Corpus-based Research in Applied Linguistics

Studies in Honor of Doug Biber

EDITED BY
Viviana Cortes
Eniko Csomay

Studies in Corpus Linguistics 66

JOHN BENJAMINS PUBLISHING COMPANY

Corpus-based Research in Applied Linguistics

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/scl*

Volume 66

Corpus-based Research in Applied Linguistics. Studies in Honor of Doug Biber
Edited by Viviana Cortes and Eniko Csomay

Corpus-based Research in Applied Linguistics

Studies in Honor of Doug Biber

Edited by

Viviana Cortes
Georgia State University

Eniko Csomay
San Diego 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.66

Cataloging-in-Publication Data available from Library of Congress:
LCCN 2014040088 (PRINT)

ISBN 978 90 272 0374 8 (HB)
ISBN 978 90 272 6905 8 (E-BOOK)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

Table of content

List of contributors vii

Foreword ix

INTRODUCTION
Douglas Biber and the Flagstaff School of corpus-based research: An introduction xv
Viviana Cortes & Eniko Csomay

CHAPTER 1
A corpus-based analysis of linguistic variation in teacher and student presentations in university settings 1
Eniko Csomay

CHAPTER 2
Telephone interactions: A multidimensional comparison 25
Eric Friginal

CHAPTER 3
On the complexity of academic writing: Disciplinary variation and structural complexity 49
Bethany Gray

CHAPTER 4
Telling by omission: Hedging and negative evaluation in academic recommendation letters 79
Mohammed Albakry

CHAPTER 5
Corpora, context, and language teachers: Teacher involvement in a local learner corpus project 99
Alfredo Urzúa

CHAPTER 6
The challenge of constructing a reliable word list: An exploratory corpus-based analysis of lexical variability in introductory Psychology textbooks 123
Don Miller

CHAPTER 7
Corpus linguistics and New Englishes 147
Chandrika Balasubramanian

CHAPTER 8
Investigating textual borrowing in academic discourse:
The need for a corpus-based approach 177
Casey Keck

CHAPTER 9
Situating lexical bundles in the formulaic language spectrum:
Origins and functional analysis developments 197
Viviana Cortes

Index 217

List of contributors

Mohammed Albakry	Middle Tennessee State University & University of Connecticut, United States of America
Chandrika Balasubramanian	Sultan Qaboos University, Oman
Viviana Cortes	Georgia State University, United States of America
Eniko Csomay	San Diego State University, United States of America
Eric Friginal	Georgia State University, United States of America
Bethany Gray	Iowa State University, United States of America
Casey Keck	Boise State University, United States of America
Michael McCarthy	University of Nottingham, United Kingdom
Don Miller	California State University, Stanislaus, United States of America
Alfredo Urzua	San Diego State University, United States of America

Foreword

This book represents the work of a quite outstanding generation of corpus linguists who all, in some degree or another, owe their present successes to their apprenticeship with Douglas Biber and his colleagues at Northern Arizona University, Flagstaff. Few centres of corpus linguistic study can lay claim to such a productive and internationally respected body of research and publications, to which the present volume is an important contribution.

I first became acquainted with Douglas Biber and his colleagues in the mid-1990s, through meetings at conferences in North America, where I and my British colleagues from the University of Nottingham (Ronald Carter and Rebecca Hughes) were attempting to disseminate our spoken corpus research to applied linguists and language teachers, in the unshaken belief that we had something new to say. It was perhaps a sign of those times that the Northern Arizona team and the Nottingham team huddled in corners pouring out mutual sympathy at the polite but under-whelming reception our ideas on things like spoken grammar and register differentiation received from teachers' conference audiences and publishers. Not that our efforts to disseminate information about spoken corpora and register differentiation went unchallenged. A healthy academic debate ensued, raising questions as to the viability of interpreting spoken data with only limited access to the highly localised factors that lead to the choices that characterise individual registers when the researcher, typically an outsider to the context, has nought but the textual record to examine, albeit in large datasets (Widdowson 2000; Carter 1998; Cook 1998).

Now, almost 20 years on, we can happily survey a landscape in which, broadly speaking, the basic battles have been won. Applied corpus linguistics has come of age, a generation of graduate students who cut their teeth on corpus analysis (see the list of present authors) are now stepping into the shoes vacated by old codgers like me, register studies based on smaller, targeted corpora are numerous and English language teaching materials are now routinely informed by corpus insights, spoken and written, both in the general English domain as well as in specialist areas such as academic speaking and business English. Römer (2008) provides an excellent survey of both direct and indirect influences of corpora on language teaching; that influence has clearly been extensive and continues to grow. What is more, the relationship between corpora and language pedagogy has become more rewardingly two-way, with corpus linguists offering insights into

language use that are seen as applicable in language teaching, and more and more language teachers becoming aware of corpora, either themselves becoming corpus users or else posing questions that prompt answers from corpus linguists. Elsewhere, I have made a call for the greater integration of corpus linguistics in teacher education programmes as a natural corollary to its growing impact on the language teaching profession (McCarthy 2008).

While on this side of the Atlantic during the 1990s and early 2000s corpus teams in the UK and mainland Europe developed their interests in areas such as spoken grammar and lexis (Svartvik 1992; Carter & McCarthy 1995; McCarthy & Carter 1997; Leech 2000; Leech et al. 2001) and, increasingly, in discourse and pragmatics (Aijmer 2002; Rühlemann 2007, 2010; Romero-Trillo 2008), in the USA it was Douglas Biber and his colleagues who were making significant progress in a wide range of corpus investigations. The special contribution of the Northern Arizona school, building on Biber's own work of the late 1980s and 1990s, was vastly to enlarge the study of register and variation using corpus-analytical techniques, a valuable counterbalance to the large-scale, homogenising lexicographical tradition that had given birth to a generation of corpus-based learners' dictionaries from the mid-1980s onwards. It is the fruits of this expansion of the study of register differentiation that the present volume represents, and its authors are worthy standard-bearers of the Flagstaff tradition.

Biber (1988), a seminal work in register analysis, showed just how complex variation can be across speaking and writing and how a simple binary division between spoken and written language is inadequate. Biber showed that multivariate analysis, judiciously applied, could exploit the key linguistic and situational factors which, taken together, account for variation among registers. Biber's work transformed the two-dimensional practice of comparing globally-based spoken and written frequency counts into a multi-dimensional prism through which to view and compare different manifestations of language use, filtered through robust statistical processing. In this work and his subsequent book on register variation (Biber 1995), there is also a clear emphasis on painstaking and principled data compilation that has become a central tenet of the legacy inherited by Flagstaff graduates, as well as the fundamental insistence upon seeing language as situated, variable and multi-levelled. Further evidence of this preference for situationally-sourced statements about language came in the massive, register-sensitive Longman English reference grammar (Biber et al. 1999) and Biber and his associates' further work on academic registers (Biber et al. 2004; Biber 2006; Biber & Gray 2010).

The growth in the international reputation of the Northern Arizona corpus linguists deservedly has come about not just through the insights they have given us into practically-grounded registers such as academic English or the language of

the media, but also because of their passionate dedication to improving the tools and methods of corpus analysis. As well as continuing to champion multivariate analysis, the Flagstaff linguists have been among the vanguard in the focus on multi-word strings (generally referred to as lexical bundles in the Flagstaff tradition) as well as single words which has now become such a taken-for-granted method in corpus investigations (see, for example, Biber et al. 2004; Biber & Barbieri 2007). Traditional studies of phraseology mainly relied on salience of expressions for their classification and typically used morpho-syntactic or lexical-semantic criteria for their identification. The Flagstaff researchers showed how even apparently fragmentary strings possessed important discourse-organising functions; thus a string such as "if you look at …", although syntactically and semantically incomplete, is shown by corpus analysis to be highly frequent in academic contexts and to function as an important focusing and topic-directing device (Biber et al. 2004). But not content simply to use existing proprietary corpus-analytical tools, the Northern Arizona team have also developed new algorithms for at least partially automating some of the processes immanent in the retrieval of such language patterns in corpus data. A further, key element of the Flagstaff tradition is the constant interplay between using discrete linguistic items and features to elucidate registers and, *pari passu*, using different registers to elucidate individual linguistic items and features. Both tendencies are on display in the present volume.

There is an understandable temptation to see corpus linguistics as an area of study dominated by quantitative analysis, its insights equated with the soulless numbers spewed out from the dispassionate maw of the computer. Yet it goes without saying, plausible interpretation and qualitative judgments informed by the statistical data are the ultimate test of the worth of any applied corpus linguistic enterprise, and the present volume shows this kind of interpretation at its best. Nonetheless, the ideal balance between quantitative and qualitative approaches to corpus linguistics is a continuing source of debate. Gries (2010), for example, takes a strongly quantitative approach to lexical semantics (see also Gries 2009), while Jones (2002), also investigating lexical semantics, adopts a more qualitative approach to his corpus data. McEnery (2001: 76–77) provides a brief summary of the tension between the two approaches to corpus linguistics.

There can be no doubt that, even with the sophistication of current corpus-analytical software, the most banal and everyday speech acts are difficult to retrieve automatically from a corpus. Investigations of phenomena such as politeness, complimenting and apologising (e.g. Holmes 2013) require immense amounts of manual sifting and interpretation alongside automated counting of discrete linguistic items and features. This is the result of the lack of a one-to-one correspondence between linguistic forms and pragmatic functions. What the present volume shows is that pragmatic features of interaction such as politeness and

respect, while not ultimately amenable to mere number-crunching, can nonetheless be powerfully spotlighted by the kind of multivariate analysis the Flagstaff school has made its own. Additionally, areas of investigation largely identified with (oftentimes controversial) qualitative interpretation (e.g. critical discourse analysis) are shown to be extensively ratified and underpinned by the addition of empirical evidence (see also Baker et al. 2008).

One of the more productive features of the qualitative-quantitative dialectic is that frameworks derived from non-corpus-based investigations have been usefully applied to corpus data and tested in the furnace of large datasets. However, qualitative analysis is not merely a question of inference and interpretation by the researcher. In stronger versions of qualitative research, importance is also placed on the inter-subjectivity provided by informant data and 'insider' insights, in addition to robust contextualisation and knowing what kinds of questions with which to interrogate the data (Chafe 1992). This type of triangulation of data is not one always associated with corpus linguistics but is manifested in an exemplary fashion in Handford's (2010) corpus-based study of business English. The growing symbiosis between quantitative and qualitative methods is evident in the present volume.

The tradition carved out by Biber (1988, 1995) and Biber et al. (1999) places value on the observation of language in individual registers. However, there is always the problem of whether macro-registers (e.g. the four major registers of *conversation*, *fiction*, *news* and *academic* adduced in the 1999 Longman grammar) are too blunt an instrument. The Flagstaff school has not lost sight of this concern and its adherents judiciously sub-divide registers such as academic writing into various sub-kinds right down to individual elements of textual artefacts such as sections of research articles. In this respect, the Flagstaff researchers continue the genre-based research in similar areas within the ESP tradition as pioneered by works such as Swales (1990). The term *genre*, in its ESP/EAP incarnation, overlaps with the Flagstaff notion of register but genre suggests a much more convention-oriented view of language use, where texts have a "predictable structure" and "linguistic regularities" (Dudley-Evans & St John 1998: xv), while register suggests a more open-ended manifestation of situated choices sensitive to users and dynamic contexts. Register, in the sense laid out by Halliday (1978) represents choices from an interacting set of repertoires accounting for field, tenor and mode of communication, a conception that would seem to sit better with the multivariate descriptions emanating from the Flagstaff school. But I would not wish to dwell here on issues of terminology: what the Flagstaff corpus linguists have done, and continue to do, sits alongside, not in competition with, genre analysis and non-corpus-based discourse- and conversation-analysis in steadily taking steps forward on the long journey towards ever greater understanding of variation in

language use. Way back, in a seminal article that should be compulsory reading for any student of language use, Mitchell (1957) showed how spoken language was sensitive to numerous situational influences. He had no computer to help him tease out how the many separate factors played simultaneously and harmoniously to create the spoken artefact, but were he engaged in research today, he would undoubtedly feel comfortably at home working on corpora with Douglas Biber.

Michael McCarthy
Cambridge, UK

References

Aijmer, Karen. 2002. *English Discourse Particles. Evidence from a Corpus* [Studies in Corpus Linguistics 10]. Amsterdam: John Benjamins. DOI: 10.1075/scl.10

Baker, Paul, Gabrielatos, Costas, Khosravinik, Majid, Krzyżanowski, Michael, McEnery, Tony & Wodak, Ruth. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19(3): 273–306. DOI: 10.1177/0957926508088962

Biber, Douglas. 1988. *Variation across Speech and Writing,* Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 1995. *Dimensions of Register Variation*. Cambridge: CUP. DOI: 10.1017/CBO9780511519871

Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. DOI: 10.1075/scl.23

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Eward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas, Conrad, Susan & Cortes, Viviana. 2004. *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405. DOI: 10.1093/applin/25.3.371

Biber, Douglas & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–86. DOI: 10.1016/j.esp.2006.08.003

Biber, Douglas & Gray, Bethany. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes* 9(1): 2–20. DOI: 10.1016/j.jeap.2010.01.001

Carter, Ronald. A. 1998. Orders of reality: CANCODE, communication and culture. *ELT Journal* 52(1): 43–56. DOI: 10.1093/elt/52.1.43

Carter, Ronald & McCarthy, Michael. 1995. Grammar and the spoken language. *Applied Linguistics* 16(2): 141–158. DOI: 10.1093/applin/16.2.141

Chafe, William. 1992. The importance of corpus linguistics to understanding the nature of language. In *Directions in Corpus Linguistics*, Jan Svartvik (ed.), 79–97. Berlin: Mouton de Gruyter.

Cook, Guy. 1998. The uses of reality: A reply to Ronald Carter. *ELT Journal* 52(1): 57–64. DOI: 10.1093/elt/52.1.57

Dudley-Evans, Tony & St John, Maggi Jo. 1998. *Developments in English for Specific Purposes: A Multi-Disciplinary Approach*. Cambridge: CUP.

Gries, Stefan. 2009. *Quantitative Corpus Linguistics with R. A Practical Introduction*. New York NY: Routledge. DOI: 10.1515/9783110216042

Gries, Stefan. 2010. Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics. *The Mental Lexicon* 5(3): 323–346. DOI: 10.1075/ml.5.3.04gri

Halliday, Michael A.K. 1978. *Language as Social Semiotic*. London: Edward Arnold.

Handford, Michael. 2010. *The Language of Business Meetings*. Cambridge: CUP. DOI: 10.1017/CBO9781139525329

Holmes, Janet. 2013. *Women, Men and Politeness*. New York NY: Routledge.

Jones, Steven. 2002. *Antonymy: A Corpus-based Perspective*. New York: Routledge.

Leech, Geoffrey. 2000. Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724. DOI: 10.1111/0023-8333.00143

Leech, Geoffrey, Rayson, Paul & Wilson, Andrew. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus.* London: Longman.

McCarthy, Michael. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41(4): 563–574. DOI: 10.1017/S0261444808005247

McCarthy, Michael & Carter, Ronald. 1997. Written and spoken vocabulary. In *Vocabulary: Description, Acquisition, Pedagogy,* Norbert Schmitt & Michael McCarthy (eds), 20–39. Cambridge: CUP.

McEnery, Tony. 2001. *Corpus Linguistics: An Introduction*. Edinburgh: EUP.

Mitchell, Terrence Frederick. 1957. The language of buying and selling in Cyrenaica: A situational statement. *Hespéris* XLIV: 31–71.

Römer, Ute. 2008. Corpora and language teaching. In *Corpus Linguistics. An International Handbook,* Vol. 1, Anke Lüdeling & Kytö Merja (eds), 112–130. Berlin: Mouton de Gruyter.

Romero-Trillo, Jesús. 2008. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. Berlin: Mouton de Gruyter. DOI: 10.1515/9783110199024

Rühlemann, Christoph. 2007. *Conversation in Context: A Corpus-Driven Approach*. London: Continuum.

Rühlemann, Christoph. 2010. What can a corpus tell us about pragmatics? In *The Routledge Handbook of Corpus Linguistics,* Anne O'Keeffe & Michael McCarthy (eds), 288–301. New York NY: Routledge.

Svartvik, Jan. 1992. Lexis in English language corpora. In *Euralex '92 Proceedings,* Hannu Tomola, Krista Varantola, Tarja Salmi-Tolonen, & Jurgen Schopp (eds), 17–31. Tampere: University of Tampere.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: CUP.

Widdowson, Henry. 2000. On the limitations of applied linguistics. *Applied Linguistics* 21(1): 2–25. DOI: 10.1093/applin/21.1.3

INTRODUCTION

Douglas Biber and the Flagstaff School of corpus-based research

An introduction

Viviana Cortes & Eniko Csomay
Georgia State University / San Diego State University

The collection of essays in this book is a tribute to Professor Douglas Biber as our teacher and mentor, and in homage to the legacy of his teachings and research at Northern Arizona University (NAU, henceforth). As his disciples, we would like to call the space in which he taught and we learnt the tricks of the trade the Flagstaff School of Corpus-based Research. Undeniably, Prof. Biber and his work throughout his tenure at NAU have inspired many students, which resulted in a range of innovative corpus-based investigations.

Prof. Biber left his mark in each and every student who studied with him at NAU, inflicting his teachings with what we now recognize as the fundamentals of systematic and principled corpus linguistic research. These fundamentals are the pillars of the Flagstaff School, and they are the ones that distinguish it from any other existing or potential programs which are simply "doing" corpus-based research. In the Flagstaff School, we learnt how to:

1. Design and implement empirically-driven corpus-based research, paying close attention to the principles of Biber's definition of corpus size, representativeness, sampling, and above all, to systematic analysis;
2. Actively engage in computer programming, allowing us not only to dare ask but to be able to answer corpus-based research questions never asked before. The reason these questions had not been asked before is that the tools available at the time did not allow the processing of texts to answer those questions. Instead of accepting that fact, however, we were challenged to design and create new tools of our own in order to satisfy our true curiosity and inquiry;

DOI 10.1075/scl.66.004cor

3. Place a strong emphasis on the combination of quantitative methods that are based on sound and innovative statistical procedures, and complemented with comprehensive qualitative functional analyses of language use.

Over the past decades, Biber's work has covered extensive grounds of linguistic inquiry, most prominently focusing on the study of register variation and corpus-based descriptions of grammar. His early work (Biber 1988) fundamentally challenged the traditional views of the dichotomy between speech and writing by providing empirical evidence through systematic descriptions of how, instead, spoken and written registers vary in a multidimensional linguistic space. This comprehensive description of register variation was applied in other languages as well, for example, Somali and Korean (Biber 1995) and most recently Spanish (Biber & Tracy-Ventura 2007), and in exploring the language of specific contexts, such as the university (Biber & Conrad 2009).

A decade after his first book in 1988, the *Longman Grammar of Spoken and Written English* (Biber et al. 1999) provided us with robust, corpus-based descriptions of English grammar never done before, and inspired many subsequent works. In addition, Biber has introduced the principles and techniques of corpus linguistics (Biber et al. 1998) and of corpus-based discourse analysis (Biber et al. 2007).

Each of the nine chapters selected for this volume was written following what he believed to be the main principles to do corpus research. The authors invited to write these chapters were, at some point in the past two decades, Prof. Biber's students in the Flagstaff School. We are now professors at different universities in the United States or other parts of the world, applying Biber's approach to corpus-based research and teachings to the current and future generations of corpus-based researchers. These authors have excelled in various areas of corpus-based research and their chapters represent each of those areas.

The first two chapters apply multi-dimensional approaches to the analysis of specific spoken registers. Eniko Csomay, in Chapter 1, investigates patterns of language use in presentations in the university setting. She focuses on two participants, teachers and students, as they present new information to an audience. She compiled a relatively small corpus of 168 teacher presentation segments in the classroom, and 76 student presentations recorded (and transcribed) at a student research symposium. Student and teacher presentations are then compared based on the dimensions of linguistic variation in university settings (Biber & Conrad 2009). Her findings show differences in language use between these two groups of presenters and these differences are attributed to the social status and the relationship between the speaker and their audience.

In Chapter 2, Eric Friginal uses multi-dimensional approaches to the description of spoken discourse, comparing telephone-based interactions in

three settings: (1) customer service transactions, (2) telephone conversations between friends, and (3) spontaneous telephone exchanges between participants discussing topics identified by fixed prompts. The findings indicate that variation in these interactions is largely influenced by the nature of conversational tasks, participants' roles in the interactions, and the use of the telephone as a medium in communicating ideas, opinions, and instructions.

The next four chapters analyze specific aspects of written discourse. Bethany Gray, in Chapter 3, uses a corpus of 270 (c. 2 million words) research articles as a single register and explores variation within that one register as it relates to the use of epistemic stance markers following the framework described in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). The articles in her corpus showcase three distinctive research types (theoretical, qualitative, and quantitative) and are from six disciplines (Philosophy, History, Political Science, Applied Linguistics, Biology, and Physics). Tracking the lexico-grammatical features of stance with a special computer program she developed, the use of these stance markers are compared across disciplines and research article types.

In Chapter 4, Mohammed Albakry explores some of the linguistic and discursive aspects of framing positive and negative information in recommendation letters, using a corpus of 114 letters of recommendation to an English Ph.D. program. The findings show consistent patterns in the way different types of modals and their associated collocates are used to hedge predictions. In addition, through the analysis, the discursive frames of the most common mitigation strategies in presenting potentially negative information about applicants also become apparent.

In Chapter 5, Alfredo Urzua challenges the misconception that corpus linguistics relates to the de-contextualized nature of corpus data. To prove his point, he has designed and built a context-specific corpus of student writing produced by Spanish-speaking English learners (mostly freshman students from Mexico) at various levels of proficiency and reflecting a variety of writing tasks collected at the University of Texas El Paso. This corpus allows researchers to examine theoretical issues while helps educators to identify key pedagogical issues as they evaluate learners' needs in relation to practices and beliefs of the local culture. The chapter illustrates the various ways in which this corpus can be used to not only conduct empirical research on second language writing, but also to establish links to teaching, learning, and assessment.

Chapter 6 by Don Miller highlights the methodological challenges inherent in reliably capturing meaningful sets of vocabulary for instructional focus. An analysis of a 3.1 million-word corpus of introductory psychology textbooks suggests that, while comparatively large, and, thus, presumably representative of the lexical variability in the target domain, this corpus was unable to capture a stable list of "important" words. Findings highlight an important issue requiring further

investigation in corpus-based vocabulary research: the extent to which corpora – and the word lists based on them – reliably represent the lexical variability of their target domains.

Chandrika Balasubramanian takes a look at new varieties of English (New Englishes) in Chapter 7. More specifically, the study is an empirical investigation of spoken and written registers of contemporary Indian English. The first part of the paper outlines the theoretical bases for corpus construction for the study of international Englishes, and describes the corpus of 1.5 million words used in this study. The second part of the paper shows, through the investigation of two grammatical features (Wh-questions, and additive and restrictive circumstance adverbials) that an international English like Indian English shows the same kind of internal variation that the more traditional "native" varieties do.

In Chapters 8 and 9, Casey Keck and Viviana Cortes take a different direction, producing essays that review the state of the art in the study of two constructs that share a lot of features in common and are closely linked to the corpus-based research methodologies that originated with Biber's work in the Flagstaff School. Keck, in Chapter 8, presents a chronological review of her own work on textual borrowing in the written production of non-native speaker university students. She emphasizes the use of tailor-made computer programs that facilitated the different stages of the research studies she conducted. Her chapter includes a careful description of the methodology used, the software designed, and the results of her analyses, as well as various implications of her findings for the teaching of academic writing to non-native speaker of English (NNSE) writers. In Chapter 9, Cortes writes about her area of specialization, lexical bundles, groups of three or more words that frequently recur in a register. She goes back in time to the origins of the use of corpus-driven methodologies to identify frequent formulaic expressions empirically rather than intuitively. The purpose of this chapter is to clearly describe the lexical bundle as a construct in the spectrum of formulaic language to avoid confusion in the method of identification and analysis of these expressions. Her chapter highlights the work of Biber et al. (1999) as a foundation for the many studies of lexical bundles that have been conducted in the past decade.

Finally, we would like to acknowledge our colleagues, who took some time from their very busy agendas to write and review for our volume. First, we would like to thank Michael McCarthy. When we first envisioned this volume, more than ten years ago, we asked Michael if he would write a preface to the volume highlighting Professor Biber's contribution to the development of corpus-based research. He agreed then and he remembered that when we contacted him again a couple of years ago with the prospectus of this collection. Michael has always been a friend of the Flagstaff School, visiting the institution on several occasions sometimes with his students. Then we want to thank a group of scholars

who worked in the reviewing and helped in the editing process (in alphabetical order) Tony Berber-Sardinha, Scott Crossley, Stephanie Lindemann, David Oakey, Trevor Shankin, Heidi Vellenga, and Camila Vazquez. They all provided detailed and critical feedback to the writers that surely contributed to enrich the chapters presented in this volume.

References

Biber, Douglas. 1988. *Variation across Speech and Writing,* Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 1995. *Dimensions of Register Variation*. Cambridge: CUP. DOI: 10.1017/CBO9780511519871

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style.* Cambridge: CUP. DOI: 10.1017/CBO9780511814358

Biber, Douglas & Tracy-Ventura, Nicole. 2007. Dimensions of register variation in Spanish. In *Working with Spanish Corpora,* Giovanni Parodi (ed.), 54–89. London: Continuum.

Biber, Douglas, Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP. DOI: 10.1017/CBO9780511804489

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Eward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–86. DOI: 10.1016/j.esp.2006.08.003

CHAPTER 1

A corpus-based analysis of linguistic variation in teacher and student presentations in university settings

Eniko Csomay
San Diego State University

This study investigates patterns of language use in professional presentations by teachers and students in a university setting. A 271,500 word corpus was compiled using 122 teacher presentation segments extracted from a previously collected large corpus of classroom discourse and 69 student presentations recorded at a student research symposium and transcribed. Student and teacher presentations were compared based on the dimensions of linguistic variation in university settings (Biber & Conrad 2009). Findings show that while presenting, teachers use significantly more features associated with oral and content-focused discourse as well as more teacher stance features. In contrast, students, use more features of literate and procedural discourse with no stance features.

Keywords: Multidimensional analysis; spoken academic corpus; participant language use

1. Background

The number of corpus-based studies investigating the language used in academic settings has increased dramatically during the past ten years. What we know today about this context's linguistic make-up is impressive, yet there are areas for further investigation. In addition to the wealth of studies discussing various aspects of the language used in this context, we have a variety of corpus-based methodologies applied. As discussed in detail below, many studies have discussed particular linguistic features in this context, or provided a comprehensive linguistic picture of the context. The work reported here fills the gap in analyzing presentation styles.

DOI 10.1075/scl.66.01cso

The present study explores variation in language use in teacher and student presentations (a sub-register in the academic setting) as it relates to the registers and dimensions of linguistic variation in speech and writing at the university. More specifically, this study takes the dimensions of variation in academic settings reported in Biber and Conrad (2009), calculates the dimension scores for these two settings and presenters in question, and places their linguistic profile among the other registers in the academic context.

In this section first, I will briefly introduce some of the previous corpus-based studies that focused on language use in the academic setting. Second, I will briefly introduce the situational parameters for the context in which the presentations take place, and from which the language samples were taken for the present study. Third, I briefly describe the basic concept behind the multi-dimensional framework, and finally, I will point to the goals and the outline of the present study.

1.1 Previous corpus-based studies on student language in the academia

Over the past three decades, a growing number of studies has explored language in the academic setting in general, or focused on the characteristics of language used in various sub-registers in this context. For example, patterns of language use were investigated in study groups, textbooks, and in classrooms. Other studies focused on ways in which teachers talk in different disciplines in university classes, or how students use language in their academic writing. Although all of the studies discussed here look at language patterns in corpora, the approach they take to carry out the analysis is vastly different. One group of researchers identifies functional categories as their starting point, and takes examples from the corpus to prove or illustrate their point. For example, they take a corpus of academic student writing to investigate how rhetorical moves are constructed by students in their master's thesis introductions (Samraj 2008), or look at how particular discourse functions, for example, hedging, are expressed (Hyland 1996, 1998) in academic prose. Also using a corpus, scholars provide detailed analyses of particular, pre-selected lexico-grammatical items they are interested in; for example, they look at how students use "attended and unattended *this*" in their writing (Swales 2005; Wulff et al. 2012) or how individual part of speech categories (e.g. nouns) are used in various disciplines in the classroom (Biber 2006).

Another group of scholars applying corpus-based methodologies take a bottom-up approach and provide comprehensive linguistic descriptions of language variation in university settings (e.g. Biber et al. 2002). These studies typically include in their analyses both written and spoken registers taken from a variety of situations and from a number of participants, and use sophisticated multivariate statistical techniques to provide a comprehensive profile of language

use (Biber et al. 2004). Based on the overall linguistic profile then, they describe disciplinary differences (Conrad 1996), distinguish between language use in different instructional levels and/or varying interactivity patterns in the classroom (Csomay 2002), or discuss how language patterns vary between the two participants in the classroom, teacher versus student talk (Csomay 2007a, 2013).

1.2 Situational characteristics of the academic presentations

Register studies analyze the context or the speech situation from which the sample texts are taken, as the basic starting point in their subsequent linguistic analyses. The notion of a communicative context has been studied and discussed by a number of researchers during the past few decades (e.g. Hymes 1972; Halliday 1978; Duranti 1985), and scholars managed to isolate the components the speech situation. Biber (1988), exploring variation across speech and writing, synthesized earlier work and introduced a framework that included many aspects of a context, and based on which he we was able to show register differences in a multidimensional linguistic space.

The assumptions behind Biber's (1988) framework, which is applied in this study, stem from the basic ideas that (a) language forms are associated with communicative functions; (b) communicative functions are related to situations; and (c) change in situational parameters is connected to variation in language use. According to Biber and Conrad (2009: 40–47), who later slightly modified the original 1988 framework, the most pertinent situational characteristics of registers include the following: Participant, relations among participants, channel, production circumstances, setting, communicative purposes, and topic. As we compare two registers, the "linguistic features will mark particular components of the situation" (Biber 1988: 28). For example, a letter to a friend and a letter to a boss will have at least two situational parameter differences: audience and the writer's relationship to the audience. Hence, the language used in these two letters can, and most probably will, vary even though another parameter, production circumstances (mode), will remain the same (i.e. both are produced in a written mode). Most pertinent to this study is the situational analysis that Biber and Conrad provided as they analyzed textbook versus classroom teaching. In this work, they outlined "key situational differences" (2009: 65) in these contexts (Appendix A).

In this study, we investigate how the language of student presentations differs from teacher presentations in the academic context. Although in both cases the same genre, "expository text", is apparent and the situational characteristics may be similar for academic presentations for both participants due to the fact that the communicative purpose is the same, there are important differences between the two situations investigated. A brief analysis of these two contexts is presented next.

As outlined in Table 1, the main differences between the two situations in this study lie in (a) participant characteristics; (b) relations among participants; and (c) production circumstances. More specifically, first, the social characteristics of the participants differ as the presenter in the classroom is an expert professional while the presenter at a student symposium is a novice, or an emerging professional.

Second, the most differences can be found in the relations between the participants. In the classroom, questions could be asked at any time during the teacher's

Table 1. Key situational differences between student presentations at a symposium and teacher presentations in the classroom

	Classroom presentation (Instructor)	**Symposium presentation (Student)**
Participants	One addressor, multiple addressees Social characteristics: **expert** professional	One addressor, multiple addressees Social characteristics: **novice** professional
Relations among participants	Interaction is possible **during** presentation Addressor has more knowledge than audience All participants have some specialist knowledge Addressor gets to **know** most or all participants Social/academic status of addressor is superior to addressee (**high status**) Power is held in **addressor's** hand	Interaction is possible but **only after** presentation Addressor has more knowledge than audience All participants have some specialist knowledge Addressor **does not know** most or all participants Social/academic status of addressor is subordinate to most of the addressee in the audience (**low status**) Power is held in **addressee's hand** (judges evaluate performance)
Channel	Spoken	Spoken
Production circumstances	Text has been planned and may have been revised or edited prior to production Text **can be** negotiated, and revised on the spot Text can be read out (mostly it isn't)	Text has been planned and may have been revised or edited prior to production Text **cannot be** negotiated, revised or edited on the spot Text can be read out (mostly it is)
Setting	Addressor and addresses are physically in the same room	Addressor and addresses are physically in the same room
Communicative purposes	Convey information potentially new to the audience Explain concepts and methods Convey personal attitudes Direct students what to do	Convey information potentially new to the audience Explain concepts and methods

presentation while in a formal presentation setting such as the symposium, set routine requires that questions are posed after the presentation. In the classroom, the teacher knows most of the participants, since the same group of people get together for the class sessions for weeks. At the symposium, the student presenter may know some people in the audience but most likely s/he would not. The teacher (addressor) in the classroom setting has a high social/academic status in the community, and is certainly superior to that of the addressee. This, however, is the opposite for the student presenting at the symposium. S/he will have a low status in the community and they play a subordinate role. Among the audience are teachers, other students, and perhaps community members with high social status. Lastly in this area, the power is in the teacher's hand in the classroom while the power is in the audience's hand at the symposium as they serve as judges of the presentation for content and performance.

Thirdly, the production circumstances are also different in the two settings in terms of the ability to revise text on the spot. Since students can ask clarification questions at any point in time in a classroom setting, there is plenty of room to negotiate and revise the text "online" or on the spot. In fact, one of the main tools teachers use is to reformulate the text for better understanding of the content. In the symposium presentation, there is no room for immediate negotiation of the text, or editing, and certainly there is a lack of spontaneous interaction.

After highlighting the various aspects of the situational circumstances, we can now see how the linguistic profiles of texts can be best characterized. For this, a multi-dimensional analytical approach is adopted and applied.

1.3 Comprehensive descriptions of linguistic variation in texts

Comprehensive descriptions of variation in language use cannot be based on investigating one single linguistic feature in isolation. In addition to other problems with such an approach is that it would be rather difficult to know a priori which feature to choose that will mark the difference in the situations we are comparing. Although earlier work can be consulted to identify functional categories and their associated features before the investigation (Biber & Conrad 2009: 63), it is problematic to perceptually determine which feature may be responsible for the differences in a text's entire linguistic profile. The linguistic characteristics of texts can be systematically described based on empirical measures and in a comprehensive way, and by documenting the relationships across a number of linguistic features and across texts. To capture these relationships among a large number of features extracted and counted in many texts at the same time, quantitative, exploratory multivariate statistical methods (factor analysis) are used. The analytical framework applying this statistical method to provide comprehensive linguistic

descriptions was developed by Biber (1988) and is coined as "multi-dimensional analysis of linguistic variation."

A number of earlier studies applied a multi-dimensional analytical framework (Biber 1988; Biber 1995; Conrad & Biber 2001; Biber & Conrad 2009) to describe language variation across registers. Depending on the motivation for each study applying this methodology, researchers have used this framework in one of two ways: either (1) run the multivariate, factor analysis from the beginning in order to identify factors and interpret them to describe novel dimensions of variation within their own dataset, or (2) use an already existing model where the dimensions had already been identified prior to the given study, and the given study is set out to examine how their own texts would place on the already existing continuum of variation. Examples of the former approach include pioneer studies that distinguish dimensions of linguistic variation across registers in both English (Biber 1988) and languages other than English (e.g. Somali and Korean by Biber 1995), variation in student and adult speech and writing (Reppen 2001), or explore dimensions of variation in language use within just one register, e.g. university classroom discourse (Csomay 2005). Examples of the latter includes studies that use an existing dimensional framework, typically using Biber's (1988) study, to look at register evolution from a historical perspective (Atkinson 2001; Biber & Finegan 2001), variation in language use as it relates to specialized domains such as, author's style (Connor-Linton 2001), disciplinary language use (Conrad 2001), intra-textual patterns in medical writing (Biber & Finegan 2001), or dialect variation (Rey 2001; Biber & Burges 2001; Helt 2001).

1.4 Outline of the present study

The present study applied the second approach to explore variation in language use in teacher and student presentations as it relates to the registers and dimensions of linguistic variation in speech and writing at the university. More specifically, this study takes the dimensions of variation in academic settings (Biber & Conrad 2009), calculates the dimension scores for these two settings, and places their linguistic profile among the other registers at the university. In the subsequent sections, I outline the methodology (2), then report on and discuss the findings (3), and finally, draw conclusions and implications (4).

2. Methodology

In this section, the design of the study and the analytical procedures are described. In the process of carrying out the study, decisions were made about the unit of analysis, the corpus of texts, and the selection of linguistic features for the analysis.

2.1 Corpus

A total of 191 text files were used in the study selected from two corpora. One corpus contains 122 teacher turns of more than one thousand and less than two thousand words each, extracted from a range of university class sessions.[1] The other corpus contains 69 student presentations extracted from the Student Research Symposium held at San Diego State University in 2009.[2] Table 2 shows the distribution of texts according to the speaker categories, and the number of words distributed across speakers.

Table 2. Text distribution based on speaker category

Speaker	Number of texts	Total number of words	Average turn length in number of words
Teacher	122	166,770	1,450
Student	69	104,730	1,367
Total	191	271,500	1,408.5

The digital recordings were transcribed following predefined transcribing conventions and all texts were tagged for grammatical features using Biber's grammatical tagger. The texts were classified according to who the presenter was and the context in which they presented in a university setting. The situational parameters were discussed in Section 1.2 above. Below are definitions and the unit of analysis.

2.2 Definitions and unit of analysis

Presentations are defined in this study as continuous talk given by one speaker standing in front of an audience in an academic setting (presentation mode). The purpose of academic presentations is the dissemination of academically focused content or information to the audience with whom the speaker (presenter) shares the same physical space.

1. Teacher turns were identified and extracted from a combined corpus of university classroom discourse. One subset of classroom discourse originates from the T2KSWAL corpus representing university language in North America collected at five universities (Biber et al. 2002; Biber et al. 2004) and the other subset originates from the MICASE corpus representing language use at the University of Michigan (Simpson & Swales 2001).

2. The corpus of student presentations at SRS was compiled as part of an ongoing large-scale international project, the purpose of which is to collect a large sample of student presentations in various cultural settings.

In a classroom situation, typically one of two speakers could potentially take the floor: the teacher, or one or more students. For the purposes of this study, only the teachers were selected as speakers, given their conventionally perceived leadership role in the classroom to present information to the students as their audience.

On the other hand, presentations at the student research symposium are determined by the general framework and the rules of the Student Research Symposium. Accordingly, each of the ten-minute student presentations was prefaced by a moderator, and followed by a five-minute question answer section. As the texts were transcribed, the exact beginnings and ends of the presentations were marked and the presentation sections were separated from the moderator's introduction and the question-answer sections.

The unit of analysis in this study is a turn. To operationalize a turn, Tao's definition was applied stating that "any speaker change will be treated as a new turn." (2003: 189) An extended turn, referred to above as continuous talk taken by one speaker with no interruptions, constitutes a presentation. Since student presentations are limited in length to the rules of the symposium, I selected teacher presentations from the classroom that were turns with similar length: more than 1,000 words but no longer than 2,000 words. The average teacher turn was 1,450 words long while the average turn-length for student presentation was 1,367 words.

2.3 Linguistic features on four dimensions of academic language use

Biber and Conrad's (2009) analysis of variation across speech and writing in the university setting served as the basis for the current linguistic investigation. Although the vast majority of the features and their statistical measures are available in the current work, not all measures overlap with Biber et al. (2004), Biber (2006), and Biber and Conrad (2009); hence, a few features are missing from the analysis in this study (see full list of features included in this study in Appendix B). Below are the linguistic features discussed in this study and as they relate to the four dimensions of linguistic variation across speech and writing in the university setting (Biber & Conrad 2009).

Among other linguistic features on one side of Dimension 1, the high occurrence of contractions, pronouns, present and past tense, mental, activity, and communication verbs, time, place, and likelihood adverbials as well as hedges and discourse particles, wh-questions, clausal coordination, stranded prepositions, conditional and causative clausal adverbials, and *wh-* and *that-* complement clauses was associated with personal, interactive discourse typical to oral discourse, hence was called *Oral Discourse*. In contrast, the opposite side of Dimension 1 contains features such as, common nouns, nominalizations, nouns classified into semantic categories such as abstract nouns, group nouns, human nouns, and mental nouns,

long words, phrasal coordination, prepositional phrases, attributive adjectives, passives, relative clauses, a variety of *to-* clauses and phrasal coordination. This set of linguistic features has been associated with literate discourse, and therefore, was called *Literate Discourse*. Hence, texts placed on Dimension 1 can be associated with oral versus literate discourse based on their linguistic characteristics.

One side of the spectrum on Dimension 2 contains linguistic features such as necessity and future (predictive) modals, causative and activity verbs, second person pronouns, group nouns, *to-* clauses with desire verbs, and conditional adverbial clauses. These features were associated with spoken discourse in the university setting where (institutional) rules and procedures are outlined, and was called *Procedural Discourse*. On the other hand, the opposite side of Dimension 2, called *Content-focused Discourse*, contains features such as rarely occurring vocabulary items in all four of the content word category, adjectives describing size, *to-* clauses with probability verbs and by-passives. These features were associated with "written academic registers" as exemplified through a graduate level natural science textbook. (Biber 2006: 236).

Dimension 3, called *Reconstructed Events*, has features such as third person pronouns, past tense, communication and mental verbs, and *that*-complement clauses especially with likelihood verbs, and where the complementizer *that* is omitted. These features were associated with discourse with a reconstructed account of events. On the other hand, the features grouping on the other side of this dimension are nouns of various kinds such as, concrete and technical nouns as well as quantity nouns.

Finally, the linguistic features grouping together on Dimension 4, called *Teacher-centered Stance*, were relative clauses with *that* as relative pronoun, stance adverbials of various semantic kinds such as, certainty, likelihood and attitudinal, adverbial clauses of condition, and *that-* clauses controlled by stance nouns. On the opposite side of this dimension denoting lack of teacher-centered stance features are *wh-* questions and stranded prepositions.

2.4 Analytical procedures

After running the texts through a grammatical tagger, I developed computer programs with Delphi Pascal to count the various linguistic features for the study. The first program was designed to identify speaker turns and turn length measures, and to select turns with the appropriate length (one to two thousand words). The second program was developed to count the frequencies of the linguistic features in each turn as outlined in the previous Section (2.3), and to write out the normalized counts (2.4.1). To write out relevant excerpts, a freely available program called AntConc was used.

2.4.1 *Counts and statistical procedures*

Several steps were taken in calculating the dimensions scores for each text (turn) based on Biber et al. 's (2004) and Biber and Conrad's (2009) dimensions (2.3), and these are outlined below. In addition, to calculate the dimension scores for the corpus in this study, the statistical measures (means and standard deviations) and vocabulary lists[3] for the entire corpus of spoken and written registers in the university setting listed in Biber et al. (2004) were used.

First, the linguistic features identified for each dimension were counted in each turn. The full list of features is in Appendix B. The raw frequency counts for each turn were tracked and were normed to 1,000 words (total feature count for a turn divided by the number of words in that turn multiplied by 1,000). This procedure allows a comparison of turns of unequal sizes, normalizing the feature counts to the point as if they were all 1,000 words.

Second, the normed counts for each feature were scaled to the entire corpus (not to individual text/turn-length). This procedure was done to compensate for those features that typically occur very often versus those that are typically rare, and therefore, to bring them under the same scale. Accordingly, z-scores were calculated for each feature outlined above by taking the normed feature count, minus the mean score of that feature for the entire corpus, and divided by the standard deviation of that feature in the entire corpus as reported in Biber et al. (2004: 61–64).

Third, the dimension scores for each text were calculated based on the sum of the z-scores per features on one side of the dimension minus the sum of the z-scores per features on the other side of the dimension. For example, Dimension 3 would have the z-scores for the features on the positive side (see list of features in Appendix B) added up, from which we deduct the z-cores for the features on the negative side (see list of features in Appendix B) added up. This is standard procedure to calculate dimension scores (Biber 1988: 94; Biber & Conrad 2009: 227–229). The same procedure is repeated for each observation, which in our case is each turn or text.

Finally, in order to identify the statistically significant differences between teacher presentations and student presentations on each of the four dimensions, the mean scores of each dimension and for the two groups were calculated and compared using an Independent Sample T-test that was run through an SPSS 20.0 software package.

---

3. The counts associated with vocabulary distribution measures such as, 'common nouns,' for example, were computed for this study based on the vocabulary lists provided in Biber et al. (2004).

3. Findings

3.1 General patterns

The linguistic characteristics of a total of 191 presentation segments, 122 teacher turns and 69 student turns, were analyzed further. Based on the linguistic features grouping together on Biber and Conrad's (2009) dimensions of language variation in the university setting, the mean scores for each dimension were calculated for each participant.

As Table 3 shows, the mean score for teachers on Dimension 1 is 13.21, and the scores are between −14.56 and 56.47. For students, the mean scores is −1.25, and the student scores are between −31.15 and 23.39 (range is 71.03, and 54.54, respectively). The standard deviation is high for both speakers on this dimension, indicating that the scores are spread rather than being close to each other. At the same time, the majority of the student scores (55 percent) are negative, which indicates that the use of features associated with literate discourse is more pertinent by students. In contrast, 90 percent of the teacher scores are positive, indicating that the majority of the teachers tend to use oral language in this situation.

Table 3. Descriptive statistics for each dimension

		N	Mean	SD
Dimension 1: Oral versus literate discourse	Teacher	122	13.21	11.23
	Student	69	−1.25	12.15
Dimension 2: Procedural versus content-focused discourse	Teacher	122	−3.08	7.52
	Student	69	.14	3.57
Dimension 3: Reconstructed account of events	Teacher	122	−.17	6.48
	Student	69	−1.02	4.59
Dimension 4: Teacher-centered stance	Teacher	122	4.56	5.72
	Student	69	2.01	4.81

The scores for Dimension 2 also vary, with a range of 36.68 for teachers and 15.77 for students. Standard deviation figures are lower, indicating that the scores cluster closer together more than it was the case in the previous dimension. For Dimension 3, both participants indicate negative dimension scores, and a much more similar range of scores is apparent than for the previous two dimensions. Teachers' scores range between −13.46 and 17.59 (31.05) and students' scores range between −10.40 and 16.13 (26.53). Finally, for Dimension 4, both means are

positive, deviations are similar, and the range for teachers is 29.28, while the range for students is 22.75.

The mean scores for each dimension and for each speaker could be charted on the four dimensions of linguistic variation in the university setting (Biber & Conrad 2009: 231–240). To visually see where these presentations place on the linguistic continuum of university language use, and in relation to other registers in this setting, see Table 4.

Table 4. Registers placing on the dimensions of linguistic variation in university settings (Biber & Conrad 2009)

Oral discourse	Procedural discourse	Reconstructed account of events	Teacher-centered stance
13 + ***teacher presentation***	| CLASSROOM MANAGEMENT		
//	|		
| SERVICE ENCOUNTERS	10 +	10 +	10 +
10 +	|	|	|
|	|	|	|
|	9 +	9 +	9 +
9 + OFFICE HOURS	|	|	|
| LABS	|	|	|
| STUDY GROUPS	8 + COURSE MANAGEMENT,	8 + STUDY GROUPS	8 +
8 +	| INSTITUTIONAL WRITING	|	|
|	| SERVICE ENCOUNTERS	|	|
|	7 +	7 +	7 +
7 +	|	| OFFICE HOURS	|
|	|	|	|
| CLASSROOM MANAGEMENT	|	6 +	6 +
6 +	|	|	|
|	6 +	|	|
|	|	5 +	5 + CLASSROOM MANAGEMENT
5 +	| OFFICE HOURS	|	| CLASSROOM TEACHING
|	5 +	|	| ***teacher presentation***
|	|	4 +	4 +
4 + CLASSROOM TEACHING	|	|	|
|	|	| CLASSROOM TEACHING	|
|	4 +	3 + SERVICE ENCUNTERS	3 +
3 +	|	|	| OFFICE HOURS
|	|	| LABS	|
|	3 +	2 +	2 + ***student presentation***
2 +	|	|	|
|	|	|	|
|	2 +	1 +	1 +
|	|	|	|
1 +	1 + CLASSROOM TEACHING	|	|

(*Continued*)

Table 4. (Continued)

Oral discourse	Procedural discourse	Reconstructed account of events	Teacher-centered stance
|	| LABS	0 + ***student presentation***	|
|	| ***student presentation***	|	|
0 +	0 +	-1 + ***teacher presentation***	0 +
| ***student presentation***	| STUDY GROUPS	| CLASSROOM MANAGEMENT	|
|	|	| COURSEPACKS	-1 +
-1 +	-1 +	-2 +	|
|	|	|	| TEXTBOOKS
| TEXTBOOKS	|	-3 + |	-2 +
-2 + COURSEPACKS	-2 +	|	| COURSE PACKS
|	|	-4 +	|
|	|	| TEXTBOOKS	-3 +
3 +	-3 +	|	|
|	| ***teacher presentation***	-5 +	|
|	|	// INSTITUTIONAL WRITING	-4 + LABS
-4 +	-4 +	|	|
|	|	-8 +	| COURSE MANAGEMENT
|	|	|	-5 +
-5 +	-5 +	-9 +	|	// INSTITUTIONAL WRITING
//	//	|	|	STUDY GROUPS
|	|	-10+	-10+	-8 +
-8 + COURSE MANAGEMENT	-8 +	|	|	|
| TEXTBOOKS, COURSPACKS	|	-11+	-11+	|
|	|	| COURSE MANAGEMENT	| SERVICE ENCOUNTERS	-9 +
-9 + COURSE PACKS	-9 + COURSE PACKS			|
|	|			|
|	|			-10+
-10+	-10+			|
|	|			|
|	| TEXTBOOKS			-11+
-11+	-11+			| SERVICE ENCOUNTERS
| INSTITUTIONAL WRITING	|			
Literate discourse	**Content-focused discourse**	**Concrete current information**		

It is evident from Table 4 that in general, teacher presentations exhibit unique language that is associated with oral and content-focused discourse, with concrete current information, and that is framed with teacher-centered stance. In contrast, students present their work in a way that is somewhat neutral on all four dimensions, but perhaps resembles more to literate and content-focused discourse closer to that of study groups, and using teacher-centered stance features closer to that of office hour language.

Descriptive statistics are generally informative, showing central tendencies in a dataset as well as patterns of distribution. However, just looking at descriptive statistics, it is difficult to determine whether, for example, two mean scores are statistically significantly different. For our study, whether teachers use statistically significant language from students when they are presenting their materials cannot be determined without statistical calculations. To compare mean scores between two independent groups, an Independent T-test was applied. In this dataset, an Independent T-test was performed on each of the four dimensions to see whether the two groups are different in those measures. Table 5 shows the Independent T-test scores for each of the four dimensions.

Table 5. Results of the Independent Samples Test for each Dimension

		Levene's test for equality of variances		t-test for equality of means				
		F	Sig.	t	df	Sig. 2-tailed	Mean diff	Std. error diff
Dimension 1: Oral versus literate discourse	Equal variances assumed	2.04	.155	8.30	189	.000	14.46	1.74
	Equal variances not assumed			8.12	132.26	.000	14.46	1.78
Dimension 2: Procedural versus content-focused discourse	Equal variances assumed	32.90	.000	−3.35	189	.001	−3.22	.96
	Equal variances not assumed			−4.00	184.51	.000	−3.22	.81
Dimension 3: Reconstructed account of events	Equal variances assumed	15.51	.000	.96	189	.339	.85	.89
	Equal variances not assumed			1.05	179.49	.294	.85	.81
Dimension 4: Teacher-centered stance	Equal variances assumed	.90	.35	3.13	189	.002	2.55	.81
	Equal variances not assumed			3.29	162.03	.001	2.55	.78

As Table 5 shows, overall, teachers use different language from students in their presentations on three of the four dimensions identified in this study:

Dimension 1, *Oral versus Literate Discourse*, Dimension 2, *Procedural versus Content-focused Discourse*, and Dimension 4, *Teacher-centered Stance*. Eta square (strength of association) is calculated to see how strong the association is between the dependent and the independent variable. The dependent variable here is the given dimension score and the independent variable is the speaker with two levels, teacher and student. In this case, the per cent value (26.7 per cent) tells us that the variation in the data on Dimension 1 is accounted for by who talks. In other words, if we know the score for Dimension 1, we can predict who the speaker is roughly one-third of the time. On Dimensions 2, eta square around 8%, and on Dimension 4 it is around 5%. In the next sections, each dimension, where the two presenters differ in their use of language statistically significantly, will be discussed separately, supported by textual examples to illustrate the constellation of the linguistic patterns (or lack thereof) as well.

