–.211
NOM	–.190	.323	–.218	.433
adv_time	.344	.081	–.023	–.473
adv	.535	.212	–.306	–.036
mod_pred	.694	.200	–.094	–.107

(*Continued*)

Pattern Matrix (Continued)

	Factor			
	1	2	3	4
mod_nec	.645	.083	.024	–.015
adv_conjunct	.475	.020	–.245	.063
v_pass_agless	–.317	–.515	.265	.023
v_pass_by	–.090	–.470	–.098	–.222
pass_postnom	–.533	–.115	.198	.129
adj_pred	.699	–.296	.082	.253
v_have	.673	.043	–.025	.171
v_presprog	.019	.424	.489	.175
tht_rel	.278	.464	–.033	.139
vcmp_nonfact	.290	.450	.181	.078
vcmp_fact	.048	–.267	.416	.023
vcmp_like	.588	–.006	.185	.173
to_ncmp_stance	.239	.346	–.037	.029
n_animate	–.145	.434	.138	–.033
n_process	–.403	–.024	.499	.384
n_cog	.566	.079	.125	.234
n_other-abstract	.140	–.060	–.102	.322
n_concrete	–.167	–.373	.095	–.054
n_technical	–.183	–.605	.176	–.035
n_quant	–.172	–.464	–.030	–.092
n_group	–.183	.486	–.105	–.019
adj_attrb_size	–.131	–.374	–.047	–.115
adj_attrb_time	–.223	.472	–.112	–.122
adj_attrb_eval	.329	–.011	–.067	.046
adj_attrb_relat	–.057	–.103	.172	.451
adj_attrb_topic	.094	.533	–.418	.145
v_activity	–.246	–.016	.601	–.034
v_comm	.082	.469	.509	–.028
v_mental	.240	–.081	.650	.195
v_caus	.340	.203	.029	.153
v_exist	.236	.000	.103	.366
v_aspect	–.224	.524	.190	–.281

(*Continued*)

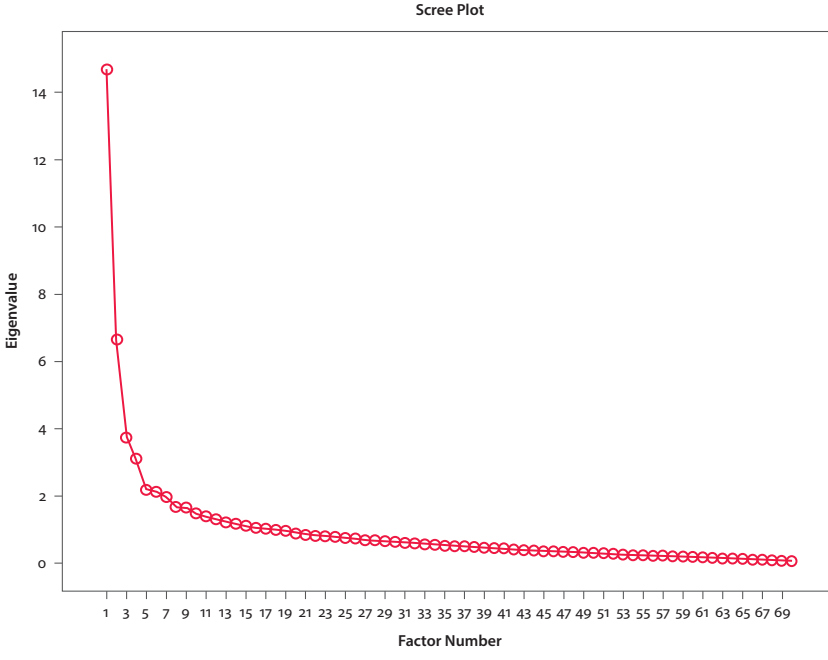
Pattern Matrix (Continued)

	Factor			
	1	**2**	**3**	**4**
conj_coor_clsl	.247	.351	.019	–.070
conj_coor_phrsl	–.184	.513	.020	.168
conj_sub_cond	.833	–.007	.025	–.058
conj_sub_other	.388	.192	.092	–.100
jcmp_fact	.482	.088	.108	–.091
jcmp_like	.259	.066	–.044	.382
ncmp_att	.466	.126	–.077	.051
ncmp_fact	.441	–.145	–.141	.019
ncmp_like	.654	.092	–.054	.062
jcmp_att	.197	.085	.157	.087
to_vcmp_speech	–.115	.266	.450	.040
to_vcmp_desire	.025	.414	.459	.078
to_vcmp_mod	.109	.574	.204	–.051
to_vcmp_prob	.321	.267	–.040	.126
SUM_to_jcmp_stance	.369	.173	.101	.329
SUM_adv_stance	.466	.258	–.115	–.242

Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

APPENDIX E

Scree plot of the four-factor solution



APPENDIX F

Significance testing for four-factor solution

Table F1. Means and standard deviations for dimension scores by discipline and register

Register		dim1	dim2	dim3	dim4
PHIL-TH	Mean	41.6655	6.6877	4.8846	–1.7465
	Std. Deviation	21.54724	7.53429	10.33706	4.54761
HIST-GEN	Mean	–6.2142	12.5904	–2.0584	–4.7721
	Std. Deviation	7.37961	6.60724	4.44655	2.47817
AL-QL	Mean	.1404	11.9700	10.6548	1.7363
	Std. Deviation	8.12020	6.82181	6.67208	2.69612
AL-QT	Mean	–6.0045	2.6060	5.2931	3.1945
	Std. Deviation	10.05967	6.54249	5.79253	2.64792
POLISCI-QL	Mean	–4.4330	12.4243	–3.9305	1.4721
	Std. Deviation	8.47495	5.28294	4.89765	3.63203
POLISCI-QT	Mean	2.1027	2.6227	–3.3812	5.2210
	Std. Deviation	9.84931	5.78123	4.46606	5.07216
BIO-QT	Mean	–15.9337	–10.4463	–4.9813	–.7671
	Std. Deviation	5.07284	5.47525	4.54063	3.17191
PHYS-QT	Mean	–9.7068	–19.8076	–3.5697	–2.2899
	Std. Deviation	10.31595	4.34117	2.96253	3.16465
PHYS-TH	Mean	–1.6164	–18.6472	–2.9114	–2.0484
	Std. Deviation	10.96577	5.58895	3.17187	2.96420
Total	Mean	.0000	.0000	.0000	.0000
	Std. Deviation	19.05492	13.70370	7.62285	4.52687

Table F2. ANOVA table for all four dimensions

		ANOVA				
		Sum of Squares	**df**	**Mean Square**	**F**	**Sig.**
dim1	Between Groups	65564.964	8	8195.621	66.624	.000
	Within Groups	32106.226	261	123.012		
	Total	97671.190	269			
dim2	Between Groups	40912.370	8	5114.046	138.987	.000
	Within Groups	9603.499	261	36.795		
	Total	50515.869	269			
dim3	Between Groups	7276.561	8	909.570	28.416	.000
	Within Groups	8354.436	261	32.009		
	Total	15630.997	269			
dim4	Between Groups	2354.885	8	294.361	24.331	.000
	Within Groups	3157.622	261	12.098		
	Total	5512.507	269			

Table F3. Assumption of normal distribution testing, showing that all dimension scores are normally distributed (α =.001)

	Register	**Shapiro–Wilk**		
		Statistic	**df**	**Sig.**
dim1	PHIL-TH	.975	30	.689
	HIST-GEN	.959	30	.294
	AL-QL	.960	30	.316
	AL-QT	.916	30	.021
	POLISCI-QL	.963	30	.375
	POLISCI-QT	.939	30	.084
	BIO-QT	.943	30	.111
	PHYS-QT	.970	30	.534
	PHYS-TH	.955	30	.233
dim2	PHIL-TH	.946	30	.131
	HIST-GEN	.969	30	.505
	AL-QL	.987	30	.969
	AL-QT	.956	30	.238
	POLISCI-QL	.972	30	.594

(*Continued*)

Table F3. (Continued) Assumption of normal distribution testing, showing that all dimension scores are normally distributed (α =.001)

	Register	Shapiro–Wilk		
		Statistic	df	Sig.
	POLISCI-QT	.979	30	.797
	BIO-QT	.980	30	.833
	PHYS-QT	.979	30	.789
	PHYS-TH	.980	30	.831
dim3	PHIL-TH	.901	30	.009
	HIST-GEN	.981	30	.851
	AL-QL	.936	30	.072
	AL-QT	.969	30	.523
	POLISCI-QL	.924	30	.034
	POLISCI-QT	.986	30	.952
	BIO-QT	.957	30	.256
	PHYS-QT	.966	30	.445
	PHYS-TH	.965	30	.410
dim4	PHIL-TH	.976	30	.720
	HIST-GEN	.929	30	.047
	AL-QL	.955	30	.233
	AL-QT	.988	30	.975
	POLISCI-QL	.992	30	.998
	POLISCI-QT	.980	30	.814
	BIO-QT	.975	30	.694
	PHYS-QT	.971	30	.575
	PHYS-TH	.971	30	.573