3.1.1 *Oral vs. Literate Discourse*

As the results show, the linguistic features grouping on Dimension 1, *Oral versus Literate Discourse*, are used statistically significantly differently by teachers and students as they present their materials. Teachers tend to use an overwhelming number of features associated with oral discourse while students tend to use language associated with literate discourse.

As outlined before, features on the positive side of Dimension 1 are, for example, contractions, pronouns, present and past tense, mental, activity, and communication verbs, time, place, and likelihood adverbials as well as hedges and discourse particles, *wh*-questions, clausal coordination, stranded prepositions, conditional and causative clausal adverbials, and *wh-* and *that-* complement clauses. The constellation of these features (bolded in Extract 1) was associated with personal, interactive discourse typical to *Oral Discourse*. Features on the negative side of Dimension 1 are, for example, common nouns, nominalizations, nouns classified into semantic categories such as abstract nouns, group nouns, human nouns, and mental nouns, long words, prepositional phrases, attributive adjectives, passives, relative clauses, a variety of *to-* clauses and phrasal coordination. These features (bolded in Extract 2) were associated with *Literate Discourse*.

The text extracts below show examples from a teacher monologue (1) and from the student presentation (2) illustrating these features.

(1) Teacher: So what **I'm suggesting** to **you** then, **is**, **is that** this second dynamic, **which accounts** for the popularity, the contemporary popularity of civilian review, **has** to **do** with money, and civil liability, and the ways in **which** the behavior of law enforcement institutions **can**, render, municipalities liable for millions and millions and millions of dollars, **uh**, in, **uh,** civil liability lawsuits. Not only **that**, usual contingency, **um**, **uh**, rules, **are** waived in these kinds of lawsuits. All right? What **that means is**

that usually, **when you pursue** a civil claim, against somebody, **you ask** for a hundred thousand bucks, **OK?** And, **you get it**, and **your** lawyer **takes** a third. All right? **What happens if you sue** a municipality and **they say yeah we think you're** right but [short laugh] the situation **was** so much more complicated, **we award**, one dollar, **OK? Is your** lawyer **gonna take** thirty three cents? Not in these kinds of lawsuits, right?

(2) Student: And we found that an **immature cynofields** resides **in** the **kidney** that's where we found the most **cells with** those **characteristics** and I interpreted that we found also oh… oh… a **relationship for** those **cynofiedls** but does were more mature. We can say that because ….The **electro- microscopy results with** that we can see the **morphology and chronology** and this is a **human cynofield with** a **transmission electronic microscopy of** the **human cynofield** and we did **with** a **zebrafish** we found very similar **morphology that granules** are round as same as the **human ones** and the **nucleus** is big at this **stages** so we found the **cell** that looks like **cynofiedls** so now we want to study its **function** we study the **function by migration of recommendation to** the **infection** and then we see they change their **morphology**. So we know that **cycles-sum in human cynofields** includes **information response** and we inject the **fish with** the **cycles-sum** we let them live for 6 hours in order **to** provide an **order response** and then **to** (syll) we sacrifice the **single cell suspension** and **within** the **facts analysis of photometry** and those are our **results**. We found we use a **control** also and we can see **in** the **control** the **populations of cynofields** are **in** not increase as dramatically **with** the one that we injected we **cycle-sum** and it was 20% more **of population of** those **cell that** we found **in** this **gate**.

As Extract (1) illustrates, teachers present informational materials in a way that resembles oral discourse. In contrast, as Extract (2) shows, student presentations display literate discourse. Closest to the latter type of discourse (Extract 2) in the academic setting are registers such as course packs and textbooks (cf., Table 4). The fact that students talk in such a way is often indicative of their reading out their papers as they present their research at the symposium. Often times, students also read off the text from the power point slides, the text of which had been prepared and edited ahead of time.

3.1.2 *Procedural vs. Content-focused Discourse*

The linguistic features grouping on Dimension 2, *Procedural* versus *Content-focused Discourse*, are used statistically significantly differently as well while teachers and students present. Teachers tend use linguistic features associated with content-focused discourse (Extract 3) while students tend to use language associated with procedural discourse (Extract 4). On the positive side of Dimension 2 are features such as, necessity and future (predictive) modals, causative and activ-

ity verbs, second person pronouns, group nouns, *to-* clauses with desire verbs, and conditional adverbial clauses. These features were associated with *Procedural Discourse*. On the negative side of Dimension 2, linguistic features such as, rarely occurring vocabulary items in all four of the content word category, adjectives describing size, *to-* clauses with probability verbs and by-passives were associated with *Content-focused Discourse*. The following two text samples illustrate how the two groups use these features.

(3) Teacher: let's talk about **elasticity**, of the population growth rate. We've mentioned this before lemme, uh say a couple more things about it and uh, show an example. for the same proportional change in each matrix element, **elasticity**, measures the **proportional** change in the growth rate **lambda**. Because we're talking about **proportional** changes in the growth rate, of each **factor**, the **elasticities** sum to one. The interpretation is that, that uh the **elasticity** reflects the **proportional** contribution of each element, to the population growth rate **lambda**. So high **elasticity**, means a big effect, of that, parameter on the population growth rate. For painted turtles, adult survival has the greatest effect on population growth rate and **fecundity** has the least. The paper by **Heppell**, sh- she, uh, added together the **elasticities** for matrix, what she called juvenile turtles, juvenile survival, which was age classes one to three, and the **subadults**, which were age classes four five and six, and, this column represents the, **elasticity** of adult survival.

(4) Student: So what happens is. **If you see** the arrow of the laser **coming** in and the [unclear] **moving** towards the laser they're actually **going to** end up absorbing at a hi-uh-lower frequency, **cause** of the [unclear] **moving** away from the laser they're **going to** absorb at a higher frequency, **so** the [unclear] at exactly the frequency **you want** [unclear]. **So** the way **you want to eliminate** this is by using the whole-grain effect, and the way this **happens**, **see** how the lower state and the lower state. Well at the lower state, they're **going to have to** be excited by [unclear]. And **so**, **you see** how the higher state is an extra amount of atoms that got excited, **so** what we're **trying to** do is hit that container with two different beams coming in, and uh, and again they term this the pumpbeam and the one [unclear] back is the probeam, and they **have to** intersect each other, and they create two different holes but the holes, when they combine [unclear] um frequency, which is known as the magnitive effect, and **so**, we're **trying to** detect that dip, the twenty-one of those dips of the 21 lines [unclear] I'm **going to** tell you about that later.

As Extract (4) shows, student presentations display procedural discourse. Closest to this type of discourse in the academic setting is discourse occurring in labs (cf., Table 4). In student presentations, where an account is given of a particular research project, it is not surprising that there may be more linguistic features associated with procedural discourse. The fact that students talk in such a way is often indicative of their presenting how things work, or how they carried out the study.

3.1.3 *Reconstructed account of events*

When it comes to linguistic features associated with reconstructed accounts of events on Dimension 4, teachers and students do not differ. Features such as, third person pronouns, past tense, communication and mental verbs, and *that*-complement clauses especially with likelihood verbs and omitting the complementizer *that* were grouped together on this dimension. Since the two presenters do not differ significantly on this measure, this dimension is not discussed any further.

3.1.4 *Teacher-centered Stance*

The linguistic features grouping on Dimension 4, *Teacher-centered Stance*, are used statistically significantly differently by teachers and students as they give their presentations. Perhaps not surprisingly, teachers tend to use many more of the features associated with stance than students. Such features are, for example, relative clauses with *that* as relative pronoun, stance adverbials of various semantic kinds such as, certainty, likelihood and attitudinal, adverbial clauses of condition, and *that*- clauses controlled by stance nouns. The following two text samples illustrate this difference.

(5) Teacher: **Typically**, uh though not exclusively but **typically** the members of possession cults are women. And this is largely a reflection not again but there's some psychological susceptibility on the part of women, but primarily it reflects **the fact that** women are in a politically and economically (support) position in **virtually** every society in the world. And **that** essentially with no means, legal, economic, political, to express their wants, their grievances, their complaints. It generates an emotional response in the form of possession. And of course the symptoms of possession will vary but they're **typically** ones that um include for example, inertia, (laxity), not wanting to do much, depression, uh basically a lack of initiative or motivation uh sense of grief or sadness.

(6) Student: With the increase in globalization of American companies more businesses are deciding to send their employees on business trips either domestically or abroad. Some of these employees are required to travel constantly for work or leave their homes for long periods of time each year. According to the center for long distance relationships there were three point six commuting couples in two-thousand five. As this numbers continue to grow. More and more families are spending days, weeks or even months away from their families. This has a large effect on the family system and may in turn influence an employee's performance at work. It is important to explore the effects of business travel on the commuter, his or her family and the organization for which the employee is working.

Extract (6) showing a segment of a student presentation illustrates a complete lack of stance features identified on Dimension 4.

3.2 Summary

The present analysis was based on the constellation of linguistic features on four dimensions of linguistic variation in the academic context. The two participants in that context differ in the way they use linguistic features on three of the four dimensions despite the fact that their communicative purpose is the same, which as discussed before, is to present new information to the audience: to present in front of an audience with the same communicative purposes, which is to convey information potentially new to the audience, and to explain concepts and methods.

4. Conclusion and implications

This study explored how teachers and students use language differently while presenting in front of an audience in an academic context. Language variation was measured on four dimensions of language use in the university. We found differences between the two participants on three of the four measures, namely, oral/literate discourse, procedural versus content-focused discourse, and teacher-centered stance.

First, the fact that teachers overuse features of oral discourse while presenting information in the classroom is almost inevitable. Teachers tend to present information in a way that is understandable to the students. Depending on how they sense the degree of student engagement in the given moment, they may spontaneously edit their own text. They may insert elaborations, explanations, and examples in their presentation expressed in a way that resembles oral discourse. They are the expert professionals with a high social status who know the audience well. In this relationship, the power is held in their hands. Given all of these characteristics, it is not surprising that a more informal kind of style may be acceptable, one that is close to the highly interactive, question-answer type of register such as, service encounters. On the other hand, the fact that students seem to use more literate features is also expected. They present at the symposium in front of an audience unknown to them, and not only are they novice professionals, but they are being judged by experts in the field. Also, the text cannot be negotiated or edited on the spot. Therefore, the language that they use tends to be closer to that of written materials such as, textbooks and course packs.

Second, differences between the two participants in the features associated with procedural versus content-focused discourse are also inevitable. Interestingly enough, even though teachers tend to overuse features of oral discourse, as discussed above, they also tend to use more of the features associated with content-focused discourse. Again, this suggests that it is not the content that the teachers are trying to negotiate or alter but it is the way they present the content that differs largely from students. This could be due to their role in the context itself, and

again, their expertise in the field and their role as pedagogues. In contrast, students use fewer features of content-focused discourse, and so, their talk resembles much more to students in study groups. This is not surprising and only shows how different the two speakers are even though in a similar professional situation.

Third, there is a difference between teachers and students using teacher-centered stance features. Students, again, use language in their presentations in terms of stance features similar to how they talk in their study group sessions. This is markedly different from teachers, whose language exhibits a higher degree of stance feature use.

The study has further implications and suggests further research in multiple areas. First, although differences were shown in presentation styles between teacher and students, an intervening variable, namely discipline, may have an effect. Further research could find out whether discipline has a greater effect or whether the two variables together (discipline and speaker) may account for the variability in the dimension scores.

Second, further research with a three-way comparison could show how students present in the classroom setting versus when they are in a symposium setting, and how they differ from teachers in a presentation situation in the classroom situation (Csomay 2013). Furthermore, a four-way comparison can show how teachers and students present in the classroom and in conferences.

Finally, the results here also demystify the misconception that the language of university classes always reflects a dense informational package. Instead, and as my earlier work has shown, university classroom talk displays a mix of literate (informational) discourse and oral discourse, therefore, places this register in the middle of an oral-literate continuum (Csomay 2006). While this is not surprising to some, it may be unexpected to those who think that classroom discourse is solely relying on difficult vocabulary and complex grammar. Earlier research has also shown that features of informational discourse change depending on discipline, interactivity, and level of instruction (Csomay 2002) as well as depending on the structure of classroom discourse (Csomay 2005; Csomay 2007b). Additional research has looked at lexical features in the classrooms as well, showing that students may have trouble with the constantly alternating style between oral and literate discourse rather than the specialized vocabulary.

References

Atkinson, David. 2001. Scientific discourse across history: A combined multidimensional/rhetorical analysis of *The Philosophical Transactions of the Royal Society of London*. In *Variation in English: Multidimensional Studies*, Susan Conrad & Douglas Biber (eds), 45–65. London: Longman.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 1995. *Dimensions of Register Variation. A Cross-Lingusitic Comparison.* Cambridge: CUP. DOI: 10.1017/CBO9780511519871

Biber, Douglas. 2006. *University Language. A Corpus-based study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. DOI: 10.1075/scl.23

Biber, Douglas & Burges, Jena. 2001. Historical shifts in the language of women and men. In *Variation in English: Multi-Dimensional Studies,* Susan Conrad & Douglas Biber (eds), 21–37. London: Longman.

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style.* Cambridge: CUP. DOI: 10.1017/CBO9780511814358

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, Conrad, Susan, Reppen, Randi, Byrd, Patricia & Helt, Marie. 2002. Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly* 36(1): 9–48. DOI: 10.2307/3588359

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Patricia, Helt, Marie, Cortes, Viviana, Csomay, Eniko & Urzúa, Alfredo. 2004. *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* [TOEFL Monograph series (MS-26)]. Princeton NJ: Educational Testing Service.

Conrad, Susan. 1996. Academic Discourse in Two Disciplines: Professional Writing and Student Development in Biology and History. Ph.D. dissertation, Northern Arizona University.

Conrad, Susan. 2001. Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In *Variation in English: Multidimensional Studies*, Susan Conrad & Douglas Biber (eds), 94–107. London: Longman.

Conrad, Susan & Biber, Douglas. (eds), 2001. *Variation in English: Multidimensional Studies*. London: Longman.

Connor-Linton, Jeffrey & Shohamy, Elana. 2001. Register variation, oral proficiency sampling, and the promise of multi-dimensional analysis. In *Variation in English: Multidimensional Studies*, Susan Conrad & Douglas Biber (eds), 124–137. London: Longman.

Csomay, Eniko. 2002. Variation in academic lectures: Interactivity and level of instruction. In *Using Corpora to Explore Linguistic Variation* [Studies in Corpus Linguistics 9], Randi Reppen, Susan Fitzmaurice & Douglas Biber (eds), 203–224. Amsterdam: John Benjamins. DOI: 10.1075/scl.9.14cso

Csomay, Eniko. 2005. Linguistic variation within university classroom talk: A corpus-based perspective. *Linguistics and Education* 15(3): 243–274. DOI: 10.1016/j.linged.2005.03.001

Csomay, Eniko. 2006. Academic talk in American university classrooms: Crossing the boundaries of oral – literate discourse. *Journal of English for Academic Purposes* 5: 117–135. DOI: 10.1016/j.jeap.2006.02.001

Csomay, Eniko. 2007a. A corpus-based look at linguistic variation in classroom interaction: Teacher talk versus student talk in American university classes. *Journal of English for Academic Purposes* 6: 336–355. DOI: 10.1016/j.jeap.2007.09.004

Csomay, Eniko. 2007b. Vocabulary-based discourse units in university class sessions. In *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure* [Studies in Corpus Linguistics 28], Douglas Biber, Ulla Connor & Thomas Upton (eds), 213–238. Amsterdam: John Benjamins. DOI: 10.1075/scl.28.11cso

Csomay, Eniko. 2013. A corpus-based analysis of student talk in the university setting. Paper presented at the annual conference by the American Association of Applied Linguistics, Dallas TX, March 16–19.

Duranti, Alessandro. 1985. Sociocultural dimensions of discourse. In *Handbook of Discourse Analysis*, Teun van Dijk (ed.), 193–230. New York NY: Academic Press.

Halliday, Michael A.K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. London: Edward Arnold.

Helt, Marie. 2001. A multi-dimensional comparison of British and American spoken English. In *Variation in English: Multidimensional Studies*, Susan Conrad & Douglas Biber (eds), 171–183. London: Longman.

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. *Hedging in Scientific Research Articles* [Pragmatics & Beyond New Series 54]. Amsterdam: John Benjamins. DOI: 10.1075/pbns.54

Hymes, Dell. 1972. On communicative competence. In *Sociolinguistics. Selected Reading* [Penguin Modern Linguistics Reading], John B. Pride & Janet Holmes (eds), 269–93. Harmondsworth: Penguin.

Reppen, Randi. 2001. Register variation in student and adult speech and writing. In *Variation in English: Multidimensional Studies*, Susan Conrad & Douglas Biber (eds), 187–199. London: Longman.

Samraj, Betty. 2008. A discourse analysis of master's theses across disciplines with a focus on introductions. *Journal of English for Academic Purposes* 7(1): 55–67. DOI: 10.1016/j.jeap.2008.02.005

Simpson, Rita & Swales, John (eds). 2001. *Corpus Linguistics in North America. Selections from the 1999 Symposium*. Ann Arbor MI: University of Michigan Press.

Swales, John. 2005. Attended and unattended "this" in academic writing: A long and unfinished story. *ESP Malaysia* 11: 1–15.

Tao, Honyin 2003. Turn initiators in spoken English: A corpus-based approach to interaction and grammar. In *Corpus Analysis. Language Structure and Language Use* [Language and Computers: Studies in Practical Linguistics 46], Pipin Leistyna & Charles Meyer (eds), 187–208. Amsterdam: Rodopi.

Wulff, Steffi, Römer, Ute & Swales, John. 2012. Attended/unattended *this* in academic student writing: Quantitative and qualitative perspectives. In *Contemporary Approaches to Discourse and Corpora*, Csomay, Eniko (ed.), Special issue of *Corpus Linguistics and Linguistic Theory* 8(1): 129–157.

Appendix A

Key situational differences between textbooks and classroom teaching (Biber & Conrad 2009: 65)

	Textbook	Classroom teaching
Participants	One author addressing un-enumerated readers	One instructor addressing relatively few students
Relations among participants	No interaction Author has more knowledge All participants have some specialized knowledge No personal relations	Interaction possible Instructor has more knowledge All participants have some specialized knowledge Instructor knows students
Channel	Written	Spoken
Production circumstances	Text carefully planned, edited, revised	Text planned but cannot be edited or revised
Setting	Unknown	Participants in same physical space
Communicative purpose	Convey information Explain concepts of methods	Convey information Explain concepts of methods Convey personal attitudes Direct students what to do

Appendix B

Dimension 1 positive: contractions, pronouns, present and past tense, mental, activity, and communication verbs, time, place, and likelihood adverbials, hedges and discourse particles, wh-questions, clausal coordination, stranded prepositions, conditional and causative clausal adverbials, and *wh*- and *that*- complement clauses

Dimension 1 negative: common nouns, nominalizations, nouns classified into semantic categories such as abstract nouns, group nouns, human nouns, and mental nouns, long words, phrasal coordination, prepositional phrases, attributive adjectives, passives, relative clauses, a variety of *to*- clauses and phrasal coordination.

Dimension 2 positive: necessity and future (predictive) modals, causative and activity verbs, second person pronouns, group nouns, *to*- clauses with desire verbs, and conditional adverbial clauses.

Dimension 2 negative: rare vocabulary items, adjectives describing size, *to*- clauses with probability verbs and by-passives.

Dimension 3 positive: third person pronouns, past tense, communication and mental verbs, and *that*-complement clauses with likelihood verbs, *that*-deletion.

Dimension 3 negative: concrete, technical, and quantity nouns.

Dimension 4 positive: relative clauses with *that* as relative pronoun, stance adverbials of certainty, likelihood and attitudinal, adverbial clauses of condition, and *that*- clauses controlled by stance nouns.

Dimension 4 negative: *wh*- questions and stranded prepositions.

CHAPTER 2

Telephone interactions

A multidimensional comparison

Eric Friginal
Georgia State University

This chapter presents the functional features of linguistic dimensions from three telephone-based interactions: (1) customer service transactions (Call Center corpus), (2) telephone conversations between friends and family members (Call Home corpus), and (3) spontaneous telephone exchanges between participants discussing topics identified by fixed prompts (Switchboard corpus). These three telephone-based corpora are then compared with data from face-to-face English conversation (American English Conversation corpus). Linguistic comparisons across these registers followed a corpus-based, multidimensional analytical approach developed by Biber (1988) using established dimensions of customer service talk from Friginal (2008). Results suggest that variation in these spoken interactions is largely influenced by the nature of conversational tasks and the use of the telephone as a medium in communicating ideas, opinions, or instructions.

Keywords: Multidimensional analysis; spoken corpora; telephone interactions

1. Introduction

Telephone interactions have been explored by applied linguists typically by looking at the flow of talk through the analysis of sociophonetic structures of speech (Orr 2003), transactional and interactional dialogues (Cheepen & Monaghan 1990; Cheepen 2000), and how speakers complete specific tasks through turn-taking and related turn features such as interruption, overlaps, and latching (Schegloff 2001; Gardner & Wagner 2004). In addition, sociopragmatic issues in telephone talk are examined with a substantial degree of interest by many discourse or conversation analysts. For example, Cameron (2008) considers top-down talk in call centers based in the United Kingdom (U.K.) and investigates the flow of speech and the

DOI 10.1075/scl.66.02fri

use of language that is highly regulated and standardized. Economidou-Kogetsidis (2005) investigates directness and politeness variables between Greek and British callers in telephone service encounters. In this study, Greek callers were found to be more direct in expressing requests or asking for specific information than British callers. In a sense, this directness in speech was accomplished through the repetitive use of *parakalo*, the Greek equivalent of *please*. A study by Silvester and Anderson (2003) compares the structure of face-to-face and telephone employment interviews focusing on interviewers' questioning strategies and applicants' causal responses and attributions. They report, in part, that applicants produce more causal attributions (i.e. responses indicating the relationship between events, outcomes and/or behaviors, and their causes) in telephone interviews, resulting in slightly higher ratings from interviewers compared to face-to-face interviews.

A large-scale study of business telephone interactions was conducted by Friginal (2008, 2009, 2010, 2011), investigating telephone-based customer service transactions using a corpus of outsourced call center texts between Filipino call-takers and callers based in the United States (U.S.). These interactants engage in various types of communicative tasks, e.g. troubleshooting a technical problem or processing orders for a wide range of products, with defined speaker roles similar to a business service encounter (i.e. server vs. servee or agent vs. customer). Friginal's primary sets of foci include the dynamics of cross-cultural communication between participants, gender of speakers, call-takers' experience in phone support and quality of service performance, and the linguistic structure of communicative tasks in customer service interactions. Many other studies of globalized call center interactions have been conducted in the past 10 years matching the growth of the outsourcing industry in the Philippines and India (e.g. Cowie 2007; Forey & Lockwood 2010; Lockwood 2012). Among these, Poster (2007) and Taylor and Bain (2005) look at labor practices in Indian call centers that require Indian agents to pose as Americans for American call centers, or British for those that serve companies located in the U.K. These two studies focus on the effects of globalization in social and national identity against the structure of English used in cross-cultural telephone service encounters.

Over the years, a methodical description of specific register features of spoken discourse has been achieved through corpus analysis. Corpus-based comparisons across transcribed texts have shown variations in the use of lexical and syntactic choices of participants in many spoken contexts. Quaglio (2009) and Alsurmi (2012), for example, identify the linguistic characteristics of speech from a television sitcom and selected soap operas for comparison with real-world conversations. These studies reveal important functional differences between television dialogues and naturally-occurring "real-world" conversation. Adolphs, Brown, Carter,Crawford and Sahota (2004) explore the application of corpus

methodologies in health care encounters in order to describe the characteristics of communication events in clinical settings. Using a corpus of staged telephone conversations between patients and clinicians, the researchers are able to show several linguistic characteristics of the strategies used by healthcare professionals in addressing caller/patient concerns. Other related studies have analyzed the distribution of linguistic features of spoken texts, e.g. stance expressions in classroom management (Biber 2006), *so* and *oh* in social interactions (Bolden 2006), or features of accommodation and involvement in class lectures (Barbieri 2008). Results from these analyses have shown unique distributional data of speech characteristics and linguistic strategies employed by speakers across spoken registers.

1.1 The focus of this chapter

This chapter presents the functional features of linguistic dimensions from three telephone-based interactions: (1) customer service transactions, (2) telephone conversations between friends and family members, and (3) spontaneous telephone exchanges between participants discussing topics identified by fixed prompts. These groups of texts are taken from a Call Center corpus collected by Friginal (2008, 2009), the Call Home corpus, and a sub-section of the Switchboard corpus, respectively. The Call Center and Switchboard corpora were obtained from the American National Corpus (ANC) (see ANC's website at: ⟨http://www.americannationalcorpus.org/⟩) and through the Corpus Linguistics Program at Northern Arizona University. The three telephone-based corpora are then compared with data from face-to-face English conversation from the American English Conversation (AmE Conversation) corpus collected by Longman. Linguistic comparisons across these registers followed a corpus-based, multidimensional approach developed by Biber (1988) using established dimensions of customer service talk from Friginal (2008). The three functional dimensions of call center talk from Friginal's original factor analysis (of only texts from his Call Center corpus) are: (1) addressee-focused, polite, and elaborated information vs. Involved and simplified narrative; (2) planned, procedural talk; and (3) managed information flow.

1.2 Multi-feature, multidimensional analytical framework

Biber's (1988) multi-feature, multidimensional analytical (MDA) framework has been applied in the study of a range of spoken and written registers and used in the interpretation of various linguistic phenomena. MDA data come from factor analysis (FA) which considers the sequential, partial, and observed correlations of a wide-range of variables producing groups of occurring factors or dimensions. According to Tabachnick and Fidell (2001), the purposes of FA are to summarize patterns of correlations among variables, to reduce a large number of observed

variables to a smaller number of factors or dimensions, and to provide an operational definition (i.e. a regression equation) for an underlying process by using these observed variables. The purposes of FA support the overall focus of corpus-based MDA which aims to describe statistically correlating linguistic features and group them into interpretable sets of linguistic dimensions. The patterning of linguistic features in a corpus creates linguistic dimensions which correspond to salient functional distinctions within a register, and allows cross-register comparison. MDAs of spoken registers have covered topics such as gender and diachronic speech (Biber & Burges 2001; Rey 2001), stance and dialects (Precht 2000), televised cross-cultural interaction (Connor-Linton 1989; Scott 1998), and job interviews (White 1994).

1.3 Friginal's (2008) dimensions of call center interactions

For the purposes of this chapter, established dimensions from Friginal (2008) were used to compare the distribution of linguistic features from three groups of telephone-based registers and one set of texts of face-to-face interactions. The composition of the tag-counted features for Friginal's FA was based primarily on prior studies, especially Biber (1988) and White (1994). Additional discourse features of telephone-based service transactions (e.g. filled-pauses, politeness markers, length of turns) were included in the dataset. Table 1 shows the complete list of tagged features (37 total lexical and syntactic features) used in this FA.

Table 1. Complete list of linguistic features used in Friginal (2008)

Linguistic features	**Description/Example**
Type/Token	Number of words occurring in the first 400 words of texts
Word Length	Mean length of words in a text (in letters)
Word Count	Total number of words per agent/caller texts
Private Verbs	e.g. *anticipate, assume, believe, feel, think, show, imply*
That Deletion	e.g. *I think* [Ø] *he's gone.*
Contractions	e.g. *can't, I'm, doesn't*
Present Tense Verbs	All present tense verbs identified by the tagging program
2nd Person Pronouns	*you, your, yours, yourself* (and contracted forms)
Verb *Do*	*do, does, did* (and contracted forms)
Demonstrative Pronouns	*that, those, this, these*
1st Person Pronouns	*I, me, my, mine, myself* (plural and all contracted forms)
Pronoun *It*	Instances of pronoun *It*
Verb *Be*	Forms of *Be* verb

(*Continued*)

Table 1. (Continued)

Linguistic features	Description/Example
Discourse Particles	e.g. *oh, well, anyway, anyhow, anyways.*
Possibility Modals	*can, could, might, may.*
Coordinating Conjunctions	*and, or, but.*
WH Clauses	Clauses with WH (*what*, *which, who*) head.
Nouns	All nouns identified by the tagging program.
Prepositions	All prepositions identified by the tagging program.
Attributive Adjectives	e.g. *the small chair.*
Past Tense Verbs	Past tense verbs identified by the tagging program.
Perfect Aspect Verbs	Verbs in perfect aspect construction.
Nominalizations	Words ending in *-tion, -ment, -ness*, or *-ity* (and plurals).
Adverb Time	Time Adverbials e.g. *nowadays, eventually.*
Adverbs	total Adverbs (not Time, Place, Downtoners, etc.).
Prediction Modals	*will, would, shall.*
Verb *Have*	*has, have, had* (and contracted forms).
Average Length of Turns	Total number of words divided by number of turns.
Filled-Pauses	*uhm, uh, hm.*
Respect Markers	*ma'am, Sir.*
Politeness Markers – *Thanks*	*thank you*, *thanks*, [*I*] *appreciate* [*it*].
Politeness Markers – *Please*	*please.*
Discourse Markers – *OK*	*ok* (marker of information management).
Discourse Markers – *I mean*	*I mean* and *You know* (marker of participation).
Discourse Markers – *Next/Then*	*next, then* (temporal adverbs).
Discourse Markers – *Because*	*because, 'coz, so* (marker of cause and result).
Let's or *let us*	Instances of *let's* or *let us.*

The final composition of the three extracted factors (i.e. linguistic dimensions) of call center interactions is presented in Table 2. Factor loadings and subsequent functional interpretations of each dimension is also presented in this table and the following sections. Discourse particles, 2nd person pronouns, average word length, total word count, length of turns, and type/token ratio loaded highly in the three factors. Friginal's (2008) FA reported that Kaiser-Meyer-Olkin Measure for Sampling Adequacy (KMO = .724, middling) and Bartlett's Test for Sphericity (Approx. Chi-Square = 13101.705, df = 667; $p < .0001$) were sufficient for exploratory FA with principal axis factoring. Results from a three-factor solution were deemed to be the most interpretable merging of features, with 34.29 cumulative percentage of Initial Eigenvalues (Total Variance Explained).

Table 2. Summary of the linguistic features of the three factors extracted from the Call Center corpus

Dimension	Features	
Dim 1:	**Dim 1 Positive: Addressee-focused, polite, and elaborated information**	
	2nd Person Pronouns	.683
	Word Length	.612
	Please	.523
	Nouns	.515
	Possibility Modals	.445
	Nominalizations	.394
	Length of Turns	.376
	Thanks	.325
	Ma'am/Sir	.309
	⇳	
	Dim 1 Negative: Involved and simplified narrative	
	Pronoun *It*	–.687
	1st Person Pronouns	–.663
	Past Tense Verbs	–.609
	That Deletion	–.506
	Private Verbs	–.439
	Perfect Aspect Verbs	–.345
	I mean/You know	–.338
	Verb *Do*	–.321
Dim 2:	**Dim 2 Positive: Planned, procedural talk**	
	Word Count	.821
	Length of Turns	.678
	Type/Token	.630
	2nd Person Pronouns	.515
	Next/Then	.417
	Word Length	.422
	Adverb Time	.409
	Prepositions	.383
	Please	.369
	Present Tense Verbs	.341
	Nominalizations	.321
	Because/So	.310
	Let's	.300
	⇳	
	Dim 2 Negative: (no title)	
	Discourse Particles	–.397
Dim 3:	**Dim 3 Positive: Managed information flow**	
	Discourse Particles	.947
	OK	.865
	Adverbs	.845
	Let's	.422
	⇳	
	Dim 3 Negative: (no title)	
	Length of Turns	–.349

2. Method

2.1 Corpora

Table 3 summarizes the composition of the corpora used for register comparison in this chapter. A brief description of these four corpora is provided below.

Table 3. Composition of corpora used in the present study

Corpora	Number of text files	Number of words	Average number of words per text file
(1) Call Center	500	553,765	1,108
(2) Call Home	120	345,237	2,876
(3) Switchboard	600	1,057,830	1,763
(4) American Conversation	200	1,166,105	5,828

2.2 Call Center corpus

The corpus of call center transactions was collected by Friginal from 2006 to 2007 in the Philippines from a sponsoring call center company that uses web-based software for storing audio files of transactions for quality monitoring and documentation of transactions. The calls in the corpus ranged from five to 25 minutes in duration. The 500 audio files that comprise the Call Center corpus have an average call duration of eight minutes and 45 seconds per transaction and have a combined length of over 73 hours of customer service interactions. Convenience sampling of audio files was done to ensure, among other considerations, a comparable number of files per task category (e.g. troubleshooting, telemarketing) or a balanced number of male and female call-takers and callers as much as possible. The audio files of customer calls were transcribed by trained Filipino transcriptionists following conventions used in the collection of the service encounter corpus of the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL), (Biber 2006).

2.3 Call Home corpus

The Call Home corpus (or "Call Home English Corpus of Telephone Speech") consists of 120 unscripted and unplanned telephone conversations between native speakers of American English. All calls, mostly lasting up to 30 minutes in length, originated in the U.S.; however, 90 of the 120 calls were directed or placed to different locations outside of North America. Most participants called family members or close friends following specific instructions and suggested topics developed during data collection. The Call Home corpus from the Linguistic Data Consortium (LDC) contains speech data files and minimal amount of

documentation needed to describe the contents and format of speech files and the software packages needed to uncompress the speech data ("Call Home American English Speech" 2004).

2.4 Switchboard corpus

The Switchboard corpus is comprised of spontaneous conversations of "telephone bandwidth speech" between American speakers. The corpus was collected by Texas Instruments and funded by the Defense Advanced Research Projects Agency (DARPA). A complete set of Switchboard CD-ROMs available from the Linguistic Data Consortium includes about 2,430 conversations averaging six minutes in length (with over 240 hours of recorded speech), and about three million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English ("Switchboard: A Users' Manual" 2004). A total of 600 files with approximately over one million words comprise the Switchboard sub-corpus used in this chapter. Interaction with the switchboard system was conducted via touch-tones and recorded instructions given to the participants. The topics for conversation (e.g. "*What do you think about dress codes at work?*" or *"How do you feel about sending an elderly family member into a nursing home?"*) were randomly identified by the system. The two speakers, once connected, were allowed by the system to "warm-up" before recording began. The speakers did not know each other personally and have no previous information about each other's personal background before the warm-up conversation.

2.5 American English (AmE) Conversation corpus

The American English Conversation corpus used in this chapter was obtained from the Longman Grammar Corpus of Spoken American English. The Longman Grammar corpus has approximately over four million words and was designed to be a representative corpus of American conversation covering a wide-range of speech types (e.g. casual conversation, service encounters, task-related interaction), locations or settings (e.g. home, classroom), geographic regions in the U.S., and speaker characteristics (e.g. age, gender, occupation). Only text files of face-to-face conversations from this corpus were used in the present study. The American Conversation sub-corpus has a total of 200 texts with approximately 1.1 million words.

2.6 Computing dimension scores

The frequencies of co-occurring linguistic features from the three dimensions (Table 2) were standardized (using z-scores) across four corpora, allowing highly

different distributions to be more comparable with each other and offering scores that reflected a feature's range of variation. Each dimension comprised linguistic features that significantly co-occurred with one another and contained both positive and negative loadings. Standardization of frequencies allowed for these complementary patterns of polarity. In other words, when a text contains frequent instances of one group of co-occurring linguistic features (positive or negative), the features from the opposite group are likely to be absent (Biber 1988). Using the composition of Friginal's dimensions, the standardized frequency data (z-scores) from the four corpora in the present study were then added to obtain dimension scores per individual text. Once scores in all four dimensions had been calculated for each text, mean scores per corpus were obtained by averaging the texts' dimension scores.

3. Results

For each of the three dimensions from Friginal (2008), four average scores comprising the corpora for the present study are shown along comparison figures below. These figures describe cross-register linguistic distributions and relationships per dimension. Text samples with high or low dimension scores are provided in the following sections to better understand the functional characteristics and significance of these distributions.

3.1 Dimension 1: Addressee-focused, polite, and elaborated information vs. Involved and simplified narrative

Eighteen (18) linguistic features comprise this dimension with nine features on each of the positive and negative sides. Positive features include politeness and respect markers (e.g. *thanks*, *please*, *ma'am* and *sir*), markers of elaboration and information density (e.g. long words and turns, nominalizations, and more nouns), and 2nd person pronouns (e.g. *you*, *your*) which indicate "other-directed" focus of talk. Possibility modals (*can*, *could*, *may*, *might*) also loaded positively on this factor. The features on the negative side of this factor, especially pronoun *it*, 1st person pronouns, *that* deletion, private verbs, WH clauses, and verb *do*, resemble the grouping in the dimension "Involved Production" identified by Biber (1988) and White (1994). These features are typical of spoken texts and generally contrast with written, informational, and planned discourse. Also on the negative side of the factor are past tense verbs, perfect aspect verbs, and the use of discourse markers *I mean* and *You know*. These elements point to an accounting of personal experience or narrative that tries to explain the occurrence of a particular situation or event. Schiffrin (1987) considers *I mean* and *You know* as markers

of information and participation; *I mean* marks speaker orientation toward the meaning of one's own talk while *You know* marks interactive transitions.

These co-occurring set of features represent the contrast between the dominant objectives of speakers' utterances. Speakers in telephone exchanges who use more positive features are likely aiming to give details, explanations, or solutions (especially in the case of customer service call-takers). In the process, these interactants use more nouns, nominalizations, and longer utterances or turns to deliver the information. The information density in these turns is high because of higher average word lengths in the texts. Participants' turns are elaborated with detailed explanations, likelihood, or risks through the use of a significant high frequency of possibility modals. The high frequency of 2nd person pronouns indicates that the transfer of information is highly addressee-focused.

Conversely, the grouping of features on the negative side of the dimension illustrates personal narrative and experiences, and simplified information. The combination of past tense verbs, private verbs, pronoun *it*, and discourse markers *I mean* and *You know* demonstrates the typical goal of utterances which is to provide a personal account on how a situation or an event happened. Involved production features (e.g. 1st person pronouns, WH clauses, verb *do*, and *that* deletion) and *I mean, You know* serve a communicative purpose to establish personal orientation (White 1994) and purposely ask for a response. Most utterances on the negative side of the dimension have fewer word counts and are significantly shorter in length. To summarize, the combination of positive and negative features of Dimension 1 differentiates between addressee-focused, polite, and elaborated information and involved and simplified narrative portraying how informational content is produced in the discourse. Figure 1 shows the range of variation across the four corpora.

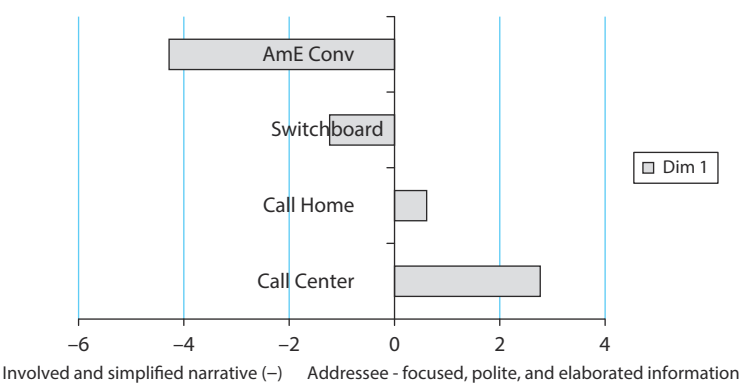

Figure 1. Comparison of dimension scores for Dimension 1: Addressee-focused, polite, and elaborated information vs. Involved and simplified narrative ANOVA: Registers, $F = 5.212$; $p < .001$

Call Home (0.611) and Call Center (2.768) texts have scores on the positive side of the dimension while AmE Conversation (−4.281) and Switchboard (−1.231) have scores on the negative. Telephone service encounters commonly allocate for courteous language and the recognition of roles; call-takers, especially, are expected to show respect and courtesy in assisting their customers (D'Ausilio 1998). In this dimension, the frequency of politeness and respect markers in the Call Center corpus is significantly higher than the other three comparison corpora. In fact, these features were seldom used in face-to-face texts. Both Biber (1988) and White (1994) characterize spoken discourse as highly involved and interactive from increased use of pronouns, private verbs, and discourse markers. Linguistic features that show spoken narratives (e.g. past tense verbs, pronouns, *that*-deletion, etc.) are also very common in these interactions especially in face-to-face conversations and also in Call Home (e.g. narratives and accounts of experiences or events).

The two text excerpts below highlight the use of past tense verbs and personal pronouns in face-to-face conversations against polite, elaborated and informational utterances from a call-taker in customer service interactions. The call center text (Text Sample 1: task – purchase Mobile Phone Minutes, Dim Score = 5.713) shows detailed explanation and additional information given to the caller. Technical information, business-related items, and politeness markers are all used by the call-taker in this excerpt. The call-taker engages the caller by using conventional customer service responses (e.g. *"I apologize for the inconvenience.."* or *"Let me just verify the charges.."*). In Text Sample 2 (setting: office/lunch time talk, Dim Score = −6.231), the two speakers discussed two overlapping set of past events (bachelor party and previous work experience in North Carolina).

Text Sample 1. Purchase mobile phone minutes (Dim Score = 5.713) (name replaced by pseudonym)

Caller: Yes, uh, when are you guys gonna go back telling us when how much time is left on these phone cards? I mean on these phones?

Call-taker: **I apologize for the inconvenience sir**, I'll, **let** me explain on that **ok**? **Please**, give me **your** cell phone number so I can check on **your** minutes

Caller: [phone number], I think it has run out because I wanted to use it but it said it didn't have enough time

Call-taker: **Ok, let me just verify the charges** at the moment, **please** give me **your** name and address on the account please

Caller: John A. Smith, 2635 […] Road, in […] Ohio

Call-taker: **Thank you** for that Mr. Smith, **let me just pull out your account to check your balance**, **ok**? Mr. Smith, **you** have now zero balance on the account and uh, ok Mr. Smith, **you** are notified of **your** balance when **you** reached below $10, below [interruption]

Caller: There never was a word said anytime, I never heard anything, how am I supposed to be notified?

Call-taker: I see, well **sir** do **you**, **ok** just a moment, while I check on **your** account. **Ok, sir** did **you** give out any e-mail address where we can send updates regarding **your** account?

Caller: Yeah I did, but I don't know, my computer is down right now

Call-taker: For the meantime Mr. Smith, **you** can also check **your** balance on **your** phone by calling [phone number], and that is a free call always. Just choose the option for **you** to receive the minutes on **your** account, either through uh, or via text message or by speaking to a live agent, ok? For the meantime Mr. Smith, **you** have a zero.

Text Sample 2. Talk during lunch time (Dim Score = −6.231)

Speaker 1: It's like Greg um, we **had** the lovely bachelor party at our house for a friend and I **was** like fumigating my house when it **was** over.

Speaker 2: Cigars?

Speaker 1: Oh that **wasn't** it. My towels **smelt** bad. But it **wasn't** the cigars, it **was**, I could **have handled** ⟨unclear⟩, it **was** bad stuff. And you know that **wouldn't have even bothered**. I mean I can handle that.

Speaker 2: Excuse me?

Speaker 1: Oh, oh that **was**, it **was** really, really and the thing **was** … well not … uh, let me look at your beer menu if that's here.

Speaker 2: We'll, we'll snooze through the movie this afternoon but hey that's okay.

Speaker 1: We'll find out so Greg **told** me … **told** me in the house right? He **told** me, this **happened** a year ago, he **told** me … I **had** uh, when I **checked** into my first duty station in the service I **was** in North Carolina and uh, fortunately probably **was** one of the finest working experiences I've **had** in my whole working life. They **were** real serious about their work, **took** it very seriously but they also **didn't** take themselves too seriously. And it **was**, senior NCO's down to the individuals **had** fun and yeah.

Speaker 2: I'm watching to see if it does. It's got the record level on.

Speaker 1: No I don't think so. Back up here. Everything's good?

3.2 Dimension 2: Planned, procedural talk

The items loading on the positive side of Dimension 2 include lexical specificity and information density features (type/token ratio, average word length), temporal adverbs (*next/then*) and specific time adverbials (e.g. *eventually, immediately*), complex and abstract information features (word count, length of turns, and nominalization), 2nd person pronouns, prepositions, cause and result discourse markers (*because/so*), politeness marker *please*, present tense verbs, and *let's* (including *let us*). Only discourse particles (e.g. *oh*, *well, anyway*) loaded on the negative side. The positive side of the dimension, thus, signifies a one-way (addressee-focused) transfer of a large amount of abstract and technical information. In this case, the

information appears to be "real-time," procedural or process-based due to the presence of temporal adverbs combining with the imperative *let's*, prepositions (e.g. *in, on, below, above*), and, especially, present tense verbs. The frequent occurrence of present tense verbs in the texts illustrates the use of directives/imperatives in utterances (e.g. "*..then hit save*"; "*..now, remove the tracking tape…*"). It appears that this form of instructional language, especially common in call center talk, is expressed through a series of directions marked by 2nd person pronouns (especially *you* and *your*), succession between steps (*next/then*) and progression through the discourse (*now*). Discourse particles, used very sparingly in this dimension, perhaps indicate that the utterances are somewhat prepared or organized, and produced with limited hesitations or tentativeness.

As shown in Figure 2, Dimension 2 differentiates call center interactions from the three comparison spoken corpora in the present study. Face-to-face interactions, Switchboard discussions of topics, and telephone interactions between family members all have negative aggregate scores. These three corpora have a higher frequency of discourse markers which are in complementary distribution with temporal adverbs, cause and result discourse markers, and especially imperative *let's*. Friginal (2008) suggests that the merging of features indicating lexical specificity, complexity, and abstraction of information helps to differentiate call center discourse from general conversation and other sub-registers of telephone talk. In typical customer service calls, longer words (based on average word lengths) and technical vocabulary are often used in extended turns during the interaction. Information-packaging in call center discourses is, therefore, somewhat more similar to written, planned texts because of the presence of features that are not commonly produced online, including nominalizations and higher type/token ratio. Biber (1988) states that these features are more common in academic written

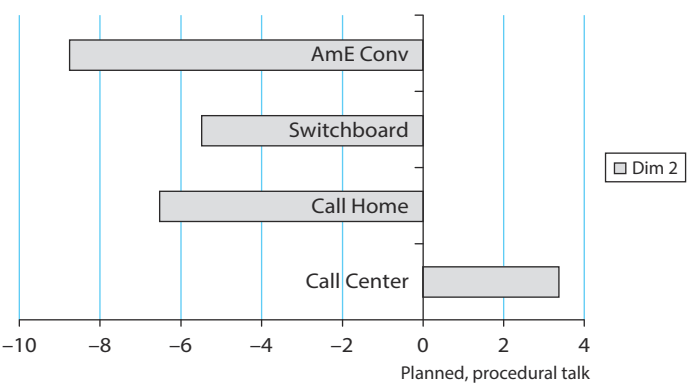

Figure 2. Comparison of dimension scores for Dimension 2: Planned, procedural talk
ANOVA: Registers, F = 30.134; p < .0001

texts and less observed in spoken texts because of the influence of production circumstances. In typical, online conversations, general topic shifts allow for the occurrence of more common words and phrases and limited complex or abstract vocabulary.

The Call Center corpus has a collective Dim Score of 3.379 compared to all negative averages from AmE Conversation (–8.774), Call Home (–6.521), and Switchboard (–5.483). Procedures and instructions are not common in face-to-face interactions unless they involve the performance of tasks. In Switchboard, there are instances of short, procedural discourse especially in the beginning of the discussions when participants talk about the instructions following the automated prompts during the recording of their conversations. However, these instructions echoed by the speakers are also limited and not extensively repeated in the exchanges. Texts from Call Home and AmE Conversations have significantly higher frequencies of discourse particles such as *oh, well,* and *anyway/s*.

In call centers, call-takers use more of the features on the positive side of Dimension 2 and predictably engage in directive, procedural talk more than other speakers across registers. Call-takers' speech in this dimension is produced online but covers a wide-range of topics and makes use of a variety of specialized terms or jargon that comprise their set scripts (see Text Sample 3 below). In a way, call-takers' utterances in giving directions and steps are planned, many of them written, because they have clear expectations about the variety of caller questions they respond to. The moves in assisting callers are well-defined, and procedures are commonly established during many on-the-job training programs. For example, memorized procedural scripts (e.g. "*… thank you for your call, first I will ask you for your account number...*") are often part of call-takers opening sequences from prescribed protocols.

Text Sample 3 (Dim Score = 8.333) shows an excerpt of planned, procedural interaction in a troubleshooting transaction from the Call Center corpus. This excerpt shows a range of new, technical words (e.g. *T1, DSL, Voice Over IP, broadband*) and nominalizations (e.g. *documentation, possibility, connection*) that are not necessarily repeated over in the text. The use of these words increases type/token ratio, average word count, and average length of turns in procedural accounts. In contrast, Text Sample 4 from Switchboard (Dim Score = –4.212) features spontaneous discussion about the weather with short turns and some highlighted use of discourse markers, especially *well* and *oh*.

Text Sample 3. Troubleshooting interaction from the Call Center corpus (Dim Score = 8.333) (caller's name is a pseudonym)

> Call-taker: **Then go ahead** and please type in "Yes" and **then hit 9**
> Caller: Ok, and then enter again?
> Call-taker: Yes, uh-huh?

Caller: Well it just says dialing

Call-taker: Uh-huh, by the way Sarah **just give me an update** whenever the message on the screen changes so that I could uh put down **documentation** here

Caller: Ok [long pause] it says "connect phone cord and press," then it says "done press enter"

Call-taker: Hmm, it, it actually means Sarah that uhm the only reasons that the **postage machine** would say connect the "connect phone cord message" is because it's not **detecting** a **dial tone** because it's connect, it's hooked up to a wrong type of **phone line** or the **phone cord** itself is defective. **Now we need a connection**, uhm since this is a brand **new postage machine** uh there's a big possibility that the **phone line** that it's hooked up to is not correct, so uhm Sarah is it ok if I get the **phone number** where you have the **postage machine** hooked up to so that I could check if uhm if it's dialing out or not?

Caller: Yeah it's the office number

Call-taker: Are you on the same line as the **postage machine**?

Caller: Uhm well it's actually connected to a connector, well there's three of them

Call-taker: Oh you mean a splitter?

Caller: Yeah

Call-taker: **Now** that's actually the reason why it's not uh going out properly. As I said earlier uhm Sarah this **postage machine** needs a **dedicated analog line**, so when we say it's a **dedicated** line it should not be sharing the line with any other **equipment**, it should not have a **rollover system**, uhm if the number has **extensions** uh we should be sure that those uh **extensions** doesn't have any **equipment** hooked up to it, and uh when we also say **analog** we **have to make sure** that it doesn't have **T1, DSL, Voice Over IP**, or even **broadband** on it. Now the best example for a **dedicated analog** line would be your **fax line**, so if we could just [interruption]

Text Sample 4. Switchboard – weather (Dim Score = –4.212)

Speaker 1: **yeah** we set a record yesterday and uh very very windy but then today the wind has dropped off and also the temperature so

Speaker 2: **oh** very cool uh i think right now it's like **oh** sixty nine

Speaker 1: hm

Speaker 2: and that's cool for **anyways**

Speaker 1: or if it it feels cool compared to yesterday **though** but very pleasant no rain in the last month i don't think ground's very dry and

Speaker 2: our yard work everything everything is in bloom **right** so our yard work's pretty tough uh ground being dry but

Speaker 1: i guess **well** it also uh brings about allergies we're having a lot of allergies down here right now

Speaker 2: um-hum

Speaker 1: everything blooming and and the weather and uh think a lot of people have contracted uh spring fever

Speaker 2: too so had a lot of people out at work **well you know** for fishing and and uh golf reasons and things like that
Speaker 1: hm
Speaker 2: the blue flu yeah
Speaker 1: **yeah** the blue flu or the white collar flu depending on where you work i guess
Speaker 2: **oh** we have had uh as i've said we've had variable weather uh
Speaker 1: hm
Speaker 2: it has been
Speaker 1: untypically wet for this time of year
Speaker 2: hm
Speaker 1: and also we have a lot of
Speaker 2: **oh** green **you know** the grass has been growing and **well**
Speaker 1: if you look outside you
Speaker 2: would like to go out and mow your lawn if you could go out and
Speaker 1: spring and **well** i guess we're still in winter and uh

3.3 Dimension 3: Managed information flow

The linguistic features on the positive side of Dimension 3 are discourse particles (e.g. *oh, well, anyway*), the discourse marker *ok*, occurrences of *let's* (and *let us*), and adverbs (any adverb form occurring in the dictionary, or any form that is longer than five letters and ends in *-ly*). The adverbs comprising this list do not include time and place adverbials and those counted as amplifiers or downtoners. The positive features in this factor are very common in spoken registers. Discourse particles are regarded as necessary for conversational coherence (Schiffrin 1994) and in monitoring the flow of information in talk (Biber 1988; Chafe 1985; Friginal 2009). *Ok* is also regularly used in conversation and purposeful interactions like service encounters, and it serves as either a marker of information management (Schiffrin 1987) or an apparent backchannel (Tottie 1991). The use of the imperative *let's* is characteristic of interactions that especially focus on the performance of tasks (Friginal 2009). The combination of discourse particles and backchannels could be interpreted as a conversational device to maintain and monitor the overall flow of transactions. More of these features emerge because the interactions are conducted over the telephone with clearly defined turns and adjacency pairs. It is possible that backchanneling through *ok* and the use of discourse particles that initiate turns are preferred by participants in telephone interactions to avoid dead air or very long pauses.

Thus, the grouping of linguistic features in Dimension 3 signifies speakers' attempt at managing the flow of information. In call center talk, for example, this dimension separates callers and call-takers in their use of discourse particles, *ok*, and adverbials intended to facilitate and monitor the transaction. Figure 3 shows

that the three telephone-based interactions all have positive average dimension scores in Dimension 3, with Call Center having the highest average frequencies of discourse markers and *ok* (but both of these features are also commonly used in Switchboard and Call Home interactions). The use of *let's* contributes to the difference in the factor scores of the Call Center corpus against the two other telephone-based corpora. There is a high frequency of *let's* and *let us* in the turns of call-takers in call centers potentially to signal the introduction of an instruction given to the caller or customer (e.g. "*Ok, sir, let's send this order to customer service and wait for their response …*").

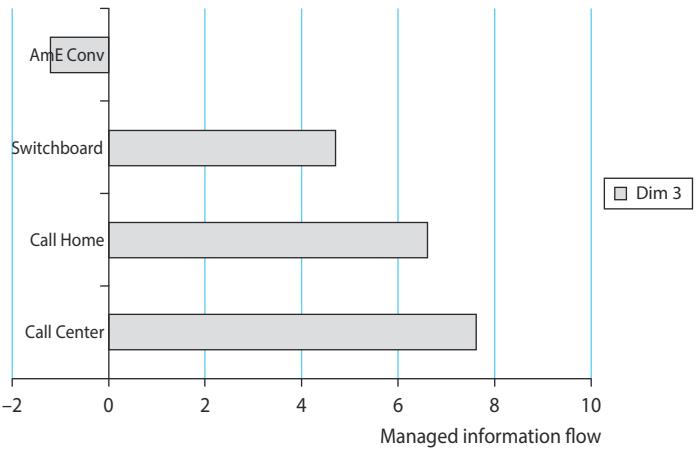

Figure 3. Comparison of dimension scores for Dimension 3: Managed information flow
ANOVA: Registers, $F = 21.852$, $p < .0001$

In call centers, the use of the positive features of Dimension 3 could be related to common conventions in customer service such as establishing rapport, avoiding dead air, as well as backchanneling to show attentiveness and focus on the customer in the transactions. In fact, Filipino agents undergo skills training in phone-handling, and some of the topics covered in many training sessions include backchanneling and providing confirmatory responses to control the flow of transactions. Some researchers have noted that Filipino agents tend to be quiet during callers' turns which may suggest to the callers limited engagement or low level of interest (Friginal 2009, 2011). Because of this awareness during training, it is possible that agents consciously backchannel in their turns. Call Home and Switchboard discussions are not primarily task-based, with limited imperative *let's/let us,* but communicative markers such as *ok*, *actually*, *basically*, *exactly*, and *anyway* are also frequent in speakers' turns. In managing the flow of information and trying to control turns in telephone-based talk, it appears that speakers are

serving three unique purposes: (1) direct management, i.e. avoiding dead air, confirming the message, initiating the turn; (2) indirect management though speaker mannerisms and speech patterns; and (3) making use of the positive features to supplement fillers to "buy thinking time" before a response (Friginal 2008, 2009).

Text Sample 5 illustrates a call-taker's use of the positive features of Dimension 3, while Text Sample 6 shows similar patterns from two speakers in Call Home. *Ok, anyway, let's,* and *well* co-occur with adverbials *actually*, *supposedly*, *exactly*, and *basically* in the call-taker's turns. *Ok* and other discourse particles often start the call-takers' turns, and sometimes are used together to mark the beginning of utterances. In several instances, *ok* is also used to signal transitions or turn endings. Adverbials often belong to different semantic categories with different discourse functions. In this context, stand-alone adverbial *exactly* is used as a direct, confirmatory response, while stance adverbial *actually* implies verification of information (e.g. *"..actually June 30"*; *"I actually checked your.."*).

Text Sample 4. Purchasing transaction (Dim Score = 4. 318)

Caller: [long pause] Uhm one of them was I believe was on I believe was on the 25th of June
Call-taker: **Ok?**
Caller: Two of them was on the 25th and one of them was on the 21st of June
Call-taker: **Ok, let's** just go ahead and check [long pause] **ok** [hold 22 seconds] the other one I believe was on the two you have **actually** won three recruits right?
Caller: Yes
Call-taker: **Ok** you have three recruits so **let** me just check [long pause] **ok** so it is here that since you recruited them just last 25th they supposedly [long pause] **ok let** me just go ahead and check on this, I'll call you back because I **actually** checked your [XX Account] and that coupon is not loaded in your [XX Account] **ok?**
Caller: [unclear] I don't see it there
Call-taker: Yes, yes and uh you know the start is **actually** June 30 **well** but **anyway** you have until the end of this month to redeem this coupon **basically**, so **whatever**, **let** me just go ahead and check why the coupon is not loaded

Text Sample 5. Call home travel schedules (Dim Score = 3.213)

Speaker 1: **Right, right**.
Speaker 2: **Right**.
Speaker 1: But **anyway**, they they live in New York City in Queens and they're **really** nice so if you get stuck like you know I could give you their number or something.
Speaker 2: **uh-huh.. yeah well** no I hopefully it will be okay
Speaker 1: **yeah**.

Speaker 2: um what was I going to say um ah so **anyway** uh I'm **basically** what I'm doing I'm I'll be **just** like only four days in Buffalo.
Speaker 1: **uh-huh.**
Speaker 2: And, of them, in fact, only the Grandma Ruth and Grandma Henning will only be there for three.
Speaker 1: **Really?**
Speaker 2: The last the last day I'll be spending with my friend, Cathy.
Speaker 1: **uh-huh**
Speaker 2: And then I'm going to Johnny and I think we have
Speaker 1: When are you getting to Johnny's?
Speaker 2: I'm flying in on the twenty-eighth.
Speaker 1: **oh**, the I'll **probably just** be leaving Michigan then.
Speaker 2: By ya
Speaker 1: Like to drive to South Carolina.
Speaker 2: **uh-huh**.
Speaker 1: I
Speaker 2: Because he's getting I me he's got off from the twenty-eighth and right now I've got a flight that I'm coming in like ten or eleven in the morning something like that.
Speaker 1: **okay**.

4. Summary and discussion

There are clear differences in the use of co-occurring linguistic features across telephone registers in Dimensions 1 to 3. Call Center interactions are more polite, highly addressee-focused, and elaborated compared to Call Home conversations and Switchboard discussions. Expectedly, there are more features of procedural language in customer service interactions than in the two sub-registers of telephone talk. In addition, there is a consistent, explicit management of information in call centers that speakers in the two comparison corpora do not necessarily observe. The three primary variable accounting for differences in dimension scores across the three telephone registers are (1) imperative *let's/let us*, (2) 2nd person pronouns, and (3) respect markers (*ma'am/sir*).

In general, telephone interactions and face-to-face conversations are statistically different in linguistic and textual composition across the three dimensions. Face-to-face conversations are predominantly involved and simplified (narrative), non-procedural, and spontaneous or unplanned. Turns are not constantly managed or monitored by speakers and topic-shifts are more frequent. The fact that face-to-face conversations in the AmE Conversation corpus also typically involve more than two speakers influences the structure of oral

exchanges and the introduction of new topics. In telephone talk (often between two individuals), turns are more defined and topic shifts are relatively more predictable.

In the spirit of customer service and personalization of support, call-takers in outsourced call centers use politeness markers frequently and try to engage customers by giving sufficient information and explanation and using discourse markers to monitor the flow of talk. Callers' discourses, on the other hand are generally involved, personal, and simplified; non-procedural; and less managed. Most of the turns originate from a first person perspective and are based on the occurrence of a past issue or concern. The various ways speakers discuss topics or events, or in customer service, how call-takers provide information and responses to caller questions are captured by the resulting first two dimensions from Friginal (2008). Specific tasks influence the tone and flow of transactions and different operational processes in customer support are illustrated by the manner in which information is delivered to the callers. Dimension 1, for example, differentiates tasks based on whether information is elaborated or simplified. The extent of explanation given to the callers is demonstrated by the co-occurring features in this dimension. Some customer service tasks may require more elucidation and repeated confirmation of understanding while others rely on direct question-answer sequences. There are also service encounters that regularly include "spiels" reminding callers about products for sale or issues with legal or monetary implications. Whenever additional selling and explanations occur in call centers, features of elaboration in the texts increase.

5. Conclusion

This chapter applied Biber's MDA procedures in describing the linguistic characteristics of three registers of telephone talk relative to face-to-face conversations. Dimension scores from three established dimensions of customer service interactions from Friginal (2008) were used to compare the linguistic preferences of speakers across four spoken registers. MDA revealed several interesting and unique characteristics of call center interactions against other types of telephone talk. The wide-range of information exchanged by call-takers and callers was described by the statistical co-occurrence of different linguistic features. Comparisons across registers in the three extracted dimensions likewise exposed marked attributes distinguishing the important influence of specific tasks in the microscopic structure of telephone talk. Specific foci on the amount of information required to be exchanged, the overall objective of the exchange, and ways of facilitating the exchange over the telephone were interpretable

through the linguistic dimensions used in this study. It would be relevant to apply the same three dimensions to parallel call center corpora and other specialized, task-based telephone interactions. Examining, for example, how the Filipino call-takers compare with Indian or American call center representatives will provide data that could be used for a more focused cross-cultural comparison.

References

Adolphs, Svenja, Brown, Brian, Carter, Ronald, Crawford, Paul & Sahota, Opinder. 2004. Applying corpus linguistics in a health care context. *Journal of Applied Linguistics* 1: 9–28. DOI: 10.1558/japl.1.1.9.55871

Alsurmi, Mansoor. 2012. Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly* 46(3): 472–495. DOI: 10.1002/tesq.38

Barbieri, Frederica. 2008. Patterns of age-based linguistic variation in American English. *Journal of Sociolinguistics* 12(1): 58–88. DOI: 10.1111/j.1467-9841.2008.00353.x

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 2006. *University Language: A Corpus-based Study of Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. DOI: 10.1075/scl.23

Biber, Douglas & Burges, Jena. 2001. Historical shifts in the language of women and men. In *Variation in English: Multi-Dimensional Studies,* Susan Conrad & Douglas Biber (eds), 21–37. London: Longman.

Bolden, Galina. 2006. Little words that matter: Discourse markers "So" and "Oh" and the doing of other-attentiveness in social interaction. *Journal of Communication* 56: 661–688. DOI: 10.1111/j.1460-2466.2006.00314.x

Call Home American English Speech. 2004. University of Pennsylvania, ⟨http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC97S42⟩ (28 March 2013).

Cameron, Deborah. 2008. Talk from the top down. *Language and Communication* 28: 143–155. DOI: 10.1016/j.langcom.2007.09.001

Chafe, Wallace. 1985. Linguistic differences produced by differences between speaking and writing. In *Literature, Language, and Learning: The Nature and Consequences of Reading and Writing,* David Olson, Nancy Torrence & Angela Hildyard (eds), 105–123. Cambridge: CUP.

Cheepen, Christine. 2000. Small talk in service dialogues: The conversational aspects of transactional telephone talk. In *Small Talk: Professional and Commercial Applications*, John Coupland (ed.), 231–249. Harlow: Pearson.

Cheepen, Christine & Monaghan, James. 1990. *Spoken English: A Practical Guide*. London: Pinter.

Connor-Linton, Jeffrey. 1989. Crosstalk: A Multi-Feature Analysis of Soviet-American Spacebridges. Ph.D. dissertation, University of Southern California.

Cowie, Claire. 2007. The accents of outsourcing: The meanings of "neutral" in the Indian call center industry. *World Englishes* 26(3): 316–330. DOI: 10.1111/j.1467-971X.2007.00511.x

D'Ausilio, Rosanne. 1998. *Wake Up Your Call Center: How to be a Better Call Center Agent.* West Lafayette IN: Purdue University Press.

Economidou-Kogetsidis, Maria. 2005. "Yes, tell me please, what time is the midday flight from Athens arriving?": Telephone service encounters and politeness. *Intercultural Pragmatics* 2(3): 253–273. DOI: 10.1515/iprg.2005.2.3.253

Forey, Gail & Lockwood, Jane (eds). 2010. *Globalization, Communication, and the Workplace*. London: Continuum.

Friginal, Eric. 2008. Linguistic variation in the discourse of outsourced call centers. *Discourse Studies* 10(6): 715–736. DOI: 10.1177/1461445608096570

Friginal, Eric. 2009. *The Language of Outsourced Call Centers: A Corpus-Based Study of Cross-Cultural Communication* [Studies in Corpus Linguistics 34]. Amsterdam: John Benjamins. DOI: 10.1075/scl.34

Friginal, Eric. 2010. Call center training in the Philippines. In *Globalization and Communication in the Workplace*, Gail Forey & Jane Lockwood (eds), 190–203. London: Equinox.

Friginal, Eric. 2011. Interactional and cross-cultural features of outsourced call center discourse. *International Journal of Communication* 21(1): 53–76.

Gardner, Rod & Wagner, Johannes (eds). 2004. *Second Language Conversations*. London: Continuum.

Lockwood, Jane. 2012. Developing an English for specific purpose curriculum for Asian call centres: How theory can inform practice. *English for Specific Purposes* 31(2): 14–24. DOI: 10.1016/j.esp.2011.05.002

Orr, Susan. 2003. Hanging on the Telephone: A Sociophonetic Study of Speech in a Glaswegian Call Center. MA dissertation, University of Glasgow.

Poster, Winifred. 2007. Who's on the line? Indian call center agents pose as Americans for U.S.-outsourced firms. *Industrial Relations* 46(2): 271–304.

Precht, Kristen. 2000. Patterns of Stance in English. Ph.D. Dissertation, Northern Arizona University.

Quaglio, Paulo. 2009. *Television Dialogue: The Sitcom Friends vs. Natural Conversation* [Studies in Corpus Linguistics 36]. Amsterdam: John Benjamins. DOI: 10.1075/scl.36

Rey, Jennifer. 2001. Changing gender roles in popular culture: Dialogue in Star Trek episodes from 1966 to 1993. In *Variation in English: Multi-Dimensional Studies,* Susan Conrad & Douglas Biber (eds), 138–156. London: Longman.

Schegloff, Emmanuel. 2001. Accounts of conduct in interaction. Interruption, overlap, and turn-taking. In *Handbook of Sociological Theory,* Jonathan Turner (ed.), 287–321. New York NY: Plenum.

Schiffrin, Deborah. 1987. *Discourse Markers.* Cambridge: CUP. DOI: 10.1017/CBO9780511611841

Schiffrin, Deborah. 1994. *Approaches to Discourse: Language as Social Interaction.* Oxford: Blackwell.

Scott, Suzanne. 1998. Patterns of Language Use in Adult, Face-to-Face Disagreements. Ph.D. dissertation, Northern Arizona University.

Silverster, Joanne & Anderson, Neil. 2003. Technology and discourse: A comparison of face-to-face and telephone employment interviews. *International Journal of Selection and Assessment* 11(2–3): 206–214. DOI: 10.1111/1468-2389.00244

Switchboard: A Users' Manual. 2004. University of Pennsylvania, ⟨http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html⟩ (28 March 2013).

Tabachnick, Barbara & Fidell, Linda. 2001. *Using Multivariate Statistics,* 4th edn. Boston MA: Allyn and Bacon.

Taylor, Phil & Bain, Peter. 2005. 'India calling to the far away towns': The call center labour process and globalization. *Work, Employment, and Society* 19(2): 261–282. DOI: 10.1177/0950017005053170

Tottie, Gunnel. 1991. Conversational style in British and American English: The case of backchannels. In *English Corpus Linguistics,* Karin Aijmer & Bengt Altenberg (eds), 254–271. London: Longman.

White, Margie. 1994. Language in Job Interviews: Differences Relating to Success and Socioeconomic Variables. Ph.D. dissertation, Northern Arizona University.

CHAPTER 3

On the complexity of academic writing

Disciplinary variation and structural complexity

Bethany Gray
Iowa State University

Building upon renewed research on the pervasive phrasal or nominal style of academic writing, I investigate the use of phrasal compression and clausal elaboration structures in research articles across six academic disciplines. Results indicate that all disciplines rely on phrasal complexity features to a much greater extent than clausal features. However, these results also show systematic patterns of variation across disciplines, with hard sciences (physics, biology) exhibiting the densest use of phrasal features, followed by social sciences (applied linguistics, political science), and then humanities disciplines (history, philosophy). Furthermore, the patterns for clausal features displayed the opposite trend: most frequent in humanities and least frequent in hard sciences.

Keywords: Complexity; clausal elaboration; phrasal compression; disciplinary writing; informational discourse; research articles

1. Introduction

Researchers have long been interested in the linguistic variation between spoken and written language, and have often turned to linguistic complexity as a way of characterizing the pervasive differences between the two general registers. Academic writing has garnered particular attention in this regard, as researchers and language teachers attempt to understand how academic writing develops, and how pedagogical materials can support the development of what is perceived as a 'complex' variety of language.

On the Complexity of Discourse Complexity, Biber (1992/2001) describes his study on discourse complexity as a register-based approach that considers a range of spoken and written text types and focuses on structural complexity, in contrast

DOI 10.1075/scl.66.03gra

to social or psycholinguistic perspectives. Building off of earlier register variation studies (e.g. Biber 1988; Halliday 1979; Wells 1960), Biber (1992/2001) uses confirmatory factor analysis to examine how 33 markers of increased and decreased linguistic complexity interact with discourse functions and situational varieties of language (i.e. registers).[1] Biber (2001: 227) concludes that texts are both complex *in different ways* and *to differing extents*. In addition to a fundamental differences between spoken and written language, Biber finds that written registers exhibit a great deal of variation in the extent and types of discourse complexity (see Biber 2001: 235–237 for a summary).

1.1 Discourse complexity in written academic language

Biber's (1992/2001) approach to complexity was informed by general patterns of register variation that documented the differing discourse styles of spoken and written language, with spoken registers relying on verbs and clausal structures, and academic writing relying on a nominal style. In particular, academic writing has been characterized as relying heavily on nouns, nominalizations, and structures added to noun phrases. This nominal style results in sentences with a relatively simple clause structures but long, complex noun phrases (Biber & Gray 2010). Thus, it is quite common to see sentences such as the following in academic writing (head nouns underlined; pre-modifiers in *italics*; post-modifiers in [brackets]:

(1) a. X deserves Y
The *distinctive* effect [of the size [of the *Asian* population]] [on *income* inequality] certainly deserves *further* research. [Political Science]

b. X provide Y
Zones [of *secondary* contact [between closely *related* species]] provide a *rare* opportunity [to examine evidence [of *evolutionary* processes [that reinforce *species* boundaries and/or promote diversification]]]. [Biology]

These patterns have been well-supported (e.g. Banks 2005, 2008; Biber 1988; Biber et al. 1999; Biber & Clark 2002; Fang et al. 2006; Halliday 1989, 2004; Schleppegrell 2001; Wells 1960), particularly with regard to scientific writing. For

1. Studies employing factor analysis to investigate the linguistic characteristics of texts have followed two major approaches. In confirmatory factor analysis, the researcher proposes several models in which linguistic features are grouped together based on theoretical rationales, and then tests the goodness of fit of those models to the text data. This is in contrast to exploratory factor analysis, in which co-occurring sets of linguistic variables are identified inductively based on quantitative co-occurrence patterns (see Biber 2001: 218–221 for further discussion).

example, Vande Kopple (1994) noted that experimental science articles had many very long subjects, and upon analysis demonstrated that these subjects were made up of noun phrases with many pre- and post-modifiers integrated into the phrase structure. Researchers exploring the nominal discourse style of academic writing interpret findings relative to informational density – that these noun phrases allow for a great deal of information to be expressed in compressed phrasal structures (Biber 1988; Vande Kopple 1994). Likewise, Halliday's work on scientific writing has focused on describing 'grammatical metaphor', whereby processes and actions typically expressed with verbs are nominalized (see Halliday 2004 for a collection of key works on nominalization and grammatical metaphor in science writing).

More recently, Biber and colleagues have built upon on this earlier research and the major patterns of register variation documented in the *Longman Grammar of Spoken and Written English* (Biber et al. 1999). They investigate the use of a range of clausal and phrasal features across registers both synchronically and diachonically (Biber & Gray 2010; Biber et al. 2011a; Biber et al. 2011b; Biber & Gray 2013a). Biber and colleagues relate clausal features to structural elaboration, and phrasal features to structural compression respectively. Based on the results of large corpus analyses, they argue research should account for phrasal complexity in addition to the more traditional notion of clausal complexity.

Biber and Gray (2013a) have shown that the dense nominal style of academic writing has developed relatively rapidly over the past century, and that science writing has adopted the nominal style to a much greater extent than non-science writing. Most previous research on nominal style has focused on science writing; however, this initial cross-disciplinary finding suggests the likelihood of disciplinary variation in the use of the syntactic features of phrasal and clausal complexity. In addition, it reflects Biber's (1992/2001) finding that texts are complex to variable extents. This possibility is taken up in the present chapter, as I investigate the degree to which writing across disciplines relies on phrasal and clausal complexity features.

1.2 Purpose and overview of the current study

Thus, the purpose of the present study is to investigate the extent to which research articles from a cross-section of academic disciplines (humanities, social sciences, and natural sciences) rely on the phrasal and clausal discourse styles that have been identified in register studies involving academic language. More specifically, the analysis utilizes the framework developed in Biber and Gray (2010) to examine two major types of structural complexity: phrasal features that function to compress discourse, and clausal features that function to elaborate discourse. In the

next section, I turn to an explanation of that framework. Section 3 briefly situates the present study within the body of research on cross-disciplinary variation. In Section 4, I describe the corpus and methods employed in the study. Section 5 presents the distributions of the features across disciplines. Section 6 concludes with an overview of the findings, along with a case study in which the distributions of the phrasal and clausal complexity feature are compared across different rhetorical sections of research articles in three of the disciplines.

2. Notions of complexity: A framework of clausal elaboration versus phrasal compression

The framework for investigating structural complexity in research articles used in the present study is adopted from Biber and Gray (2010) and Biber et al. (2011a). Building upon the substantial research which has established the distinctive structural characteristics of spoken and written English, Biber and Gray (2010) set out to document the differing nature of structural complexity in spoken versus written language. While many paradigms of grammatical complexity define complexity based on the use of embedded clauses, Biber and Gray (2010) and Biber et al. (2011a) 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 illustrated in the introduction above.

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), exemplified in Table 1 below.

Biber and Gray (2010) link these clausal features to structural elaboration; that is, the embedded clauses incorporate additional information into the main clause. Finite relative clauses are an example of such elaboration, where the relative clause offers additional information to either describe or specify the referent of the head noun (Excerpts 2a, b).

(2) a. On the other hand, it may have been the case that this freedom was more threatening to the high *achievers*, *who* were used to succeeding within known and comfortable boundaries and perhaps felt they had more to lose. [Applied Linguistics]

b. These small zooplanktivores are in turn likely consumed by the smallest piscivore *species* *which* may in turn be prey of the apex predator. [Biology]

Table 1. Features associated with structural elaboration and compression (see Biber & Gray 2010; Biber et al. 2011a)

'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 religion.*
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*

Likewise, adverbial clauses 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), as in (3):

(3) *Because* retailers such as Wal-Mart depend heavily on cheap Chinese imports, this measure aims to capture constituency support for MFN. [Political Science]

Complement clauses, on the other hand, are not optional elements as they typically fill the slot of a required clause element; yet Biber and Gray point out that they are elaborating because the information from an entire clause occurs in a syntactic slot often filled by a noun phrase. This embedded clause results in a greater amount of information being included into the main clause (Excerpt 4):

(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. [Physics]

In contrast, Biber and Gray (2010) link embedded phrasal features to structural compression, in which information is added to noun phrases in optional phrases that can be considered more condensed alternatives to fuller clausal structures. Features like prepositional phrases and nouns as nominal pre-modifiers convey meanings that could be more explicitly stated through elaborating clausal structures. For example, the noun phrase "a further rationale *for pension privatization*" could be paraphrased as "a rationale *that supports pension privatization*," and the noun phrase "*recovery* time" can be paraphrased as "the time *that it takes for something to recover*." Thus, the phrasal features that can be embedded in noun phrases function to express exlaborate and varied meanings in a highly compressed manner.

Biber and Gray (2010, 2013a) have shown that the use of these phrasal modifiers has increased over the past 100 years in written registers. Published academic research articles, with their informational purpose and highly specialized audience, have shown particularly dramatic increases. Furthermore, they find that writing in the natural sciences has exhibited an increase in the use of these features that is markedly higher than in non-science writing. While Biber and Gray distinguish only between science and non-science disciplines, the clear difference between the two types of disciplines is indicative of potential variation between individual disciplines.

3. Disciplinary variation in academic writing

In addition to large-scale studies that have sought to describe the linguistic characteristics of academic writing more generally, there has also been increased interest to the way that academic writing exhibits variation across disciplines. Silver (2006) attributes this increased attention to factors such as the development of English for Specific Purposes (ESP), a refocusing on 'communication-based' models of language use, and the growing importance being placed on studies of language use within naturally-occurring contexts. In addition, the rise of movements such as Writing across the Curriculum (WAC) has led to increased awareness that

language varies in systematic and meaningful ways across academic disciplines (e.g. see Russell 2002 for a review of the development of WAC).

Research on disciplinary variation has considered a range of registers written and read by students and professionals in academic settings. The development of the British Academic Written English Corpus (BAWE) (Nesi & Gardner 2012; Gardner & Nesi 2012) and the Michigan Corpus of Upper-Level Student Papers (MICUSP) (Römer & Swales 2010) has enabled investigation of undergraduate and graduate student writing across 30 and 16 different disciplines respectively. These large-scale corpora representing many disciplines have complemented research that has provided detailed investigations focused on a smaller number of disciplinary comparisons for L1 and L2 writers (e.g. Aktas & Cortes 2008; Harwood 2005a; Hewings & Hewings 2002; Hyland 2002, 2008; Lee & Chen 2009).

Much research, however, has focused on the published research article as the primary mode through which disciplinary knowledge is created and transmitted. Investigations of the linguistic characteristics of research articles has taken a range of approaches, varying along the parameters of the type of linguistic features investigated, the registers focused on, and the number of disciplines considered.

For example, one approach has been to focus on detailed analyses of particular lexical, grammatical, or rhetorical devices in a single register and in a single discipline (e.g. Afros & Schryer 2009; Hemais 2001; Hyland 1996; Warchal 2010). A second approach has been to compare the use of linguistic features across multiple academic registers within a single or small number of disciplines. These have included comparisons to student/learner writing (e.g. Harwood 2005a; Hewings & Hewings 2002; Koutsantoni 2006), textbooks (e.g. Biber et al. 2002; Conrad 1996; Hyland 1999), book reviews (e.g. Diani 2008; Groom 2005), editorials (e.g. Webber 1994); and popular science texts (e.g. Hyland 2010).

A third approach has provided comparisons of research articles across two or more disciplines. While comparing a small number of disciplines is common (e.g. Harwood 2005b; Peacock 2006), some studies investigate a broader range of disciplines (e.g. Hyland's work on stance and engagement in eight disciplines).

Across these studies, research articles have been described according to their use of specific lexical items, grammatical features, rhetorical structures, realizations of discourse functions, and phraseological patterns. This body of research has allowed us to gain an in-depth understanding of particular features of writing within the disciplines; however our knowledge of disciplinary variation has been constrained in two ways. First, most cross-disciplinary comparisons of academic research articles have provided focused, comprehensive analyses of the distributions and discourse functions of individual (or a small number of) specialized linguistic features (e.g. explanations of how noun + *that*-complement clauses are used to construct evaluation in research article abstracts in Hyland & Tse 2005).

In contrast, we do not have good understanding of how the use of more pervasive register features (see Biber & Conrad 2009: Chapter 3) might vary across academic disciplines – such as the extent to which the discourse is verbal or nominal in style. As discussed above, Biber and Gray (2013a) uncover the potential for such variation. But in fact, research on nominal style has typically focused on science writing (e.g. Banks 2008; Vande Kopple 1994) or has considered a more general construct of academic writing that represents a range of disciplines but does not distinguish between disciplines (e.g. Biber 1988; Biber et al. 1999; Biber & Gray 2010). Thus, the purpose of the present study is to investigate one set of such 'register features': markers of clausal and phrasal complexity.

The second constraint is that relatively few corpus builders have systematically considered variation within disciplines into the design of corpora of research articles. Inquiries that do consider intra-discipline variation typically focus on differences across major register categories (e.g. textbooks or book reviews versus research articles) or between student/learner writing and published research articles. In contrast, research articles are defined broadly with little consideration to differences in *types* of research articles. A few studies acknowledge likely linguistic differences in types of articles (e.g. Williams 1996 on clinical vs. experimental research articles; Vande Kopple 1994 on experimental vs. theoretical science articles) and incorporate considerations of article types in their corpus design (either through deliberate inclusion or exclusion of article types).

Gray (2011, 2013) is a one of the first more recent studies to design a balanced corpus of research articles that represents multiple sub-registers corresponding to research paradigms. The corpus includes nine sub-corpora encompassing six disciplines and three types of research articles (quantitative, qualitative, and theoretical research; see Section 3 below). On the one hand, this corpus design allows for comparison within and across disciplines, as in Gray (2011, 2013). On the other hand, this same corpus also serves as a strong foundation for studies focused primarily on differences across disciplines (the focus of the present study); variation within disciplines is accounted for in its design, thus increasing the external representativeness of the corpus.

4. Methods

4.1 The corpus

Table 2 describes the corpus used in the present study, which includes research articles from six disciplines from the humanities, social sciences, and natural sciences: philosophy, history, political science, applied linguistics, biology, and physics. As

mentioned above, the corpus was designed to enable within and cross-disciplinary comparisons, and distinguishes between major types of research published within the six disciplines represented: theoretical, qualitative, and quantitative research. For the purposes of this study, results are presented for each discipline without distinguishing between article types.[2] A full description of the corpus, including how journals and articles were selected for inclusion, can be found in Gray (2011: Chapter 5 and Gray 2013).

As Table 2 shows, each discipline is represented by the types of articles typically published in that field (based on a survey of journals in each field and in consultation with disciplinary experts). The full corpus comprises 270 research articles (30 articles per register-discipline combination), and about 2 million words.

Table 2. Corpus description in number of words (30 texts per discipline/category)

	Types of research represented	Number of texts	Number of words
Philosophy	Theoretical	30	280,826
History	Qualitative	30	282,898
Political Science	Qualitative & Quantitative	60	422,177
Applied Linguistics	Qualitative & Quantitative	60	439,960
Biology	Quantitative	30	154,824
Physics	Theoretical & Quantitative	60	377,308
Total		*270*	*1,957,993*

4.2 Analytical tools and procedures

The corpus was annotated for parts of speech and syntactic information using the Biber tagger (see Biber 1988; Biber et al. 1999). The accuracy of the tags was investigated in a subset of fifteen text samples, which were first hand-coded for tagging

2. In a multi-dimensional analysis of the corpus, Gray (2013) finds the linguistic characteristics of research articles vary along multiple parameters that encompass both disciplinary differences as well as differences that can be attributed to types of research articles regardless of discipline. These findings support a corpus design that accounts for multiple types of research reports in a representative corpus of disciplinary writing, a task which has been largely neglected in the current body of research. While findings support the need for such balanced corpora, results from Gray (2011: Chapter 7), show that the variation within disciplines is comparatively small for these particular complexity features. Thus, the findings have been reported by discipline in the present chapter. Readers are directed to Gray (2011) for results broken down by research type.

errors. A computer program was then created to automatically re-tag systematic errors. Each sample was then coded a second time to calculate rates for precision and recall (a more detailed description of the tag-checking process and the reliability rates can be found in Gray 2011: Chapter 5 and Appendix B).

A second specialized computer program was used to analyze the full corpus and identify the 'elaborating' and 'compression' features summarized in Table 1 above. This program 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, 2013a, 2013b; Biber et al. 2011a). Most features (attributive adjectives, nouns as nominal pre-modifiers, adverbial clauses, relative clauses, noun + of prepositional phrases) could be identified based on the grammatical tags assigned to each word in the corpus. Complement clauses were identified through a combination of grammatical tags and lexical information, which allowed for a more reliable identification of the features of interest. That is, *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).

Several features that have been investigated within this same complexity framework (i.e. appositive noun phrases, prepositional phrases as post-nominal modifiers) represent functional relationships between the head noun and the post-modifying structure that cannot reliably be identified automatically. Thus, appositive noun phrases are not included in the present study, and noun + of-phrases, which can be identified automatically, are used to represent prepositional phrases as post-nominal modifiers.

In the final part of this study, the possibility that these elaboration and compression features are used to differing extents in different rhetorical sections of research articles is explored. To enable this comparison, articles were split into the following sections: abstract, introduction (including literature reviews), methods, results, discussion, and conclusion – what is commonly referred to as Introduction-Method-Results-Discussion (IMRD) structure.[3] However, as the analysis in Gray (2011: Chapter 6) showed, this structure is by no means universal to all research articles, and is particularly not applicable to theoretical articles, or to articles in

3. IMRD refers to a frequently-used scheme to characterize the major organizational pattern of research articles, in which each section represented by the acronym (Introduction, Method, Results, Discussion) carries out particular discourse functions, such as introducing and motivating a topic, describing methods of analysis, presenting results, and providing a discussion of what those results mean.

history and qualitative political science. The internal structure of these non-IMRD studies are being explored in an on-going study; for the purposes of the present study, four groups of texts that typically used the IMRD structure were selected for comparison: quantitative political science, quantitative applied linguistics, quantitative biology, and quantitative physics. Using internal headings within the articles to indicate the content of major sections, each text in these sub-corpora was coded for rhetorical section. [4] Then, the 'complexity' program was used to obtain counts for the complexity features within each section.

Normalized (to 1,000 words) rates of occurrence for each feature were calculated for each text in the corpus (and for each section of each text), and the mean rate of occurrence was calculated for each discipline and register combination in the corpus.

5. Elaboration and compression across disciplines

In this section, I describe the use of elaboration and compression features across the six disciplines represented by the corpus. First, I consider the use of embedded clauses that serve to elaborate discourse, followed by a consideration of phrasal modifiers to nouns that function to compress information into dense, information-laden phrases. I then turn to clausal noun post-modifiers, arguing that these structures exhibit characteristics of both clausal elaboration (they are clausal in structure), but also phrasal compression (they add information within noun phrases).

5.1 Clausal elaboration

Figure 1 shows the frequency of use of finite and non-finite complement clauses, along with finite and non-finite adverbials. Two trends are apparent from Figure 1. First, and perhaps most noticeably, is that there is a general pattern of greater use of these elaborating features in humanities disciplines (i.e. the 'soft' disciplines

4. Not all articles in the sub-corpora, however, followed this organization. For the purposes of the case study, only those texts which included the sections were used to calculate the rates of occurrence for the sections. For example, all 30 quantitative applied linguistics texts included an abstract, introduction, and methods section. Twenty-four of these contained separate results sections, 21 contained separate discussion sections, and 6 contained combined results/discussion sections. Thus, the figures displaying these results below contain separate columns for 'results' and 'combined results and discussion'. The biology texts rarely contained conclusion sections, and thus no results are presented for biology conclusions.

philosophy and history), much less use of these clausal features in the natural sciences (i.e. the 'hard' disciplines biology and physics), and the social sciences (political science and applied linguistics) fall in between. This trend is particularly observable for finite and non-finite complement clauses, and for finite adverbial clauses to a somewhat lesser extent.

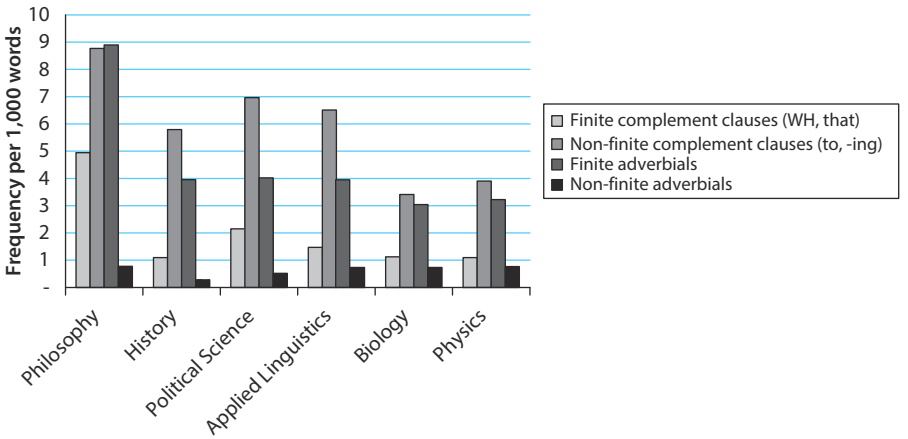

Figure 1. Complement clauses and adverbials (finite and non-finite) across disciplines

Apart from this general trend, however, it is interesting to note that the relative distributions of the four clausal elaboration features are generally parallel across disciplines. In all disciplines, non-finite complement clauses are the most frequent (except for theoretical philosophy, in which finite adverbials are equally common), followed by finite complement clauses and then finite adverbials. Non-finite adverbials are relatively rare in all disciplines.

Figure 2 displays the frequencies for non-finite complement clauses, the most frequent elaborating feature, broken down by controlling word (ing-clauses and to-clauses are combined). Non-finite verb complement clauses are the most frequent type utilized in the humanities and social sciences, but are comparatively rare in the hard sciences. In fact, non-finite complement clauses controlled by adjectives are more frequent than other types of clauses in the hard sciences, and are the only structures which are generally consistent across all disciplines and registers.

The higher rate of occurrence of verb-controlled clauses in the humanities and social sciences likely correspond to the overall higher prevalence of verbs in these disciplines. However, it appears that the more frequent use of nouns overall in biology and physics (see Gray 2011: Chapter 6) does not correspond to a higher reliance on noun-controlled complement clauses in the hard sciences (in fact, non-finite noun

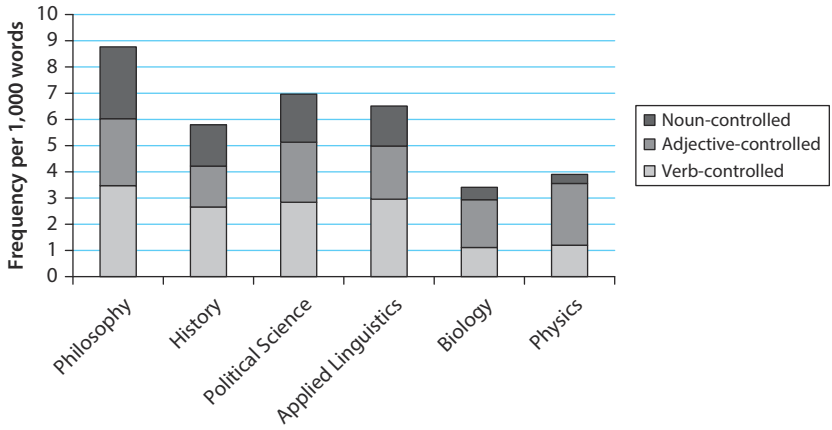

Figure 2. Types of non-finite complement clauses across disciplines

complement clauses are more than twice as frequent in all other disciplines). This lower reliance on noun-controlled clauses in biology and physics is likely related in part to the nature of nouns that are common in these disciplines. That is, nouns that control non-finite complement clauses are often cognition nouns (Examples 5 a–b), process nouns (Examples 6 a–b), and other abstract nouns (7 a–b) – all of which are more common in philosophy, history, political science, and applied linguistics, as illustrated below. Biology and physics, in contrast, rely to a greater extent on concrete, technical, and quantity nouns, which do not often take complementation (see Gray 2011: Figures 7.2 and 7.3).

(5) a. The ability *to isolate important causal forces* is important and experiments offer the opportunity. [Political Science]

b. Several actively disliked the thought *of learning more* but expressed the knowledge of benefits derived from acquiring a certain level of linguistic ability. [Applied Linguistics]

(6) a. In Apr. 1503, Fabyan was ordered by the court of aldermen to fulfill his agreement *to be alderman* 'upon payne of enprisonemet'. [History]

b. 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* [Philosophy]

(7) a. the Islamic Republic provides a significant opportunity to Moscow *to expand its influence and interest in both regions.* [Political Science]

b. 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'. [Philosophy]

Adjective-controlled non-finite complement clauses, however, are used fairly consistently across the six disciplines to convey personal stance, or evaluations and attitudes towards propositions, as in Examples (8a–c):

(8) a. This was unexpected, as it is difficult *to detect intracellular ZO-1 pools immunohistochemically.* [Biology]

b. 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.* [Physics]

c. This did not end factional strife in the branch, but it was impossible *to distinguish political from personal motives.* [History]

Finite adverbial clauses also exhibit the general trend of being more frequent in softer disciplines, although the pattern is much less dramatic (see Figure 1). Figure 3 shows that overall, finite adverbials are about twice as common in theoretical philosophy articles than in any other discipline; the differences between the hard sciences and history, political science, and applied linguistics is much smaller. Figure 3 also displays the extent to which specific adverbial subordinators are used. Adverbial clauses beginning with *if* are most common in all disciplines

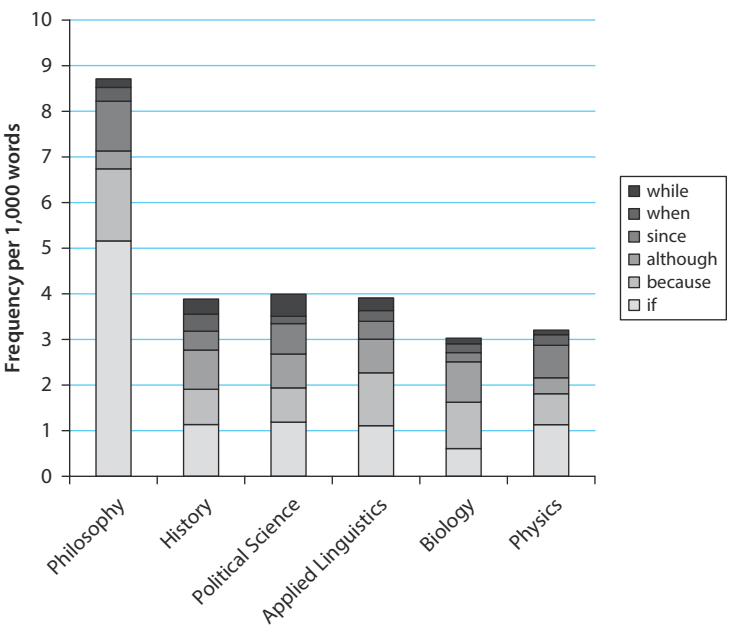

Figure 3. Common adverbial subordinators across disciplines

and registers, but *if* is extremely common in theoretical philosophy, accounting for a large portion of the difference in adverbial subordination between philosophy and the other disciplines.

The following excerpt from theoretical philosophy illustrates the dense use of adverbial subordination, showing that adverbial subordination in philosophy is used to explore possibilities and logical relationships; adverbial subordinators explicitly mark relationships between propositions:

(9) 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. [Philosophy]

The four features described in this section all elaborate at the clausal level. The patterns of use show that these elaboration features occur much more extensively in philosophy than in any other discipline, and that the hard sciences of biology and physics are marked in the relatively low reliance on clausal elaboration. However, it is also useful to attend to the scale in which these results are reported. On one hand, all disciplines rely on these elaboration features to a lesser extent than spoken language does. For example, even in philosophy, which exhibited the most reliance on finite complement clauses of all the disciplines, finite complement clauses occur at a rate of about 5 times per 1,000 words. As a frame of reference, Biber and Gray (2010: Figure 1) report a rate of occurrence of almost 14 times per 1,000 words for conversation – nearly a three-fold difference. Non-finite complement clauses and finite adverbial clauses are likewise more frequent in spoken language.

On the other hand, it's also important to keep the scales of Figures 1, 2 and 3 in mind as we turn to phrasal complexity features, where we find that these phrasal features are much more common than clausal features in all disciplines.

5.2 Phrasal compression: Complex noun phrases

The rates of occurrence for three types of phrasal structures that can be embedded within noun phrases are displayed in Figure 4. The most frequent phrasal modifiers in all disciplines and registers are adjectives as noun pre-modifiers, ranging from between 60 to 75 times per 1,000 words but showing no systematic pattern across the range of disciplines (and note that the use within disciplines is quite consistent). Prepositional phrases are the next most frequent, occurring

between 30 and 40 times per 1,000 words.[5] Both features are relatively frequent in all disciplines, particularly when compared with the rates of occurrence for clausal structures reported in Section 4.1. These phrasal features also exhibit less variation across the disciplines than the clausal features; nor is there a systematic trend of increase or decrease in frequency from the humanities to the social sciences, and from the social sciences to the hard sciences.

The third feature explored in this section, however, does exhibit a systematic trend in this regard. As shown in Figure 4, nouns as nominal pre-modifiers exhibit the opposite trend that we saw with the clausal features in Section 4.1: they increase systematically along the soft-to-hard dimension. While noun + noun sequences are relatively rare in philosophy (occurring about 15 times per 1,000 words), they are more than three times as common in physics (occurring more than 50 times per 1,000 words). And there is a clear and gradual increase as we move from the humanities (philosophy and history) to the social sciences (political science and applied linguistics), and then again from the social sciences to the hard sciences.

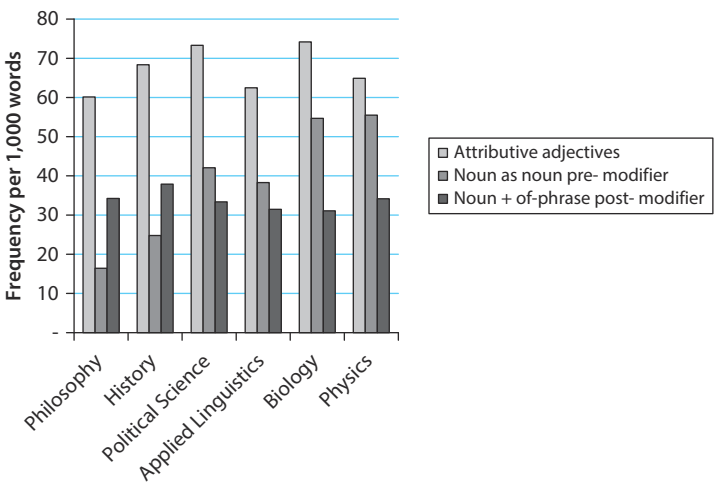

Figure 4. Phrasal modifiers across disciplines

In summary, noun + noun sequences are quite characteristic of writing in the hard sciences, and to a lesser degree in the social sciences. As the least common

5. The total number of prepositional phrases as noun post-modifiers is expected to be higher than this, due to the operationalization here as of-phrases as post-nominal modifiers. For example, counting prepositional phrases as noun modifiers beginning with *in, on, with, for,* and *of,* Biber and Gray (2010: Figure 2) report a frequency of just over 50 per 1,000 words.

type of phrasal modifier in philosophy and history, noun + noun sequences are less characteristic of the humanities disciplines. The differing densities of these nouns as noun modifiers are exemplified in the following text excerpts from physics, applied linguistics, and history, where nouns as noun modifiers are *italicized* (head nouns are underlined).

(10) The *cloud size distribution*, representing the fraction of total clouds within a finite *size range*, varies with *pixel resolution*. *Cloud size* is represented as *the area-equivalent diameter* of a cloud. Figure 4 shows the *size distribution* of clouds calculated using the 15 m *resolution data*, for a variety of *domain sizes*. Reduction of the domain of the observed *cloud field* may result in partitioning of a single cloud into several smaller clouds, if that particular cloud crosses the *subdomain boundaries*. [Physics]

(11) Reading also entails the use of linguistic knowledge. One type of competence contributing to *text processing*, comprehension, and *vocabulary acquisition* is *vocabulary knowledge* associated with the texts, hereafter referred to as *passage sight vocabulary*. [Applied Linguistics]

(12) Phoenix experienced the same kind of spatial and demographic growth after the war. Between 1940 and 1960 the *city's population* increased from approximately 65,000 to 439,000 while the municipality expanded from 9.6 to 187 square miles. The arrival of Motorola in 1949 and other electronic firms, as well as *military bases* and a booming *tourist industry* propelled the expansion. [History]

This section has demonstrated the reliance on highly compressed noun phrases containing phrasal modifiers in academic writing generally, and the hard sciences particularly. This is in contrast to the trends observed in Section 5.1, where clausal elaboration features were most common in philosophy and least common in the biology and physics. Two final features which are inherently connected to both clausal elaboration and phrasal complexity are considered in Section 5.3.

5.3 Clausal modifiers within the noun phrase

As discussed above in Section 2, relative clauses provide structural elaboration in the sense that they add additional, often optional information to noun phrases. However, relative clauses can be said to be 'intermediate' complexity features because they are embedded at the phrasal level, creating complex noun phrases. In this section, I propose that finite relative clauses can be considered as 'elaborating' features, while non-finite relative clauses can be considered as contributing to syntactic compression on theoretical grounds. The quantitative trends observed here add further empirical support that such an interpretation may be warranted.

Finite relatives contain full clauses with subjects, verb phrases marked for modality, tense, and aspect, and there is typically a clear relationship between the head noun and the relative clause. For example, in (13a–b), the *italicized* relative clauses add substantial information to noun phrases, containing detailed information like tense and modality, and clear subjects (in these examples, the head noun fills the subject gap in the relative clauses). Thus, the relative clause contributes fully-specific information about the head noun.