Table F4. Assumption of homogeneity of variances testing, showing that dimensions 1, 3 and 4 violate the assumption of homogeneity of variance and indicating the use of Games-Howell post-hoc procedures

dim1	Test of Homogeneity of Variances			
	Levene Statistic	df1	df2	Sig.
	11.557	8	261	.000
dim2	.973	8	261	.457
dim3	7.170	8	261	.000
dim4	2.923	8	261	.004

Table F5. Post-hoc comparisons (Games-Howell) for Factor 1. Mean differences marked with * are significant at the p < .05 level

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Philosophy – Theoretical	---								
2. History – Qual	**47.88***	---							
3. Political Science – Qual	**46.10***	–1.78	---						
4. Political Science – Quant	**39.56***	**–8.32***	–6.54	---					
5. Applied Linguistics – Qual	**41.53***	–6.35	–4.57	1.96	---				
6. Applied Linguistics – Quant	**47.67***	–0.21	1.57	8.11	6.14	---			
7. Biology – Quant	**57.60***	**9.72***	**11.50***	**18.04***	**16.07***	**9.93***	---		
8. Physics – Quant	**51.37***	3.49	5.27	**11.81***	**9.85***	3.70	–6.23	---	
9. Physics – Theoretical	**43.28***	–4.60	–2.82	3.72	1.76	–4.39	**–14.32***	–8.09	---

Table F6. Post-hoc comparisons (Games-Howell) for Factor 2. Mean differences marked with * are significant at the p < .05 level

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Philosophy – Theoretical	---								
2. History – Qual	**–5.90***	---							
3. Political Science – Qual	**–5.74***	0.17	---						
4. Political Science – Quant	4.07	**9.97***	**9.80***	---					
5. Applied Linguistics – Qual	–5.28	0.62	0.45	**–9.35***	---				
7. Biology – Quant	**17.13***	**23.04***	**22.87***	**13.07***	**22.42***	**13.05***	---		
8. Physics – Quant	**26.50***	**32.40***	**32.23***	**22.43***	**31.78***	**22.41***	**9.36***	---	
9. Physics – Theoretical	**25.33***	**31.24***	**31.07***	**21.27***	**30.62***	**21.25***	**8.20***	–1.16	---

Table F7. Post-hoc comparisons (Games-Howell) for Factor 3. Mean differences marked with * are significant at the p < .05 level

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Philosophy – Theoretical	---								
2. History – Qual	**6.94***	---							
3. Political Science – Qual	**8.82***	1.87	---						
4. Political Science – Quant	**8.27***	1.32	−0.55	---					
5. Applied Linguistics – Qual	−5.77	**−12.71***	**−14.59***	**−14.04***	---				
6. Applied Linguistics – Quant	−0.41	**−7.35***	**−9.22***	**−8.67***	**5.36***	---			
7. Biology – Quant	**9.87***	2.92	1.05	1.60	**15.64***	**10.27***	---		
8. Physics – Quant	**8.45***	1.51	−0.36	0.19	**14.22***	**8.86***	−1.41	---	
9. Physics – Theoretical	**7.80***	0.85	−1.02	−0.47	**13.57***	**8.20***	−2.07	−0.66	---

Table F8. Post-hoc comparisons (Games-Howell) for Factor 4. Mean differences marked with * are significant at the p < .05 level

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Philosophy – Theoretical	---								
2. History – Qual	3.03	---							
3. Political Science – Qual	−3.22	**−6.24***	---						
4. Political Science – Quant	**−6.97***	**−9.99***	**−3.75***	---					
5. Applied Linguistics – Qual	**−3.48***	**−6.51***	−0.26	**3.48***	---				
6. Applied Linguistics -Quant	**−4.94***	**−7.97***	−1.72	2.03	−1.46	---			
7. Biology – Quant	−0.98	**−4.00***	2.24	**5.99***	**2.55***	**3.96***	---		
8. Physics – Quant	0.54	**−2.48***	**3.76***	**7.51***	**4.03***	**5.48***	1.52	---	
9. Physics – Theoretical	0.30	**−2.72***	**3.52***	**7.27***	**3.78***	**5.24***	1.28	−0.24	---

Index