(13) a. Even those *who do not, in general, regard the claims of the worse off as having any special weight* may worry about unfairness in the case of particular competitions. [Philosophy]

b. The fourth question dealt with factors *that should be considered in designing the syllabus for nursing and midwifery students.* [Applied Linguistics]

Figure 5 displays the distribution of finite and non-finite relative clauses, and shows that like many of the clausal elaboration structures explored in Section 4.1, relative clauses are most common in the humanities (philosophy and history), slightly less common in the social sciences of political science and applied linguistics, and least common in the hard sciences (biology and physics). Unlike finite adverbials, however, the trend is a more gradual one.

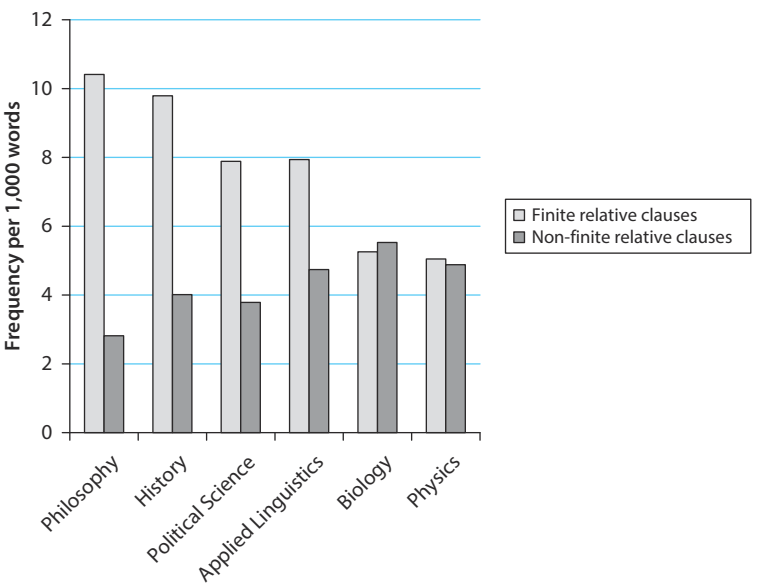

Figure 5. Finite and non-finite relative clauses across disciplines

Non-finite relative clauses are also clausal in nature, but the relationship between the clause and the head noun is less explicitly stated, as there is no overt

subject (or overt marker, such as a relative pronoun, that indicates that the head noun is the subject). In addition, the non-finite verb is not marked for elaborating information such as verb tense, modality, or aspect. For Example, (14a–b) illustrate the compression of information that results from the use of non-finite relative clauses. Both examples contain object gaps, the subjects are additionally omitted in non-finite relatives, and there is no marking for tense, modality, or aspect. Thus, information about the subject and the verb are all absent from the clauses, leading to less elaborating information and less explicit statements of these aspects of the discourse. The lack of explicitly stated information is further demonstrated by comparing the non-finite relative to an equivalent finite relative clause, which more explicitly marks relationships:

(14) a. Salmonella enterica, the cause of food poisoning and typhoid fever, has evolved sophisticated mechanisms *to manipulate host cell functions.* [Biology]
compare: *mechanisms that manipulate/manipulated host cell functions*
compare: *mechanisms that salmonella can use to manipulate functions*

b. The t-tests indicated significantly shorter reading times for formulaic sequences *used idiomatically* [Applied Linguistics]
compare: *sequences that participants* (*had*) *used idiomatically*

Like the patterns seen for finite relative clauses, the distribution of use of non-finite relatives also supports an interpretation in which non-finite relatives are considered 'compression' features. Going back to Figure 5, we see that non-finite relative clauses also reflect earlier patterns, this time mimicking the trend seen with noun + noun sequences: non-finite relatives are most common in biology, physics, and applied linguistics, slightly less frequent in political science and history, and least common in philosophy. That is, the use of non-finite relative clauses generally *increases* in frequency as we move from soft to hard disciplines.

It is thus supported on both theoretical and empirical grounds to group finite relative clauses with other elaborating clausal structures, and non-finite relative clauses with other phrasal compression features, as I do in the next section.

6. Summary: Overall patterns of elaboration and compression

The analyses presented in Sections 5.1–5.3 have shown that even within the construct of academic writing, structures which result in elaborated and compressed discourse styles vary across disciplines. Figure 6 summarizes the specific trends, showing nearly opposite patterns of variation between features of elaboration and compression. Specifically, features of structural elaboration are generally more common in the humanities disciplines represented here (particularly philosophy)

than in the social sciences, and even less common in the hard sciences. In contrast, features of structural compression in which information is packed into noun phrases are most frequent in hard disciplines. Figure 6 also demonstrates, however, that despite the variation, all disciplines maintain the nominal style of academic writing, relying on phrasal features of compression to much greater extents than clausal embedding.

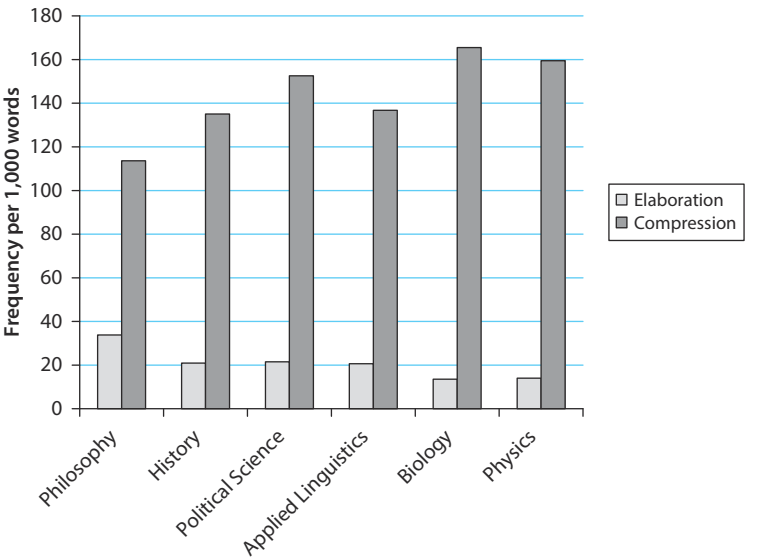

Figure 6. Summary of the use of elaboration (including finite relative clauses) and compression (including non-finite relative clauses) features

The following excerpts exemplify these trends. Excerpts 15 and 16 illustrate the differing degrees to which compression features are used through examples from biology and history. In these excerpts, head nouns of complex noun phrases are **bolded** (for purposes of illustration, nouns that head phrases with no pre- or post-modification are not marked), nominal pre-modifiers are *italicized*, and nominal post-modifiers are underlined (except for finite relative clauses, an elaborating feature). For reference, the main verbs in the passage are presented in SMALL CAPS, in both main clauses and embedded clauses.

Excerpt (15) illustrates an extremely dense use of noun modifiers in quantitative biology, and further illustrates one head noun being modified by multiple pre- and post-modifiers (e.g. *species sorting* **hypothesis,** *spatial* ***structure*** *in metacommunities*). There are relatively few main verbs and no finite relative clauses, but there are two non-finite relative clauses (**degrees**…resulting from and **points** located in).

(15) *Biology:*

Spatial **structure** in metacommunities and their **relationships** to *environmental* **gradients** HAVE BEEN LINKED to *opposing* **theories** of *community* **assembly**. 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**. Here we STUDY the *relative* **influence** of *environmental* **gradients** and *spatial* **structure** in *bird* **assemblages** of the *Chilean temperate* **forest**. We CARRIED OUT *bird* and *vegetation* **surveys** in *South American temperate* **forests** at 147 **points** located in nine *different protected* **areas** in *central* **Chile**

In contrast, excerpt (16) from a qualitative history article exhibits fewer compression features relative to the density seen in (15), and an increased density of elaboration features. There are more main clause verbs, in addition to several finite relative clauses (those whom he followed, those who succeeded him).

(16) *History:*

As a **result** of the *staggering* **amounts** of *potential* **income** involved, the *contemporary* **view** of Henry as a *greedy* **monarch** largely AROSE from his **pursuit** of these bonds. However, it MIGHT BE ARGUED that, given the needs endemic to the **job** of kingship, he was nothing more than perspicacious. Historically, governments, whether English or otherwise, HAVE BEEN WELL AWARE that *effective* **rule** REQUIRED *positive cash* **flow**. Henry VII WAS no different from those whom he FOLLOWED and those who SUCCEEDED him, except that he BECAME **one** of the few *solvent English* **kings** since 1066.

Clausal elaboration features can be compared in the same manner. Excerpts (17) and (18) illustrate the differing degrees to which elaboration features are used in philosophy and quantitative physics. In these excerpts, 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. Every sentence in Excerpt (17) contains an embedded clausal structure, including 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.

(17) *Philosophy:*

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 (18), from a quantitative physics article, contains only two embedded elaboration clauses (although there are several non-finite relative clauses: mixtures excited by electric discharges, state being populated, processes… correlated to).

(18) *Physics:*

> 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 5d–6p and 6p–6s 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 Kr–Xe 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.

These extended text excerpts illustrate the overall patterns of variation: disciplines utilize the nominal compression features (typically associated with academic writing) to different extents, and this variation largely complements the trends for clausal elaboration features. However, a further area of interest is how the use of these features might vary *within* articles. A great deal of research, mostly focused on functional moves, has described the differing characteristics of specific sections in research articles. This research has documented the differing functions of these sections within articles, and it is possible that the communicative purposes of those sections can also be related to the use of compression and elaboration features. (For example, Biber and Finegan (2001) used multi-dimensional analysis to demonstrate variation within medical research articles – although that variation was small relative to the full range of variation exhibited across registers).

To begin to investigate this possibility, a case study of a sub-corpus of quantitative research articles in four disciplines is reported here. Figure 7 displays the rates of occurrence for elaboration features, while Figure 8 displays patterns for compression features across sections of articles in quantitative political science, applied linguistics, biology and physics.

While these figures, on the one hand, reinforce the general patterns observed in earlier sections, it also becomes apparent that a fair amount of variation occurs within research articles as well (but as Biber & Finegan 2001 found, this variation is smaller than the overall patterns of variation). Figure 7 shows that regardless of the overall prevalence of elaboration features, methods sections use the fewest elaboration features within a discipline. And this relative lack of elaborating features is characteristic of methods sections across all four disciplines. In contrast, discussion and conclusion sections rely more heavily on elaboration features as authors comment on, interpret, and discuss the implications of research findings.

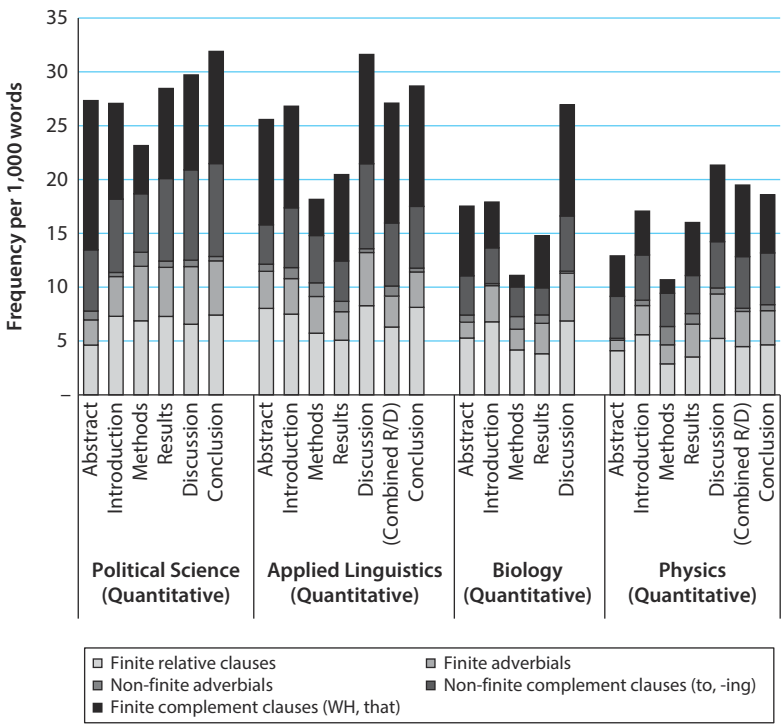

Figure 7. Distribution of elaboration features across sections within research articles in quantitative political science, applied linguistics, biology, and physics

Figure 8 also illustrates systematic patterns across disciplines, largely mirroring the patterns observed in Figure 7. That is, sections which relied heavily upon elaborating features do not rely as heavily on compression features, and vice versa. For example, in all four disciplines, abstracts exhibit the densest use of phrasal compression features. Abstracts generally have very constrained word limits, and

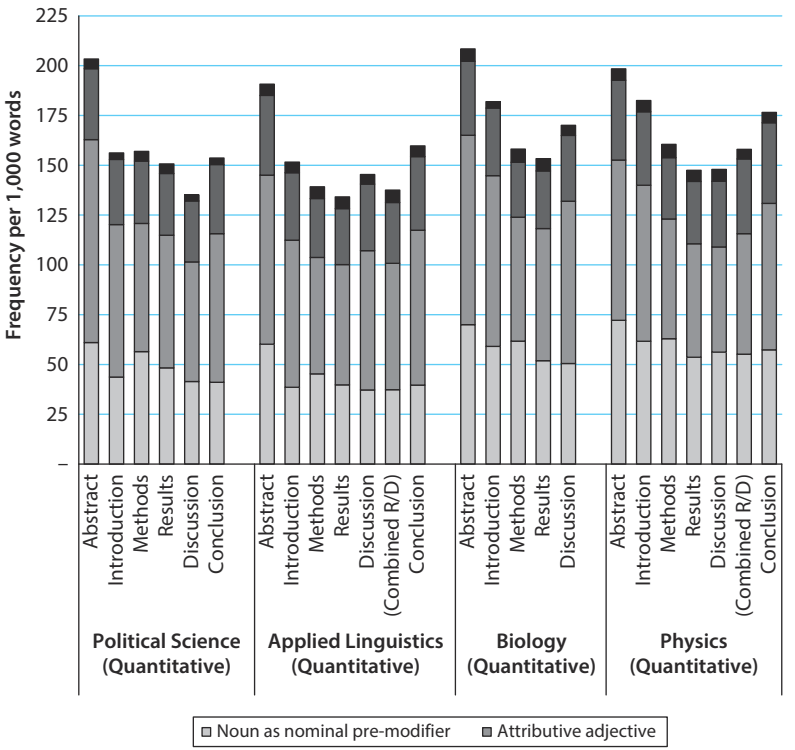

Figure 8. Distribution of compression features across sections within research articles in quantitative political science, applied linguistics, biology, and physics

authors are tasked with presenting a lot of relevant information in a short amount of space. Thus, compression features are utilized to produce highly informational, less elaborated abstracts.

While Figures 7 and 8 provide a brief overview of the quantitative patterns for this brief case study, the next step is to analyze each section functionally to identify the discourse functions that link these rhetorical sections with elaboration versus compression features.

7. Conclusion and future directions

This analysis has documented variation across disciplines in the extent to which texts rely on two different types of structural complexity features: clausal elaboration features and phrasal compression features. However, it's also important to keep the scope of these differences in mind. That is, the variation seen across

different academic disciplines is smaller than the variation seen across spoken and written registers more generally. All academic writing maintains its reliance on phrasal structures when compared to spoken language; but writing within the disciplines is also variable.

Several of the complexity features exhibited clear and systematic variation, increasing or decreasing along the parameter of hard-to-soft disciplines. This cline of variation illustrates what Biber (1992/2001) found in terms of texts differing not categorically in terms of their use of particular features, but rather in the relative extent to which those features are used. This finding is also theoretically important as it represents a systematic pattern of variation for *register features*. Biber and Conrad (2009: 53) define register features as "words or grammatical characteristics that are (1) pervasive – distributed throughout a text from the register, and (2) frequent – occurring more commonly in the target register than in most comparison registers". The consistency with which all texts in the corpus relied upon the phrasal complexity features provides further evidence of the validity of these features as being characteristic of academic writing regardless of discipline. Based on this and the previous research (cited in Sections 1 and 2 above), the explicit teaching of phrasal compression features is well-motivated in English for Academic Purposes classrooms, particularly at upper levels where research articles may be a common reading task for students, as well as potentially a target production register. Little research to date has specifically focused on teaching these structures, although attention is building (e.g. Musgrave & Parkinson 2014, Parkinson and Musgrave 2014).

At the same time, the systematic patterns of variation seen across disciplines demonstrates that register features are not monotonic: phrasal complexity features are used to differing extents across disciplines. As register features are interpreted as *functional* characteristics of a variety of language (Biber & Conrad 2009: 54–55), this differing frequency of use in turn reflects the differing extents to which research article authors compress information into highly compact, informational units.

In addition to quantitative differences in the extent to which phrasal and clausal complexity features are used across disciplines, it's highly likely that qualitative differences also exist in terms of the specific functions of these features. For example, are the types of meanings expressed by noun + noun sequences similar across disciplines? How do non-finite structures function to both elaborate and compress information? While the focus in the present study has been on establishing these quantitative patterns of use, much remains to explain the functional reasons behind these quantitative patterns in specific disciplines. Features such as appositive noun phrases, and all prepositional phrases as noun post-modifiers, need to be incorporated into these analyses as well. Such qualitative and functional

analyses of disciplinary variation are surely needed to inform pedagogical materials that can help L1 and L2 writers produce these complex phrasal structures.

Finally, although the focus in this conclusion has so far been on the frequency and pervasiveness of phrasal complexity measures, this study has also demonstrated that clausal complexity features are used to a certain extent across academic disciplines. Although they are less frequent in academic writing than in many other registers, their systematic variability across registers is also likely related to the discourse functions that they enable writers to carry out. Extending our knowledge of when these clausal structures are used can help us to further understand how writers construct texts to create and transmit knowledge in particular disciplinary communities.

References

Afros, Elena & Schryer, Catherine. 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 & 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

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.

Biber, Douglas. 1988. *Variation across Speech and Writing.* Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 1992/2001. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes, 15,* 133–163. Reprinted in In *Variation in English: Multi-dimensional Studies,* Susan Conrad & Douglas Biber (eds), 215–240. London: Longman. DOI: 10.1080/01638539209544806

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 Linguistic Theory 223], Teresa Fanego, Javier Pérez-Guerra & María José López -Couso (eds), 43–66. Amsterdam: John Benjamins. DOI: 10.1075/cilt.223.06bib

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre and Style.* Cambridge: CUP. DOI: 10.1017/CBO9780511814358

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 emerging 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. 2013a. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2): 104–134.

Biber, Douglas & Gray, Bethany. 2013b. Discourse characteristics of writing and speaking task types on the *TOEFL iBT Test: A Lexico-grammatical Analysis* [TOEFLiBT Research Report (TOEFL eBT-19)]. Princeton NJ: Educational Testing Service.

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, Gray, Bethany & Poonpon, Kornwipa. 2011a. 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, Gray, Bethany, Honkapohja, Alpo & Pahta, Päivi. 2011b. Prepositional modifiers in early English medical prose: A study *ON* their historical development *IN* noun phrases. In *Communicating Early English Manuscripts,* Päivi Pahta & Andreas Jucker (eds), 197–211. Cambridge: CUP.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Conrad, Susan. 1996. Academic discourse in two disciplines: Professional writing and student development in Biology and History. Ph.D. dissertation, Northern Arizona University.

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

Fang, Zhihui, Schleppegrell, Mary & Cox, Beverly. 2006. Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research* 38(3): 247–273. DOI: 10.1207/s15548430jlr3803_1

Gardner, Sheena & Nesi, Hilary. 2012. A classification of genre families in university student writing. *Applied Linguistics* 34(1): 1–29.

Gray, Bethany. 2011. Exploring Academic Writing through Corpus Linguistics: When Discipline Tells Only Part of the Story. Ph.D. dissertation, Northern Arizona University.

Gray, Bethany. 2013. More than discipline: Uncovering multi-dimensional patterns of variation in academic research articles. *Corpora* 8(2): 153–181. DOI: 10.3366/cor.2013.0039

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, Michael A.K. 1979. Differences between spoken and written language: Some implications for literacy teaching. In *Communication through reading: Proceedings of the 4th Australian Reading Conference,* Vol. 2, Glenda Page, John Elkins & Barrie O'Connor (eds), 37–52. Adelaide SA: Australian Reading Association.

Halliday, Michael A.K. 1989. *Spoken and Written Language.* Oxford: OUP.

Halliday, Michael A.K. 2004. *The Language of Science.* London: Continuum.

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

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

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. 1999. Talking to students: Metadiscourse in Introductory coursebooks. *English for Specific Purposes* 18(1): 3–26. DOI: 10.1016/S0889-4906(97)00025-2

Hyland, Ken. 2002. Directives: Argument and engagement in academic writing. *Applied Linguistics* 23: 215–239. DOI: 10.1093/applin/23.2.215

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(2): 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

Koutsantoni, Dimitra. 2006. Rhetorical strategies in engineering research articles and research theses: Advanced academic literacy and relations of power. *Journals of English for Academic Purposes* 5: 19–36. DOI: 10.1016/j.jeap.2005.11.002

Lee, David & Chen, Sylvia. 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.2009.05.004

Musgrave, Jill & Parkinson, Jean. 2014. Getting to grips with noun groups. *ELT Journal* 68(2): 145–154. DOI: 10.1093/elt/cct078

Nesi, Hilary & Gardner, Sheena. 2012. *Genres across the Disciplines: Student Writing in Higher Education.* Cambridge: CUP.

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

Römer, Ute & Swales, John. 2010. The Michigan Corpus of Upper-level Student Papers (MICUSP). *Journal of English for Academic Purposes* 9: 249. DOI: 10.1016/j.jeap.2010.04.002

Russell, David. 2002. *Writing in the Academic Disciplines, 1870–1990: A Curricular History.* Carbondale IL: Southern Illinois University Press.

Schleppegrell, Mary. 2001. Linguistic features of the language of schooling. *Linguistics and Education* 12(4): 431–459. DOI: 10.1016/S0898-5898(01)00073-0

Silver, Marc. 2006. *Language across Disciplines: Towards a Critical Reading of Contemporary Academic Discourse.* Boca Raton FL: Brown Walker Press.

Vande Kopple, William. 1994. Some characteristics and functions of grammatical subjects in scientific discourse. *Written Communication* 11(4): 534–564. DOI: 10.1177/0741088394011004004

Warchal, K. 2010. Moulding interpersonal relations through conditional clauses: Consensus-building strategies in written academic discourse. *Journal of English for Academic Purposes* 9(2): 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(3): 257–268. DOI: 10.1016/0889-4906(94)90005-1

Wells, Rulon. 1960. Nominal and verbal style. In *Style in Language*, Thomas Sebeok (ed.), 213–22. 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(3): 175–197. DOI: 10.1016/0889-4906(96)00010-5

CHAPTER 4

Telling by omission

Hedging and negative evaluation in academic recommendation letters

Mohammed Albakry
Middle Tennessee State University & University of Connecticut

This corpus-based study explores some of the linguistic and discursive aspects of framing positive and negative information – mainly modals, evaluative adjectives, and mitigation strategies – in recommendation letters. The corpus is comprised of 114 letters of recommendation spanning three years of applications to an English Ph.D. program, approximately 46,000 words. The results reveal consistent patterns in the way different types of modals and their associated collocates are used to hedge predictions, and the analysis identifies the discursive frames of the most common mitigation strategies in presenting potentially negative information about applicants. The study illustrates the need to combine both corpus-based quantitative and qualitative methods for a more robust and fine-grained analysis of evaluative language in this occluded genre.

Keywords: Recommendation letters; evaluative language; modals; negative presentation; mitigation strategies; occluded genres

1. Introduction

Letters of recommendation are a commonly accepted component of almost any application process to an academic institution. Most graduate programs in North America, for example, often require two or more letters as part of their program admission requirements. Recommendation letters (henceforward LRs), along with an applicant's curriculum vitae, test scores and experience, are used to evaluate the applicant's past performance, work ethics, personality, and academic aptitude (Aamodt & Bryan 1993; Lopez et al. 1996).

DOI 10.1075/scl.66.04alb

In spite of the fact that there are not many studies of LRs, the general genre has been examined from different perspectives. While some studies have investigated the effectiveness of LRs as predictors of future performance, their relevancy to core competencies of a particular program, or their validity and reliability in admissions processes (Aamodt & Bryan 1993; Baxter et al. 1981; Blechman & Gussman 2008), the majority of the studies take on psychological or sociological perspectives with fewer studies adopting linguistic, rhetorical or intercultural approaches.

Most of the studies of LRs to date, however, tend to focus mostly on the issue of gender differences and equality among applicants. Bell et al. (1992), for example, analyzed seventy-eight letters of recommendation and concluded that men and women wrote letters differently, but their letters also varied by the gender of the applicant. Biernat and Eidelman (2007) reported that the participants in their study interpreted equivalent letters as indicating lesser ability and qualifications in female than male applicants, while Colarelli et al. (2002) – who looked at the appeal and tone of LRs – found that male recommenders are more likely to write favorable letters for female than male applicants.

Trix and Psenka (2003) examined over three hundred letters of recommendation for medical faculty positions and concluded that letters written for female applicants differ systematically from those written for male applicants in terms of length and percentage of doubt raisers, among other features. Their results reveal that more letters written for females as compared to males included language related to gender (10% versus 5%), doubt (24% versus 12%), and a higher degree of what they termed "grindstone adjectives" (e.g. hardworking, effort, conscientious; 34% versus 23%). They also commented that letters for male applicants made more reference to "his research," or "his career," as opposed to "her teaching," or "her training" for female applicants. Still related to language use and gender, Schmader et al. (2007) used text analysis software to compare letters written to recommend male and female applicants for tenure track faculty positions in chemistry and biochemistry at a large research university. They found out that most referees who choose to use more "standout words" (e.g. *remarkable, unmatched, unparalleled*, etc.) describing applicants also included more words referencing "ability" (e.g. *talent, skill, capacity*, etc.). However, reporting gender discrepancies being their focus, the researchers also found that recommenders used significantly more standout adjectives to describe male as compared to female candidates. Overall, this line of research suggests that the existence of gender bias and stereotyping could influence how applicants are evaluated.

Finally, using contrastive rhetoric and intercultural perspectives, Bouton (1995) and Precht (1998) focused on the different expectations and conventions for structure and content of LRs in the wider international and EFL academic

community. Based on a corpus of sixty-five letters of reference written by American referees and 65 letters written by referees from five Asian cultures, Bouton (1995) found that, in spite of the many similarities, Asian referees tended to use direct recommendations more frequently than their American counterparts. Adopting a similar approach but focusing on the western academic context (U.S, UK, Germany, and Eastern Europe), Precht (1998) also concluded that the LR format showed cross-cultural similarities. However, she also found some distinct regional patterns in her quantitative analysis of features such as linearity, data integration, advance organizers, and sentence types as well as qualitatively in terms of the use of different types of evidence (e.g. personal, factual, and narrative) and methods of support for applicants.

One of the main reasons for the relative paucity of research on the LR is the fact that it is an "occluded genre" hidden from public view (see Feak 2009; Swales 1996, 2004). Besides recommendation letters for students, other similarly "occluded" academic genres may include submission letters to journal editors, reviewer reports, research and grant proposals, and evaluation letters for tenure and promotion, to name a few. These important formal documents typically remain on file and rarely, if ever, become part of the public record (see Swales 1996: 46–47). This is all the more reason to study how such genres get structured and interpreted and how they function within the academic system.

This study focuses on the particular genre of LRs and some of its salient linguistic and discursive features without consideration of the variables of gender or culture. The study addresses the question of evaluative linguistic resources, mainly modals, adjectives and mitigation strategies, and how they encode appraisal information and indicate a recommender's level of confidence, or lack thereof, in an applicant's ability. Since both parties – the writer/recommender and the reader/evaluator – are often in similar fields, the LR tends to develop common patterns in its formatting and content. LRs, therefore, could be described as a genre of a "typified communicative action" characterized by similar substance and patterns in response to recurrent situations (Yates & Orlikowski 1992: 301). These patterns and the deviations from them can be used to draw conclusions about both applicants and writers of the letters.

Thus, it is important to investigate the language of LRs in different academic programs, particularly in English programs. As Bruland (2009: 406) notes, while it might be impossible or at least unproductive to try to prove "authorial intention," "it is arguable that a body of recommendation letters authored by and for English professors would exhibit more carefully theorized and intentional uses of language than letters from another field." With this in mind, looking at a body of recommendation letters written specifically for a discourse community that is presumably particularly sensitive to subtle language use could make the study of

recommendation for English programs a fruitful site for investigating the role, conventions, and interpretation of LRs in the humanities.

The chapter is organized in the following manner: I first provide a description of the corpus of LRs and the general framework for analysis. Next, I present the analysis and discuss some of the major findings on the distribution and use of central modals as well as the common frames used in mitigating the potentially negative information. Finally, I close the chapter with some concluding remarks.

2. Corpus, methodology, and data analysis

The study is based on 114 letters of recommendation. The letters, spanning three years of applicants, were collected from the Ph.D. English program of a large comprehensive university.[1] The total word count across all the letters comes out to 46,381 words. Letters about male applicants make up 44.4% of the corpus (48 letters), while 55.6% of the corpus (66 letters) are about female applicants (see Table 1). The gender of the applicants could be determined by the usage of pronouns in reference to the applicants as the actual names were redacted and replaced with numerical notations assigned to each letter in the corpus. For example, the first letter in the corpus is assigned the label "Student 1," the second "Student 2," and so on. As noted previously, the information regarding applicants' gender is provided for data purposes only since gender is not an analytical factor for this study.

Table 1. Breakdown of the LRs by gender and number of words

	Total	Male	Female
Number of letters	114 (100%)	48 (42.1%)	66 (57.9%)
Number of words	46,381 (100%)	20,601 (44.4%)	25,780 (55.6%)
Average letter length	406.85 words	429.19 words	390.61 words

The letters were initially examined using WordSmith concordance to investigate the lexical profile through word lists and their frequency order. I then noted

1. For full disclosure, I was the Director of Graduate Admission in English in my current institution (Middle Tennessee State University) from July 2010 until July 2013, a position that greatly facilitated my access to the "occluded" recommendation letters examined in this study. I am grateful to my department for giving me such a great opportunity, which sparked my interest in the current study.

the frequency of common lexical items that could signal as markers of appraisal expressing positive/negative attitude (Martin & White 2005: 2) as well as the frequency of the eight central modals: (*will, would, may, might, can, could, should, must*) and their contextual usage within the corpus (see Biber et al. 1999: 483). The corpus was not grammatically tagged, thus most of the identification of words and word classes was done manually. It should be stated, moreover, that the appraisal framework of Martin and White (2005) serves only as an overall interpretive guidance and I do not necessarily follow its terminology or its full taxonomy that covers the attributes and wide range of nominal and verbal appraisal features.

For the items in the lexical profile, the focus, besides the modals, was more on the most common content words, exclusively adjectives. Interestingly, very few explicitly negative content adjectives, if any, appeared in the lexical profile. For instance, words such as "poor," "inadequate" or "unsatisfactory" were totally absent from the corpus. Since the corpus search method proved insufficient in capturing the presentation of salient words with negative or potentially negative orientation, I also relied on close reading and qualitative/pragmatic analysis that entailed underlining and coding of positive (adjectives signaling praise) as well as negative orientation. Negative orientation is operationalized here as any statement or presentation which could potentially raise doubt about the recommendee's ability, level of intelligence, academic preparedness, and work ethics informed by the general category of doubt raisers (Trix & Psenka 2003). In their corpus-based study, Trix and Psenka (2003: 203) classify such presentation into semantic categories that include: negative, potentially negative, hedges, unexplained, irrelevancy, and faint praise. In this study, the markers of the above-mentioned categories are combined in a broad category of (potentially) negative information and the focus is more on identifying the common syntactic formulas that encode the expression of any negative/potentially negative presentation.

For a more fined-tuned analysis of praise-signaling adjectives, these adjectives were divided into two categories: markers of excellence and markers of adequacy. Markers of excellence are the words, which imply superiority in an applicant's ability and set the applicant apart from others. Such words are distinguishable from others in that they also imply a standard, which has been surpassed, or a difference from other possible applicants (e.g. *superior, superb*, *extraordinary*, etc.) These markers can be defined in a more general category as being words which imply "better than good" (good+) or "better than the standard" (standard+). The word *extraordinary* could perhaps provide a good illustration: There is a "norm" or baseline (*ordinary*) and the indication of being above the norm level (*extra*). The markers of adequacy, on the other hand, could be represented as "equal to standard or good" (standard= or good=), e.g. *good*, *fine*, and *solid*. These adjectives are examined in the contexts in which they occur.

3. Results and discussion

3.1 Modal usage

Because they are a closed set, central modals lend themselves easily to corpus-based investigation (given that simple searches with a concordancer can produce clear frequency counts). Modal verbs are one of the major grammatical devices that a writer/speaker uses to convey stance. They are often used to hedge or reassert confidence in a writer's commitment to an idea or mark the level of certainty in a statement (Biber et al. 1999:972–3). As hedges, they have both epistemic and affective functions in suggesting probability or mitigating potential damage of critical comments (Hyland 1998). The data on the frequency of the different modal verbs is presented in Figure 1.

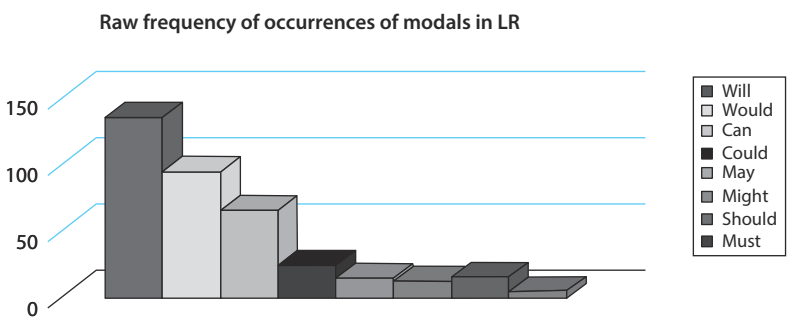

Figure 1. Frequency of occurrence of modal verbs in LRs

Recommenders – the majority of whom are faculty members in this case – sometimes exhibit the need to mitigate potential face threats in their endorsements of students' academic aptitude (see Brown & Levinson 1987). This becomes particularly clear in the use of certain modals, primarily those of possibility or necessity, to hedge their statements in regard to a student's ability or potential. On the reverse side, a faculty member who is confident in the applicant's ability can use modals to reinforce the confidence in that student's ability to contribute to and thrive in an academic environment.

When examining the letters of recommendation, certain patterns about the usage of modals emerged. The modals of prediction were found to be most prevalent. The items "will" and "would" occur 230 times across the corpus: "Will" occurs 135 times and "would" 95 times. This higher usage of the modals of prediction conforms to the understood purpose of LRs as a means of assuring evaluators about applicants' chances of success. With a vested interest in the acceptance of

the student, recommenders are more likely to express confidence in their statements about the student's performance in a graduate program. The following are examples of the usage of the prediction modal "will":

(1) "I am confident that she *will* quickly establish herself as an excellent teacher at both the undergraduate and graduate level." (Student 3)

(2) "In recommending [Student 4] to your attention, I have no reservations, and I am certain you *will* find him a promising and avid learner."

(3) "I feel confident that [Student 12] *will* be a diligent and productive participant in our graduate student community in English and undoubtedly a valuable teaching assistant."

(4) "Without a doubt, I believe that [Student 2] *will* be successful in whatever graduate program he ends up attending."

(5) "[Student 66] has the intellectual ability, diverse interests, personality, and drive to make an outstanding teacher of literature, one who *will* undoubtedly inspire his students to broaden their cultural perspectives."

In looking at these five examples of the usage of the modal "will," certain patterns become clear. Every example of this usage occurs near some reference to the level of confidence the writer has in the applicant. This can be expressed through a direct reference to confidence or certainty or by pointing out to the reader that there are "no doubts" in these statements. In Examples 1 and 2, the writer directly references this idea by stating explicitly that he or she is "confident" or "certain" of how the other educational institution "will" find the applicant. This "bald on-record" strategy (Brown & Levinson 1987:94) not only lets the reader know directly that the writer has confidence in the student but by expressing it in the "I am *confident/certain*" frame, it shows a high level of commitment. In other words, modals are not used in an effort to hedge predictions in these first two examples; instead we have explicit statements of the writer's state of mind in regard to the applicants' teaching and learning qualities.

Examples 3, 4, and 5 refer to doubt, or rather lack thereof. By using such enhancers as "undoubtedly" or "without a doubt", the recommenders hope to convey their strong conviction by asking the reader to trust in their judgments about the applicant. Such pattern of use brings a certain tone of finality to the discourse between reader and writer. However, note the distinction between "feeling" confident (as in Example 3) and "being" confident. The statement of "feeling" confident could perhaps be interpreted as a type of quality hedge where the writer does not endeavor to take responsibility for his/her statement, unlike the statement of "being" which equates confidence more with the writer's own self.

Continuing the examination of modals of prediction, below are some examples of the usage of the modal "would":

(6) "Overall, I *would* rate him certainly in the top ten percent of our students, perhaps in the top five percent." (Student 77)

(7) "If I had to choose one student I feel I will remember in ten years, [Student 18] *would* be that student.

(8) "Among the roughly 40 graduate students I've taught at [Student 102's University], I *would* place [Student 102] in my top 5, in terms of overall intellectual achievement, and in my top 2 in terms of literary, research and argumentative skills."

(9) "I *would* rank him one of the top 2 students that I have taught during this past decade." (Student 6)

As the examples indicate, "would" collocates more closely with indications of numerical rank than "will" and does not collocate as closely with explicit references to certainty or confidence. Expression of supposed rank of the applicant in regard to other applicants brings about the possibility of the phenomenon of inflation in letters of recommendation, where a writer exaggerates the abilities of achievements of applicants in order to give them a better chance at gaining acceptance to the institution (see Ryan & Martinson 2000; Miller & Van Rybroek 1988). The collocation of the word "would" with the different expressions of rank indicates a willingness to ascribe a quantitative number or category to an applicant. The level of confidence the writer has in the applicant is, in these cases, supported by the statistical data provided by the writer.

Moving on from the modals of prediction, I discuss the modals of possibility: *can, could, may,* and *might*. These modals occur 119 times across the corpus, making up a much smaller percentage of the modals used. Here are two examples of the use of "can" and "could":

(10) "On the whole, I think she *could* be as good a doctoral student as many in the program right now." (Student 55)

(11) "I believe that she *could* do equally solid work at the doctoral level." (Student 55)

While these examples are still supportive, they can be considered as instances of faint praise or "apparent commendation" (Trix & Psenka 2003), praise that is *less* positive than the previous examples involving modals of prediction. Here we no longer see the references to a level of certainty, but instead we have collocates that involve "thinking" or "believing" – verbs that also serve as another type of hedge. While the modals of prediction are often accompanied by opinions stated as if

they were facts, these examples with modals of possibility are presented as opinions to be shared. The writer is letting the reader know that he or she believes these claims to be true but does not offer up the same level of personal commitment found in the use of modals of prediction.

In addition to the lack of stated confidence, these modals are usually located near words that could be described as markers of adequacy as opposed to excellence (see more details in Section 3.2). As previously defined, markers of adequacy indicate the reaching of a standard without implying a surpassing of that standard. In Example 10, the referenced applicant is thought to be "as good as" other graduate students by the writer. This may seem like praise for the applicant, but the applicant is just being equated to other students without being placed in a category above them. In other words, the nature of the statement speaks of an "average" competency – not a particularly high praise in a genre where most recommendees often turn out to be well "above average" (Liberman 2010). Example 11 gives the impression of a state of stasis on the part of the applicant. The applicant's work is considered to be "solid" by the recommender, a positive adjective that could pale in comparison with the many markers of excellence often found in LRs. The implicature is that the student is "merely" satisfactory.

Finally, if modals in LRs are to be arranged in a hierarchy of confidence implicature, then modals of necessity (*must* and *should*) may represent the lowest level of that confidence. These modals are used most often to hedge statements about an applicant's future actions in an effort to protect the writer's face needs:

(12) "He *should* be able to function at your school and would be a great research assistant." (Student 22)

(13) "Her maturity and self-motivation *should* also be assets in a doctoral program." (Student 25)

(14) "As [Student 41] continues to reflect and grow as a teacher and scholar, he *should* become even better."

In looking at these examples, we can see the evidence of hedging on the part of the writer. Example 12 paints the picture of a student who may or may not be able to function at the reader's school. Even though the applicant *would* be a great research assistant, this positive outcome is dependent on the outcome of whether or not the student is able to *function* at the intended institution. Reading between the lines, perhaps the applicant has had a previous problem with transitions from one environment to another that the writer is aware of but chooses not to disclose – in other words, what is said here could be less important than what is implied. Example 14 concerns a beginning teacher who is just starting out in his career. The recommender uses the brief window of past performance as a

prediction that this growth *should* continue. Instead of committing to the certainty or near-certainty of this continued development, the writer merely suggests that history supports the idea of this growth continuing, a less personal investment on the confidence scale.

It is clear from the discussion here that *will* is used with positive semantic prosodies and these prosodies/preferences tend to get more neutral and more negative as the frequency of the modals analyzed goes down. As a concept, the term "semantic prosody" arises from the "phraseological" tradition of corpus linguistics associated with the focus on the typical behavior of individual lexical items in their co-text. The notion of negative and positive evaluation, however, as Hunston (2007: 256) argues, may be over-simplistic since the attitudinal meaning of a word could be altered by its co-text. In other words, it might be more useful to conceptualize "semantic prosody [as] as discourse function of a sequence rather than a property of a word" (2007: 258). The alternative terms "semantic preference" or "attitudinal preference" are suggested as perhaps a better way to refer to "frequent co-occurrence of a lexical item with items expressing a particular evaluative meaning of a lexical item (Hunston 2007: 266).

3.2 Lexical markers of positive evaluation

Now I turn to examining in more detail some of the content words and positive orientation adjectives that were attested in the lexical profile of the corpus. As previously explained, these words were divided into the two categories: markers of excellence and markers of adequacy. The following are some illustrative examples of both categories:

(15) "In fact, her scores were quite *extraordinary*: on all eight graded exercises in the two courses, she earned A's. Please understand that I am a virulent opponent of grade inflation." (Student 17)

(16) "I give [Student 44] my highest recommendation for admission into a doctoral program because her scholarship is *exceptional* and her teaching is *extraordinary*."

The word *extraordinary* in Example 15 means exactly that: The applicant did something that by the writer's exacting standards was not merely ordinary. The adverbial modifier "quite" serves as a type of vague hedging, but perhaps it does not have a strong impact because of the "standout" quality of the adjective used. Interestingly, the writer here, assuming it is a male recommender, seems keen on raising his own credibility as well as that of the applicant. He manages to cater to his own positive self-image while letting the reader know that the applicant's accomplishment was atypical of other students and thus noteworthy. Example 16 employs both the chosen example word *extraordinary* as well as another word from the

same category: *exceptional*. If this student is "exceptional", then that means she surpasses the normal scholarship of her peers. The two markers of excellence in such a short space combine to "intensify the force" and "sharpen the focus" (see Martin & White 2005: 38). Other similar words attested in the corpus that belong to this category include (*superb, outstanding, superior, excellent, best,* and *finest*).

Now, on the medium end of the praise spectrum, the corpus has a few other words that also serve positive evaluation but arguably for a lesser degree. To distinguish them from the first category, these are grouped together as markers of adequacy and include such words as *good, fine, solid, competent, satisfactory*, and of course the adjective *adequate* itself. This appraisal category may have some superficial similarities to what Trix and Psenke (2003: 207) term "grindstone adjectives" (e.g. *hard working, dependable*, and *industrious*, etc.) except that markers of adequacy as defined here do not necessarily focus on diligence and dependability. The following examples offer some contextual illustration:

(17) She is a *good* public speaker and a *competent* teacher. (Student 83)

(18) He did *solid* work in that class, and ended up getting a B+. (Student 110)

(19) [Student 68's] written work was pretty *good*.

The generic adjective "good" can be used to qualify a range of domains and it is inarguably positive in its semantic content. However, again when we take into account the paradigmatic dimension of evaluative adjectives available to LR writers and attested in the corpus, the use of "good" does not equal the more desirable levels of skill or ability signaled by markers of excellence. There is nothing inherently negative about words such as *good, competent, solid, fine* and such adjectives, but when examined in relation to the possible words commonly used to describe applicants in this genre, these restrained words suddenly appear less praising. The effect, of course, depends mainly upon readers' conception of these words and their subjective reaction, but I argue that it is also conditioned by the overall presence of more lavish praise markers. In the context of the usually celebratory tone particular to the LR genre, the interpretation of markers of adequacy is likely to be that of evaluation that is more "critical". Interestingly, many words in this category also tend to be preceded by adverbial modifiers such as *pretty*, *somewhat,* etc. (see Example 19) which can down-scale their pragmatic force and weaken their positive effect even more.

3.3 Common frames in mitigating the negative

Because of the more subtle nature of doubt casting and negative expression, close reading was necessary as a complement to the corpus analysis. Guided by the use of modals of possibility and necessity as well as markers of adequacy discussed above, I further examined the letters to identify any statements of (potentially)

negative information. All the identified statements were then selected and coded for their common frames.

As Figure 2 illustrates, out of the 114 recommendation letters, potentially negative information was found to be present in 24 of them, i.e. 21% of the letters as a whole.

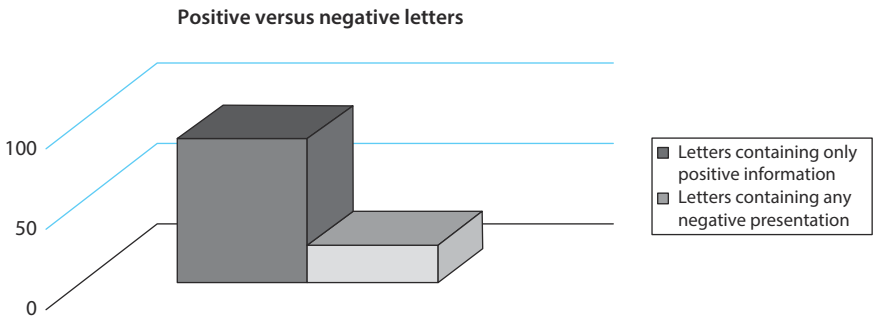

Figure 2. Comparison of positive versus negative observations in the LRs

Within these 24 coded letters, 34 instances of negative information were identified: 16 of the letters include only one instance of potential negativity each; six include two instances each; and the remaining two letters contain three separate instances each. Figure 3 gives a visual illustration of the findings.

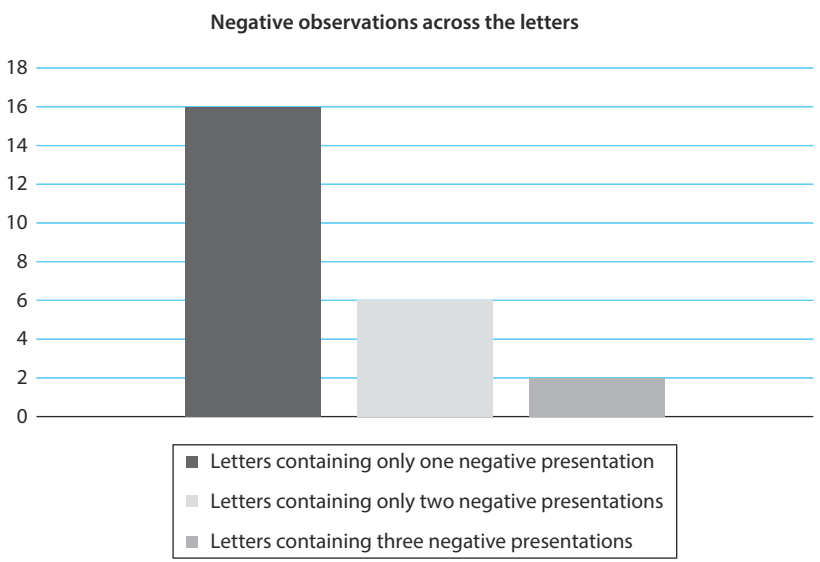

Figure 3. Frequency of incidence of negative observations in the LRs

Notably, with the exception of only one letter, every single one of the 24 negatively coded LRs was also shorter in length than the total average of 406.85 words. From a quantitative point of view, length of letters, it seems, could serve as a good indicator/predictor of the recommendee's quality and the recommender's high confidence.[2] Most strikingly, the identified negative statements are almost always bookended by mitigation strategies (see Table 2 for a summary). Only two out of the 34 instances of potentially negative presentation were found to stand on their own in one letter as two independent statements (Example 20).

(20) "[Student 87] seems a bit unsure of the direction he wants to take. I also did have some report that his work as a research assistant was not always done on time."

While there are a total of eight letters in the corpus that contain two or more instances of negative information, it is rare that a recommender will actually offer up two negative remarks consecutively with no intervening mitigation. This letter contains some of the most damning commentary in any of the LRs reviewed. It calls into question the applicant's sense of academic direction as well as his time management skills (notice the multiple hedges: *seems, a bit*). In a following short paragraph, the recommender mentions some non-specific positive traits and concludes the letter by stating that:

(21) "[Student 87] possesses traits that are worthy of acceptance into your program. I recommend him to your Ph.D. program in English."

The ending to this short and weak letter, while still commendatory, is also telling in what it does not contain. None of the usual collocates or enhancers of "recommend" are used such as "I *strongly*, *highly,* or *wholeheartedly* recommend" or "I recommend *without reservation*." Clearly, the applicant here does not have the recommender's *full* support.

The other remaining 32 negative instances are presented in a formulaic way: The negative information is adjacently paired with a positive or mitigating statement. This discursive strategy seeks to get credit for honest appraisal of the applicant's qualifications while at the same time softening the potentially negative impact. The common formulas or discursive frames of combining negative and positive presentations that evolved from the data analysis are summarized in Table 2.

---

2. For an interestingly similar case where quantity or word length may equal quality and confidence, see Liberman's blog entry on the length of wine reviews (2012) and Colarelli et al. (2002).

Table 2. Summary of the five major discursive frames of negative/positive pairings in the LRs

Frame	Example
Frame I Subordinator + "*Good*", "*Bad*" (or Subordinator + "*Bad*", "*Good*")	Frame I While the idea is an intriguing one, [*Student 1*] did not always, in these papers, provide close readings that fully substantiated his claims.
Frame II Specific "*Good*" in contrast to the unspecified "*Bad*"	Frame II I will say that [*Student 2*'s] stellar GRE scores impressed us all, although there was some skepticism about other elements of his application.
Frame III Initially "*Bad*". "However" applicant undergoes transformation = "*Good*"	Frame III During this semester, he rarely spoke, handed in at least one paper late, and, although clearly bright, was generally an unremarkable student. After two years of teaching in a local middle school, however, [*Student 2*] returned to apply to [*Student 2's School*] program with a new sense of commitment and intellectual curiosity.
Frame IV "*Good*", but "*Bad*" or "*Bad*", but "*Good*"	Frame IV I found him intelligent and outgoing, but his work proved to be less than first-rate.
Frame V Like "them," *not particularly Good*	Frame V [*Student 80*] like almost all of our students was not able to have a perfect 'A' record."

3.3.1 *Frame I*

Subordinator + "*Good*", "*Bad*" or Subordinator + "*Bad*", "*Good*"

1. "While the idea is an intriguing one, [Student 1] did not always, in these papers, provide close readings that fully substantiated his claims."
2. "While there are certainly areas in which she needs to grow as a scholar, I have been impressed by her effort…" (Student 55)

The first example presents the reader with positive information followed by a rather cautious presentation of negative information. The words "not always" and "in these papers" aim to give some context to indicate that the writer is referring to very specific incidents, but perhaps the applicant has the capability to accomplish the task of "close reading" in other assignments. The second example presents the negative information first and foregrounds the "honesty" concern that has been discussed in other studies (Aamodt & Bryan 1993). In this example, a legitimate concern is presented, followed by a positive assessment in a vague area

("effort") in which the applicant is supposedly "impressive" and does not need to "grow as a scholar". "Effort," however, seems to be one of those "code" words that can be taken to mean "the student is trying but she is certainly not trying hard enough".

3.3.2 *Frame II*

Specific "*Good*" in contrast to the unspecified "*Bad*"

3. "I will say that [Student 2's] stellar GRE scores impressed us all, although there was some skepticism about other elements of his application."

The example of this formula makes use of two of what could be described as "standout" words: "stellar" and "impress" (see Schmader et al. 2007: 514). This positive presentation is followed by a contrastive word, which signifies a shift in thought. The negative information is then presented, but it should be noted that the writer – unlike the specific mention of GRE scores – does not go into detail about these "other elements" in the application package. This lack of specificity may diffuse the impact of the negative presentation, but then again it may exacerbate it for some readers.

3.3.3 *Frame III*

Initially "*Bad*". "However" undergoes transformation = "*Good*"

4. "During this semester, he rarely spoke, handed in at least one paper late, and, although clearly bright, was generally an unremarkable student. After two years of teaching in a local middle school, however, [Student 2] returned to apply to [Student 2's School] program with a new sense of commitment and intellectual curiosity."
5. "Initially a bit shy, she developed confidence over the two semesters as demonstrated through improved class participation." (Student 38)

The formula involving the applicant undergoing some form of personal or professional transformation, as we can see from both examples given, often relates to public persona or "shyness." In a field where the applicants are often required to present their own opinions verbally and support them in front of a room full of people, as is the case with English Ph.D. students, an applicant's "shyness" could be perceived as a negative factor. The transformation in the first example is the result of experience gained in a teaching position, or time (in the second example) that helped the applicant to overcome her perceived shortcoming.

3.3.4 *Frame IV*
"*Good*", but "*Bad*" or "*Bad*", but "*Good*"

6. "I found him intelligent and outgoing, but his work proved to be less than first-rate." (Student 51)
7. "He is not absolutely the brightest graduate student I've known, but he is one of the brightest, and his writing has always demonstrated more intellectual curiosity and a little more ambition than is the norm here." (Student 77)

The interpretation of the word "but" will always create the conventional implicature of a sense of contrast and almost half the examples of negative/positive pairing depend on this common frame. The last two examples give us a view of the more basic means of pairing positive information with negative information. One set of information is stated, then the other is offered as a counterpoint either for downgrading (Example 6) or upgrading (Example 7). The statements here differ from the examples of Frame I & II primarily in their consistent use of the coordinating conjunction "but". Example 6 operates on the basis that while the qualities initially presented are considered valuable in an academic community, the institution being applied to supposedly wants its applicants to be of a higher quality. Note in this example, we have another hedge which is the avoidance of the direct expression of the negative. The circumlocutory choice of "less than first-rate" is clearly a euphemism for "inadequate" or at least "not good enough". This semantic point illustrates the wide range of possible linguistic resources for delivering negative evaluation and the potential insufficiency of corpus methods alone in capturing language nuances.

Speaking of semantics, the second example presents a case of ambiguity in regard to the use of the word "absolutely." Does the writer intend this sentence to be read as "He is not the [absolute] brightest graduate student I have ever known…" or "He is [absolutely] not the brightest graduate student I have ever known…"? The actual intentions of the writer may not matter in the reader's reception of the sentence. The reader is going to perceive this sentence one way or the other, but the polarity effect of "not" (to use Martin and White's terms) is still potentially negative. The recommender, however, does rank the student higher "than is the norm here" in the areas of ambition and intellectual curiosity.

3.3.5 *Frame V*
Like "them," *not particularly Good*
There are two instances of the potentially negative being presented in a different but still formulaic fashion given below.

8. "Like most of the students in the 'Images of Africa' course, he came to it with little or no knowledge of African literature." (Student 4)

9. "[Student 80], like almost all of our students, was not able to have a perfect 'A' record."

The potentially negative is presented here without any explicit positive information. Instead, it is paired with an excuse to avoid singling out the student. In both examples, the excuse is that the applicant exhibits some type of academic limitation, but so are the peers. The perceived weakness, therefore, is somewhat ameliorated. Although unsaid, Example 8 does imply that even though the student initially had no knowledge of the subject matter when he "came" to the class, he "departed" with the required knowledge he previously lacked.

4. Conclusion

The problem of interpreting letters of recommendation is that some LR writers tend to be superlative, while others are more reserved. This can complicate candidate calibration, but it does not necessarily get in the way of successful communication (Liberman 2010). By their very nature, LRs tend to have a "superiority bias"; they tend to overstate positive qualities and understate negative ones. While this awareness should be taken as an element in the interpretive and evaluative considerations we apply to these texts, it may also lead us to judge LRs according to an unfairly higher standard (see Miller & Rybroek 1988). Paradoxically, in a genre where most applicants receive "whole-hearted" and "enthusiastic" recommendations, even words and comments that normally present positive information can be seen negatively in the company of other more "glowing" letters in the same group. Certain words can also become present by their absence, and what is not explicitly stated can be as salient as (or even more salient than) what is. Letters of recommendations, for better or for worse, are telling in both commission and omission.

The interpretation of any remarks in LRs, however, is conditioned not only by the context of reception and readers' subjectivity, but also by the readers' experience with the conventions of the genre. It might, therefore, be beneficial for educational institutions to have workshops for their faculty to raise their awareness about the issues involved in writing and interpreting LRs. After all, it is a rare professor who is never asked to write or read them. It is hoped that the findings here can offer some guidance especially to junior faculty who can be at a disadvantage when expected to write or evaluate recommendation letters.

This study is limited by the relatively small size of its corpus and its focus on only a small part of the evaluative linguistic resources. For example, the different nominal and verbal aspects of evaluation, important as they are, were not included in this investigation. The study is also focused only on analyzing LRs of graduate

admission in one discipline within the humanities, and thus the results may not be generalizable to other contexts and disciplines in the academy. The strength of the study, however, lies in its combining of the corpus method and the qualitative interpretation to broaden research on evaluative language to the relatively under-studied genre of recommendation letters. This combination of both methods makes the identification of more subtle patterns of language use possible, and allows for a more fine-tuned analysis.

Such occluded genres as recommendation letters are typically "out of sight to outsiders and apprentices" (Swales 2004: 18). However, given their evaluative and administrative importance for both students and professors, further linguistic and discourse studies in different contexts are needed.

References

Aamodt, Michael & Bryan, Devon. 1993. Predicting performance with letters of recommendation. *Public Personnel Management* 22(1): 81.

Baxter, James, Brock, Barbara, Hill, Peter & Rozelle, Richard. 1981. Letters of recommendation: A question of value. *Journal of Applied Psychology* 66(3): 296–301. DOI: 10.1037/0021-9010.66.3.296

Bell, Susan, Cole, Suzanne & Floge, Liliane. 1992. Letters of recommendation in academe: Do women and men write in different languages? *The American Sociologist* 23(3): 7–22. DOI: 10.1007/BF02691910

Biber, Doublas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Written and Spoken English*. Harlow: Longman.

Biernat, Monica & Eidelman, Scott. 2007. Translating subjective language in letters of recommendation: The case of the sexist professor. *European Journal of Social Psychology* 37(6): 1149–1175. DOI: 10.1002/ejsp.432

Blechman, Andrew & Gussman, Debra. 2008. Letters of recommendation: an analysis for evidence of Accreditation Council for Graduate Medical Education core competencies. *Journal of Reproductive Medicine* 53(10): 793–797.

Bouton, Lawrence. 1995. A cross-cultural analysis of the structure and content of letters of reference. *Studies in Second Language Acquisition* 17: 211–211. DOI: 10.1017/S0272263100014169

Brown, Penelope & Levinson Stephen. 1988. *Politeness: Some Universals in Language Usage*. Cambridge: CUP.

Bruland, Holly. 2009. Rhetorical cues and cultural clues: An analysis of the recommendation letter in English studies. *Rhetoric Review* 28(4): 406–424. DOI: 10.1080/07350190903185064

Colarelli, Stephen, Hechanova-Alampay, Regina & Canali, Kristophor. 2002. Letters of recommendation: An evolutionary psychological perspective. *Human Relations* 55(3): 315–344. DOI: 10.1177/0018726702553002

Feak, Christine. 2009. Negotiating publication: Author's responses to peer review of medical research articles in thoracic surgery. *Revista Canaria de Estudios Ingleses* 59: 17–34. ⟨http://publica.webs.ull.es/upload/REV%20RECEI/59%20%202009/02%20Feak.pdf⟩

Hyland, Ken. 1998. *Hedging in Scientific Research Articles* [Pragmatics & Beyond New Series 54]. Amsterdam: John Benjamins. DOI: 10.1075/pbns.54

Hunston, Susan. 2007. Semantic prosody revisited. *International Journal of Corpus Linguistics* 12(2): 249–268. DOI: 10.1075/ijcl.12.2.09hun

Liberman, Mark. 2010. Lying by telling the truth? *Language Log*, 5 January 2010. ⟨http://languagelog.ldc.upenn.edu/nll/?p=2023⟩

Liberman, Mark. 2012. The quality of quantity. *Language Log,* 25 April 2012. ⟨http://languagelog.ldc.upenn.edu/nll/?p=3922⟩

Lopez, Shane, Oehlert, Mary & Moberly, Rebecca. 1996. Selection criteria for American Psychological Association-accredited internship programs: A survey of training directors. *Professional Psychology: Research and Practice* 27(5): 518–520. DOI: 10.1037/0735-7028.27.5.518

Martin, James & White, Peter. 2005. *The Language of Evaluation: Appraisal in English*. Houndmills: Palgrave Macmillan.

Miller, Rodney & Rybroek, Gregory. 1988. Internship letters of recommendation: Where are the other 90%? *Professional Psychology: Research and Practice* 19(1): 115–117. DOI: 10.1037/0735-7028.19.1.115

Precht, Kristen. 1998. A cross-cultural comparison of letters of recommendation. *English for Specific Purposes* 17(3): 241–265. DOI: 10.1016/S0889-4906(97)00012-4

Ryan, Michael & Martinson, David. 2000. Perceived effects of exaggeration in recommendation letters. *Journalism & Mass Communication Educator* 55(1): 40–52. DOI: 10.1177/107769580005500105

Schmader, Toni, Whitehead, Jessica & Wysocki, Vicky. 2007. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles* 57(7–8): 509–514. DOI: 10.1007/s11199-007-9291-4

Swales, John. 1996. Occluded genres in the academy: The case of the submission letter. In *Academic Writing: Intercultural and Textual Issues* [Pragmatics & Beyond New Series 41], Eija Ventola & Anna Mauranen (eds), 45–58. Amsterdam: John Benjamins. DOI: 10.1075/pbns.41.06swa

Swales, John. 2004. *Research Genres: Explorations and Applications*. Cambridge: CUP. DOI: 10.1017/CBO9781139524827

Trix, Francis & Psenka, Carolyn. 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society* 14(2): 191–220. DOI: 10.1177/0957926503014002277

Yates, Joanne & Orlikowski, Wanda. 1992. Genres of organizational communication: A structural approach to studying communication and media. *Academy of Management Review* 17(2): 299–326.

Acknowledgements

I would like to thank the University of Connecticut's Humanities Institute (UCHI) and its director Sharon Harris for their support for my research and writing for the year 2013–2014. I would also like to thank my former students David LeDoux and Christopher Hand for their help in the initial stages of data coding. Finally, I am grateful for the helpful editorial comments I received from an anonymous reviewer.

CHAPTER 5

Corpora, context, and language teachers

Teacher involvement in a local learner corpus project

Alfredo Urzúa
San Diego State University

The language teacher is an often neglected figure in learner corpora projects, including those whose aim is to apply corpus findings to second language pedagogy. Even though teacher mediation is critical to the potential success of corpus-informed instructional practices, the literature seldom addresses specific ways to get classroom teachers involved in the process of designing, collecting, and exploring learner corpus data. This chapter describes a learner corpus project in which the participation of local English language teachers was actively recruited throughout the project. The author describes ways in which teachers were involved in the project and illustrates the benefits of such a process with examples from a corpus-based study he conducted in the same local language teaching context.

Keywords: Learner corpus; teacher involvement; contextualization

1. Introduction

By the turn of the 21st century, corpus linguistics has established itself as a field that has made unique contributions to our understanding of how language is used in *real* situations, i.e. natural language produced by people in authentic settings. Its scope is broad, from uncovering patterns of language variation to describing the lexico-grammatical profile of texts types, and from analyses of multi-word units to the development of instructional materials, to name just a few. Despite some early criticisms, the value of corpora is by now recognized in most linguistic circles; however, as in any area of inquiry, controversial issues emerge recurrently and some goals remain elusive or present challenges that prove difficult to overcome. A case in point refers to the 'de-contextualized' nature of corpus data (McEnery, Xiao & Tono 2006), a notion which in turn plays an important role

DOI 10.1075/scl.66.05urz

in fulfilling the potential pedagogical applications of corpus-based approaches (Aijmer 2009; Kaltenböck & Mehlmauer-Larcher 2005). In the present chapter, I discuss these issues in relation to the design and use of a local learner corpus of general academic English and the various ways in which such a corpus can benefit from teacher involvement. The ultimate goal, in this instance, relates to the implementation of a corpus-informed second language program that can be more conducive to bridging the gap between corpora and pedagogy. In essence, this is a case study of the process of building a learner corpus within the context of a particular English language program, the various ways in which teachers can be involved in such a process, and how this approach can positively affect the use of corpus-informed data within a specific instructional program.

Before describing the project and the language program associated to it, the next section presents an overview of the literature in relation to the role of context in corpus-related work and pedagogical applications of corpus information, with a focus on learner corpora.

2. Background

2.1 Corpus data and context

A well-known criticism of corpus linguistics was made by Henry Widdowson in his article 'On the Limitations of Linguistics Applied.' Widdowson claimed that corpus data, despite its value in revealing aspects of language use not accessible via intuition or introspection, "cannot account for the complex interplay of linguistic and contextual factors whereby discourse is enacted; […] corpus analysis deals with the textually attested, but not with the encoded possible, nor the contextually appropriate" (2000: 7). Widdowson considered this limitation rather problematic whenever corpus data are applied to pedagogical situations, for example, the specifications of language content. Texts in a corpus, he argued, constitute instances of decontextualized language which, in order to be used in a classroom, would need to be reconstituted or recontextualized so that it could be made 'real' for learners, and thus appropriate from a pedagogical standpoint. Widdowson agreed that there are important and valuable applications of corpus descriptions for the language classroom, so he did not dismiss the pedagogical potential of corpus studies. Instead, he cautioned against using corpus data without paying attention to teachers' and learners' characteristics, experiences, goals, and attitudes, as well as the range of situational conditions that impact how language is taught and learned, processed and produced, in context (see also Cook 1998).

More recently, in relation to large-scale general corpora, Flowerdew (2005, 2012) has stated that we must accept that these data are basically divorced from

its original context. However, the extent to which a lack of contextualization may affect pedagogical applications of corpora, and the ways such application could be better achieved through pedagogic mediation, continues to be discussed among corpus linguists. The question of 'how to use corpora in language teaching' – to use the title of Sinclair's (2004) well known volume – remains at the crux of the debate.

2.2 Learner corpora and context

In recent years, the compilation of learner corpora has added a new dimension to discussions of corpus research design, goals, relationship to contextual factors, and classroom applications. Given that learner corpora focus on language generated by non-native speakers (NNS), the criteria for organizing a learner corpus is somewhat different from that of native-speaker (NS) corpora. For example, in the former, in addition to criteria such as text type (e.g. newspaper editorial), mode (e.g. written, spoken), or communicative situation (e.g. service encounters), language may be organized by level of proficiency (e.g. beginning, advanced) or the type of classroom situation in which the language is generated (e.g. collaborative work, testing conditions).

Nesselhauf (2004) mentions two other important differences between NS and learner corpora. First, texts in a learner corpus are not considered samples of 'naturally-occurring language,' at least not in the strict sense, as it is normally the case with NS corpus data. In learner corpora, "what comes closest to naturally occurring texts … are [texts] produced for pedagogical reasons" (p. 128). Because of this, the classroom circumstances under which oral or written texts are generated become crucial to an understanding of why texts produced by language learners look the way they do. Thus, in addition to typical descriptors included in most corpora (e.g. participants' characteristics, text types), it is important to gather information about learners' sociocultural and pedagogical context, which inevitably influence the process of language acquisition and levels of ultimate attainment (Lantolf 2000).

Research goals when analyzing learner corpora data are also unique in that they not only include the identification of typical difficulties or comparisons regarding the use of particular language features by NNS of a language in relation to their NS counterparts, but also because of increasing attempts to explore second language development using corpus-based approaches (Granger 2004). Even though most studies of learner corpora are cross-sectional, i.e. they include data collected from learners at a single point in time, the importance of conducting longitudinal studies to investigate learners' interlanguage using a corpus approach has been underscored by many researchers in the last decade.

It is not too surprising, then, that longitudinal studies of second language development are becoming increasingly visible, as evidenced by a recent issue

of *The Modern Language Journal* devoted to this topic. In relation to the issue of context, in her introduction to this special issue, Hasko (2013) highlights the importance of paying attention to contextual factors in any quasi-longitudinal or longitudinal investigations of second language development via learner corpora, which she calls a "a shift in paradigms in itself" (p. 5). She considers that these studies allow the possibility of "prolonged tracking of *contextualized* indices of L2 development" (p. 6, italics added), thus allowing more insightful analyses that move away from an over-reliance on repeated cross-sectional comparisons. Hasko agrees with other researchers in the field that there is a pressing need to compile learner corpora annotated not only for learner and text variables, but also for contextual and instructional ones. Such information is crucial in explorations of the relationship between pedagogies and their effect on L2 learning, and this provides support to the compilation of corpora at the learners' and researchers' home institution, which "makes it possible to get access to and consider the nuanced variables describing the learner, speech events, larger communities, and pedagogical context." (p. 7).

Hasko (2013) also highlights the important role that contextual information plays in implementing effective pedagogical applications of learner corpora, or any corpora for that matter. This is, indeed, no trivial matter. One can only wonder how descriptions resulting from analyses of a large, decontextualized learner corpus can be effectively related to groups of learners whose characteristics and situational conditions may be quite dissimilar to those reflected in the corpus. This goes back to the notion that, in a teaching context, 'authenticity' is not an inherent characteristic of texts but one which needs to be reconstituted "on account of the impossibility of replicating the original contextual conditions of the language in the classroom" (Mauranen 2004: 93).

The argument supported in this paper is that, in order to make effective use of corpus data, teacher mediation is needed and, for this to happen, teachers need to develop not only technical, content, and pedagogical knowledge in corpus linguistics, but also the beliefs, attitudes, and motivational drive that can compel them to incorporate information from corpora into their teaching, as well as explore their possible advantages over more traditional methods.

2.3 Corpora, teachers, and pedagogical applications

At present, two main types of pedagogical applications of corpora have emerged. Indirect applications, which refer to the use of corpus descriptions to inform the production of pedagogical materials, such as dictionaries, reference grammars and ELT textbooks; and direct applications, also known as data-driven learning (DDL), which broadly refer to cases in which students analyze a corpus (or

examples from a corpus) in a classroom setting in order to discover, and thus better understand, aspects of language previously unnoticed (Bernardini 2004). It must be said, however, that the role of the teacher in direct applications of corpora has not been discussed extensively. Few scholars have paid attention to what this role is or should be, although the figure of the teacher has slowly started to appear in recent work (e.g. Tsui 2004; Römer 2006; Reppen 2010).

In terms of their pedagogical applications, the use of learner corpora to improve pedagogic materials appears to be the area with more potential, as more and more corpus-based studies looking at different aspects of learner language are investigated. However, their impact has not been very significant yet, except perhaps in the production of learner dictionaries. In addition, attempts have also started to emerge to explore the potential use of DDL activities using learner corpora; for instance, by presenting students with samples of negative evidence (typical or frequent mistakes) for them to identify, analyze, and correct in light of positive evidence from native-speaker corpora. Nonetheless, it may seem as if the most we can hope for is that corpus-informed teaching materials become increasingly available and that teachers adopt them because they are sufficiently knowledgeable or well-informed about the potential pedagogical applications of corpora in comparison to more 'traditional' materials. Römer (2009), however, has documented that many English language teachers do not see using language corpora as an alternative or supplement to traditional teaching materials.

Aijmer (2009), in her introduction to 'Corpora and Language Teaching' states that, despite the enthusiasm generated by the potential pedagogical applications of corpus linguistics research, "the use of corpora in the EFL classroom is a rare occurrence and teachers are still unwilling to or lack the skill to use corpora as an aid to get new insights into English." (p.1) Part of the problem, Aijmer indicates, lies on the challenges of establishing appropriate relationships with practicing language teachers so that information from corpus studies is not only well received and understood, but applied in positive, successful ways, from the point of view of the teacher as well as that of the linguist-researcher.

Similarly, in her recent volume 'Corpora and Language Education', Flowerdew (2012) also comments on the by-now persistent gap between corpus linguistics and language pedagogy. She agrees with various practitioners (e.g. Frankenberg-García 2010; McCarthy 2008; O'Keefe & Farr 2003) that a crucial factor is "the lack of training in how to use corpora by the teachers themselves" (p. 221). This training, ideally, should include information about corpora, how to use corpora, and how to teach using corpora (see O'Keefee & Farr 2003; Farr 2010). In addition, Sinclair (2004) has also argued that more attention needs to be paid to training teachers on how to evaluate information retrieved from corpora.

Various attempts have been made in recent years to try to bridge the gap between teachers and corpus researchers. Some applied linguists, for instance, provide teachers with practical suggestions on how to use available corpora to examine linguistic features and create instructional materials and teaching activities appropriate for their students (e.g. Reppen 2010). Others advocate for including courses or modules on corpus linguistics and corpus-based materials design in language teacher preparation programs (e.g. Farr & O'Keefee 2011). And yet others focus on the need to develop more user-friendly tools to facilitate the inclusion of corpus-based activities in the classroom (e.g. Romer 2006). Nonetheless, Aijmer (2009) concludes, the impact of corpora on syllabus and materials design has not been nearly as dramatic as expected.

Despite these attempts, it is not difficult to see that, for many if not most practicing language teachers, the possibility of getting first-hand experience with corpus research is still somewhat remote. Mukherjee & Rohrback (2006) contend that, very often, "teachers are confronted with suggestions of corpus-based activities which […] are difficult (if not impossible) to put into practice" (p. 209). They believe that, in order for corpora to become part of a teacher's pedagogical repertoire, the use of such corpora must have "a surplus value within a given language-pedagogical framework" (p. 212). It is thus not sufficient to present teachers with general, abstract, and often decontextualized corpus descriptions, but to make corpus data relevant to teachers' concerns and to their teaching situations.

The need to offer specialized learner training in corpus linguistics in order to successfully implement pedagogical interventions in second language classrooms is often mentioned in the literature. However, something that is readily apparent when reading reports on ways in which corpora are used for instructional purposes is a narrow focus on the need to train learners. Very little can be found on the role that teachers play or should play in this process. For example, Bernardini (2004) mentions that students can use corpora to develop a 'researcher' attitude towards data, searching for information needed to complete a task, analyzing the results, choosing appropriate solutions, and adapting these to their needs. This is helpful, it is noted, because in this way students can move away from simply trusting the authority of the teacher on what an appropriate solution to a communicative task might be. Bernardini then proposes a 'learning as discovery' approach by which learners develop "capacities and competences so that their [corpus] searches become better focused, their interpretation of results more precise, their understanding of corpus use and their language awareness sharper" (p. 23). Teachers are seldom mentioned in her proposal, although she does say, in passing, that they might also benefit from this approach as they could draw on their own learning strategies and experience of difficulties to model discovery learning for students.

Flowerdew (2009) also mentions that more strategy training of both students and teachers is needed, but the emphasis remains on the learner rather than on teachers. Even though I agree that learner strategy is vital, I believe the role of the teacher has been neglected in discussions of pedagogical applications of corpus information. Therefore, when the learner corpus described below was first envisioned, special attention was paid to teacher involvement.

3. The ULCAE project: A case study

3.1 The local context

The ULCAE (UTEP's Learner Corpus of Academic English) project started in 2009 and the corpus is still being compiled within the context of the ESOL (English for Speakers of Other Languages) program in the Department of Languages and Linguistics at the University of Texas at El Paso (UTEP), a mid-size public university located along the US-Mexico border in Western Texas. The majority of the student population at UTEP is Hispanic, which reflects the university's mission as much as its location and the demographics of the region. The ESOL program at UTEP aims to help students develop their general academic English language competence, especially in reading and writing, so that they can develop the skills needed to succeed in the context of the university.

The great majority of students in this ESOL program (97%) are Spanish-speaking students from Mexico, and most graduate from high schools in their home country. They are not required to have a minimum level of English language proficiency to be admitted to the university as degree-seeking students since they have the opportunity to develop their second language skills in the ESOL program, before or while taking content courses taught in English. As with mainstream students at the university, ESOL students must take two-semesters of freshman-level composition as part of the core curriculum block in their degree plans.

At present, enrollment in the ESOL program fluctuates between 400 and 500 students per semester, distributed in five levels. Students' average age is 19 years old, and the proportion of male students is only slightly higher than that of female students. ESOL instructors are either full-time or part-time lecturers who hold an M.A. degree (mostly in language-related fields, but not necessarily in TESOL), as well as graduate teaching assistants (MA students in Linguistics). Most of the instructors have taught in the ESOL program for at least five years, and there is a low teacher turnover, so the teaching faculty is relatively stable, with the exception of teaching assistants who usually teach in the program for one year only. In a typical semester, about 30 sections of ESOL courses are offered, and these are taught by

about 10–12 instructors and 2–3 teaching assistants. Instructors teach one to five courses per semester, with a maximum of 15 credit hours per semester depending on the type of appointment they have, and class sizes typically range from 10 to 25 students, with an average of 19 students per class.

The ESOL program comprises five different levels, starting from a low-beginning, integrated four-skill course and ending with the two composition courses that are part of the core curriculum. The first three levels can be classified broadly as general English language courses, while levels four and five include more academically-oriented reading and writing courses. The sequence of courses and corresponding credits are shown in Figure 1.

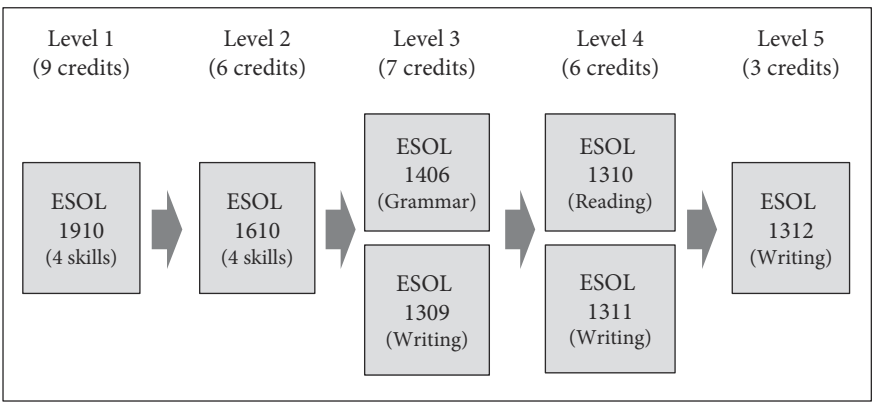

Figure 1. Sequence of courses in ESOL program

4. Rationale for the ULCAE project

The ESOL program at UTEP has the dual task of helping students develop their communicative competence in English as well as fulfill their first-year composition requirement. To do this, priority is given in the ESOL program to the development of literacy skills. Even though spoken English is an important component of the curriculum at the beginning levels, the curriculum centers on reading and writing for general academic purposes as students advance through the program. Given these objectives, ESOL students start writing essays at level two in the program, and continue doing so in subsequent levels. In their composition courses, students write different types of texts, from classification to argumentative essays and from analytical papers to research reports and they take a departmental final writing exam at the end of each course. And yet, despite the program's emphasis on writing, information regarding students' writing ability in English, at each level, could not be answered with certainty before the ULCAE project started. The

only information available was through indirect measures such as class grades or instructors' perception of students' knowledge and abilities. The learner corpus project described here aimed at providing more direct means of exploring and assessing students' writing development and composition skills.

Compiling a corpus, as it is well-known, constitutes a time-consuming and arduous process that involves much more than just collecting samples of language. A well-designed corpus requires a systematic process of data collection that pays attention to issues of size, balance, representativeness, and authenticity or 'naturalness' (Biber, Conrad & Reppen 1998). Building a learner corpus involves concerns similar to those in NS corpus building, although some additional variables need to be considered. For example, when collecting learner language, corpus compilers should try to include learner data representing the various levels of proficiency of the target population, as well as a representative sample of the different types of texts learners produce, which in turn should reflect the most typical communicative situations in which those texts are generated.

In order to build a representative learner corpus of ESOL writing, and one that allowed the possibility of tracking students' development, it was necessary to collect samples of writing from cohorts of students in each level as well as from individual students across levels. Second language programs where students remain for an extended period of time, and where they take sequential courses that make possible to track development, are not very common or they are relatively small so that building a large corpus of students' writing could take a very long time. The ESOL program at UTEP constitutes one of those rare cases with a relatively large number of students enrolled in the program and where many of them take more than one course, often in consecutive semesters. In addition, the teaching faculty is relative stable, and there was a clear emphasis on second language writing development that made the compilation of a local learner corpus of general academic written English a feasible enterprise.

5. Designing and building the corpus

The ULCAE corpus was designed, after much deliberation and discussion (see below), to include information from general English and academic writing courses from levels 2 to 5 in the ESOL program. Students' texts representing major writing assignments are collected, including drafts and final versions, as well as the final exam essays produced at the end of each semester (see Figure 2). Within each course, texts produced in response to at least three major assignments are collected. Usually, students work on these assignments throughout the semester under the guidance of their instructor.

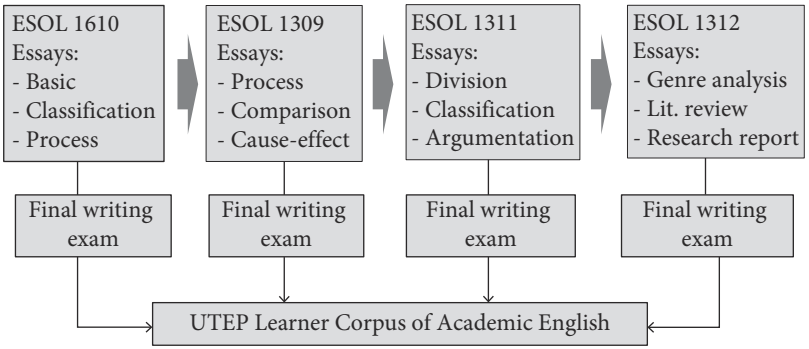

Figure 2. Design of ULCAE and types of ESOL texts collected[1]

Texts from a given class are considered to represent cross-sectional data in the ULCAE corpus, even though they are collected at different intervals during the semester. One reason for this is that there is no fixed timeframe to work on essays as each instructor can decide how much time to spend in any individual assignment. Secondly, even though it can be assumed that students' writing might change from assignment to assignment within one semester, one of the goals of building the corpus was to provide information to describe what students can and cannot do across levels in the program. Consequently, tracking change over time constituted a major goal for the project and, from the onset, the ULCAE corpus was designed to include quasi-longitudinal and longitudinal data.

Quasi-longitudinal data refers to data collected at a single point in time but from learners at different levels of proficiency; for example, comparisons of data from first and third-year students to determine progress or lack of it (Granger 2002, 2004).[2] As the ULCAE corpus includes data from sequential writing courses (from ESOL 1610 to ESOL 1312), these data can be considered as representing quasi-longitudinal data. Furthermore, as some students in the program are initially placed in the lower ESOL levels, and as they are required to move through the established sequence of courses that culminates with the second semester of

1. The writing assignments in ESOL 1311 were changed after the ULCAE project started. The syllabus was modified to reflect a more genre-based approach to teaching and new writing tasks involving problem-based writing were included.

2. Quasi-longitudinal data, which are somewhat easier to collect (in contrast to longitudinal data), can be used to suggest possible patterns of language development, as the information is assumed to reflect learners' language at different levels in a proficiency continuum. Any patterns determined with this type of data can then be checked using longitudinal corpus data in which the progress of individual students can be tracked (Granger 2004)

freshman-level ESOL composition, it was possible to collect longitudinal data, that is, texts from individual students who move from course to course in a sequential manner. Therefore, the ULCAE corpus also includes texts generated from individual learners during a period of two or more semesters.

When proposing the project to the Institutional Review Board for approval, permission was secured to collect data from students at multiple points during each semester and across semesters, in essence during the time a participating student is enrolled in ESOL courses. This was extremely important in order to collect longitudinal data. Thus, consent forms included such provision, although students could stop their participation in the project at any time and request that their texts be excluded from the corpus. As it is customary in corpus building, any identifying information included in the essays submitted by students was deleted.[3]

6. Teacher involvement

Teachers are naturally interested in the language used by their own students in their own classrooms and in finding ways to help them, so a local, context-specific learner corpus would be amiss if it didn't attempt to involve teachers in the project. Moreover, building a learner corpus in the context of a particular English language program would be extremely difficult without the cooperation and collaboration of course teachers. Even if it were possible to implement a data collection procedure directly from students, it would make little sense from a pedagogical point of view to bypass or exclude instructors from the process, especially if one of the goals for building the corpus is to find ways of applying corpus information in those instructors' classrooms. To this end, Mukherjee and Rohrbach (2006) have argued that it is as important to rethink language pedagogy from a corpus perspective as it is to rethink corpus linguistics from a language pedagogical perspective. Corpus linguists can thus be expected to take into account the needs, views, experiences, and working conditions of language teachers as much as language teachers are expected to familiarize themselves with what corpus-based information and materials can offer them.

In the case of the ULCAE corpus, given the goals and objectives of the whole enterprise, teacher involvement became a priority from the beginning of the project. However, it had to be acknowledged that most teachers knew very little about

3. During the first year of data collection, the project received an intramural University Research Institute (URI) grant to hire undergraduate students to help with the process of data collection and text processing.

corpus linguistics or about corpora and its uses and applications before the ULCAE project started. Therefore, the first steps involved providing teachers, graduate students, and teaching assistants with some basic understanding of the field, particularly in terms of its relation to classroom pedagogy. To this end, a series of activities were conducted during the first and second year of the project, starting with information sessions and moving on to hands-on teacher workshops.

6.1 Information and planning sessions for and with teachers

6.1.1 *Information sessions*

In these sessions, the field of corpus linguistics was broadly introduced (goals, aims, methodology) along with some description of studies relevant to the ULCAE project and English language teaching and learning. Material from various books and articles was used to introduce basic information about corpus linguistics to ESOL teachers, as well as to illustrate studies relevant to the ULCAE corpus (e.g. Biber & Conrad 2001; Biber & Reppen 2002; Fan 2009; Guilquin & Paquot 2008; Hinkel 2003; O'Keefe, McCarthy & Carter 2007). During the first session, teachers were also presented with some questions about language that could be appropriately answered on the basis of corpus data (e.g. most frequent verbs in spoken versus written English, most frequent vocabulary in general English versus academic English), followed by a brief discussion about the roles of speaker intuition and corpus data in descriptions of language use. The second session focused on differences between NS and NNS corpora, along with a brief description of existing learner corpora, their most common characteristics, as well as examples of typical questions that can be explored using learner corpora.

6.1.2 *Planning sessions*

These sessions focused on designing, discussing, and socializing a preliminary plan for building the ULCAE project. Teachers were presented with an outline of the project, including a possible design, as well as a proposal for the data-collection procedure. The session emphasized general goals for building the project and the unique characteristics of the proposed learner corpus. In addition, various aspects of the project were presented and discussed, such as courses to be targeted, role of teachers in the process, types of written assignments to be collected, and a possible procedure for collecting texts and including them in the corpus. Important goals for these sessions were to respond to teachers' questions and concerns as well as incorporate their suggestions into the proposal.[4]

4. During these sessions, for instance, the adequacy of including texts from assignments such as journal writing in the corpus (from ESOL 1309 and ESOL 1311) was discussed. Teachers

6.1.3 *Research-oriented sessions*

Issues related to learners' privacy and the confidential nature of research data were also presented and discussed so that instructors would have a clear idea not only of the type of data to be collected, but also the various ways in which students' privacy and confidentiality would be protected. The role of IRB committees in research was briefly explained, and drafts of consent forms were presented and discussed. The role of student-assistants in the project was also described. Also included in these sessions were discussions on the role of teachers in classroom research and the need to connect research and pedagogy, with teachers' comments, questions, and concerns addressed along the way.

6.2 Hands-on corpus-oriented workshops

After the aforementioned sessions, and once IRB approval had been secured, the initial data collection process began. During the first years of data collection, workshops for teachers and teaching assistants were conducted so that they could familiarize themselves with both the way texts are processed and stored electronically as well as to get acquainted with basic concordance software used to explore corpus data.

6.2.1 *Text processing*

The first corpus-oriented workshops focused on describing the path that a text or essay would follow, from being submitted by the student to deleting any identifying information, and from the characteristics of text-only files to the way file names were used to encode information about each file or text. Examples from the initial texts collected for the corpus were shown and the processing of handwritten final exam essays was also discussed.

6.2.2 *Concordance software*

The second workshop was devoted to learning some of the basic features of the concordance program 'MonoConc Pro' (Barlow 2004). Again, using some of the texts already included into the ULCAE, teachers were taught how to generate frequency lists and how to search for specific words and phrases. In pairs, teachers were given basic tasks so that they could search for and analyze some common lexical and grammatical items, especially items that represent typical areas of difficulty for low and high intermediate Spanish-speaking learners of English (e.g. using *people* as

---

commented that such assignments should not be included as there was little standardization across sections of the same class in terms of the requirements for such texts. Thus, journal entries were not included in the corpus.

a plural noun, choice of *job* vs. *work*, *do* vs. *make*). Teachers were able to compare the use and frequency of these items between groups of students at two levels of proficiency, as well as explore concordance lines vertically and horizontally. In addition, they were introduced to the notion of normed frequencies, and asked to generate these to compare corpus data from different groups and levels.

6.2.3 *Formulating research questions*

Finally, after the two workshops described above, teachers were asked to formulate questions that could be investigated using learner corpus data. They were encourage to think about their students, areas of difficulty, and aspects of language use that they believe characterize students' writing at different levels in the program. Even though this was the area that proved most challenging for teachers, it provided the basis for reflecting about their students' writing and possible ways to analyze ULCAE data that could inform their knowledge of learner language in their classrooms.

6.3 Project progress reports

A third type of sessions conducted with teachers focused on providing them with progress reports on the ULCAE project. These progress reports served to show how the corpus was growing, to highlight gaps in it, i.e. areas where more data was needed, as well as to thank teachers for their contributions. The sessions were also used to check if the data collection procedure in place was working effectively, and to encourage the electronic submission of texts via the university's web-based course management system in order to avoid having to scan documents.

7. Benefits of promoting teacher involvement

Getting teachers involved in a learner corpus project provides an excellent opportunity to build the type of teacher-researcher partnerships that are commonly encouraged in classroom research methodology books but that, in reality, do not occur as often as they should. In addition, when it comes to building a learner corpus within the context of a specific English language program, teacher involvement becomes not only commendable but crucial to the success of the project. There are various reasons for this. To begin with, teachers can participate more actively in discussions of the project and contribute to the process of data collection knowing why the corpus is being built and what the goals of the project are. This can also promote a sense of ownership that would benefit not only the project but also the way teachers think about research within their institutional units. Secondly, when

teachers are knowledgeable about the project (its characteristics, aims, progress) they are in a better position to address learners' questions and concerns and to become advocates of the project, rather than mere intermediaries. Furthermore, if teachers are involved in the project from the onset, they have a better sense of its scope and the fact that it would be a long process that is likely to take several years. Without a good understanding of the project, teachers might question why texts keep being collected semester after semester and year after year. Finally, teachers who get actively involved in research projects are more likely to participate in other research and scholarly activities, as they can see themselves as being members of a research community perceived as less exclusive.

7.1 Teachers' roles and levels of involvement

In addition to the aforementioned benefits, there are other important and specific reasons for getting teachers involved in a learner corpus project. First, teachers can point towards areas that can benefit from empirical investigation or identify pressing instructional concerns and thus suggest possible areas of inquiry. In addition, they can also help researchers interpret results from corpus-based studies in light of what they know about their students, the instructional materials used, and their own classroom practices, i.e. interpret results on the basis of contextual factors. Finally, teacher input can also be very valuable when making curriculum decisions based on results from analyses of their own learners' corpus data. These areas will be illustrated next using some of the data first analyzed as part of the ULCAE project: patterns of pronominal choice.

7.2 Defining areas to explore

The issue of how ESOL students use pronouns in their writing emerged in various discussions with teachers regarding the adoption of a first-person authorial stance in academic writing. A review of the literature revealed opposing views about this issue, and teachers in the program also had differing opinions about it. In addition, some teachers commented that they were unsure about the feedback they provide to students regarding the use of 'I' in comparison to other alternatives, e.g. using 'we' or avoiding first-person mentions altogether. Finally, some teachers commented on the students' common strategy of addressing the reader directly in their essays by means of the second person pronoun.

A preliminary analysis of first and second subject person pronouns in ESOL students' writing at different levels of proficiency (Mendoza & Martínez 2011) revealed great differences in the frequency of occurrence of 'I', 'we,' and 'you' in students' texts (Table 1).

Table 1. Frequencies (per 1000 words) of I, WE, and YOU in course essays

Course	# essays	# of words	I	WE	YOU
ESOL 1309	362	178,298	5.00	3.34	29.95
ESOL 1311	345	253,565	3.29	6.10	8.63
ESOL 1312	279	285,503	2.02	3.87	2.38

As Table 1 shows, students in ESOL 1309 tend to use the second person subject pronoun in their essays with extremely high frequency, almost 30 times per 1000 words, in comparisons to pronouns 'I' and 'we' which are used 5 and 3.3 times per 1000 words, respectively. In contrast, in the next writing courses, ESOL 1311 and ESOL 1312, the frequency of occurrence and distribution of the three pronouns is not as high or disparate.

Issues of pronominal choice were subsequently investigated in a more in-depth manner in a separate study that also included a qualitative pragmatic analysis of the relationship between subject pronouns and students' self-positioning strategies across texts types and courses (Urzúa 2013).[5] Such analyses indicated that the preference for one pronoun over another was not tied merely to the students' level in the program, an indirect measure of language proficiency, but also to particular patterns of self-positioning in each specific text type at each level. In ESOL 1309, for example, a great deal of variation in the use of first and second person subject pronouns was found across different writing assignments (Table 2).

Table 2. Frequencies (per 1000 words) of I, WE, and YOU in ESOL 1309 essays

Essay	# essays	# of words	I	WE	YOU
Process	49	24,693	0.9	1.9	43.1
Comparison	45	24,504	4.0	5.3	13.3
Cause-effect	16	8,623	3.7	3.6	20.8
Total	110	58,090	2.6	3.6	27.1

As with the overall usage of pronouns, frequencies of occurrence in subsequent courses, i.e. ESOL 1311 and ESOL 1312, were quite different from those

5. In this chapter, only some of the frequency data generated from the analyses of quasi-longitudinal and longitudinal data conducted in Urzúa (2013) is presented. The reader is referred to the original paper for more information on the qualitative, pragmatic analysis and the discussion of corresponding results.

found in the writing of students in ESOL 1309, but internal variation across text types was also evident (Tables 3 and 4).

Table 3. Frequencies (per 1000 words) of I, WE, and YOU in ESOL 1311 essays

Essay	# essays	# of words	I	WE	YOU
Division	41	36,259	2.9	4.3	5.0
Classification	34	29,070	0.3	3.1	8.8
Argumentation	13	13,180	0.1	4.9	8.0
Total	88	78,509	1.4	4.0	6.9

Table 4. Frequencies (per 1000 words) of I, WE, and YOU in ESOL 1312 essays

Essay	# essays	# of words	I	WE	YOU
Genre analysis	31	33,470	0.8	2.4	1.6
Lit. review	23	36,137	0.3	3.0	0.5
Research reports	19	35,819	3.0	4.0	1.6
Total	73	105,426	1.3	3.1	1.2

In essence, texts produced in the context of ESOL 1309, the first writing intensive course in the program, show a very high frequency of occurrence of the pronoun 'you' in comparison to 'I' and 'we,' but this situation is strikingly different in the writing students do in ESOL 1312, the third writing-intensive course in the program, in which the frequency of occurrence of all pronouns decreases substantially and where the pronoun 'we' becomes the preferred choice.

An important advantage of a localized, context-specific learner corpus such as the ULCAE is the possibility of looking at these changes using longitudinal data. Even though not all students in the data so far discussed (Tables 1 to 4) progressed from course to course without interruption, data from the ULCAE project allowed the identification of individual students whose progress could be tracked from one class to the next, in the same academic year (Urzúa 2013).

The corresponding frequency analysis of first and second person pronouns in a subset of these longitudinal data shows that the overall pattern determined for intact groups of students at each level occurs also in texts composed by individual students, as shown in Tables 5 and 6. [6]

6. In the original study, the analyses of longitudinal data included a close reading of each one of the texts generated by individual students in a period of two consecutive semesters. Therefore, only a small group of students was included in such analysis. Individual students

Table 5. Frequencies (per 1000 words) of I, WE, and YOU in individual students' texts from ESOL 1309 to ESOL 1311 (2009–2010)

Student	ESOL 1309 – Fall 2009				ESOL 1311 – Spring 2010			
	# of words	I	WE	YOU	# of words	I	WE	YOU
Dora	1,625	0.06	0.18	3.13	1,538	0.19	0.45	0.84
Mario	1,739	0.00	0.23	1.78	2,172	0.00	0.36	0.18
Francisco	1,460	0.13	0.20	4.17	1,546	0.25	1.81	0.38
Ana	2,465	3.0	0.29	2.31	1,953	0.10	0.87	2.56

Table 6. Frequencies (per 1000 words) of I, YOU, and WE in individual students' texts from ESOL 1311 to ESOL 1312 (2009–2010)

Student	ESOL 1311 – Fall 2009				ESOL 1312 – Spring 2010			
	# of words	I	WE	YOU	# of words	I	WE	YOU
Mateo	2,156	0.18	0.23	1.71	4,557	0.00	0.13	0.02
Laura	2,631	0.00	0.76	1.02	4,579	0.00	0.30	0.02
Elena	2,562	0.00	0.50	0.93	3,678	0.00	1.14	0.54
Pedro	2,360	0.25	0.97	0.04	3,961	0.30	0.12	0.42

The patterns of pronominal choice uncovered by the quasi-longitudinal analyses were confirmed by the analysis of longitudinal data. In additional, the qualitative exploration of the texts generated by individual students provided specific examples of how students change their use of subject pronoun over time. For instance, excerpts (1) and (2) were both written by the same student (Francisco). The first one is found in his comparison-contrast essay (written in ESOL 1309), while the second one comes from his argumentative essay (written the following semester in ESOL 1311).

(1) First let me tell **you** about the engine. The gasoline vehicles have a combustion engine which uses unleaded fuel, while the hybrid car has two engines: an electric and a gasoline engine. The electric engine runs on battery power. The batteries store energy to move the car. The gasoline engine is used as the last option. In the hybrid cars **you** can choose which engine **you** want to use.

were selected randomly from those who had generated at least five major papers in one academic year. Student names in Tables 5 and 6 are pseudonyms.

(2) The study show us that ASARCO air emissions decrease the life quality of the people who lives near the company, provoking a lot of illnesses, like multiple sclerosis, affecting the central nervous system. **I** consider that the re-open of the ASARCO Company is a big mistake; **we** can notice that it has a bunch of dangerous effects in the people, especially in children.

Excerpts (1) and (2) illustrate specific self-positioning strategies used by Francisco. In ESOL 1309, he often addressed the reader explicitly, establishing a closer and more personal connection between writer and reader, while in ESOL he tended to write in a more impersonal way and, when subject pronouns were used, he showed a clear preference for first person pronouns. As a result, Francisco's stance as a writer became more distanced while at the same time he strategically manages to align himself with the reader in order to be more convincing as the text author. Thus, in addition to confirming the overall pattern, the analyses of the longitudinal data were helpful to uncover pragmatic strategies related to pronominal choice and tied to particular modes of self-presentation and reader-writing relationships.[7]

7.3 Interpreting corpus-based information

Being able to interpret learner corpus data in light of the learners' instructional program and based on local teachers' curriculum content knowledge and their experiences in the classroom represents one of the most important advantages of context-specific learner corpora. In the case of the aforementioned patterns of pronominal usage among ESOL students, when teachers were presented with such information, they suggested a number of possible intervening factors that might explain the results. To begin with, teachers confirmed that there was a lack of clear and explicit information in textbooks regarding the use of personal pronouns in college-level academic writing, and that they often hesitated about whether to focus their corrective feedback on these elements, particularly in ESOL 1309, as it seemed to them that, at this point in students' language development, there are other language features that demanded more of their attention, both grammatical (e.g. verb tenses, subject-verb agreement, word forms) and rhetorical (e.g. paragraph development, organization of ideas, supporting evidence). Furthermore, one teacher pointed out that a couple of sample essays included in the textbook used in ESOL 1309 are composed with explicit reference to readers by means of the second person subject pronouns, and thus it seemed to her that addressing

7. A more extended discussion of these patterns and strategies, including possible interpretations and implications, is presented in Urzúa (2013).

the reader using the pronoun 'you' should be allowed. On the other hand, even though there may be legitimate uses of second person subject pronouns in some types of general academic writing found in second language textbooks, such as 'process' essays, studies on pronoun usage in academic writing indicate that 'you' is not commonly found in professional academic writing. Instead, when writers explicitly insert themselves into their texts, they tend to use first person pronouns (Kuo 1999; Hyland 2001; Tang & John 1999).

ESOL 1311 and ESOL 1312 teachers expressed that they paid more attention to pronominal choices in these courses (especially in the feedback they provided to students during the drafting process) and discouraged students from using the pronoun 'you,' a trend that can be seen in the data presented in Tables 3 to 6. Moreover, they also reported putting more emphasis on the importance of an impersonal, objective tone in academic writing, particularly in the more research-oriented texts composed in ESOL 1312, e.g. literature reviews, research proposals and reports. However, many novice writers feel uncomfortable using the latter because of the wide-spread notion that academic writing is impersonal and thus such forms should be avoided, or because their use denotes a position of authority that novice writers may not want to adopt (Hyland 2002). In addition, some teachers commented that, even if they allowed some first-person pronouns in their students' texts, when deemed appropriate, their students often reported that tutors in the University Writing Center strongly advise against using such pronouns, and thus these ESOL instructors felt it was important to provide a consistent message to students.

7.4 Evaluating the curriculum

Discussions about students' pronominal choices held with teachers also provided an excellent opportunity to share information from relevant corpus studies on the topic. This information helped instructors compare results from previous studies (e.g. Hyland 2001; Kuo 1999; Tang & John 1999) with those yielded by the analysis of their own students' patterns of pronominal choice and usage. In turn, these conversations were helpful in discussing the type of essays that students write in each course, and whether these assignments were the most appropriate ones or not. Obviously, changes to the curriculum cannot be based on a sole aspect of language use, but the conversations regarding the use of pronouns led to further questions as to the adequacy of asking students to write certain essays if these reinforced aspects of language use that may not reflect some of the more conventional characteristics of academic genres. In the case of the ESOL program at UTEP, the discussion of whether or not the 'process' essays should be kept in ESOL 1309, especially since it was an essay that students were asked to write in

more than one course, led to a proposal to consider eliminating the 'process' essay from the syllabus.

More importantly, discussions on the role of authorial self-positioning in students' academic writing provided teachers with an opportunity to reflect on their own teaching and feedback practices, as well as expand their understanding of their students writing, based on empirical data yielded by the ULCAE and not only on intuitive notion of what students did or did not do in each course. Furthermore, this reflective experience contributed to underscore the value of building a corpus comprised of one's own student texts.

8. Final remarks

Based on the case just presented, I believe that teacher involvement in a learner corpus project not only facilitates the process of compiling a corpus but can also lead to building a more solid foundation for, eventually, establishing more concrete, direct, and teacher-led pedagogical applications of corpus-based information. In addition, teacher involvement can enrich the design of the corpus and facilitate the process of data collection. Teachers can also share with researchers their perceptions of what students can and cannot do at different levels of proficiency, problems areas, typical errors, challenges, and so forth, and in this way help identify or suggest potential areas on inquiry, particularly those that reflect their own concerns about their students' language development and usage.

Another important aspect that can benefit from increased teacher involvement is the interpretation of results yielded by local corpus-based investigations. Descriptions of language use are most useful when they can be interpreted in light of contextual factors, and teachers are in an ideal position to contextualize corpus data. Researchers and teachers working together can analyze results from a local learner corpus in light of the curriculum, materials used, student behavior, testing conditions, instructional foci, teaching activities, time spent on specific areas, sequence of presentation of linguistic information, and many other situational factors. The analyst can thus "act as a kind of mediating ethnographic specialist informant to shed light on the corpus data" (Flowerdew 2005: 329). Finally, another area in which teachers can participate, together with corpus researchers and analysts, is in discussions of syllabi and curriculum changes based on corpus-based information, as illustrated above.

In sum, teachers can define, perhaps better than any other group of people, the areas whose investigation are most likely to impact both the way they think about their students' linguistic competencies and skills as well as the way they view possible pedagogical interventions. It has been noted that most language

programs are not 'corpus-oriented' and thus teachers and tutors are not always aware of the benefits of building and using a learner corpus (Granger et al. 2007). A corpus-oriented program can greatly contribute to bridge the gap between teachers and corpus researchers. However, it must also be said that most language teachers in university-based language programs, especially part-time instructors and adjuncts, tend to have heavy teaching loads and many teach multiple courses at various levels of proficiency, not always under the best circumstances. Many are overwhelmed and underpaid, so it is understandable that they may not have the time or the disposition to actively participate in corpus projects or any other research studies.

To the extent that teachers can get involved in research and corpus activities such as those described above, they can become more actively involved in making decisions that affect them, their students, and their own professional development. A language program that becomes corpus-oriented, one in which teachers can examine corpus data to understand their students' language development and the complexities involved in this process, as well as to reflect on their own teaching practices, it is also a program in which teachers are in a better position to see research from a less distanced perspective, to conduct research for their own purposes, to gain new appreciation of their role as teachers, to participate in disciplinary conversations and, ultimately, to grow as professionals in the field.

References

Aijmer, Karin. 2009. *Corpora and Language Teaching* [Studies in Corpus Linguistics 33]. Amsterdam: John Benjamins. DOI: 10.1075/scl.33

Barlow, Michael. 2004. *MonoConc Pro*. version 2.2. Houston TX: Athelstan.

Bernardini, Silvia. 2004. Corpora in the classroom: An overview and some reflections of future developments. In Sinclair (ed.), 15–36.

Biber, Douglas & Conrad, Susan. 2001. Quantitative corpus-based research: Much more than bean counting. *TESOL Quarterly* 35(2): 331–336. DOI: 10.2307/3587653

Biber, Douglas & Reppen, Randi. 2002. What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition* 24: 199–208. DOI: 10.1017/S0272263102002048

Biber, Douglas, Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP. DOI: 10.1017/CBO9780511804489

Cook, Guy. 1998. The uses of reality: A reply to Ronald Carter. *ELT Journal* 52(1): 57–63. DOI: 10.1093/elt/52.1.57

Fan, May. 2009. An exploratory study of collocational use by ESL students: A task based approach. *System* 37: 110–123. DOI: 10.1016/j.system.2008.06.004

Farr, Fiona. 2010. *The Discourse of Teaching Practice Feedback: A Corpus-based Investigation of Spoken and Written Modes*. London: Routledge.

Farr, Fiona & O'Keefee, Anne (eds). 2011. *International Journal of Corpus Linguistics* 16(3). Special issue on *Teacher Education*. DOI: 10.1075/ijcl.16.3.01far

Flowerdew, Lynne. 2005. An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: Countering criticisms against corpus-based methodologies. *English for Specific Purposes* 24: 321–332. DOI: 10.1016/j.esp.2004.09.002

Flowerdew, Lynne. 2009. Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics* 14(3): 393–417. DOI: 10.1075/ijcl.14.3.05flo

Flowerdew, Lynne. 2012. *Corpora and Language Education*. Houndmills: Palgrave-Macmillan.

Frankenberg-Garcia, Ana. 2010. Raising teachers' awareness of corpora. *Language Teaching* 45 (4): 475–489. DOI: 10.1017/S0261444810000480

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* [Language Learning & Language Teaching 6], Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds), 3–33. Amsterdam: John Benjamins. DOI: 10.1075/lllt.6.04gra

Granger, Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In *Language and Computers: A Multidimensional Perspective*. Ulla Connor & Thomas A. Upton (eds), 123–145. Amsterdam: Rodopi.

Granger, Sylviane, Kraif, Olivier, Ponton, Claude, Antoniadis, Georges & Zampa, Virginie. 2007. Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness. *ReCALL* 19(3): 252–268. DOI: 10.1017/S0958344007000237

Guilquin, Gaëtanelle & Paquot, Magali. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1(1): 41–61. DOI: 10.1075/etc.1.1.05gil

Hasko, Victoria. 2013. Capturing the dynamics of second language development via learner corpus research: A very long engagement. *The Modern Language Journal* 97: 1–10. DOI: 10.1111/j.1540-4781.2012.01425.x

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

Hyland, Ken. 2001. 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. 2002. Authority and invisibility: Authorial identity in academic writing. *Journal of Pragmatics* 34: 1091–1112. DOI: 10.1016/S0378-2166(02)00035-8

Kaltenböck, Gunther & Mehlmauer-Larcher, Barbara. 2005. Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL* 17(1): 65–84. DOI: 10.1017/S0958344005000613

Kuo, Chih-Hua. 1999. The use of personal pronouns: Role relationships in scientific journal articles. *English for Specific Purposes* 18(2): 121–138. DOI: 10.1016/S0889-4906(97)00058-6

Lantolf, James P. 2000. *Sociocultural Theory and Second Language Learning*. Oxford: OUP.

McCarthy, Michael. 2008. Accessing and interpreting corpus information in the teacher education context. *Language Teaching* 41(4): 563–574. DOI: 10.1017/S0261444808005247

Mauranen, Anna. 2004. Spoken corpus for an ordinary learner. In Sinclair (ed.), 89–105.

Mendoza, Laura E. & Martínez, Mónica. 2011. 'How miners pick their pronouns.' Pronoun usage in general academic writing among second language learners. Talk presented at the Graduate Research and Arts Symposium, New Mexico State University, March 1.

McEnery, Tony, Xiao, Richard & Tono, Yukio. 2006. *Corpus-based Language Studies*. New York NY: Routledge.

Mukherjee, Joybrato & Rohrbach, Jan-Marc. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In *Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop*, Bernhard Ketteman & Georg Marki (eds), 205–232. Frankfurt: Peter Lang.

Nesselhauf, Nadja. 2004. Learner corpora and their potential for language teaching. In Sinclair (ed.), 125–152.

O'Keeffe, Anne & Farr, Fiona. 2003. Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly* 31(3): 389–418. DOI: 10.2307/3588397

O'Keefe, Anne, McCarthy, Michael & Carter, Ronald. 2007. *From Corpus to Classroom*. Cambridge: CUP. DOI: 10.1017/CBO9780511497650

Reppen, Randi. 2010. *Using Corpora in the Language Classroom*. New York NY: CUP.

Römer, Ute. 2006. Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift Anglistik und Americanstik* 54(2): 121–134.

Römer, Ute. 2009. Corpus research and practice: What help do teachers need and what can we offer? In Aijmer (ed.), 83–98.

Sinclair, John. 2004. *How to Use Corpora in Language Teaching* [Studies in Corpus Linguistics 12]. Amsterdam: John Benjamins. DOI: 10.1075/scl.12

Tang, Ramona & John, Suganthi. 1999. The 'I' in identity": Exploring writer identity in student academic writing through the first person pronoun. *English for Specific Purposes* 18: 23–39. DOI: 10.1016/S0889-4906(99)00009-5

Tsui, Amy B.M. 2004. What teachers have always wanted to know – and how corpora can help. In Sinclair (ed.), 39–61.

Urzúa, Alfredo. 2013. Pronominal choice and self-positioning strategies in second language academic writing: A pragmatic analysis using learner corpus data. In *Technology in Interlanguage Pragmatics Research and Teaching* [Language Learning & Language Teaching 36], Naoko Taguchi & Julie M. Sykes (eds), 121–152. Amsterdam: John Benjamins. DOI: 10.1075/lllt.36.07urz

Widdowson, Henry. 2000. On the limitations of linguistics applied. *Applied Linguistics* 21(1): 3–25. DOI: 10.1093/applin/21.1.3

CHAPTER 6

The challenge of constructing a reliable word list: An exploratory corpus-based analysis of lexical variability in introductory Psychology textbooks

Don Miller
California State University, Stanislaus

This study highlights the methodological challenges inherent in reliably capturing meaningful sets of vocabulary for instructional focus. An analysis of a 3.1 million-word corpus of introductory psychology textbooks suggests that, while comparatively large, and, thus, presumably representative of the lexical variability in the target domain, this corpus was unable to capture a stable list of "important" words. Findings highlight an important issue requiring further investigation in corpus-based vocabulary research: the extent to which corpora – and the word lists based on them – reliably represent the lexical variability of their target domains.

Keywords: Corpus representativeness; lexical diversity; word list reliability

1. Introduction

For well over half of a century, researchers have expended considerable time and energy in pursuit of lists of "important" vocabulary – vocabulary that is frequently and widely used in English – in order to help learners, teachers, and materials developers focus and maximize the efforts of language learning. This robust tradition of word list development research has produced some extremely influential lists. Perhaps the most widely known and commonly used of these lists is the General Service List (GSL) (West 1953) of 2,000 word families, which accounts for approximately 80% word coverage of most texts (Nation 2001). More recent studies have sought to identify lists of important vocabulary for more narrowly defined domains, particularly with regard to language at varying levels of specificity within academic English (e.g. University Word List, or UWL; Xue & Nation 1984; the Academic Word List or AWL; Coxhead 2000).

DOI 10.1075/scl.66.06mil

Even more recent studies have in fact begun to question the efficacy of a single, "one-size-fits-all" list of general academic vocabulary (e.g. Hyland & Tse 2007). Noting some inconsistencies in coverage provided by the AWL across disciplines, as well as some specialized uses of vocabulary across disciplines, these studies have led researchers to make a case for more targeted, discipline-specific vocabulary lists. These researchers have proposed changes reflecting disciplinary variation, ranging from the need for modest modifications to the AWL, such as removing items from and/or adding items to the AWL (e.g. Martinez et al. 2009; Chen & Ge 2007), to the need for completely new discipline-specific lists (e.g. for public health: Millar & Budgell 2008, for medicine: Wang, Liang & Ge 2008).

Designing word lists makes intuitive sense: doing so can help focus program curricula or individual efforts toward those lexical items that students will encounter most often in their target use domain, thus, presumably, increasing the return on their efforts. Such was the expressed goal of Coxhead's (2000) research which resulted in the AWL. Noting that among "the most challenging aspects of vocabulary learning and teaching in English for academic purposes (EAP) programs is making principled decisions about which words are worth focusing on during valuable class and independent study time" (Coxhead 2000: 312), she proposed that her word list "might be used to set vocabulary goals for EAP courses, construct relevant teaching materials, and help students focus on useful vocabulary items" (p. 227). Additionally, she expressed the hope that "authors will undertake to write […course books specifically designed to teach academic vocabulary…] based on the AWL" (ibid.). Indeed, the AWL has since figured prominently in EAP syllabi and popular published teaching materials. Many course books have been entirely based on her list (e.g. Burgmeier & Zimmerman 2007; Huntley 2005; Schmitt & Schmitt 2005) or have drawn significantly from this list to inform the vocabulary component of instruction (e.g. Upton 2004).

Without question, word list research has allowed for great strides in our understanding of "important" vocabulary for a variety of purposes, and, as noted, findings have been applied directly to the development of curricula and instructional materials. Teachers and learners no doubt take comfort in – and have benefited from – the empirical basis for their increasingly focused efforts. Simply stated, a great deal of faith has been placed in word lists. However, it may be time to consider whether this faith should be tempered until practitioners have more thoroughly examined the extent to which corpora are able to capture a stable representation of lexical variability in their target domain.

A core assumption in the design of word lists has been that they are based upon truly representative corpora. In the many studies that have produced word lists, researchers describe – often in great detail – their attempt to design corpora which "mirror the experience of" eventual list users (Schmitt 2010). For example, they often note the range of topics (e.g. disciplines, sub-disciplines) and text categories

(e.g. genres, registers, text types) encountered in a target domain, and then demonstrate how the corpus composition reflects these characteristics. In addition, they often evidence the quality of the sampled texts (e.g. source, relevance) and the size of the corpus (e.g. number of texts included, number of total running words). The corpus that West (1953) based the GSL on contained five million words and included a wide variety of texts from a variety of topics and registers, so he felt it represented general English. A more recent undertaking to identify words which are frequent across a range of texts and topics in a general English corpus (i.e. general English vocabulary) is Nation's (2004) analysis of the 100,000,000-word British National Corpus (BNC), a corpus designed to be a more accurate representation of contemporary spoken and written English in proportions reflecting "their representation of everyday English use" (Leech et al. 2001: 1). An even more recent undertaking by Davies & Gardner has produced *A Frequency Dictionary of Contemporary American English*, which is a list that they describe as "the 5,000 most frequently used words[1] in the language" (2010: cover) This list was based on the Corpus of Contemporary American English (COCA), an impressive corpus of approximately 450 million words described by Davies as "the first reliable monitor corpus of English" (2010: 447).

With regard to more specialized corpora, Coxhead's (2000) Academic Corpus contained 3.5 million words from 414 texts from 28 different disciplines, so she felt that it, and the resulting AWL, represented academic writing. Hyland and Tse's Academic Corpus included more contemporary texts, included student writing, and, according to the researchers, "more systematically represent[ed] a range of key genres in several fields" (2007: 239), thus even better representing academic writing. Wang, Liang, and Ge's (2008) corpus of just one narrow genre, medical research articles, contained 218 complete research articles, an equal number from each of 32 different medical fields. Because of the careful, principled design of their corpus, they argued that it provided an accurate representation of the vocabulary their target domain.

What is extremely surprising, however, is that, despite the tremendous amount of care, time, and thought that has been put into the design and compilation of corpora and the development of word lists, a critical question related to representativeness – of both corpora and word lists – has not been asked:

> *To what extent do the corpora upon which word lists are based – and indeed the word lists generated from them – reliably represent the lexical variability in their domains of interest?*

---

1. Unlike with many previously proposed word lists, Davies and Gardner used the lemma as their unit of analysis, so 5,000 "words" means 5,000 lemmas.

For teachers and learners, understandably, the mention of "corpus linguistics," "lexical diversity," and "lexical variability" may appear entirely theoretical, far removed from the actual practice of vocabulary teaching and learning. In the end, what teachers and learners want is simply a meaningful list of words that merit focus so that they can make the most efficient and productive use of their time. However, it is impossible to know whether instructional and learning time is being well spent without asking very fundamental questions about the source of these lists and the assumptions upon which they have been based.

The present study directly addresses the issue of whether additional analysis might be included in estimations of the degree to which our corpora – and word lists based on them – truly represent the lexical variability and distributions possible in a given target domain. Such analysis would add validity evidence to claims of corpus representativeness and, potentially, increase the reliability of word lists produced from these corpora.

2. Review of the literature

2.1 Corpus representativeness

> "A [word list] is only as good as the corpus it is based upon, and every corpus has limitations. Firstly, no corpus can truly mirror the experience of an individual person; rather it is hopefully representative of either the language across a range of contexts… or of a particular [domain] of language." (Schmitt 2010: 67)

Corpus linguistics manuals and methodological papers (e.g. Atkins et al. 1992; Biber 1993; Biber et al. 1998; Bowker & Pearson 2002; McEnery & Wilson 1996; McEnery et al. 2006) discuss a number of important considerations for the purpose of achieving representativeness through corpus design. Several of the most commonly noted considerations, including topic and register coverage and relevance of included texts, are discussed in the following sections.

2.1.1 *Domain topic coverage*

Most texts on corpus linguistics note the importance of "topic" or "subject" coverage in corpora upon which lexical studies are conducted (e.g. Bowker & Pearson 2002; Biber 1993; Biber et al. 1998). Clearly, this consideration is a key component of representativeness, as "subject matter is especially important for lexicographic studies, since the frequency of many words varies with the subject matter" (Biber et al. 1998: 248).

In corpus-based investigations of academic vocabulary, domain "topic" coverage is perhaps the most often noted evidence for representativeness, and it is

typically discussed in great detail. With academic corpora design, "topic" has been operationalized at varying levels and combinations of specificity, for example, macrodiscipline (e.g. science), discipline (e.g. biology) or subfields within a discipline (e.g. hematology, hepatology, oncology). A number of methods have been used as evidence for topic coverage. For example, both Coxhead and Hirsh (2007; a pilot science-specific wordlist) and Durrant (2009; a collocation list for general EAP) ensured topic coverage in their academic corpora by designing their corpora based on the disciplinary makeup of their schools. Mudraya (2006) culled a list of engineering lexis from materials representing the nine courses required of the target list users. Wang et al. (2008) constructed their list of academic medical words from a corpus based on a survey of medical subfields represented in a database of academic medical journals.

2.1.2 *Domain text type/register coverage*

Corpus linguistics manuals also note the importance of including the range of text categories (i.e. genres, registers, text types) that are found in target domains (e.g. Bowker & Pearson 2002; Kennedy 1998; Sinclair 1991). Depending on the ultimate goal, a corpus designer may try to balance spoken and written texts or even varying types of spoken encounters or written texts. For example, for its spoken component, the designers of the BNC were careful to include both unscripted conversation (40%) and more formal, often pre-planned, "task-oriented" oral language such as lectures, sermons, and television or radio broadcasts (60%) (Leech et al 2001). Such considerations are also illustrated in the careful design of specialized corpora, including the Education Testing Service's TOEFL 2000 Corpus of Spoken and Written Language (T2K-SWAL; Biber et al. 2004), which selected register categories from "the range of spoken and written activities associated with academic life…" (p. 7), Hyland and Tse's academic corpus designed to represent "the range of sources students are often asked to read at university…" from the "main academic discourse genres…" (2007: 238–239), or the Hong Kong Financial Services Corpus (HKFSC), whose 25 text types were felt to represent "a comprehensive picture of the written discourse in the financial services industry in Hong Kong" (Li & Qian 2010). Clearly, researchers have striven to account for the important role of text/register type in representing a target domain.

2.1.3 *Quality/relevance of texts sampled*

The quality and relevance of sampled texts, that is, the degree to which included texts are actually encountered and/or commonly used in the target domain, is yet another corpus-design consideration typically noted. Once again, various methods have been used as evidence of text quality and relevance, from disciplinary expert provision or recommendation of texts (e.g. Coxhead & Hirsh 2007) to

claims regarding the reputation of the database (Wang et al. 2008) or journals (Vongpumivitch et al. 2009) from which texts were selected.

2.1.4 *Corpus size*

Academic corpus designers also note the size of their corpora with respect to number of texts and number of total running words compiled, essentially reflecting the maxim that the larger the corpus the better. Bowker & Pearson however, acknowledge that "there are no hard and fast rules that can be followed to determine the ideal size of a corpus" (2002: 45). The lack of standard "rules of thumb" regarding size is likely influenced by two issues: (a) the distributional characteristics of the features of interest (e.g. the frequency or rarity of occurrence), and (b) the scope of the domain to be represented (e.g. NY Times sports section articles or academic writing). In general, larger corpora are required to capture less frequently occurring features (e.g. many specialized vocabulary or low-frequency vocabulary) and to represent broader domains. Thus, corpus designers are provided somewhat, though understandably, vague guidance on corpus size, e.g.:

> "…a corpus should be as large as possible" (Sinclair 1991: 18)
> "…a corpus needs to contain many millions of words" (Sinclair 1991: 19)
> "…it is important to have a substantial corpus if you want to make claims based on statistical frequency" (Bowker & Pearson 2002: 48)
> "…lexicographic work requires the use of very large corpora…" (Biber et al. 1998: 25) comprising "…many millions of words" (249)

2.1.5 *Additional considerations*

Many other representativeness considerations have been noted and applied to corpus design, including authorial diversity (i.e. the wider the diversity the better) and completeness of sampled texts. Additionally, a critical consideration often noted is the balance of many of the variables noted above, including balance of total running words or texts per topic, discipline, genre (Hyland & Tse 2007: 8) or even a balance of texts of varying length (Coxhead 2000: 221).

2.2 What evidence for representativeness is missing?

Once the important corpus-design issues noted above have been considered and applied, researchers tend to jump strait to the creation of word lists, satisfied that, because of the careful attention to corpus design, the corpora, and, indeed, the lists based on them, are representative of their target domain. While all of these corpus design issues are, arguably, crucial, they all share at least one common characteristic: they are primarily *external* criteria. That is, while they help to ensure some

degree of ecological validity (i.e. textual, topical, and register representativeness), they may not ensure what corpora are ultimately designed to achieve: representativeness of *lexical variability* in our target domain.

Biber notes that "Representativeness refers to the extent to which a sample [i.e. a corpus] includes the full range of variability in a population" (1993: 243), and that determining the representativeness of a corpus should be a recursive endeavor based on corpus-internal evidence. That is, while external criteria such as topic coverage may guide the initial design of a corpus, there should be "discrete stages of extensive empirical investigation" (Biber 1993: 256) of a pilot corpus, and the corpus design should be revised as necessary. Unfortunately, this important step in validation of corpus representativeness – validation based on evidence that a corpus indeed represents "naturally occurring linguistic feature [e.g. lexical] distributions" (Biber 1993: 243) – does not often occur.

According to Atkins et al. (1992: 5), "…a corpus selected entirely on external criteria would be liable to miss significant variation among texts since its categories are not motivated by textual (but by contextual) factors". Biber (1993) concurs, noting that, while it is certainly crucial to consider situational variables which may have an effect on feature distribution (e.g. topic, register), and to use these variables to inform our corpus design, ultimately, the variability we are interested in is not simply variability in these *external* variables (e.g. topic or register). Rather, we are interested in representing the distribution of linguistic features (e.g. lexical variability) within our "population" (i.e. target language use domain). How do we know that our corpora have indeed captured this "full range of [lexical] variability" (Biber 1993: 243) without testing this assumption?

2.3 The current study

The goal of the current study is to examine lexical distribution within a target domain in order to directly assess the assumption of corpus representativeness. Specifically, the current paper details an attempt to identify the corpus composition required to capture a stable, reliable list of "important" words from one restricted register (i.e. introductory textbooks) in one academic discipline (i.e. psychology). The guiding research question for this study was: What size and composition of corpus is required to capture a stable, reliable list of "important" words from undergraduate introductory psychology textbooks?

Such an understanding is crucial for the assessment of the reliability of any word lists based on this corpus. More importantly, findings may allow other word list designers and users to make inferences regarding the lexical representativeness of other corpora used to produce word lists. Inferences can then also be made regarding the reliability of the word lists themselves.

3. Methodology

3.1 The undergraduate introductory psychology textbook (PSYTB) corpus

For the purpose of the current study, the target domain was operationalized as undergraduate introductory psychology textbooks. Two complementary methods guided the selection of the 10 textbooks comprising the PSYTB corpus: (a) a survey of textbooks used in introductory psychology classes at 28 tertiary academic institutions in the United States; and (b) a survey conducted by College Board's College-Level Examination Program (CLEP) of psychology textbooks commonly used in colleges and universities (The College Board 2010). Of the 10 books selected, five books were identified by both surveys, and five were identified by only one of the surveys.

Each chapter from each book was saved as a separate file, and all files were part-of-speech tagged using the Biber tagger (1988) to facilitate lemmatization. Front matter (e.g. publication information, tables of contents, forwards, and introductions) and appendices, indexes, and bibliographies from the textbooks were not included in the text files.

Table 1. Design of the introductory psychology textbook corpus (PSYTB)

Textbook	Chapters	Total running words
PSY_1	15	324,200
PSY_2	14	310,120
PSY_3	15	291,900
PSY_4	16	268,810
PSY_5	16	403,590
PSY_6	18	302,880
PSY_7	14	227,130
PSY_8	17	351,710
PSY_9	18	341,860
PSY_10	14	282,690
Average	15.7	310,489
TOTAL Corpus (PSYTB)	157	3,104,890

Table 1 outlines the design of the PSYTB corpus. To contextualize this corpus in relation to other academic corpora, Table 2 compares the PSYTB corpus with Coxhead's (2000) Academic Corpus. As can be seen from these tables, at 3.1 million words, the PSYTB corpus is nearly as large as the 3.5 million-word

Academic Corpus. If we assume that the Academic Corpus was balanced among macrodiscipline, it would have consisted of approximately 104 texts, 875,000 words, representing all 7 academic disciplines in the "Arts," one of which was psychology. Further breakdown into subdiscipline would suggest approximately 15 texts comprising approximately 125,000 words representing the discipline of psychology. It is important to note that this represents not only introductory psychology textbooks, but the entire discipline of psychology. As can be seen through this surface comparison, at least in terms of size, it is reasonable to suspect that the PSYTB corpus should be at least as representative of its comparatively narrow target use domain as the Academic Corpus was of its much broader target use domain.

Table 2. Comparison of the PSYTB with the academic corpus (Coxhead, 2000)

Point of comparison	PSYTB	Academic corpus (Coxhead, 2000)
Target Domain	Writing in Introductory Psychology Textbooks	Academic Writing encountered by university students in New Zealand
Corpus Design	10 complete contemporary introductory psychology textbooks	414 texts (mixture of whole texts and 2,000-word text samples) from 28 academic disciplines
Total Words	3.1 million words	3.5 million words

For all analyses outlined in the next section, the target domain was operationalized as the 10 complete textbooks comprising the PSYTB corpus. It is important to note that no claim is being made that the PSYTB corpus 10 textbooks is a perfect representation of lexical distributions in introductory psychology textbooks; however, the results of the analyses will provide insights into the size corpus required to represent lexical distributions in a target domain.

3.2 Procedures

3.2.1 *Vocabulary analysis program*

A vocabulary analysis program written for this study was capable of producing output almost identical to that produced by Heatley and Nation's *Range* program (1994), in that it produced the frequency of every word in every text in the corpus. In the current study, frequencies were provided by textbook and by chapter, and textbook range (out of 10) and chapter range (out of 157) totals were provided as well. Unlike *Range*, the analysis program designed for this study was based on the lemma, operationalized using Francis and Kucera's definition: "a set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling" (1982: 1). That is, each lemma, or each base

word and all inflectional variants (e.g. propose, v. = propose, proposes, proposed, proposing) were considered one lexical item. Derived variants (e.g. through affixation: propose > proposal; through conversion: increase as a verb > increase as noun) were considered different lexical items. The vocabulary analysis program was capable of reading the Biber tagger's (1988) part of speech tags and of grouping all occurrences of inflectional variants of a word into a single lemma. For example, all occurrences of *walk* (v.), *walks* (v.), *walked* (v.), and *walking* (v.) were combined and noted as occurrences of the lemma *walk* (v.), whereas *walk* (n.) and *walks* (n.) were combined and noted as occurrences of the lemma *walk* (n.).

3.2.2 *The analyses*

The goal of the study was to assess the degree to which different size samples are able to produce "important" word lists which reliably reflect the lexical distribution in my target domain: introductory psychology textbooks. Toward this goal, the first step was to decide which criteria for "importance" would be appropriate. The criteria that Coxhead (2000) used in selecting AWL words were first considered. As noted above, Coxhead proposed that a word merited inclusion on the AWL (i.e. it was deemed worthy of instructional focus) if it occurred at least 100 times (approximately 28 times/million words) in her corpus, at least 10 times in each of the four main macro-disciplines in her corpus, and in approximately one half of the subdisciplines represented in her corpus.

Though not a perfect equivalent, an introductory psychology book could be seen as a disciplinary overview, with each chapter representing a different "field" – or at least focus of study – within the discipline of psychology. Thus, as a place to begin for the first set of experiments, a word was deemed "important" if it occurred in one half of the chapters in the corpus (whether the corpus be the entire set of 10 textbooks, or a smaller set of textbooks sampled from the corpus for comparison). This single criterion had the added benefit of, in effect, "forcing" two additional criteria. First, if a word occurred in one half of the chapters in the corpus, it was, as a rule, also found in at least one half of the textbooks in the corpus. Additionally, the chapter range requirement forced a minimum frequency of approximately 22 occurrences per million words.[2] The experiment, then, was

---

2. By comparison, the minimum frequency requirement for words on the AWL, 100 occurrences in the corpus, norms to approximately 29 occurrences per million words. Though there is a difference in minimum frequency criterion between Coxhead's study and this set of experiments (i.e. 29 occurrences/million vs. 22 occurrences/million), it is important to keep in mind that the AWL criterion was based on the frequency of occurrence of word family members combined, rather than the combined frequency of lemma members. Thus, it is to be expected that frequencies for lemmas would be lower than they would for the word

conducted as follows. First, the criterion for "importance" (i.e. occurrence in 50% of chapters) was applied to the whole PSYTB corpus of 10 books, generating a list of "important" lemmas. Then, different size subsamples from the corpus, from single whole textbooks through a set of nine whole textbooks, were assembled. These different size samples might be thought of as representing different sampling rates. A sample of one textbook (out of 10 textbooks) could be considered a 10% sampling rate, samples of two textbooks a 20% sampling rate, etc. Then, the same range criterion (i.e. occurrence in 50% of chapters) was applied to each sample, and lists of important words were generated for each sample. Each sample list was then compared with the word list generated by the whole corpus (i.e. the PSYTB list). The comparisons between the lists were made as follows. First, a list of words meeting the criteria of "importance" was identified from the whole PSYTB. Then, the same criteria were applied to a sample from the corpus, creating another list of words. The two word lists were then compared, as illustrated in the Venn diagram in Figure 1.

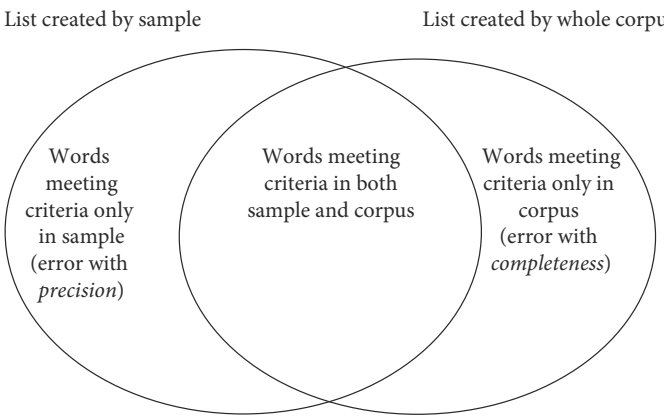

Figure 1. Comparing lists of important words from the whole corpus with the lists identified in the samples

Essentially, there can be two types of difference between sample lists and the whole corpus list. First, some words might meet the criteria of "importance" in the sample (i.e. will occur in ≥50% of the chapters in the sample), but not in the whole corpus. These "additional" words constitute error with *precision*, as they

families to which they belong. If anything, my inclusion criteria are somewhat stricter than those used by Coxhead.

are incorrectly identified as "important" when we can see that they are not in fact "important" in the target domain (operationalized as the whole PSYTB corpus of 10 textbooks).

Conversely, some words *will not* meet the criteria of "importance" in the sample (i.e. will occur in fewer than 50% of the chapters in the sample), but *will* in the whole corpus. These "missing" words constitute error with *completeness*, as their absence makes the sample lists incomplete.

Interpretation of these comparisons was made in the following way. If a subset of textbooks, for example, a sample of three or four complete textbooks, provided a word list comparable to one generated by the whole corpus of 10 complete textbooks, it could be argued that the subset of textbooks sufficiently represents the lexical distributions in the domain (represented by the set of 10 textbooks), and that a larger sample (i.e. additional textbooks) is unnecessary. Alternatively, if there were still notable differences between a list produced from a sample and the list produced from all 10 textbooks, it could be argued that the sample did not adequately represent the lexical distribution in the domain (i.e. the PSYTB corpus).

4. Results

Table 3 provides the results of an experiment investigating whether one whole psychology textbook provides a sufficient sample of the domain, allowing for the creation of a stable list of "important" words. Specifically, it summarizes the comparison between lists produced from a sample of single individual whole textbooks (i.e. 1 TB) with the list produced from the whole corpus of 10 textbooks. Five "important" word lists were generated, one for each of five individual, randomly selected textbooks (i.e. Textbook 1, Textbook 2…), in order to account for between-book diversity.

In the first row of Table 3, we can see how many words met the criteria of importance in each of the five textbooks samples. 1,745 words occurred in at least 50% of the chapters in Textbook 1, 1,771 words met this criterion in Textbook 2, etc. These lists, then, were compared with the list of 1,532 words that were found in at least 50% of the chapters in the whole corpus.

The difference in list size alone indicates that there is indeed a difference between the sample lists and the whole corpus list. For example, 1,745 lemmas were identified as important in Textbook 1, whereas only 1,532 lemmas were identified as important across the whole corpus. However, this surface observation only begins to tell the story of the difference between the two lists. The word list culled from Textbook 1 is not necessarily only 213 words (1,745–1,532 = 213) different from the whole corpus word list. Rather, the Textbook 1 word list and the

Table 3. Comparison of list produced by one whole textbook with lists produced by whole corpus

Size of sample compared with corpus	Number of words meeting criteria in different samples						Average number and % of words not meeting criteria in both sample and whole corpus			
	Textbook 1	Textbook 2	Textbook 3	Textbook 4	Textbook 5	Whole corpus	Only in sample	Only in whole corpus	Total difference	SD
1 TB*	1,745	1,771	1,895	1,470	2,176	1,532	429.4	163.0	592.4	143.9
							28.0%	10.6%	38.7%	9.4%

*_Note_: 1 TB = one complete textbook. Results of a comparison between larger samples – two complete textbooks (2 TBs) through nine complete textbooks (9 TBs) – and the whole PSYTB can be seen in Table 4.

whole corpus word list share some words, but there are some words identified as important only in the sample, and some words identified as important only in the whole corpus. Again, Figure 1 illustrates this phenomenon.

In Table 3, we can see *precision* error in the column headed "Only in Sample." On average, there were approximately 429 words that met the criteria of importance only in the samples of one textbook that did not maintain the ≥50% chapter range across the whole corpus. These words account for, on average, 28.0% of the difference between these sample and whole corpus lists. Error with *completeness* is noted in Table 3 in the column titled "Only in whole corpus." From this table, we can see that, on average, the whole corpus identifies 163 words as "important" that the samples do not. This set of missing words, *completeness* error, accounts for 10.6% of the difference between the sample lists and the whole corpus list. In sum, there are, on average, 592.4 (*SD* 143.9) words that are not shared by both lists. So, we can say that lists produced from samples of one textbook are, on average, 38.7% (*SD* 9.4%) different from a list identified by the whole corpus of 10 books.

Thus, it would be reasonable to conclude that a sample of one whole textbook (i.e. a sampling rate of 10%) does not provide an adequate representation of lexical distributions in this target domain. We can see this because these samples produce word lists that are, on average, nearly 40% different than the list generated by a corpus of 10 textbooks. Stated simply, a sample of one whole textbook (i.e. a sampling rate of 10%) is too small.

So what size sample *can* capture the "important" words identified from the target domain (i.e. the PSYTB corpus of 10 textbooks)? To answer this question, the comparisons were repeated with random sample sets of two through nine whole textbooks taken from the corpus of 10 textbooks. Again, five samples of each size were taken to account for between-book lexical variability. The ≥50% range criterion was applied to each sample, and the word lists generated were compared with the whole corpus list of 1,532 "important" words. Results of these comparisons can be seen in Table 4.

It is a logical necessity that, as samples reflect a greater proportion of the domain that they are designed to represent, lists of "important" lemmas that they produce would more closely match the list representing "important" words in the whole domain. On average, this is the case here, as can be seen in Table 4. As the samples continue to grow (e.g. from 2 TBs to 3 TBs, to 4 TBs, etc.), the difference between word lists (i.e. the average percentage of additional "important" lemmas and missing "important" lemmas) decreases. However, at what point might we conclude that lists from our sample reflect a reasonably equivalent set of important words?

Many factors must be considered here. Practical considerations such as the time, effort, and expense of acquiring texts are certainly relevant. In other words,

Table 4. Comparison of lists produced by samples of two through nine textbooks with lists produced by whole corpus

Size of sample compared with corpus	Number of words meeting criteria in different samples						Average number and % of words not meeting criteria in both sample and whole corpus			
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Whole corpus	Only in sample	Only in whole corpus	Total difference	SD
2 TBs	1,564	1,814	1,506	1,431	1,612	1,532	183.2	142.8	326.0	37.6
							12.0%	9.3%	21.3%	2.5%
3 TBs	1,529	1,797	1,519	1,558	1,822	1,532	187.4	87.4	274.8	72.7
							12.2%	5.7%	17.9%	4.7%
4 TBs	1,784	1,621	1,521	1,658	1,555	1,532	146.8	64.0	210.8	35.7
							9.6%	4.2%	13.8%	2.3%
5 TBs	1,685	1,438	1,513	1,675	1,718	1,532	119.8	59.0	178.8	35.2
							7.8%	3.9%	11.7%	2.3%
6 TBs	1,603	1,594	1,605	1,669	1,565	1,532	100.0	37.8	137.8	23.4
							6.5%	2.5%	9.0%	1.5%
7 TBs	1,567	1,555	1,576	1,613	1,506	1,532	61.4	43.0	104.4	9.3
							4.0%	2.8%	6.7%	0.6%
8 TBs	1,546	1,552	1,556	1,528	1,606	1,532	47.6	35.0	82.6	11.8
							3.1%	2.3%	5.4%	0.8%
9 TBs	1,569	1,553	1,524	1,582	1,553	1,532	32.4	21.2	53.6	9.2
							2.1%	1.4%	3.5%	0.6%

the fewer textbooks needed, the better. However, we must also consider the reliability of the lists we produce. Ideally, a list produced from a sample would maximally concur with a list of "important" words in the domain. That is, the sample list would be maximally *precise* (i.e. not include additional words that do not hold their currency across the whole domain) and *complete* (i.e. identify all "important" words from the domain).

As we look at Table 4, it is important to reflect for a moment on the relationship between sample size and reliability of "important" word lists generated from samples of different sizes. As an example, consider a comparison between word lists culled from a sample of three textbooks and the word list culled entire corpus of 10 textbooks. As a point of reference, a sample of three textbooks accounts for approximately 45–48 chapters (i.e. texts) and 1,000,000 running words. This sample is equal to nearly one third of the total running words in the entire Academic Corpus from which Coxhead (2000) generated the AWL. Thus, by comparison, three complete textbooks might seem a reasonable sample size to represent the narrow domain of introductory psychology textbooks. In addition, in this experiment, three textbooks represents a 30% sampling rate of the target domain.

Despite the corpus size and sampling rate, however, the comparison suggests that a sample of 3 textbooks is not adequate. On average, the lists from samples of three textbooks do indeed capture approximately 94% of the "important" words generated from a corpus of 10 textbooks, but more than 12% of the words on these sample lists do not hold their currency across the entire corpus (i.e. target domain). In real terms then, a list generated by a corpus of three textbooks has, on average, 275 *different* lemmas than does a list generated by the whole PSYTB corpus. This is a total difference of approximately 18%.

To further illustrate this difference, Table 5 provides examples of actual words that would not be shared by a list generated from one sample of three textbooks and a list generated by the whole corpus of 10 books. From this table, we can get an idea of some words that would be identified as important in either the sample of three textbooks or across the entire corpus, but not in both. These lists illustrate the possible lexical variability existing even in a restricted domain, and, thus, further highlight the importance of assessing the representativeness of corpora and the reliability of conclusions drawn from them. If representativeness were determined based solely on the size of the sample (i.e. three whole textbooks; approximately 1 million words), we might conclude that our corpus would be sufficiently representative, and that the word list it generates was worth instructional focus. Such a conclusion might be tempered, however, when we consider that valuable time might then be spent on 75 words that do not hold currency across the whole corpus, such as *activate, adjust*, or *accomplish*. Conversely, and perhaps more importantly, perhaps no focus would be given to the 86 words that actually do

prove "important" across the entire corpus but did not meet the criteria of importance in the sample (e.g. *process*(n.), *creative*, or *sensation*).

Table 5. Comparison of words meeting "importance" criteria in a sample of 3 textbooks and words meeting "importance" criteria in the whole corpus

Lemmas that meet the criteria only in the sample (i.e. error with precision)		**Lemmas that meet the criteria only in the whole corpus (i.e. error with completeness)**	
activate	*interview* (n.)	*process* (v.)	*unique*
adjust	*trigger* (n.)	*sensation*	*responsible*
accomplish	*manage*	*creative*	*description*
neutral	*underlying*	*threat* (n.)	*design* (n.)
violence	*substantial*	*theme* (n.)	*contain*
moderate (adj.)	*recognition*	*forget*	*encounter* (n.)
diagnose	*service* (n.)	*video* (n.)	*insight*
discrimination	*team*	*norm* (n.)	*progress* (n.)
design (n.)		*surface* (n.)	*estimate* (n.)
		advantage	

Table 6. Comparison of words meeting "importance" criteria in samples of 5 textbooks and words meeting "importance" criteria in the whole corpus

Comparison made	**Words only "important" in…**	**Number of shared "important" words**	**Words only "important" in…**
sample 1 (5 textbooks) **vs. corpus**	**…sample 1:** 173 words *e.g. core, phase, theoretical, manipulate, objective, complexity, selective*	1,512	**…the whole corpus:** 33 words *e.g. variable, reinforce, dimension, adaptation, regulate, function, generate, minimize*
sample 2 (5 textbooks) **vs. corpus**	**…sample 2:** 18 words *e.g. accuracy, criterion, design, ethnic, evident, integrate*	1,420	**…the whole corpus:** 125 words *e.g. consume, exhibit, sufficient, survey, prediction, reject*
sample 1 vs. sample 2	**…sample 1:** 298 words *e.g. crucial, enormous, summarize, portion, undergo, accompany*	1,387	**…sample 2:** 51 words *e.g. marriage, total, close, representation, wife, extreme, essential*

Next, a sample of five textbooks – a 50% sampling rate and over 1.5 million words on average – was considered. Table 6 presents a comparison of two word lists culled from half of the corpus with the word list produced from the whole PSYTB corpus (sample 1 vs. corpus and sample 2 vs. corpus). In addition, it

presents a comparison between two halves of the corpus (sample 1 vs. sample 2). Interestingly, from this latter comparison, we can see that there are 349 words (i.e. 298 + 51) that are considered "important" in one sample or the other, but not in both. This suggests a notable amount of unreliability in word lists produced from five whole textbooks – or approximately 1.5 million word corpora. And referring back to Table 4, we can see that no sample – even a sample of nine out of 10 textbooks (a 90% sampling rate) – is able to produce a word list that perfectly mirrors the list produced from the whole corpus (i.e. the target domain).

The following sections discuss these findings with regard to their implications for vocabulary researchers as well as for users of vocabulary research.

5. Discussion of findings

The experiments detailed above have sought to determine the degree to which different size samples (i.e. different sampling rates) could represent the lexical distributions (i.e. capture the important words identified) in the PSYTB corpus. Based on the findings from these experiments, we can reasonably conclude that there is a tremendous amount of lexical variability in undergraduate introductory psychology books, and that it takes a very large corpus to reasonably capture this variability. Indeed, no samples culled from the PSYTB corpus, even samples representing 90% of the target domain, produced completely reliably representations of the lexical distributions (i.e. reliably capture the "important" words) in the PSYTB corpus.

This finding leads to two possible conclusions. One possibility is that a sample of 10 introductory psychology textbooks is simply too small to represent the target domain. However, this possibility only further highlights the problem. If 10 whole textbooks, over three million words, do not represent this narrow domain, how many additional textbooks would be needed to do so? A second possibility is that there is just so much variability in academic writing that any list of "important" words for this domain, even a domain as narrowly defined as introductory psychology textbooks, is far more restricted in size and/or reliability than we have considered previously.

Either way, the findings from this study have important implications with regard to decisions that are made based on corpus-based vocabulary research. Currently, assessments of word list *stability* are not standard practice. Instead, faith in word lists has rested primarily on faith in the careful attention paid to external corpus design issues (e.g. size of corpus, representation of the types of topics and texts that occur in a given domain). This evidence, it appears, has been considered sufficient support for conclusions that have been drawn based on corpora (e.g. lists of "important" words).

Regarding the word lists themselves, a good deal of evidence has been put forth in order to demonstrate their validity. For example, distributional characteristics, namely range, frequency, and dispersion, have been used to demonstrate the "importance" of word lists. That is, if words appear frequently, widely, and evenly throughout a corpus, there is evidence of their usefulness and, thus, justification for their inclusion on lists. Post-hoc analysis has also been used to validate word lists. For example, Coxhead (2000) determined that her AWL provided a great deal more coverage of academic texts (i.e. 10% on average) than it did of more general English texts, and thus concluded that the AWL indeed represented important *academic* vocabulary. Nation (2004) suggested that his 1,000-word bands from the BNC were properly ordered because the first 1,000 words together provided higher coverage of the BNC and other general English corpora than did the second 1,000 words.

Word list research for more specialized domains has looked at the reliability of, for example, the AWL in these specialized domains (e.g. Chen & Ge 2007; Li & Qian 2010; Martinez et al. 2009; Wang et al. 2008). These studies have demonstrated varying degrees of difference between the AWL and word lists produced from specialized corpora, and have thus concluded that modifications to the AWL (or completely new, specialized lists) are necessary. As with previous word list studies, however, they have not provided direct evidence of the stability of these revised (or new) lists. Despite expressed confidence that these word lists better reflect the lexical distributions of their specialized target domains, there is still a lack of corpus-internal evidence that the corpora used – or the lists produced – are reliably representative.

So what does this mean for corpus-based vocabulary researchers and for those who rely on the conclusions these researchers draw? Two key considerations arise:

5.1 Size may not be the whole story

Brysbaert and New (2009) note that, with the increasing ease of compiling electronic corpora, corpora in the hundreds of million, or even billion running words will become increasingly common. Compilers of these new corpora note how the larger size and increased consideration of topic or register coverage and balance are evidence of their corpora's increased representativeness, and, in turn, the validity of conclusions drawn from them (e.g. BNC: Leech et al. 2001; Corpus of Contemporary American English: Davies 2009, 2010; Davies & Gardner 2010). While these corpora may indeed be more representative of lexical distributions than their smaller, sometimes less-principled predecessors were, evidence to this effect has simply not been produced. As well, no evidence has been produced to demonstrate the reliability of conclusions drawn from these corpora (e.g. lists of words

meriting instructional focus). This is surprising when we consider the painstaking effort and care that has gone into designing and compiling these corpora and creating lists from them. It must be strongly noted that evidence from the current study does not demonstrate any lack of reliability of lists drawn from these contemporary corpora or from others with more specialized focus. However, it does suggest the benefit of additional analysis in order to better understand lexical variability in a given target domain and to assess the degree to which corpora capture this variability.

5.2 Word list users must understand what word lists are and what they are not

"Important" word lists are simply lists of words meeting predetermined distributional characteristics in the corpora upon which they were based. And as Schmitt (2010) noted, a word list is only as good as the corpus upon which it is based. Thus, there is a limit to the generalizability of word lists to other texts, even to other texts within the same domain. It is crucial, therefore, that list users understand this and exercise caution in applying word lists to their given context. For example, while the AWL consists of words that meet Coxhead's (2000) pre-determined distributional characteristics within her Academic Corpus, the AWL is not necessarily reliable across all academic texts. As has been noted, this lack of generalizability has been demonstrated on numerous occasions and with domains of various scope (e.g. Hyland & Tse 2007; Martínez et al. 2009; Vongpumivitch et al. 2009; Wang et al. 2008). The present study has further highlighted and extended this concern by demonstrating the challenge of identifying a stable, reliable list of "important" words that is generalizable even across very narrow domains (e.g. undergraduate, introductory psychology textbooks). This is not to say that there do not exist sets of words that *are* generalizable within a domain or even across domains. However, as demonstrated in the current study, it is likely that such lists are much more restricted – either in terms of size or in terms of reliability – than has been realized or acknowledged.

6. Conclusion

This study has had one primary goal: to assess the size and composition of the corpus required to reliably capture the important words in undergraduate, introductory psychology textbooks. Findings demonstrate the significant challenge of producing a stable, reliable list of "important" words for this domain. There simply appears to be far greater lexical variability in this target domain than the PSYTB

corpus represents. Even a 3.1 million-word corpus of 10 whole introductory psychology textbooks was unable to yield such a list.

These findings have much broader implications for corpus-based vocabulary research that seeks to identify lists of words meriting valuable teaching and learning time. Specifically, this case study has provided further support for Biber's (1993) contention that corpus-internal analysis must be included in operationalizations of corpus representativeness. Without such analysis, it is impossible to ensure that a corpus indeed captures the naturally occurring distributions of target features it has been designed (or used) to represent. And, as the present study illustrates, this may be particularly important in attempts to understand lexical distributions and, ultimately, to produce reliable word lists from corpora.

References

Atkins, Sue, Clear, Jeremy & Ostler, Nicholas. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7: 1–16. DOI: 10.1093/llc/7.1.1

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: CUP. DOI: 10.1017/CBO9780511621024

Biber, Douglas. 1993. Representativeness in corpus design. *Literary & Linguistic Computing* 8: 243–257. DOI: 10.1093/llc/8.4.243

Biber, Douglas, Conrad, Susan & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Structure and Use.* Cambridge: CUP. DOI: 10.1017/CBO9780511804489

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Pat, Helt, Marie, Clark, Victoria, Cortes, Viviana, Csomay, Eniko & Urzúa, Alfredo. 2004. *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* [ETS TOEFL Monograph Series, MS-25]. Princeton NJ: Educational Testing Service.

Bowker, Lynne & Pearson, Jennifer. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routlege. DOI: 10.4324/9780203469255

Brysbaert, Marc & New, Boris. 2009. Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41(4): 977–990. DOI: 10.3758/BRM.41.4.977

Burgmeier, Arline & Zimmerman, Cheryl Boyd. 2007. *Inside Reading 1 Student Book Pack: The Academic Word List in Context.* Oxford: OUP.

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: 502–514. DOI: 10.1016/j.esp.2007.04.003

The College Board. 2010. CLEP® Introductory psychology: At a glance. ⟨http://clep.collegeboard.org/clep-introductory-psychology-glance⟩

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213–238. DOI: 10.2307/3587951

Coxhead, Averil & Hirsh, David. 2007. A pilot science word list for EAP. *Revue Française de Linguistique Appliquée* XII(2): 65–78.

Davies, Mark. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics* 14: 159–90. DOI: 10.1075/ijcl.14.2.02dav

Davies, Mark. 2010. The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25(4): 447–65. DOI: 10.1093/llc/fqq018

Davies, Mark & Gardner, Dee. 2010. *A Frequency Dictionary of Contemporary American English*. New York NY: Routledge.

Durrant, Philip. 2009. Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes* 28: 157–169. DOI: 10.1016/j.esp.2009.02.002

Francis, W. Nelson & Kucera, Henry. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston MA: Houghton Mifflin.

Heatley, A. & Nation, Paul. 1994. *Range*. Victoria University of Wellington, NZ. Sofware. ⟨http://www.vuw.ac.nz/lals/⟩

Huntley, Helen. 2005. *Essential Academic Vocabulary: Mastering the Complete Academic Word List*. New York NY: Houghton Mifflin.

Hyland, Ken & Tse, Polly. 2007. Is there an "academic vocabulary?" *TESOL Quarterly* 41(2): 235–253.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Leech, Geoffrey, Rayson, Paul & Wilson, Andrew. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. London: Pearson.

Li, Yongyan, & Qian, David. 2010. Profiling Academic Word List (AWL) in a financial corpus. *System* 38: 402–411. DOI: 10.1016/j.system.2010.06.015

Martínez, Illiana, Beck, Silvia & Panza, Carolina. 2009. Academic vocabulary in agricultural research articles: A corpus-based study. *English for Specific Purposes* 28(3): 183–198. DOI: 10.1016/j.esp.2009.04.003

McEnery, Tony & Wilson, Andrew. 1996. *Corpus Linguistics*. Edinburgh: EUP.

McEnery, Tony, Xiao, Richard & Tono, Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York NY: Routledge.

Millar, Neil & Budgell, Brian. 2008. The language of public health: A corpus-based analysis. *Journal of Public Health* 16: 369–374. DOI: 10.1007/s10389-008-0178-9

Mudraya, Olga. 2006. Engineering English: A lexical frequency instructional model. *English for Specific Purposes* 25(2): 235–256. DOI: 10.1016/j.esp.2005.05.002

Nation, Paul. 2001. *Learning Vocabulary in Another Language*. Cambridge: CUP. DOI: 10.1017/CBO9781139524759

Nation, Paul. 2004. A study of the most frequent word families in the British National Corpus. In *Vocabulary in a Second Language* [Language Learning & Language Teaching 10], Paul Bogaards & Bahtia Laufer (eds), 3–14. Amsterdam: John Benjamins. DOI: 10.1075/lllt.10.03nat

Schmitt, Norbert. 2010. *Researching Vocabulary: A Vocabulary Research Manual*. Houndmills: Palgrave Macmillan. DOI: 10.1057/9780230293977

Schmitt, Diane & Schmitt, Norbert. 2005. *Focus on Vocabulary: Mastering the Academic Word List*. New York NY: Pearson.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*: *Describing English Language*. Oxford: OUP.

Upton, Thomas. 2004. *Reading Skills for Success: A Guide to Academic Texts*. Ann Arbor MI: The University of Michigan Press.

Vongpumivitch, Viphavee, Huang, Ju-yu, & Chang, Yu-Chia. 2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AL content words in applied linguistics research papers. *English for Specific Purposes* 28(1): 33–41. DOI: 10.1016/j.esp.2008.08.003

Wang, Jing, Liang, Shao-lan & Ge, Guang-chun. 2008. Establishment of a medical academic word list. *English for Specific Purposes* 27(4): 442–458. DOI: 10.1016/j.esp.2008.05.003

West, Michael. 1953. *A General Service List of English Words*. London: Longman.

Xue, Guo-yi & Nation, Paul. 1984. A university word list. *Language Learning and Communication* 3(2): 215–229.

CHAPTER 7

Corpus linguistics and New Englishes

Chandrika Balasubramanian
Sultan Qaboos University

The rising status of English as a world language has resulted in the emergence of new varieties of English that have been legitimized by such expressions as New Englishes and New Varieties of English. Accepting the idea of New Englishes has allowed much-needed movement away from the previously accepted notions of nativeness and non-nativeness (Mesthrie 2010), and today, they are seen as systems unto themselves as opposed to deviant forms of traditional native varieties (Jenkins 2003). The current study investigates spoken and written registers of contemporary Indian English and demonstrates, through the investigation of WH-questions, and the circumstance adverbials *also* and *only* that Indian English shows the same kind of internal variation present in more traditional "native" varieties.

Keywords: Indian English; register; circumstance adverbials; wh-questions

1. Introduction

The way regional and social factors have influenced the growth of New Varieties of English and fostered change has formed the subject matter of sociolinguistics and dialectology from both theoretical and practical standpoints, and today, nobody would deny the fact that "World English exists as a political and cultural reality" (Crystal 2003: xii). Further, as Schneider (2003: 233) puts it, "present-day English as a global language is more than the world's predominant lingua franca – it is also a language which is currently growing roots in a great many countries and communities around the world, being appropriated by local speakers, and in that process it is diversifying and developing new dialects…" Gargesh (2006: 90) claims that the "nativization of English has enriched English as well as the indigenous languages through processes of borrowing and coinage of new words and expressions, and through semantic shifts."

DOI 10.1075/scl.66.07bal

Initial studies of New Englishes focused on what contributed to such nativization, mainly at the phonological and lexical levels. Indeed, as Bolton (2006: 255) explains, "dictionaries are profoundly important for the recognition of world Englishes" and noted that "it is only when a world variety of English is supported by codification (chiefly expressed through national dictionaries) that one can make a strong claim that such a variety is "institutionalized." Other studies on New Englishes included those that were concerned with establishing how these Englishes differed from traditional native speaker varieties, and these studies focused on identifying characteristic features – lexical, phonological, and grammatical – of the variety.

Schneider (2003) outlined a process of development that he claimed all new Englishes go through. The five stages of his process include "Foundation, Exonormative Stabilization, Nativization, Endonormative Stabilization, and Differentiation" (p. 243). Schneider explains further that the first process, Foundation, is the initial phase where "English begins to be used on a regular basis in a country that was not English-speaking before;" he characterizes this phase as a "complex contact situation" (p. 244). In this phase, contact between the two language groups remains restricted, with cross-cultural communication being achieved by just a few people. Further, during this phase, indigenous languages do not influence the English spoken by the settlers. During Phase 2, Schneider (2003: 245–247) explained, the external norm, "usually written and spoken British English as used by educated speakers, is accepted as a linguistic standard of reference." Also, during this phase, Schneider notes that Structural Nativization occurs where "as soon as a population group starts to shift to a new language, some transfer phenomena at the level of phonology and structure are bound to occur." According to Schneider, Phase 3 is "the most important, the most vibrant one, the central phase of both cultural and linguistic transformation in which both parties realize that something fundamental has been changing". It is during this phase of Nativization that the New English starts to construct its identity independent of the "native" English. It is during this phase, then, that characteristic "features" of the new English emerge. Endonormative Stabilization is "marked by the gradual adoption and acceptance of an indigenous linguistic norm, supported by a new, locally rooted self-confidence…" (Schneider 2003: 249), while during the fifth phase, Differentiation, "the focus of an individual's identity construction narrows down, from the national to the immediate community scale…consequently, new varieties of the formerly new variety emerge as carriers of new group identities within the overall community" (ibid: 253).

Early studies on New Englishes focused largely on identifying the features that arise during the Nativization phase of the new variety's development, and did not

focus on variation within each international variety, treating the variety, instead, as a homogeneous entity. Studies such as these abound in literature on New Englishes (Ahulu 1995; Bamiro 1995; Bansal 1976; Bakshi 1991; Banjo 1997; Barbe 1995; Bauer 1989; Baumgardner 1996; Coelho 1997; D'Souza 1997; Gisborne 2000; Huber 1995; Kallen 1989; Mazzon 1993; Setin 1997; Skandera 1999; Watermeyer 1996; Youssef 1995; and Zhiming 1995). It is clear that in the late 1980s and 1990s, the study of New Englishes, with a view to determining that they were indeed their own varieties, was tremendously popular.

1.1 Corpus linguistics and the study of New Englishes

More recently, advances in corpus linguistics and the development of a methodology that utilizes both qualitative and quantitative research traditions has allowed us to study variation within these new diversified varieties of English with a depth not formerly possible. The use of corpus linguistics methodology to study variation within dialects by studying register variation is perhaps best exemplified by Biber et al. (1999), who strongly advocated its use; Biber et al. (1999) argued that using corpus linguistics methodology would allow a researcher to provide a linguistic analysis of the whole range of spoken and written registers in English, something that dialectologists had hitherto not done. As Biber and Finegan (1991: 209) explain, such studies are significant, among other ways, in that "they analyze particular constructions in naturally occurring discourse rather than made-up sentences." Biber et al. (1999) claim, "the use of computer-based corpora provides a solid empirical foundation for general purpose language tools and descriptions, and enables analyses of a scope not otherwise possible" and that "corpus-based analyses of linguistic variation have provided fresh insights into previously intractable issues" (p. 257).

Today, numerous corpus-based studies of both New Englishes and less-studied Englishes such as Australian and New Zealand Englishes exist, that draw from the relevant sections of the International Corpus of English (Greenbaum 1996) for their analyses. Others rely on existing corpora of their national variety, like the corpus of Australian English, and the corpus of New Zealand English. Examples of such studies include those by Banjo (1997), on syntactic features of Nigerian English, and Schmied's (1994) study on syntactic features of Indian English; Starks et al. (2007), on Niuean English; Mukherjee and Gries (2009) on verb construction associations in the International Corpus of English; Mukherjee and Hoffman (2006) on verb complementation in Indian English; and Peters (2009) on Australian English, to mention just a few. In addition to these shorter studies on different national varieties, the field has more recently seen several

books on corpus-based investigations of new varieties. These include Hundt & Gut (2012); Mesthrie & Bhatt (2008); Schneider (2007); Schneider et al. (2004); and Kirkpatrick (2010).

1.2 English in India

As described by Bolton et al. (2011:468), "the history of English in India is as old as that in North America, dating from the very beginning of the 17th century." English was introduced into the Indian education system in Macaulay's famous "minute" of 1835. Since then, it has become the language of the Indian education system, and has officially and unofficially assumed the position of lingua franca in the country. The language has since undergone "a process of Indianization in which it has developed a distinctive national character comparable to that of American or Australian English" (Jenkins 2003:7).

The English used in India has long been the subject of inquiry from theoretical and sociolinguistic perspectives. As Bolton et al. (2011) put it, "Indian English is one of the earliest recognized and documented of the new Englishes" (p. 468), with Hobson-Jobson's dictionary of Indian English being first published as early as 1896. Kachru (1969) began studying what makes Indian English Indian in the 1960s, with the focus of much of his work being the establishment of the *Indianness* of Indian English. Kachru (1969) pointed out that studies of Indian English considered "linguistic interference and the Indian cultural context as essential for the understanding and description of the Indianness in this variety of English" (p. 5).

There are many studies concerning the nature of the English used in India, and how this differs from other "standard" varieties of the language, such as standard British or American English. Most studies have focused on features that have almost become stereotypical of the English used in India, including stative verbs in the progressive (one only need think of characters on popular television shows like *The Simpsons* to understand just how widespread this stereotypical feature is), the use of prepositions (Kachru 1994; Verma 1980; Hosali 1991), and the use of articles (Bakshi 1991), to name just a few. While many of these studies have been based on anecdotal evidence, others have used different types of data for their analyses, with a few studies drawing from the Kohlapur Corpus, and the Indian section of the International Corpus of English (ICE India). Table 1 provides a summary of several of the studies on Indian English conducted thus far.

It is clear from Table 1 below that Indian English has been extensively studied for several decades. It is also important to mention that while earlier studies tended to be more anecdotal in nature, latter studies have been more data driven. What is also clear from the table, however, is that thus far, few studies have focused on register differences in the variety; few studies (even those that are more

recent), therefore, have focused on the variation within Indian English. If we are to determine whether Indian English is truly gaining (or has gained) an identity distinct from other varieties of English, and, therefore, whether it is in Schneider's Endonormative Stabilization phase, or Differentiation phase (phases 4 and 5), we need to determine how it varies internally.

Table 1. Previous studies on Indian English

Author(s)	Aim of study	Data used
Davidova (2012)	To develop a framework for the systematic investigation of institutionalized varieties of English around the world	No specific database; register differences not investigated
Sedlatschek (2009)	To study variation and change in Indian English	Sections of the Kohlapur Corpus and ICE-India
Sailaja (2009)	A description of Indian English at the phonological, lexical, morpho syntactic, and discourse levels	No specific database; register differences not investigated
Balasubramanian (2009)	To study the distribution of grammatical and lexical features across registers of Indian English	A corpus of contemporary Indian English and sections of ICE-India
Lange (2007)	To determine the syntactic and semantic uses of the focus markers *itself* and *only* in registers of Indian English	ICE India; register differences investigated
Mukherjee & Hoffman (2006)	Verb complementation patterns	ICE India; register differences not investigated
Sharma (2005)	To apply the principles of language transfer and discourse universals to the use of articles in Indian English; studied the use or absence of articles in different linguistic environments	12 sociolinguistic interviews; register differences not investigated
Sand (2004)	Article use in Indian English and other international varieties of English	Different ICE corpora; discussion does focus on distribution of articles across different registers.
Schneider (2004)	Particle verbs in different international varieties of English, including Indian English	ICE India (in addition to 4 other ICE corpora). Register differences not investigated
Olavarria de Ersson & Shaw (2003)	Verb complementation patterns	Corpus of Indian and British newspaper archives. Patterns of verb complementation determined in one register
Shekar & Hegde (1996)	Phonological, Morphological, and grammatical features of Indian English	No specific data; register differences not investigated
Shastri (1996)	Two types of verb and adjective complementation in Indian English	Kohlapur Corpus of Indian English; register differences not investigated

(*Continued*)

Table 1. Previous studies on Indian English (Continued)

Author(s)	Aim of study	Data used
Schmied (1994)	To examine the patterns of occurrence of six grammatical features in the Kohlapur Corpus of Indian English. A contrastive study, contrasting patterns of occurrence of the features examined in Kohlapur Corpus, results compared with patterns of occurrence of the features in "native varieties"; features examined included Sentence complexity, Verb order in questions, Invariant tag – *isn't it*, Progressive forms of stative verbs, Use of articles, and Relative constructions	The Kohlapur Corpus of Indian English; register differences not investigated
Hosali (1991, 1992)	To describe "Butler English," a sub variety of English in India	22 extracts from a sub corpus of Butler English; register differences not investigated
Leitner (1991)	To analyze the Kohlapur Corpus for patterns of occurrence of a few grammatical features including the Subjunctive, Complex prepositions, and Modal verbs	The Kohlapur Corpus of Indian English; register differences not investigated
Agnihotri & Khanna (1984)	Article use in Indian English	Responses from 366 Hindi/Punjabi speaking undergraduates; register differences not investigated

1.3 Aims of the present study

The aims of the present study are to show, through the investigation of two grammatical features (WH-questions, the additive and restrictive circumstance adverbials *also* and *only*), that an international English like Indian English shows the same kind of internal variation that the more traditional "native" varieties do.

2. Methodology

The following section outlines the methodology behind corpus compilation. This description is followed by a description of the two linguistic features examined in this study, and the specific steps taken for their analyses.

2.1 Corpus design: The corpus of contemporary Indian English

The Corpus of Contemporary Indian English (CCIE) consists of eight large registers each with several sub-registers. The overall aim in compiling this corpus was

to gather a set of spoken and written registers that together attempt to represent the range of settings and functions of the English used in India. After the CCIE was compiled, sections of ICE India, which was compiled at the same time as the CCIE was, were added to the CCIE for the current study to make the corpus more representative of the English used in India. The sections of ICE combined with the CCIE are discussed below.

The design of the CCIE was greatly influenced by comparable large-scale corpora (prior to the development of ICE) that were developed for the quantitative investigation of linguistic characteristics of different varieties of language in different settings. Examples of existing corpora that have influenced the design of the CCIE include the Brown Corpus, the LOB corpus, and the Kohlapur corpus of Indian English. New registers that are absent in other corpora, such as Correspondence, differentiate the corpus from several existing corpora; further, it includes a substantial spoken component while other corpora (previous to ICE) do not. Also, given the fact that English in India is not homogeneous, with both the written and the spoken components, it was important to get as varied a population of contributors as possible. Thus the contributors for the CCIE range from students to drivers, from store keepers and housewives to professional writers and journalists. Table 2 below shows all the registers of the CCIE and provides the word counts of the sub-registers.

Table 2. The corpus of contemporary Indian English and its registers

	Registers	**Sub-registers**	**# of files**	**# of words**
News	Written News	Editorials	29	101,759
		Features	42	92,800
		Regional News	64	142,375
		Business News	36	110,612
			TOTAL: 171	**TOTAL: 447,546**
	Spoken News	Spoken News	12	45,304
		Spoken Political Discussions	5	12,770
			TOTAL: 17	**TOTAL: 58,074**
Academic English	Spoken Academic English	Office Hours	1	4,041
		Oral Presentations	2	11,765
		Lectures	4	19,678
			TOTAL: 7	**TOTAL: 35,484**

(*Continued*)

Table 2. The corpus of contemporary Indian English and its registers (Continued)

	Registers	Sub-registers	# of files	# of words
Conversational English	Conversational English	Conversation	9	65,324
		Oral Interviews	28	54,577
		Service Encounters	21	15,404
		Interviews	26	94,404
		Spoken Entertainment	3	4,203
			TOTAL: 87	**TOTAL: 233,912**
Fiction	Fiction	Indian Fiction	27	95,993
		English Fiction	36	95,804
			TOTAL: 63	**TOTAL: 191,797**
Entertainment	Written Entertainment News	Written Entertainment News	38	86,378
			TOTAL: 38	**TOTAL: 86,378**
Correspondence	Business Correspondence	Letters to the Editor	59	23,339
		Dear Abby Letters	72	45,219
			TOTAL: 131	**TOTAL: 68,558**
	Personal Correspondence	Emails	25	24,340
			TOTAL: 25	**TOTAL: 24,340**
Sports	Written Sports	Written Sports News	19	54,823
			TOTAL: 19	**TOTAL: 54,823**
	Spoken Sports	Spoken Sports Reporting	2	5,713
			TOTAL: 2	**TOTAL: 5,713**
Travel	Written Travel News	Written Travel News	6	81,393
			TOTAL: 6	**TOTAL: 81,393**
TOTAL FILES IN CCIE: 566				
TOTAL WORD COUNT: 1,288,018				

2.2 Combining CCIE with ICE-India

It is clear from Table 2 above that several registers of the CCIE are small. As mentioned earlier, the CCIE and ICE-India were compiled at about the same time; after compilation of the CCIE, several sections of ICE-India were added to it to

make the smaller registers better represented, as well as to make the corpus better representative of the range and functions of the English used in India. The following section outlines the registers of ICE-India that were added to the CCIE. No details are provided about the compilation of ICE-India, as these details are available in Greenbaum (1996).

Registers enlarged by files from ICE-India:

1. Spoken News
   a. Files S2B-001 to S2B-020 were added. This added a word count of 40,000.
2. Spoken Academic English
   a. Files S1B-001 to S1B-020 (Class lessons) were added. This added a word count of 40,000.
3. Written Academic English
   a. Files W2A-001 to W2A-010 (Humanities);
   b. W2A-011 to W2A-020 (Social Sciences);
   c. W2A-021 to W2A-030 (Natural Sciences);
   d. W2A-031 to W2A-040 (Technology);
   e. W1A-001 to W1A-010 (Unpublished Student essays);
   f. W1A-011 to W1A-020 (Unpublished Examination scripts)

   This added a word count of 120,000.
4. Correspondence
   a. Files W1B-001 to W1B-015 (Social Letters) were added to Personal Correspondence. This added a word count of 30,000 words.
   b. Files W1B-016 to W1B-030 (Business Letters) were added to Business Correspondence. This added a word count of 30,000 words. It is important to note that these business letters were both published and unpublished, with the majority being unpublished, and therefore, possibly less formal.

Table 3 shows the registers of the combined CCIE and ICE-India used for this study as well as the word counts for each register. As evident from Table 3 below, the only register that is substantially smaller than the others is Spoken Sports Reportage.

2.3 Features analyzed in the current study

The following section describes the two grammatical features analyzed in the current study.

2.3.1 *Absence of subject-auxiliary inversion in WH-question formation*

The absence of subject-auxiliary inversion is frequently mentioned in early literature on features of Indian English, although previous literature mentions questions in general, rather than WH-questions specifically. The current analysis

Table 3. Combined corpus

Register	# of files	Word count	Total word count for register
Written News	171 files from CCIE	447,546	447,546
Spoken News	17 files from CCIE	58,074	98,074
	20 files from ICE	40,000	
Written Academic English	60 files from ICE	120,000	120,000
Spoken Academic English	7 files from CCIE	35,484	75,484
	20 files from ICE	40,000	
Conversational English	87 files from CCIE	233,912	233,912
Fiction	63 files from CCIE	191,797	191,797
Written Entertainment	38 files from CCIE	86,378	86,378
Business Correspondence	131 files from CCIE	68,558	98,558
	15 files from ICE	30,000	
Personal Correspondence	25 files from CCIE	24,340	54,340
	15 files from ICE	30,000	
Written Sports News	19 files from CCIE	54,823	54,823
Spoken Sports Reportage	2 files from CCIE	5,713	5,713
Travel Writing	6 files from CCIE	81,393	81,393
TOTAL WORD COUNT FOR SPOKEN CORPUS: 413,183 WORDS			
TOTAL WORD COUNT FOR WRITTEN CORPUS: 1,134,835 WORDS			
TOTAL WORD COUNT FOR ENTIRE CORPUS: 1,548,018 WORDS			

focuses exclusively on WH-questions since an absence of subject-auxiliary inversion in yes-no questions is common in other international varieties of English, even traditional native varieties such as British or American English. It is not uncommon, then, to hear questions like "You are on your way?" with rising intonation indicating that it is a question. Whether such constructions are more common in registers of Indian English than they are in registers of British or American English is an interesting one, but one that will not be addressed in the current analysis.

The current analysis focuses on the absence of subject-auxiliary inversion in WH-questions. Due to the potentially huge number of questions to examine in the entire corpus, this analysis is restricted to just the spoken corpus and unpublished Written Academic English. The perhaps surprising addition of this small section of this written register to the analysis is because of results obtained in a previous study (Balasubramanian 2009), where it was suggested that Indian features seemed to occur with a greater than expected frequency in the English produced by younger users of Indian English. As described in Greenbaum (1996), contributors to the academic English sections of the corpus were among the youngest contributors, being university students at the undergraduate level. Further, all the files analyzed in this study in this register consist of unpublished writing.

For the current analysis, Indian variants for this study would, therefore, include constructions such as the following:

Where you are going?
Who you are going out with?

2.3.2 *Circumstance adverbials "also" and "only"*

This investigation of the circumstance adverbials *also,* and *only* includes two separate analyses. The first analysis of circumstance adverbials deals with determining the positions of the two circumstance adverbials in sentences, i.e. medial position versus initial or final position. Biber et al. (1999) claim that in British and American English, circumstance adverbials show a marked preference for the medial position in a sentence, and with this analysis, I wished to determine if this preference is true for circumstance adverbials in question in registers of Indian English, too.

The second analysis deals with determining the relationship between the position of the circumstance adverbials *also* and *only in* a clause and the part of the clause that is semantically in focus. Biber et al. (1999) claim that while prescriptive grammar dictates that the circumstance adverbials be placed immediately before the element in the clause that is semantically in focus, in reality, this is frequently not the case. However, they explain that the position of the circumstance adverbial is important in providing meaning to the sentence. This analysis, then, deals with determining how frequently circumstance adverbials occur in positions other than immediately preceding the focused element in registers of Indian English, and whether this contributes to ambiguity in the meaning of the sentence.

3. Results

The following sections present the results of the analyses conducted for this study. First is a discussion of the distribution of the circumstance adverbials in initial, medial, and final positions across registers of the corpus. Next is a discussion of

the circumstance adverbials and the part of the sentence/clause they occur in that is semantically in focus. Finally, the results of WH-questions with no subject-auxiliary inversion across registers of Indian English are presented.

3.1 Results on position of circumstance adverbials in initial, medial, final positions

As mentioned previously, Biber et al. (1999) show that in British and American English, additive and restrictive adverbials show a "marked preference" for the medial position. For the current analysis, the proportions of the adverbials *also* and *only* in the medial position versus in the initial or final positions were determined. The following section describes the distribution of these circumstance adverbials in different positions in different registers of spoken and written Indian English.

3.1.1 *Also*

Table 4 on page 160 shows that there are considerable differences between the different registers of Indian English with respect to the positions of the circumstance adverbial *also*. A pattern that emerges with both written registers and spoken registers is that the more informal and unscripted the language in the register, the greater the tendency for the circumstance adverbial not to be in medial position (the favored position in American and British English, as described in Biber et al. 1999). Thus Conversational English and Spoken Academic English (not necessarily an informal register, but one (in this corpus) where the language is almost entirely unscripted) have high proportions of the circumstance adverbial occurring in final and initial positions. While Spoken Sports does show a high frequency of *also* in final position, I will not comment on this finding, given the very small number of adverbials examined in this register for this analysis. Example sentences from various sub-registers include the following:

1. That's the feeling I got also. (Conversation 9)
2. The owner took him soup and bread, but he's not eating also. (Conversation 2)
3. …it may be possible by emails also. (Interviews 3)
4. So you must be looking forward to this Silver Jubilee lunch also. (Miscellaneous 1)
5. So many times we did night work also. (Oral Interview 10)
6. It's a good environment, good friends, and good environment also. (Oral Interview 13)
7. We have one like this also. (Service Encounters 9)
8. They've got TV and phone also. (Service Encounters 15)

9. It is muddy state and also it is mobile state. (S1B 001)
10. Also most of the nitrates seem to be released. (S1B 013)

Among the written registers, this pattern is most apparent in Personal Correspondence (PC) and Business Correspondence (BC). Of all the written registers, these are the most informal (see a description of ICE India to determine how Business Correspondence has less formal language than the other written registers do). Examples of sentences with initial *also* from these two registers include the following:

11. Also please inform them that Varsha Pendse is no longer at TekEdge. (PC: Email13)
12. Also I would like to know if you have received my previous mail. (PC: Email15)
13. Also I wanted to do my engineering, but am forced to do a BA. (BC: FdearAb11)
14. Also my molars are very deformed and one is totally embedded in the gum. (BC: FdearAb19)

3.1.2 *Only*

The pattern we saw for the occurrence of *also* in non-medial position across registers of Indian English also hold more or less true for the distribution of *only* across the registers. As expected, Conversational English has the highest proportion of non-medial occurrences, with 12.4% occurring in the initial position and 35.2% occurring in the final position. Further, as determined with *also*, Spoken Academic English had 14.2% in the initial position and 9.46% in the final position. An interesting finding with *only* was the distribution of the adverbial in Spoken News, which had 15.8% in initial position and 9.86% in final position. Upon returning to the files in this register and examining the non-medial occurrences of the adverbial, I noticed, however, that the majority of the non-medial positions occurred in non-scripted language.

With the written registers, Personal Correspondence had 12.1% in initial position and 7.6% in final position. The interesting finding with the written register was with Written Academic English, in which 10.3% and 11.3% of only studied occurred in the initial and final position respectively.

Examples of sentences with non-medial *only* in spoken registers include the following:

1. That means it must have been here only. (Conversation9)
2. I don't take water outside. I carry my bottle only. (Conversation2)
3. It will happen like that only. (Oral Interview3)
4. All over India is like this only. (Oral Interview4)
5. Then just IT has to pay it back only. (Miscellaneous1)

6. Victim of burns can get algumin only. (S2B-004)
7. Only Alex Stuart showed good form. (S2B-015)
8. Only through coins we came to know the existence of… (S2B-022)

Examples of sentences with non-medial only in written registers include the following:

1. So she was kept in the house only. (W1A-001)
2. They grow important trees only. (W1A-012)
3. …and men should work to earn money only… (W1A-004)
4. It will not depend upon the ruling party only. (W1A-005)
5. …dated 15/03/94 reached me today only. (W1A-009)
6. I am likely to be in India for about three weeks only. (W1A-015)
7. The bills should be for that period only. (Email16)
8. Thanks, right now I am in Bangalore only. (Email17)

Table 4. *Also*: Distribution across registers of Indian English

	Register	Total *also*	# of medial *also*	% of medial *also*	# of initial *also*	% of initial *also*	# of final *also*	% of final *also*
Spoken	Conversational English	569	316	55.3%	41	7.2%	212	37.3%
	Spoken Academic English	251	173	68.9%	15	6.0%	63	25.1%
	Spoken News	424	370	87.3%	8	1.9%	46	10.8%
	Sports (Spoken)	6	4	66.7%	0	0	2	33.3%
Written	Written News	1,044	1,002	95.9%	31	2.9%	11	1.1%
	Written Academic English	400	351	87.8%	24	6%	25	6.3%
	Fiction	184	159	86.4%	5	2.7%	20	10.9%
	Written Entertainment News	143	129	90.2%	14	9.8%	0	0
	Business Correspondence	200	143	71.5%	54	27%	3	1.5%
	Personal Correspondence	137	95	69.3%	31	22.6%	11	8.0%
	Sports (Written)	96	90	93.7%	4	4.2%	2	2.1%
	Travel	358	308	86%	34	9.5%	16	4.5%

Table 5. *Only*: Distribution across registers of Indian English

	Register	Total *only*	# of medial *only*	% of medial *only*	# of initial *only*	% of initial *only*	# of final *only*	% of final *only*
Spoken	Conversational English	579	303	52.3%	72	12.4%	204	35.2%
	Spoken Academic English	169	129	76.3%	24	14.2%	16	9.46%
	Spoken News	152	113	74.3%	24	15.8%	15	9.86%
	Sports (Spoken)	5	5	100%	0	0	0	0
Written	Written News	668	620	92.8%	35	5.2%	13	1.9%
	Written Academic English	194	136	70.1%	20	10.3%	22	11.3%
	Fiction	362	313	86.5%	43	11.8%	6	1.6%
	Written Entertainment News	111	104	93.7%	5	4.5%	2	1.8%
	Business Correspondence	180	163	90.5%	7	3.88%	10	5.55%
	Personal Correspondence	66	53	80.3%	8	12.1%	5	7.6%
	Sports (Written)	70	70	100%	0	0	0	0
	Travel	6	6	100%	0	0	0	0

3.2 Circumstance adverbials and focus

The second analysis of circumstance adverbials, as explained earlier, deals with determining the relationship between the position of the circumstance adverbials *also* and *only in* a clause and the part of the clause that is semantically in focus The following sections first discuss *also* and focus, followed by a discussion of *only* and focus across the registers of Indian English.

3.2.1 *"Also" and focus*

This section outlines the results on the relationship between the position of *also* and the element in the sentence or clause that is semantically in focus. Biber et al. (1991) explain that although prescriptive grammar says that *also* should immediately precede the element that is semantically in focus (which would lead to no ambiguity of meaning), this is frequently not the case. They add that "unlike many other adverbials, these cannot be moved without affecting their meaning in the

clause. The position of the adverbial is important in determining what element of the clause is the focus of the addition or restriction" (p. 781).

Table 6 on page 163 presents the results of the analysis of the frequency with which *also* does occur immediately before the element in focus, and the frequency with which it doesn't across spoken and written registers of Indian English. Table 6 also provides several example sentences from all the spoken and written registers. We see from Table 6 below that Conversational English is the most different from both the other registers in Indian English and from British and American English with 77.9% of the *also* not preceding the focused element in the sentence or clause.

An examination of the example sentences reveals that the position of the circumstance adverbial does make the meaning of the clause unclear. The following example illustrates this:

15. And my brother-in-law also is lecturer. (Oral Interview 22)

In this sentence, it is unclear whether the speaker means that her brother-in-law is a lecturer in addition to something else or whether her brother-in-law, in addition to someone else, is a lecturer.

Spoken Academic English had a number of examples of also not preceding the focused element, as is indicated by the following example:

16. Sound is also digitized before transmitting. (Lecture 3)

The sentence is a clear example of the lack of clarity that the position of the circumstance adverbial can contribute to the sentence. The focus of the sentence, based on the position of *also*, should be *digitized*. In other words, the sentence should mean that more than one thing happens to the sound before it is transmitted – it is digitized and something else. However, the actual focus of the sentence is *sound* (and this was made clear by studying a larger stretch of discourse). In other words, the speaker means that two things are digitized before they are transmitted, *sound*, and something else.

With the written registers, Fiction and Written Academic English had the highest proportion of *also* not preceding the focused element. These two registers are followed by the two Correspondence registers. With Written Academic English, it is interesting to note that almost all the examples of *also* not preceding the focused element come from unpublished papers. With Fiction, most examples come from representations of dialog, indicating that the feature more prevalent in how conversation is imagined in fiction. Example sentences from all these registers, sentences where, once again, the meaning is less clear to a non-Indian audience by the position of the adverbial are included in Table 6.

Table 6. *Also*: Focus

	Register	# of *also*	# of *also* not preceding focused element	% of *also* not preceding focused element	Examples
Spoken	Conversational English	569	443	77.9%	– He makes *chaat* also. (Conversation2) – I think mummy sent me chutney powder also. (Conversation4) – No I don't like. After marriage also I want to be like this only. (Oral Interviews11) – And my brother in law also is lecturer. (Oral Interview22) – Actually now studying B. Com also nobody is encouraging. (Oral Interview13) – Even my uncle also, he was in the middle. (Service Encounters6) – I have not been eating anything outside also. (Service Encounters2) – In Orissa also the BJP has problems with the Biju Janata Dal. (Rint13) – And confrontation also is inevitable. (Rint14)
	Spoken Academic English	251	64	25.5%	– Sound also is digitized before transmitting. (Lecture3) – You know in the transmitter circuit also you can have problems. (Lecture3) – So I feel there could be more than this point also, but I concentrates on this point. (Miscellaneous3) – I've seen this in my previous companies also. (Miscellaneous2) – Your father is also a scientist? (Office Hours1) – Even that also can be called as industry. (S1B-008) – …rules of society also is there. (SiB-017)
	Spoken News	424	88	20.8%	– I mean even growth funds also have the same amount of transparency. (News11) – Were the Indian intelligence agencies involved in bringing Abdul Majid Dhar across from Pakistan so that his ceasefire also could have been made on Indian soil? (Politics5) – The reply which has been given is the next year, next session also I will put the same question, same reply will come. (Politics3)
	Spoken Sports Reportage	6	3	50%	– I think that makes all the difference, so the encouragement also, from whatever level you have… (Sports1) – Ronny Ronny, because of his erratic potting, was a little erratic also. (Sports1)

(*Continued*)

Table 6. *Also*: Focus (Continued)

	Register	# of *also*	# of *also* not preceding focused element	% of *also* not preceding focused element	Examples
Written	Written News	1,046	54	5.2%	– As and when the laws of the country allow for depository instructions to be accepted electronically, we will offer that facility also. (DHBus3) – Russia also has been playing this role for a long time. (Hed2) – The only Indians with empire-building proclivities also are eying this oil for their great refinery… (DCEd3) – CAD/CAM accelerates the entire process, reducing its errors also. (DHf1) – Gecko, frogs, and lizards all reside in their cavities and wasp and stingless bees also live there. (DHf1) – Mr. Byre Gowda said that letters were being sent to them also. (Dhreg16) – Power has been delegated to the lower command also. (Dhreg17)
	Written Academic English	400	145	36.3%	– Now also she is not given the superior position. (W1A-011) – Population also goes on increasing… (W1A-001) – These measures also had helped in augmenting savings. (W2A-014)
	Fiction	184	79	42.93%	– Yes, she also went with them. (IFBatti) – When she understood the joke, she also started laughing. (IFBracelet) – But a mistake can be rectified also! (Daughter)
	Written Entertainment News	143	4	2.79%	– *Kuch kuch hota hai* is also one of my favorite films. (DHEnt5) – I had the privilege of working with Meena Kumari in many films. She also hailed from my hometown. (DHEnt14)
	Business Correspondence	200	50	25%	– Some of his friends also seem to be misleading him. (FdearAb22) – The same analogy holds good especially for Indian administration also. (DHMail18) – They did this because they knew that Jinnah was dying and they knew also that if he died before independence… (Hmail6) – Your packings, packagings have won award also. (W1B-016) – Now also when enquired, same answer is given. (W1B-022)

(*Continued*)

Register	# of *also*	# of *also* not preceding focused element	% of *also* not preceding focused element	Examples
Personal Correspondence	137	35	25.5%	– Also we had it in our records that Rajinder Singh was supposed to arrive this weekend also. (Email16) – I got your mail id from Yassar who is my colleague at Infosys. He also has applied to your consultancy. (Email19) – By the way, Sindhu is also working in Biocon, but I haven't met her for the past few days. (Email21) – My wife also is required to go to Pune, Kolhapur, etc. (W1B-002) – This is the attitude also toward those who are going to retire. (W1B-007)
Written Sports News	96	4	4.16%	– The PCB interim chief said the Pakistani team will also participate in the tri-series in Australia. (DHSp1) – Railways controlled the match in the second half also. (DHSp1)
Travel Writing	358	42	11.73%	– But our *khaana* is too good. People come only for our food. The rates also are reasonable. (REt2) – The chicken *farcha* is also Rs.46. (Ret2) – And you may order the bread also with them, as also later with the rest of your meal. (Ret2) – OTDC-owned guesthouses also are there. (ITGt1)

3.2.2 *"Only" and focus*

Table 7 on page 167 shows the frequency of the restrictive adverbial *only* not immediately preceding the focused element in the sentence. As Lange (2007) pointed out, in registers of Indian English, *only* also has non-contrastive focus marking properties. Lange explains that in addition to focus marking, words like *only* and *itself* are used as reflexive pronouns or intensifiers. The current analysis does not focus on the different uses of *only*. Rather, it is restricted to a description of the proportion of *only* preceding and not preceding the focused element. The example sentences in Table 7 illustrate that the meaning of the sentence is indeed unclear, irrespective of whether *only* is used as a non-contrastive focus marker, or an intensifier or reflexive marker. The current analysis, then, focuses on how frequently *only* is used in registers of Indian English in a manner that is different from how it occurs in registers of British or American English.

3.3 WH-questions without subject-auxiliary inversion

As mentioned earlier, results of an earlier study (Balasubramanian 2009) prompted an analysis of WH-questions in the spoken registers as well as unpublished Written Academic English for the current study. Table 8 on page 169 shows the results of the current analysis. Conversational English has a large number of WH-questions (29.2% of the questions studied) without subject-auxiliary inversion, and this finding is not surprising. The other spoken registers follow expected patterns, with fewer questions with no subject-auxiliary inversion than Conversational English. As with the previous analysis, in Spoken Academic English, 7.65% of the questions studied lacked subject-auxiliary inversion. The most interesting finding, however, was in Written Academic English (unpublished student writing), where four of the ten questions studied lacked subject-auxiliary inversion. Although this is a small sample, it is worth noting that all the examples came from different files, showing that the findings are not idiosyncratic to a particular user. An added note here: as a matter of interest, files in the published academic writing sections of ICE-India were also analyzed. It is interesting to note that in the 80,000 words that published writing supplied to this register, there were no instances of WH-questions with a lack of subject-auxiliary inversion.

One limitation of the current study is the small number of WH-questions studied in this register; in all the files studied in unpublished Written Academic English, there were only 10 WH-questions. Of these, 4 showed a lack of subject-auxiliary inversion. All the examples, however, came from different files, showing that the findings are not idiosyncratic to a particular user.

Table 7. *Only* and focus

	Register	# of *only*	# of *only* not preceding focused element	% of *only* not preceding focused element	Examples
Spoken	Conversational English	578	111	19.2%	– Remember the day before yesterday you only bought. (Conversation9) – Now Joshua only talks for a longer time. (Conversation2) – No he didn't come, that day only he came. (Conversation2) – Tom is there only in America, no? (Conversation3) – No, it is, you know, thirty, forty kilometers only, and we can go and do that. (Conversation6) – He won't come here, from starting only, he is doing there only. (Oral Interviews3) – We were supposed to shift long back only. (Oral Interviews8) – Yesterday I made avial in that only. (Service Encounters4) – Before that and after that there is single road only. (Service Encounters6) – This is regular armhole cut only. (Service Encounters11)
	Spoken Academic English	190	39	20.5%	– It is based on the principle of drill only. (Lecture5) – Can there be a structure with verb only? (Lecture5) – It is limited to the abusive context only. (Lecture3) – Most of the people are from rural only. (Miscellaneous2) – Again that would be the procedure only. (S1b-003) – It reacts with the oxygen two times only. (S1b-004) – It is by downward displacement of water only. (S1b-016)
	Spoken News	152	12	7.9%	– Six took place in Nainital district only. (News4) – They are wearing cotton casuals only. (News11) – It is question regarding UP only. (Politics1) – Victim of burns can get algumin only. (S2b-004) – It will however be made available for two weeks only. (S2b-013)
	Spoken Sports Reportage	4	0	0	

(*Continued*)

Table 7. *Only* and Focus (Continued)

	Register	# of *only*	# of *only* not preceding focused element	% of *only* not preceding focused element	Examples
Written	Written News	668	24	3.6%	– It will be restricted to 50 million only. (HBus10) – It is priced at Rs. 159 only. (STf5)
	Written Academic English	194	18	9.27%	– So she was kept in the house only. (W1A-011) – Therefore individual only suffer the loss. (W1A-016) – Men should work to earn money only. (W1A-004) – It will not depend on the ruling party only. (W1A-005)
	Fiction	362	9	2.48%	– Yes, she ended up to save herself in self-defense only. (EFDouble) – He waited for their turn about two hours only. (EFLamb) – It will run its prescribed course only. (IFHarambe) – I'm talking about life only. (IFLife)
	Written Entertainment News	111	3	2.7%	– It is constricted to the Mumbai district only. (FEnt18)
	Business Correspondence	173	6	3.46%	– It is exclusively for hike in diesel price only. (DHmail8) – It is limited to actual income only. (W1b-018)
	Personal Correspondence	66	6	9.09%	– She only conducted interview for me with Mittal. (Email16) – Thanks, I am right now in Bangalore only. (Email17) – It reached me today only. (W1b-009) – I am likely to be in India for three weeks only. (W1b-015)
	Written Sports News	70	0	0	
	Travel Writing	86	5	5.8%	– This piece is on Noorani's pulaos and biriyanis only. (Ret1)

Table 8. WH-Questions without subject-auxiliary inversion across registers of Indian English

Register	# of *WH-Questions*	# of *WH-Questions without Subject-verb inversion*	% of *WH-Questions without Subject-verb inversion*	Examples
Conversational English	339	99	29.2%	– Why people call wife better half? (Conversation9) – How many things you have made? (Conversation2) – What Sunil sent? (Conversation2) – Where you are going? (Oral Interview1) – Why she wants to stay there? (Oral Interview1) – How much this is? (Service Encounters9) – What madam wants? (Service Encounters6) – What you would like to see, madam? (Service Encounters6)
Spoken Academic English	379	29	7.65%	– Then why the name has been given as aorta? (S1b-003) – Now how it smells? (S1b-004) – What word he has used for that? (S1b-005) – Then how this industrial age started? (S1b-008) – Why the author showed the sympathy of the grasshopper? (S1b-012) – How a single animal cell looks like? (S1b-015) – Why a magnet gains the property of attraction? (S1b-019)
Spoken News	165	6	3.6%	– How the money can be utilized by the state government? (Politics4) – Where they will come? Where they will stay? (Politics3)
Sports (Spoken)	1	0	0	
Unpublished Written Academic English	10	4	40%	– How that particular problem will be solved? (W1a-016) – When we are getting rid of this religion? (W1a-003)

4. Discussion

The results of this study clearly indicate that Indian English varies internally, just as British and American English do. Further, the study shows that register is most certainly a source of internal variation. As might be expected, Conversational English had the highest frequency of Indian features:

1. 44.7% of the *also*s studied occurred in initial or final position, as opposed to the favored medial position in British or American English;
2. 47.7% of the *only*s examined occurred in initial or final positions;
3. 77.9% of the *also*s studied did not precede the focused element in the sentence or clause;
4. 19.2% of the *only*s studied did not precede the focused element in the sentence or clause;
5. 29.2% of the WH-questions studied lacked subject-auxiliary inversion

It is clear from these findings, therefore, that there is truth to the idea that innovations start in the informal unscripted registers and move from there to other more formal registers. This is also borne out by the fact that in the current study, Personal Correspondence, a relatively informal written register, did have higher frequencies of Indian features than other written registers did. Moving away from issues of formality, however, the most interesting finding in this study was, perhaps, the relatively high frequency of Indian features in Spoken and Written Academic English. As mentioned earlier, this finding might be considered surprising, given the assumed formality of academic English. It is also important to mention here that within Written Academic English, all the examples of sentences with Indian usage came from the unpublished student work (see Section 2.1.1 for details about the files in this register), and not the published written files. A factor that possibly accounts for this surprising finding, and one that most definitely warrants further exploration, is the fact that of all the corpus contributors, those that contributed to the spoken and written academic English registers were the youngest. They were mostly university students between the ages of 18 and 22, while the other contributors were older than 30. Further, one can also assume (based on these results) that spoken (and unscripted) Indian English is not evaluated based on the external British or American norms that written Indian English is. This finding, naturally, warrants further investigation.

The results of this study, then, possibly suggest that the younger the user of Indian English, the greater the frequency of Indian features, particularly within the spoken registers. This result is interesting, as it suggests that the innovations that originated in informal conversational English (where, one might argue, the

identity of Indian English begins to form), they are spreading to the more formal registers by younger users of the language, those who grew up with an Indian variety of English as the norm. This finding was suggested in an earlier study (Balasubramanian & Balasubramanian 2012). This "adoption and acceptance of an indigenous linguistic norm, supported by a new locally rooted self-confidence" (Schneider 2003: 249) indicates that the English in India has at least entered the Endonormative Phase, at least in the spoken registers.

5. Conclusion and implications for future research

Based on the results of this study, it is clear that the English used in India is far from homogeneous; rather, it represents a cluster of varieties. It is also clear that one source of variation is register – with more formal registers (usually written) featuring fewer Indian features and the less formal ones featuring more Indian features. In addition to register, however, the age of the user of Indian English also seems to be a variable in determining the Indianness of the Indian English. Based on the results of the analyses described in this study, one could possibly go so far as to say there seem to be three clusters of users of English in India: the first includes educated speakers of Indian English, those who use a variety of English, both spoken and written, which most closely resembles more traditional "native" varieties. Lange (2007) claimed that the proficient speaker-users of English are those who grew up with a definite external norm, and whose English corresponds to Schneider's first two phases – Foundation and Exonormative Stabilization. As was pointed out by Balasubramanian and Balasubramanian (2012), the Indianness of the English in this educated group of English users is more likely due to phonological features than to lexical or grammatical features. As shown by the current analyses, English users in this first cluster are represented in the more formal written registers (like Written News) in addition to the more formal spoken registers (like the scripted Spoken News).

Next is a group of English users who are in transition; while many still have an external norm, the English they use is becoming more Indianized (Schneider's Nativization), to use Kachru's (1983) term. The registers represented in this study with speakers and writers in this group include Business Correspondence, Written Entertainment, Written Sports News, and Written Travel News.

The last group of English users represents younger users of Indian English, those whose English has definitely become Nativized, and is further evolving into a more distinct and differentiated varieties – i.e. a variety which shows internal variation, much as any more traditionally "native" variety does. Further research is needed to determine just how standardized the Indian features become in the

spoken and written English of this last group of Indians, the young Indians who are growing up with their own internal norm. Whether there is a certain norm for spoken registers that is more tolerant of Indian features, and another norm for written registers that is still a more traditional "native" speaker norm needs to be determined with further research, particularly research based on diachronic corpus data.

This new model of different groups of users of Indian English, while constructed based on the study of variation in the English used in India, could serve as a model for enhancing our understanding of the principles that underlie language variation, contact, and change.

References

Agnihotri, Rama & Khanna, A.L. 1984. Evaluating the readability of school textbooks: An Indian study. *Journal of Reading* 35(4): 282–88.

Ahulu, Samuel. 1995. Hybridized English in Ghana. *English Today* 11(4): 31–36. DOI: 10.1017/S0266078400008609

Bakshi, Raj. 1991. Indian English. *English Today* 7(3): 43–46. DOI: 10.1017/S0266078400005757

Balasubramanian, Chandrika. 2009. *Register Variation in Indian English* [Studies in Corpus Linguistics 37]. Amsterdam: John Benjamins. DOI: 10.1075/scl.37

Balasubramanian, Chandrika, & Balasubramanian, Tyagaraja. 2012. More on concentric circles: A new framework for analyzing variation within new varieties of Englishes. *The Journal of English as an International Language* 7(2): 25–56.

Bamiro, Edmund. 1995. Syntactic variation in West African English. *World Englishes* 14: 189–204. DOI: 10.1111/j.1467-971X.1995.tb00349.x

Banjo, Ayo. 1997. Aspects of the syntax of Nigerian English. In *Englishes Around the World,* Vol. 2: *Caribbean, Africa, Asia, Autralasia. Studies in Honor of Manfred Görlach* [Varieties of English around the World G19], Edward Schneider (ed.), 85–95. Amsterdam: John Benjamins. DOI: 10.1075/veaw.g19.10ban

Bansal, R.K. 1976. *The Intelligibility of Indian English* [Monograph Series 4]. Hyderabad: Central Institute of English and Foreign Languages.

Barbe, Pauline. 1995. Guernsey English: A syntax exile? *English World Wide* 16: 1–36. DOI: 10.1075/eww.16.1.02bar

Bauer, Laurie. 1989. The verb *have* in New Zealand English. *English World Wide* 10: 69–83. DOI: 10.1075/eww.10.1.05bau

Baumgardner, Robert. 1996. Pakistani English: Acceptability and the norm. *World Englishes* 14: 261–271. DOI: 10.1111/j.1467-971X.1995.tb00355.x

Biber, Douglas & Finegan, Edward. 1991. On the exploitation of computerized corpora in variation studies. In *English Corpus Linguistics,* Karen Aijmer & Bengt Altenberg (eds), 204–220. London: Longman.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bolton, Kingsley, Graddol, David & Meiercord, Christiane. 2011. Toward developmental world Englishes. *World Englishes* 30(4): 459–480. DOI: 10.1111/j.1467-971X.2011.01735.x

Bolton, Kingsley. 2006. World Englishes Today. In *The Handbook of World Englishes,* Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds), 240–269. Oxford: Blackwell. DOI: 10.1002/9780470757598.ch15

Coelho, Gail. 1997. Anglo-Indian English: A nativized variety of Indian English. *Language in Society* 26: 561–589. DOI: 10.1017/S0047404500021059

Crystal, David. 2003. *English as a Global Language*. Cambridge: CUP. DOI: 10.1017/CBO9780511486999

Davidova, Julia. 2012. Englishes in the outer and expanding circles: A comparative study. *World Englishes* 31(3): 366–386. DOI: 10.1111/j.1467-971X.2012.01763.x

D'Souza, Jean. 1997. Indian English: Some myths, some realities. *English World Wide* 18: 91–105. DOI: 10.1075/eww.18.1.05dso

Gargesh, Ravinder. 2006. South Asian Englishes. In *The Handbook of World Englishes,* Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson (eds), 90–113. Oxford: Blackwell. DOI: 10.1002/9780470757598.ch6

Gisborne, Nikolas. 2000. Relative clauses in Hong Kong English. *World Englishes* 19: 357–371. DOI: 10.1111/1467-971X.00184

Greenbaum, Sidney. 1996. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

Hosali, Priya. 1991. Some syntactic and lexico-semantic features of an Indian variant of English. *Central Institute of English and Foreign Languages Bulletin* 3(1–2): 65–83.

Hosali, Priya. 1992. Function reduction in Butler English. *Indian Journal of Applied Linguistics* 18: 59–70.

Huber, Magnus. 1995. Ghanaian Pidgin English: An overview. *English World Wide* 16: 215–249. DOI: 10.1075/eww.16.2.04hub

Hundt, Marianne & Gut, Ulrike. 2012. *Mapping Unity and Diversity World-Wide: Corpus-Based Studies of New Englishes* [Varieties of English around the World G43]. Amsterdam: John Benjamins. DOI: 10.1075/veaw.g43

Jenkins, Jennifer. 2003. *World Englishes.* New York NY: Routledge.

Kachru, Braj. 1976. Indian English: A sociolinguistic profile of a transplanted language. Eric Document # ED132854.

Kachru, Braj. 1983. *The Other Tongue*. Oxford: Pergamon.

Kachru, Braj, Kachru, Yamuna & Nelson, Cecil (eds). 2006. *The Handbook of World Englishes*. Oxford: Blackwell.

Kachru, Braj. 1994. Englishization and contact linguistics. *World Englishes* 13: 135–154. DOI: 10.1111/j.1467-971X.1994.tb00303.x

Kachru, Braj.1976. *Indian English: A Sociolinguistic Profile of a Transplanted Language.*

Kachru, Braj.1969. The Indianness of Indian English. *Journal of the International Linguistic Association* 21(2): 391–423.

Kallen, Jeffrey. 1989. Tense and aspect categories in Irish English. *English World Wide* 10: 1–39. DOI: 10.1075/eww.10.1.02kal

Kirkpatrick, Andy. 2010. *The Routledge Handbook of World Englishes*. London: Routledge.

Lange, Claudia. 2007. Focus marking in Indian English. *English World Wide* 28(1): 89–118. DOI: 10.1075/eww.28.1.05lan

Leitner, Gerhard. 1991. The Kohlapur Corpus of Indian English – Intravarietal description and/or intervarietal comparison. In *English Computer Corpora,* Stig Johansson & Anna-Brita Stenström (eds), 215–232. Berlin: Mouton de Gruyter.

Mazzon, Gabriella. 1993. English in Malta. *English World Wide* 14: 171–208. DOI: 10.1075/eww.14.2.02maz

Mesthrie, Rajend 2010. New Englishes and the native speaker debate. *Language* Sciences 32: 594–610. DOI: 10.1016/j.langsci.2010.08.002

Mesthrie, Rajend & Bhatt, Rajesh. 2008. *World Englishes. The Study of New Linguistic Varieties.* Cambridge: CUP.

Mukherjee, Joybrato & Gries, Stefan T. 2009. Collostructional nativization in New Englishes: verb-construction associations in the International Corpus of English. *English World Wide* 30(1): 27–51. DOI: 10.1075/eww.30.1.03muk

Mukherjee, Joybrato & Hoffman, Sebastian. 2006. Ditransitive verbs in Indian English and British English: A corpus-linguistic study. *AAA, Arbeiten aus Anglistik und Amerikanistik* 32(1): 5–24.

Olavarria de Ersson, Eugenia & Shaw, Philip. 2003. Verb complementation patterns in Indian standard English. *English World-Wide* 24(2): 137–161. DOI: 10.1075/eww.24.2.02ers

Peters, Pam. 2009. The survival of the subjunctive: Evidence of its use in Australia and elsewhere. *English World Wide* 19: 87–103. DOI: 10.1075/eww.19.1.06pet

Sailaja, Pingali. 2009. *Indian English*. Edinburgh: EUP.

Sand, Andrea. 2004. Shared morpho-syntactic features in contact varieties of English: Article use. *World Englishes* 23(2): 281–298. DOI: 10.1111/j.0883-2919.2004.00352.x

Schmied, J. 1994. Syntactic style variation in Indian English. *Anglistentag 1993 Eichstatt: Proceedings*, 217–232.

Schneider, Edgar. 2007. *Postcolonial English: Varieties around the World.* Cambridge: CUP. DOI: 10.1017/CBO9780511618901

Schneider, Edgar. 2004. How to trace structural nativization: Particle verbs in world Englishes. *World Englishes* 23: 227–249. DOI: 10.1111/j.0883-2919.2004.00348.x

Schneider, Edgar. 2003. The dynamics of new Englishes: From identity construction to dialect birth. *Language* 79(2): 233–281. DOI: 10.1353/lan.2003.0136

Schneider, Edgar, Burridge, Kate, Kortmann, Bern, Mesthrie, Raj & Upton, Clive. 2004. *A Handbook of Varieties of Englsih,* Vol. 1: *Phonology.* Berlin: Mouton de Gruyter.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and Change* [Varieties of English around the World G38]. Amsterdam: John Benjamins. DOI: 10.1075/veaw.g38

Setin, Peter. 1997. The English language in Maurutius: Past and present. *English World Wide* 18: 65–89. DOI: 10.1075/eww.18.1.04ste

Sharma, Devyani. 2005. Language transfer and discourse universals in Indian English article use. *Studies in Second Language Acquisition* 27(4): 535–566. DOI: 10.1017/S0272263105050242

Shastri, S.V. 1996. Using computer corpora in the description of language with special reference to complementation in Indian English. In *South Asian English: Structure, Use, and Users,* Robert Baumgardner (ed.), 70–81. Urbana IL: University of Illinois Press.

Shekar, Chandra & Hedge, M.N. 1996. Cultural and linguistic diversity among Asian Indians: A case of Indian English. *Topics in Language Disorders* 16(4): 54–64. DOI: 10.1097/00011363-199608000-00007

Skandera, Paul. 1999. What do we really know about Kenyan English? A pilot study in research methodology. *English World Wide* 20: 217–236. DOI: 10.1075/eww.20.2.02ska

Starks, Donna & Thompson, Laura. 2007. Niuean English: Initial insights into an emerging variety. *English World Wide* 28: 133–146. DOI: 10.1075/eww.28.2.02sta

Verma, Shivendra. 1980. Swadeshi English: Form and functions. *Indian Linguistics* 41(2): 73–84.

Watermeyer, Susan. 1996. Afrikaans English. In *Focus on South Africa* [Varieties of English around the World G15], Vivian de Klerk (ed.), 99–148. Amsterdam: John Benjamins. DOI: 10.1075/veaw.g15.08wat

Youssef, Valerie. 1995. Tense-aspect in Tobogonian English: A dynamic transitional system. *English World Wide* 16: 195–213. DOI: 10.1075/eww.16.2.03you

Zhiming, Bao. 1995. *Already* in Singapore English. *World Englishes* 14: 81–88. DOI: 10.1111/j.1467-971X.1995.tb00348.x

CHAPTER 8

Investigating textual borrowing in academic discourse

The need for a corpus-based approach

Casey Keck
Boise State University

Over the past few decades, corpus-based investigations have contributed greatly to our understanding of academic discourse. One important domain of academic language use, however, has yet to be fully explored from a corpus-based perspective: textual borrowing. Though it is widely recognized that much of what we write in the academy is in some way based upon what has been written before, little is known about when, how often, and in what ways academic writers re-use the language of others. In this paper, I describe my own attempts to provide corpus-based descriptions of student paraphrasing, I highlight the ways in which this research has challenged assumptions about student source text use, and I outline possible directions for future textual borrowing research.

Keywords: Academic discourse; textual borrowing; paraphrasing; corpus linguistics

1. Introduction

Over the past few decades, corpus-based investigations have contributed greatly to our understanding of academic discourse. Large-scale comparisons of general university registers (e.g. Biber et al. 2002; Biber et al. 2004; Conrad 1999; Csomay 2006) have revealed important ways in which the use of grammatical features, multiword phrases, and discourse structures varies according to the communicative demands of particular contexts. Corpus-based studies of disciplinary writing (e.g. Charles 2006; Cortes 2004; Hyland 1999, 2008) have demonstrated that writers' linguistic choices are greatly influenced by the disciplinary values and expectations surrounding such practices as knowledge sharing, inquiry, and argumentation. Researchers have also become increasingly interested in applying corpus-based methods to the

DOI 10.1075/scl.66.08kec

study of L2 writing and academic literacy development, by comparing the ways in which writers of different language backgrounds, levels of proficiency, and educational experience use particular linguistic features in their academic prose (e.g. Altenberg & Granger 2001; Cortes 2004; Hinkel 2002; Hyland 2004; Hyland & Milton 1997; Hyland & Tse 2004; Upton & Connor 2001).

One important domain of academic language use, however, has yet to be fully explored from a corpus-based perspective: textual borrowing. Though it is widely recognized that much of what we say and write in the academy is in some way based upon what has been said or written before, little is known about when, how often, and in what ways academic writers re-use the language of others. This is not to say that textual borrowing has not received attention within the field of applied linguistics. It has received considerable attention, particularly in regards to university student writing and plagiarism (Currie 1998; Flowerdew & Li 2007a, 2007b; Johns & Mayes 1990; Pecorari 2003; Pennycook 1996; Polio & Shi 2012; Liu 2005; Shi 2004, 2006, 2010, 2012; Sowden 2005; Tardy 2010; Yamada 2003). To date, however, few efforts have been made to provide a comprehensive, corpus-based account of the ways in which writers borrow or paraphrase source text language when composing their own written work. Thus, while there has been much debate over how often students copy from source texts, which students are more likely to copy than others, and what "counts" as textual plagiarism, many of our assumptions regarding textual borrowing in academic discourse have yet to be empirically tested.

The present paper highlights ways in which corpus-based methodologies might enrich our understanding of student source text use, using examples from my own research (Keck 2006, 2010, 2014) on university student summarization practices. The paper begins with a review of the previous research on student textual borrowing, and highlights the need for corpus-based studies of student source text use. Following this discussion, the paper highlights ways in which corpus-based research has challenged three common beliefs about student textual borrowing: (1) that L2 writers copy from source text more frequently than L1 writers, (2) that students copy from source texts because they do not understand what they are reading, and (3) that students should be taught how to paraphrase so that they can avoid plagiarism. The paper concludes by outlining possible directions for future textual borrowing research.

2. Student textual borrowing

Because reading plays such a major role in advanced academic writing tasks (Belcher & Hirvela 2001; Carson & Leki 1993; Leki & Carson 1997; Spack 1997, 2004), educators have become increasingly concerned with the ways in which

developing writers attempt to integrate source texts into their writing. This concern has led to a number of recent investigations of student textual borrowing strategies, or instances in which students select a particular excerpt from a source text and either copy the excerpt exactly, or paraphrase the excerpt by making changes to lexis and syntactic structure. Students' inappropriate source text use, in particular, has been the focus of much discussion and debate. Though typically, in the context of higher education, student plagiarism is associated with cheating and dishonesty (Pecorari 2001; Yamada 2003), educators who work with developing writers argue that, for many students, plagiarism represents not an intention to deceive, but rather their developing competence in text-responsible writing (Chandrasoma et al. 2004; Currie 1998; Flowerdew & Li 2007a; Howard 1995; Liu 2005; Pecorari 2003; Polio & Shi 2012; Sowden 2005). In these cases, most educators agree that instances of student copying should be addressed through pedagogy, rather than through disciplinary actions (Casanave 2004; Valentine 2006; Pecorari 2001, 2003). Some have even questioned whether such instances of textual borrowing should be labeled as a type of "plagiarism": Students, language teachers, and university professors have all been found to disagree about what counts as textual plagiarism (Deckert 1993; Pennycook 1994, 1996; Rinnert & Kobayashi 2005; Roig 1997, 2001; Shi 2006, 2010, 2012), and, in recent years, the idea of authorial ownership (and thus plagiarism itself) has been challenged (Howard 1995; Pennycook 1996; Scollon 1994). Nevertheless, most agree it is important to consider why students might copy from source texts when completing academic assignments, as such investigations may help us to better understand not only students' attitudes about textual borrowing, but also the role that such borrowing might play in their academic development.

A number of factors have been identified that might explain why developing writers – both students writing in their native language, and students writing in a second language – copy from source texts. In the case of second language (L2) writers, differences in cultural attitudes regarding the use of source texts have been cited as possible explanations for students' copying. A number of discussions (e.g. Matalene 1985; Pennycook 1996; Shi 2006; Sowden 2005) have focused on non-Western, primarily East Asian students, and how cultural practices such as text memorization might help to explain the textual borrowing strategies these students employ when writing in English. Surveys of students from China, Japan, and Korea (Rinnert & Kobayashi 2005; Shi 2006) have also found that, when studying English in their own countries, these students receive limited exposure to writing from sources, and little, if any, instruction in summary, paraphrase, and citation. In comparison, the U.S. students interviewed in these studies reported that writing from source texts received a great deal of attention in their academic courses.

Some have pointed out, however, that cultural differences are likely not the only, or even best, explanations for student textual plagiarism (Flowerdew & Li 2007a; Liu 2005; Pecorari 2003). Because plagiarism has been the topic of discussion in not only English L2 contexts, but also in English L1 (first language) contexts (Howard 1995; Hull & Rose 1989; Valentine 2006), many have suggested that the demands of adjusting to a new academic discourse community also play an important role in students' decisions to copy from source texts. These educators argue that copying often represents students' efforts to learn and practice the academic language that their professors expect them to use. For example, Howard (1995: 788) uses the term "patchwriting" to refer to instances of students "copying from a source text and then deleting some words, alternating grammatical structures, or plugging in one-for-one synonym substitutes" (p. 788). She argues that patchwriting is "an important transitional strategy in the student's progress toward membership in a discourse community." Similarly, Currie (1998) found that the subject of her case study, Diana, used copying as a strategy for learning the language of the academic discipline she was studying. Researchers interested in L2 writers have also suggested that textual borrowing can be seen as a language learning strategy; that there are "useful things to be learned from reusing the structures and words from others' texts" (Pennycook 1996: 225).

Though it is unclear how both cultural differences and language competence may help to explain student textual plagiarism, most researchers agree that for both L1 and L2 academic writers, copying and close paraphrasing are phases through which many developing writers pass before they acquire more sophisticated ways of integrating sources into their writing (Brown & Day 1983; Campbell 1990; Chandrasoma et al. 2004; Howard 1995; Hyland 2001; Johns & Mayes 1990; Pecorari 2003; Shi 2004; Sowden 2005; Winograd 1984). Over the past few decades, a number of text-based studies have documented such borrowing. For example, early summary studies found that novice L1 (eighth grade) writers copied or closely paraphrased individual sentences from a source text more frequently than expert (adult) writers (Winograd 1984); that "underprepared" U.S. university students copied and paraphrased source text excerpts more frequently than "adept" (more academically prepared) U.S. university students (Johns 1985); and that low-proficiency English L2 university students copied excerpts of the source text more frequently than high-proficiency students (Johns & Mayes 1990). Shi (2004) found that L2 writers in the early stages of their university study copied and closely paraphrased source text excerpts more frequently than English L1 university students. These studies, taken together, provide empirical evidence for the notion that copying and close paraphrase are strategies that many developing writers use to summarize or synthesize what they have read.

3. A corpus-based approach to textual borrowing research

A number of concerns regarding student textual borrowing, however, have yet to be sufficiently investigated. Though many feel that student strategy use likely varies across cultural background, language proficiency, and years of academic study, few large-scale studies have compared patterns of use across different student populations. And though many researchers agree that copying and close paraphrase represent initial stages in the development of academic writing skills, much is yet unknown about how students' use of these strategies changes over time. Even less attention has been paid to the rhetorical functions that copying and paraphrasing fulfill in student academic writing. Though student textual borrowing is often classified into categories such as intentional or unintentional, appropriate or not appropriate, student textual borrowing strategies are rarely described in terms of their communicative function in an academic text. The limited scope of textual borrowing research to date has been due, in large part, to the fact that this research has been carried out almost entirely by hand. Such work, which involves time-consuming comparisons of student and source text language, seriously limits the number of essays that can be analyzed, and the range of strategies that can be identified.

This paper argues that corpus-based methodologies could greatly broaden the scope of textual borrowing research. A corpus-based approach, according to Biber, Conrad, and Reppen (1998), makes use of computer technology in order to efficiently analyze large collections of naturally occurring texts (corpora), which are carefully constructed to represent specific domains of language use. One of the major advantages of a corpus-based approach is that it allows for text-based analyses that cannot be conducted without the aid of computers. For example, while it is nearly impossible to identify, by hand, every instance in which a student borrows language from a source text, computer technology allows for reliable, automatic identification of borrowed words and phrases. As a result, corpus-based studies of textual borrowing can address important questions about student source text use that have yet to be investigated.

4. Developing a corpus-based methodology

In an attempt to address this gap, I have devoted much of my research (Keck 2006, 2010, forthcoming) to exploring how computer technology might aid in the identification of student textual borrowing strategies. As a starting point, I chose the stand-alone summary as the unit of analysis. I asked university students to read a 1,000-word source text and to "explain the most important main ideas

(or arguments) of the essay in your own words." Three source texts (Meyrowitz 1982; Miller 1980; Samuelson 1985) were randomly distributed to the student participants, so that each student only summarized one text. All source texts used in the study were argumentative texts with similar reading levels and rhetorical structures.

A total of 227 university students contributed to the corpus (124 L1 writers and 103 L2 writers). Over 20 first languages were represented within the L2 writer group. The majority of L2 writers (n = 76) were enrolled in high-intermediate and advanced Writing courses within an Intensive English Program (IEP) at a U.S. university. The remaining 27 L2 writers were enrolled in credit-bearing ESL Composition courses at a U.S. university.

The student summaries were converted to text files so that computer programs could be used to aid in analysis. Descriptive statistics for the summary corpus are displayed in Table 1.

Table 1. The summary corpus

	Source 1		Source 2		Source3		Total	
	Summaries	Words	Summaries	Words	Summaries	Words	Summaries	Words
L1	41	7,376	38	6,542	45	7,039	124	20,957
L2	41	6,199	33	5,076	29	4,974	103	16,249
Total	82	13,575	71	11,618	74	12,013	227	37,206

Using Delphi software, I developed a series of computer programs which compared each student's summary against the original source text. The first program extracted individual words, two-word phrases, three-word phrases, and so on, from the original source text and stored these individual words and multiword strings in a database. The second program extracted individual words and multiword strings from the summaries and then searched for these same words and multiword strings in the original source text database. When a match was found, the program annotated the word/phrase in the student summary to indicate what line numbers in the original it occurred on.

Once these *shared words* were identified, trained coders examined the lines in the original text and the lines in the summary where these shared words occurred and identified an excerpt in the source text (usually a complete sentence or a series of sentences) which had been either exactly copied or paraphrased by the student. An Exact Copy was defined as an excerpt selected from the source text and reproduced in the summary, without the use of quotation marks, and with no linguistic changes made. A Paraphrase was defined as an instance in which a student selected

an excerpt from the source text, made at least one word-level linguistic change to the selected excerpt, and attempted to convey the meaning of that excerpt. That is, while Paraphrases of source text excerpts could contain copied strings of language, they also contained language composed by the student, while Exact Copies were full reproductions of source text excerpts. The coders then inserted brackets into the text files to mark Exact Copy and Paraphrase boundaries. The coders agreed on paraphrase location and boundaries 94% of the time (Cohen's kappa = .90). (See Keck 2006, for more details on paraphrase coding methods).

I then developed a computer program which could automatically classify the identified paraphrases into a Taxonomy of Paraphrase Types. This program analyzed each paraphrase and computed the number of shared words. It also made a distinction between *unique links* (a word or phrase in the paraphrase that also occurred in the original excerpt, but which did *not* occur at any other point in the source text) and *general links* (a word or phrase in the paraphrase that occurred in the source text multiple times). The program then classified paraphrases into a Taxaonomy based on the percentage of the paraphrase made up of unique links. The cut-off points for each category were determined through a series of qualitative analyses of a smaller set of paraphrases (see Keck 2010).

Table 2 summarizes the linguistic characteristics of each Paraphrase Type. As can be seen in this Table, the Taxonomy represents a continuum of textual borrowing, moving from Near Copies (which make use of long copied strings from the original) to Substantial Revisions (which make a number of lexical and grammatical changes to the original excerpt).

After each paraphrase was annotated to indicate its Paraphrase Type, I developed another computer program which analyzed each summary in the corpus and computed the following: (1) the number of Exact Copies, Near Copies, Minimal Revisions, Moderate Revisions, and Substantial Revisions that occurred and (2) the number of words in the summary made up of Exact Copies and each Paraphrase Type. The program also kept track of which lines in the original source text each paraphrase was based upon, and this information was used to describe to what extent students followed the order of ideas in the source text when composing their summaries.

5. Key findings

Once the summary corpus was built and Exact Copies and Paraphrase Types were identified, I carried out a series of qualitative and quantitative analyses to investigate patterns of use within and across student subgroups. In Keck (2006), I compared the rate of copying and paraphrasing observed within the L1 and L2

Table 2. The taxonomy of paraphrase types

Paraphrase type	Lexical criteria	Linguistic characteristics	Examples
			Original Excerpt **Children speak more like adults, dress more like adults and behave more like adults than they used to (Meyrowitz, 1982, p. 94).**
Near Copy	50% or more words contained within unique links	– Copied strings of 5 or more words – Simplification through synonym substitution and deletion.	Nowadays, children's behavior **more like adults than they used to.**
Minimal Revision	20–49% words contained within unique links	– Copied strings of 3–4 words – Multiple synonym substitutions	Children are acting more and **more like adults** everyday.
Moderate Revision	1–19% words contained within unique links	– Borrowing of 1–2 word phrases – Combination of synonym substitution and the revision of clause structures (e.g. *ing* → *to clause*)	Modern children seem to be behaving, through **dress** and speech, **like adults** at an alarmingly young age.
Substantial Revision	No unique links	– Borrowing of individual words – Revision of clause structures – Use of synonymous constructions, often in the form of complex noun phrases	It seems like the things that children do and even the clothes that they wear are more adult-like than ever before.

*Note: Unique links, or word strings that could be traced to only one place in the original text, are bolded and underlined. Words shared by both the paraphrase and the original excerpt, but which occurred multiple times in the source text, are underlined.

writer groups; in Keck (2010) I described the grammatical strategies that both L1 and L2 writers used to avoid exact copying; and in Keck (2014) I compared the copying and paraphrasing strategies of students in their first year of university study with those who had been studying in a U.S. university for more than one year.

As with many corpus-based investigations of language use, the findings of my research on paraphrasing has challenged many assumptions that both teachers and researchers have about student source text use. In the remainder of this paper, I highlight some of these assumptions, with a focus on beliefs regarding which students copy more than others, why students copy in the first place, and what should be done to address plagiarism in the writing classroom.

5.1 Assumption 1: L2 writers copy from source text more frequently than L1 writers

As Deckert (1993) and others (e.g. Leask 2006; Liu 2005; Pecorari 2001) have noted, international students writing in English as a second language are often characterized as "persistent plagiarizers" (Deckert 1993: 131). And although few second language researchers would encourage teachers and administrators to make assumptions about L2 writers based solely on their own perceptions of cultural difference, text-based studies of L2 textual borrowing practices (e.g. Campbell 1990; Moore 1997; Shi 2004; Keck 2006) typically compare L2 writers with native speakers, with an emphasis on how much more L2 writers copy than their L1 counterparts. Shi (2004), for example, compared the textual borrowing practices of L2 students studying English in China with L1 writers of English enrolled in first-year composition courses, and found that the Chinese writers copied larger strings of source text language without attribution more frequently than the L1 writers. In Keck (2006), I reported similar findings: the L2 writers, as a group, used more Exact Copies and Near Copies than the L1 writer group. While the aim of much of this research is to raise educators' awareness of unintentional or "non-prototypical" plagiarism (Pecorari 2003: 318) and to encourage pedagogical, rather than punitive, responses, such findings also seem to reinforce what some educators feel they have known all along – that plagiarism is a much bigger problem among international students than it is among L1 writers.

To address this concern, in Keck (2014), I revisited the L1 and L2 summaries from my 2006 study and explored variation *within* the L1 and L2 writer groups, with a focus on whether university students in their first year of U.S. university study differed in their strategy use from students studying in the U.S. for more than one year. In this investigation, I found that extensive copying occurred only within summaries composed by L2 students in their first year of university study. Within this group, summaries could be found which essentially copied and pasted source text excerpts, with little linguistic modification and few, if any, self-composed sentences. In contrast, no students in the U.S. for more than one year (neither the L1 nor the L2 writers) used copying as a primary summary-writing strategy.

The variation observed in Exact Copy use among L2 writers in their first year mirrored the Paraphrase use of L1 college freshmen: L2 Exact Copy use among students in their first year ranged from 0–12 per summary; Paraphrase use among L1 freshman ranged from 0–14. In contrast, L2 writers in the US for more than one year had a much more limited range of Exact Copy use (0–4), and sophomores and juniors had a more limited range of Paraphrase use (0–7). This suggests that, at least in the case of summary writing, novice university student writers – both

those writing in an L1 and those writing in an L2 – rely more on selected source text excerpts than university students with more experience. It may be the case that, once students develop an ability to identify key passages in a source text, they move to learning how to alter those passages linguistically so that they are not copied exactly. For some students in their first year of U.S. university study, every sentence in their summary was a paraphrased excerpt from the original; essentially, this is one step removed from the copy and paste strategy. As students gain more experience, more sentences in the summary become invented or gist sentences, and only key excerpts (e.g. the author's thesis) are paraphrased.

It should also be noted, however, that the range of copying for the L2 first-year writer group (0–12 Exact Copies per summary) shows a considerable amount of variation. In fact, within this group, the majority of L2 students used no Exact Copies, as shown in Figure 1.

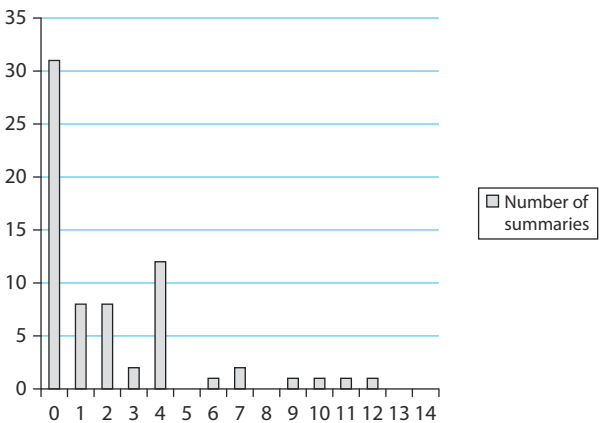

Figure 1. Exact copy use by L2 writers in the U.S. for one year or less (Keck 2014: 15)

As can be seen in Figure 1, 30 of 59 summaries composed by L2 writers in their first year of U.S. university study contained no Exact Copies, while a handful of students used 9 or more. A similar pattern was found in a recent study carried out by Weigle and Parker (2012), who analyzed 63 source-based essays written by L2 speakers of English. Weigle and Parker report that, overall, students copied source text language into their essays very infrequently, with a mean of less than 3 copied strings per essay.

> However, a few students borrowed substantially more than average, skewing these mean figures…. In particular, two students on the Globalization topic and one on the Computer topic had borrowing percentages that exceeded 25% of the essay. (Weigle & Parker 2012: 124)

These findings suggest that extreme caution should be used when making generalizations about the source text use of particular student groups. In both research and teaching, we are often drawn to extreme examples of student copying. This is understandable, considering the severe consequences often associated with inappropriate source text use. However, as corpus linguists have long argued, what is most noticeable is often not what is most typical – we are drawn to infrequent features of text precisely because they are different from the norm and they stand out (Biber et al. 1998). In the case of L2 source text use, we can easily see examples of copying, but often overlook other, much more frequent, paraphrasing strategies. This not only distracts our attention from students' effective use of textual borrowing, but also may lead to unfair stereotypes about particular student populations (Leask 2006; Liu 2005; Sowden 2005). While it is important to continue to investigate cases of student plagiarism, why they occur, and how they can be addressed pedagogically, equally important are efforts to describe the many cases in which students do successfully paraphrase and integrate source text language in their own academic work.

5.2 Assumption 2: Students copy from source texts because they do not understand what they are reading

When educators encounter a student summary of a source text that has been largely copied, they often assume that the student did not understand the text and thus was unable to explain it in their own words (Howard 1995). Language that is not understood is copied; language that is easier to understand can be paraphrased or summarized. A look at what excerpt students chose to copy or paraphrase in their own written work suggests, however, that failed reading comprehension is not always the best explanation for student textual borrowing practices. In many cases, students chose to copy or paraphrase excerpts that were not linguistically challenging. For example, as I report in Keck (2010), one of the most frequently copied and closely paraphrased excerpts (for both the L1 and the L2 writers) was made up primarily of high frequency vocabulary:

> Children speak more like adults, dress more like adults and behave more like adults than they used to. The reverse is also true: adults have begun to speak, dress and act more like overgrown children. (Meyrowitz 1982: 94)

It is possible that, in this example, the use of everyday language (*children*, *adults*, *speak*, *dress*, *behave*) makes it difficult for students to judge whether substantial changes in lexis can or should be made. In a number of cases, students re-used the words *speak*, *dress*, *behave*, and *more like adults*. For many students, it is likely unclear as to whether this type of borrowing is acceptable, or whether some or all

of these words need to be replaced. When making choices about what can or cannot be "borrowed," students must judge which words or phrases are considered to be unique or technical, and thus must be quoted or paraphrased; which words are so commonly used that they need not be quoted; and which words are so essential to the text's main idea that they should not be replaced with synonyms. This is not an easy task, and, as Shi's (2010) study reveals, teachers and students within and across academic disciplines do not always make the same judgments.

Analyses of the paraphrases identified in the summary corpus also suggest that grammatical competence is important to consider when examining student textual borrowing practices. In Keck (2010), I found that students (both L1 and L2 writers) who were able to avoid long copied strings of source text language did so through sophisticated grammatical modifications of the original excerpt. When composing Near Copy paraphrases, students typically used the strategies of deletion and substitution, leaving strings of 5 words or more unchanged. In contrast, when composing Moderate and Substantial Revisions, students revised at the constituent level, rather than at the level of the individual word. That is, students identified larger structural units (e.g. the subject noun phrase, a to-clause as direct object) and replaced them with new structures that fulfilled a similar syntactic function, as shown in Figure 2. (The original excerpt shown in this Figure comes from Samuelson 1985: 97.)

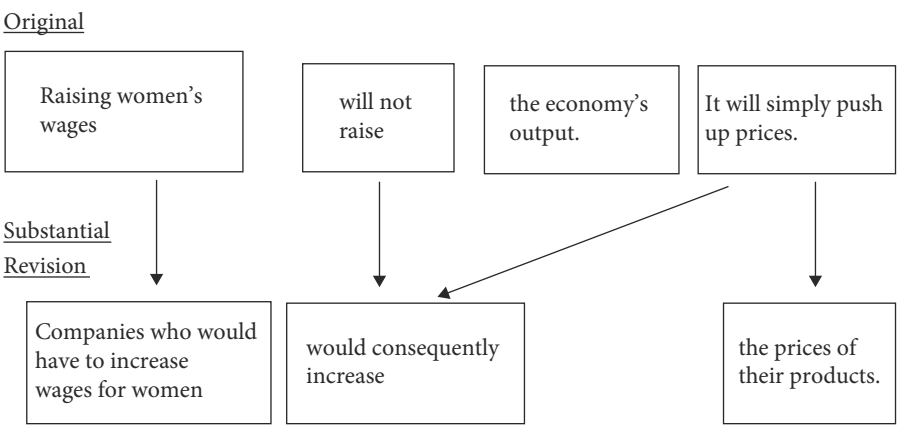

Figure 2. Paraphrasing at the constituent level (Keck 2010: 214)

These findings suggest that summarizing a source text takes a great deal of linguistic work: Students must form a mental representation of the key ideas in the source text, must identify key excerpts in the source text that help the author to convey these ideas, and must decide how those excerpts might be transformed,

linguistically, so that they can become part of their own written summary (Kennedy 1986; Kirkland & Saunders 1991). The summarization practices of novice and experienced students observed in Keck (2014) suggest that, initially, students focus on identifying key passages in a source text, and the strategies of copying and close paraphrasing help them to demonstrate to their ability to do this. Ironically, it was most often the case that copied excerpts within the summary corpus indicated that students had understood the main ideas, or at the very least, knew which ideas were more important than others (Keck 2010, 2014).

5.3 Assumption 3: Students should be taught how to paraphrase so that they can avoid plagiarism

Typically, summary, paraphrase, and quotation are presented as a triad of strategies that students can use when writing from sources (Barks & Watts 2001), and resources abound which recommend paraphrasing as a tool for avoiding plagiarism. These resources (e.g. Purdue University Online Writing Lab 2014; see also Yamada 2003) juxtapose "unacceptable" close paraphrases with "acceptable" paraphrases that make more substantial changes to the original excerpt and urge students to use synonyms, to avoid long copied phrases, and to change the grammar of the original excerpt as much as they can.

These resources, however, do little to help instructors and students understand how strategies like paraphrasing are used in the context of authentic writing assignments (Tomas 2010). Rather, advice on paraphrasing is presented in a highly decontextualized fashion. Short original excerpts are provided, but the larger texts from which these excerpts were taken are not. Sample paraphrases are shown, but in isolation; students almost never see these paraphrases in the context of an academic paper. As a result, while students are encouraged to use paraphrasing as a strategy for avoiding copying, many students do not have a clear idea of when, how, and for what purposes they might integrate paraphrases into their own written work.

Although summary and paraphrase are often treated as two separate strategies, almost all of the students who contributed to the summary corpus used paraphrases *within* their written summaries. Both the L1 and L2 writers used, on average, 5 paraphrases within a one-paragraph summary. Though avoiding plagiarism was certainly a concern of these students (their frequent use of paraphrasing suggests that they are aware that extensive copying is not acceptable), it was also the case that these students used paraphrases to accomplish important rhetorical moves within the summary. Two paragraphs in particular elicited a large proportion of the L1 and L2 Exact Copies and Paraphrases. These paragraphs were (1) an early source text paragraph that defined the concept to be

discussed, and (2) a concluding source text paragraph that stated the author's thesis. Most students (both the L1 and L2 writers) selected excerpts from the source text sequentially. Excerpts which helped to define the problem in focus occurred early in the student summary, examples from the source text which helped to illustrate this problem followed, and the author's concluding thesis appeared as a paraphrased excerpt at the end of the student summary. As Sherrard (1986) explains, this strategy may suggest that students understand the nature of expository texts and feel the need to preserve the author's own approach to building a logical argument. Whether this use of paraphrasing is always appropriate and effective depends largely on genre and disciplinary context. For example, Howard et al. (2010: 187) observed a similar strategy within research papers written by second-year university students, remarking that "these students are not writing from sources; they are writing from sentences selected from sources." For Howard et al. this was not a particularly positive strategy, as they felt it left the writer "in a position of peril … always in danger of plagiarizing." This concern suggests that paraphrasing does not always protect a student against accusations of plagiarism. If paraphrases are used too often, or for purposes not recognized as appropriate by the academy, then students may be penalized for using them. Few instructional resources, however, provide information about how to use paraphrases within the context of particular academic assignments. The current decontextualized nature of paraphrasing instruction may in part be due to the dearth of empirical research on paraphrasing in academic discourse. It is difficult to describe for students the important rhetorical functions that paraphrases fulfill if no such descriptions yet exist.

6. Directions for future research

In this paper, I have reviewed my own research on university student summarization practices, to illustrate how corpus-based approaches might challenge previously held assumptions about student source text use. Clearly, what I present here is a somewhat limited and narrow view. My work in this area has focused on only one type of academic writing, an in-class, timed, one-paragraph summary of a source text. The observations made here about student copying and paraphrasing strategies cannot be generalized to other assignment types. For example, it is likely the case that the frequency with which students paraphrase source text excerpts varies according to the nature of the assignment prompt, the number of source texts the student has consulted, and the length of those source texts. It is my hope, however, that the methodologies and findings described here will lead

others to embark on their own corpus-based investigations of textual borrowing practices.

In this spirit, I conclude the paper by making specific recommendations for future research. Two major avenues for research are outlined: research which seeks to develop methodologies for describing textual borrowing practices, and research which seeks to address the pedagogical concerns of educators who work with developing academic writers.

6.1 Methodologies for the study of textual borrowing

Much of my research on summary writing has involved the development of computer programs which can be used to help identify a wide range of textual borrowing strategies. It will be important to continue to refine these methods, so that they can be used to investigate the strategies of a variety of writer groups, in a variety of discourse contexts. While previous research has focused on describing citation practices in academic writing (see, e.g. Hyland 1999; White 2004), few studies have attempted to systematically describe the ways in which writers use textual borrowing strategies to achieve particular communicative goals. A number of questions about textual borrowing in academic discourse have yet to be explored: Is the writer's purpose for using a textual borrowing strategy related to the extent to which the writer borrows language from the original source? To what extent do writers borrow source text language to explain ideas, and to what extent do they borrow formulaic expressions, in order to conform to the expectations of academic genres? Because so many academic tasks involve writing from sources (as well as the reproduction of linguistic forms and discourse structures), it is difficult to imagine a comprehensive description of academic language that does not take into account textual borrowing practices. Continued development of both manual annotation techniques (e.g. the marking of paraphrase boundaries, the coding of communicative function) and automatic techniques (e.g. the classification of strategy types based on their linguistic characteristics) is greatly needed, so that we can begin to describe textual borrowing practices across a variety of text types and academic disciplines.

6.2 Pedagogic concerns

In addition to developing methodologies for the study of textual borrowing, it will be important to establish a clear research agenda, one that addresses the concerns of educators who work with developing academic writers. For many educators, understanding how to help student writers requires an understanding of how student writers approach their academic assignments. Educators concerned

with helping students to develop effective textual borrowing strategies, in particular, can benefit from descriptions of the types of strategies that students use and the factors that help to explain why students prefer some strategies over others.

As the present paper has demonstrated, corpus-based analyses can help to test common assumptions about student strategy use and can help to identify student strategies that have not yet captured the attention of educators. More large-scale studies of student textual borrowing may also provide educators with important information regarding the ways in which student textual borrowing practices vary across language backgrounds, years of academic study, and language proficiency.

In addition to understanding student textual borrowing strategies, it is also important for educators to have a clear picture of target strategy use. How, when, and for what purpose, do skilled writers borrow language from source texts? While countless writing handbooks and web resources provide students with examples of "acceptable" and "unacceptable" paraphrases, this information is not based on empirical studies of effective strategy use. One possible future direction for textual borrowing research, then, is to begin to describe the strategies used by expert writers. Might it be the case, for example, that paraphrases are far less frequent in published writing than in student writing? Do student writers use paraphrases for different purposes than published writers? Comparisons of successful students (i.e. those who consistently receive high grades on their written assignments) and less successful students, as well as descriptions of how instructors and professors respond to student textual borrowing, are also needed. Do particular strategies consistently draw negative feedback from teachers? Are there particular strategies that consistently receive praise? Such investigations may help both educators and students to understand the expectations and values surrounding textual borrowing in their specific academic contexts.

References

Altenberg, Bengt & Granger, Sylviane. 2001. The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics* 22: 173–195. DOI: 10.1093/applin/22.2.173

Barks, Debbie & Watts, Patricia. 2001. Textual borrowing strategies for graduate-level ESL writers. In *Linking Literacies: Perspectives on L2 Reading-Writing Connections,* Diane Belcher & Alan Hirvela (eds), 246–270. Ann Arbor MI: The University of Michigan Press.

Belcher, Diane & Hirvela, Alan. 2001. *Linking Literacies: Perspectives on L2 Reading-writing connections*. Ann Arbor MI: The University of Michigan Press.

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. DOI: 10.1017/CBO9780511804489

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

Brown, Anne & Day, Jeanne. 1983. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior* 22: 1–14. DOI: 10.1016/S0022-5371(83)80002-4

Campbell, Cherry. 1990. Writing with others' words: Using background reading text in academic compositions. In *Second Language Writing: Research Insights for the Classroom,* Barbara Kroll (ed.), 211–230. Cambridge: CUP. DOI: 10.1017/CBO9781139524551.018

Carson, Joan & Leki, Ilona. 1993. *Reading in the Composition Classroom: Second Language Perspectives*. Boston MA: Heinle & Heinle.

Casanave, Christine Pearson. 2004. *Controversies in Second Language Writing: Dilemmas and Decisions in Research and Instruction*. Ann Arbor MI: The University of Michigan Press.

Chandrasoma, Ranamukalage, Thompson, Celia, & Pennycook, Alastair. 2004. Beyond plagiarism: Transgressive and nontransgressive intertextuality. *Journal of Language, Identity, and Education* 3: 171–194. DOI: 10.1207/s15327701jlie0303_1

Charles, Maggie. 2006. 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

Conrad, Susan. 1999. The importance of corpus-based research for language teachers. *System* 27: 1–18. DOI: 10.1016/S0346-251X(98)00046-3

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

Csomay, Eniko. 2006. Academic talk in American university classrooms: Crossing the boundaries of oral-literate discourse? *Journal of English for Academic Purposes* 5: 117–135. DOI: 10.1016/j.jeap.2006.02.001

Currie, Pat. 1998. Staying out of trouble: Apparent plagiarism and academic survival. *Journal of Second Language Writing* 7: 1–18. DOI: 10.1016/S1060-3743(98)90003-0

Deckert, Glenn. 1993. Perspectives on plagiarism from ESL students in Hong Kong. *Journal of Second Language Writing* 2: 131–148. DOI: 10.1016/1060-3743(93)90014-T

Flowerdew, John & Li, Yongyan. 2007a. Plagiarism and second language writing in an electronic age. *Annual Review of Applied Linguistics* 27: 161. DOI: 10.1017/S0267190508070086

Flowerdew, John & Li, Yongyan. 2007b. Language re-use among Chinese apprentice scientists writing for publication. *Applied Linguistics* 28: 440–465. DOI: 10.1093/applin/amm031

Hinkel, Eli. 2002. *Second Language Writers' Texts: Linguistic and Rhetorical Features*. Cambridge: CUP.

Howard, Rebecca. 1995. Plagiarisms, authorships, and the academic death penalty. *College English* 57: 788–806. DOI: 10.2307/378403

Howard, Rebecca, Serviss, Tricia & Rodrigue, Tanya. 2010. Writing from sources, writing from sentences. *Writing and Pedagogy* 2(2): 177–192. DOI: 10.1558/wap.v2i2.177

Hull, Glynda & Rose, Mike. 1989. Rethinking remediation: Toward a social-cognitive understanding of problematic reading and writing. *Written Communication* 6: 139–154. DOI: 10.1177/0741088389006002001

Hyland, Fiona. 2001. Dealing with plagiarism when giving feedback. *ELT Journal* 55: 375–381. DOI: 10.1093/elt/55.4.375

Hyland, Ken. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20: 341–367. DOI: 10.1093/applin/20.3.341

Hyland, Ken. 2004. Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* 13: 133–151. DOI: 10.1016/j.jslw.2004.02.001

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 & Milton, John. 1997. Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing* 6: 183–205. DOI: 10.1016/S1060-3743(97)90033-3

Hyland, Ken & Tse, Polly. 2004. Metadiscourse in academic writing: A reappraisal. *Applied Linguistics* 25: 156–177. DOI: 10.1093/applin/25.2.156

Johns, Ann. 1985. Summary protocols of "underprepared" and "adept" university students: Replications and distortions of the original. *Language Learning* 35: 495–512. DOI: 10.1111/j.1467-1770.1985.tb00358.x

Johns, Ann & Mayes, Patricia. 1990. An analysis of summary protocols of university ESL students. *Applied Linguistics* 11: 253–271. DOI: 10.1093/applin/11.3.253

Keck, Casey. 2006. The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing* 15: 261–278. DOI: 10.1016/j.jslw.2006.09.006

Keck, Casey. 2010. How do university students attempt to avoid plagiarism? A grammatical analysis of undergraduate paraphrasing strategies. *Writing & Pedagogy 2, Special topics issue on Plagiarism and the Academy*: 193–222.

Keck, Casey. 2014. Copying, paraphrasing, and academic writing development: A re-examination of L1 and L2 summarization practices. *Journal of Second Language Writing* 25: 4–22.

Kennedy, M.L. 1986. The composing process of college students writing from sources. *Written Communication* 2: 434–456. DOI: 10.1177/0741088385002004006

Kirkland, Margaret & Saunders, Mary. 1991. Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly* 25: 105–121. DOI: 10.2307/3587030

Leask, Betty. 2006. Plagiarism, cultural diversity and metaphor – implications for academic staff development. *Assessment and Evaluation in Higher Education* 31(2): 183–199. DOI: 10.1080/02602930500262486

Leki, Ilona & Carson, Joan. 1997. "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly* 31: 39–69. DOI: 10.2307/3587974

Liu, Dillin. 2005. Plagiarism in ESOL students: is cultural conditioning truly the culprit? *ELT Journal* 59: 234–243. DOI: 10.1093/elt/cci043

Matalene, Carolyn. 1985. Contrastive rhetoric: An American writing teacher in China. *College English* 47: 789–809. DOI: 10.2307/376613

Meyrowitz, Joshua. 1982, August 30. Where have the *children* gone? *Newsweek* 94: 13.

Miller, R.K. 1980, July 21. Discrimination is a virtue. *Newsweek* 92: 15.

Moore, Thomas. 1997. From text to note: Cultural variation in summarization practices. *Prospect* 12: 54–63.

Pecorari, Diane. 2001. Plagiarism and international students: How the English-speaking university responds. In *Linking Literacies: Perspectives on L2 Reading-Writing Connections*, Diane Belcher & Alan Hirvela (eds), 229–245. Ann Arbor MI: The University of Michigan Press.

Pecorari, Diane. 2003. Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing* 12: 317–345.

DOI: 10.1016/j.jslw.2003.08.004
Pennycook, Alastair. 1994. The complex contexts of plagiarism: A reply to Deckert. *Journal of Second Language Writing* 3(3): 277–284. DOI: 10.1016/1060-3743(94)90020-5
Pennycook, Alastair. 1996. Borrowing others' words: Text, ownership, memory, and plagiarism. *TESOL Quarterly* 30: 201–230. DOI: 10.2307/3588141
Polio, Charlene & Shi, Ling. 2012. Perceptions and beliefs about textual appropriation and source use in second language writing. *Journal of Second Language Writing* 21: 95–101. DOI: 10.1016/j.jslw.2012.03.001
Purdue University Online Writing Lab. 2014 Paraphrase: Write it in your own Words, 4 March 2014. ⟨https://owl.english.purdue.edu/owl/owlprint/619/⟩
Rinnert, Carol & Kobayashi, Hiroe. 2005. Borrowing words and ideas: Insights from Japanese L1 writers. *Journal of Asian Pacific Communication* 15: 31–56. DOI: 10.1075/japc.15.1.05rin
Roig, Miguel. 1997. Can undergraduate students determine whether text has been plagiarized? *The Psychological Record* 47: 113–122.
Roig, Miguel. 2001. Plagiarism and paraphrasing criteria of college and university professors. *Ethics & Behavior* 11: 307–323. DOI: 10.1207/S15327019EB1103_8
Sameulson, Robert. 1985, April 22. The myths of comparable worth. *Newsweek* 97: 13.
Scollon, Ron. 1994. As a matter of fact: The changing ideology of authorship and responsibility in discourse. *World Englishes* 13: 33–46. DOI: 10.1111/j.1467-971X.1994.tb00281.x
Sherman, Jane. 1992. Your own thoughts in your own words. *ELT Journal* 46: 190–197. DOI: 10.1093/elt/46.2.190
Sherrard, C. 1986. Summary writing: A topographical study. *Written Communication* 3: 324–343. DOI: 10.1177/0741088386003003003
Shi, Ling. 2004. Textual borrowing in second-language writing. *Written Communication* 21: 171–200. DOI: 10.1177/0741088303262846
Shi, Ling. 2006. Cultural backgrounds and textual appropriation. *Language Awareness* 15: 264–282. DOI: 10.2167/la406.0
Shi, Ling. 2010. Textual appropriation and citing behaviors of university undergraduates. *Applied Linguistics* 31: 1–24. DOI: 10.1093/applin/amn045
Shi, Ling. 2012. Rewriting and paraphrasing source texts in second language writing. *Journal of Second Language Writing* 21: 134–148. DOI: 10.1016/j.jslw.2012.03.003
Sowden, Colin. 2005. Plagiarism and the culture of multilingual students in higher education abroad. *ELT Journal* 59: 226–233. DOI: 10.1093/elt/cci042
Spack, Ruth. 1997. The acquisition of academic literacy in a second language: A longitudinal case study. *Written Communication* 14: 3–62. DOI: 10.1177/0741088397014001001
Spack, Ruth. 2004. The acquisition of academic literacy in a second language: A longitudinal case study. In *Crossing the Curriculum: Multilingual Learners in College Classrooms,* Vivian Zamel & Ruth Spack (eds), 19–46. Mahwah NJ: Lawrence Erlbaum Associates.
Tardy, Christine. 2010. Writing for the world: Wikipedia as an introduction to academic writing. *English Teaching Forum* 1: 12–27.
Tomas, Zuzana. 2010. Addressing pedagogy on textual borrowing: Focus on instructional resources. *Writing & Pedagogy* 2: 223–250. DOI: 10.1558/wap.v2i2.223
Upton, Thomas & Connor, Ulla. 2001. Using computerized corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes* 20: 313–329. DOI: 10.1016/S0889-4906(00)00022-3
Valentine, Kathryn. 2006. Plagiarism as literacy practice: Recognizing and rethinking ethical binaries. *College Composition and Communication* 58: 89–100.

Weigle, Sara & Parker, Keisha. 2012. Source text borrowing in an integrated reading-writing assessment. *Journal of Second Language Writing* 21: 118–133. DOI: 10.1016/j.jslw.2012.03.004

Winograd, Peter. 1984. Strategic difficulties in summarizing texts. *Reading Research Quarterly* 19: 404–425. DOI: 10.2307/747913

Yamada, Kyoko. 2003. What prevents ESL/EFL writers from avoiding plagiarism?: Analyses of 10 North-American college websites. *System* 31: 247–258. DOI: 10.1016/S0346-251X(03)00023-X

CHAPTER 9

Situating lexical bundles in the formulaic language spectrum

Origins and functional analysis developments

Viviana Cortes
Georgia State University

If Douglas Biber and his collaborators in the Longman Grammar of Spoken and Written English (Biber et al. 1999) had not devoted a great deal of work to replicate the corpus-driven methodology used by Bent Altenberg (1993) in the identification and analysis of recurrent word combinations, chances are lexical bundles and the dozens of studies of lexical bundles conducted in the last decade would not have come to exist. This chapter outlines the development of the study of these expressions, which have generated a strong area of research for discourse analysis, particularly analyses of academic prose in a wide variety of text types: research articles, dissertations and theses, and textbooks, among many others.

Keywords: Lexical bundles; formulaic language; move analysis

1. Introduction

Formulaic language has been of great interest to the applied linguistic field for several decades and formulas have been studied from many different linguistic perspectives and using a wide variety of research methodologies (Coulmas 1981; Firth 1951; Hakuta 1974; Wang Filmore 1979; Yorio 1980). Early studies of collocations and lexical co-occurrence were conducted using a rather impressionistic methodology but the developments that have been taking place in the analysis of language corpora since the end of the 1970s brought about important changes in the methodologies used in the identification and study of recurrent word combinations. It is undeniable that the advancements introduced by corpus-based research methodologies have made an invaluable contribution to the study of this linguistic phenomenon. An important body of research made up of corpus-based studies that identified, defined, and classified different types

DOI 10.1075/scl.66.09cor

of fixed expressions originated in the last decade of last century in studies like those by Sinclair (1991), Renouf & Sinclair (1991) Kjellmer (1991), Nattinger & DeCarrico (1992), Alternberg (1993), Butler (1997), DeCock (1998), Hunston & Francis (1998), and Moon (1998), to mention only a few of them in chronological order. Biber et al. (2004a) explained that these studies differ in the way they identified and described these formulaic expressions (structurally and/or functionally), the methods that were used for their identification (perceptual importance, frequency, or some other methodology), the type of expressions that was the focus of each of those studies (continuous sequences of words, frames or collocational frameworks, lexico-grammatical patterns, two-word collocations or longer word combinations, obscure idioms, etc.) and the corpus used in each investigation that ranged from small corpora of less than 100,000 words to mega corpora of more than 100 million words from various registers in the language (Biber et al. 2004a: 372).

One of the constructs that emerged in the last decade of last century and developed in the first years of the new millennium is that of lexical bundles. Lexical bundles are recurrent word combinations; groups of three or more words that frequently recur in a particular register (Biber et al. 1999: 990). They are identified empirically rather than intuitively through a strict corpus-driven methodology. Lexical bundles often found in conversation are expressions such as *I don't know what to do*, *you won't be able to*, *do you want to go*, *how do you know*, and *let's have a look*, to mention only a few. In academic prose, frequent lexical bundles are word sequences like *as a result of, on the other hand, it is unlikely that*, and *as shown in figure*. The work of Biber et al. (1999) in the analysis and structural classification of these expressions was groundbreaking for the applied linguistics field. Their investigation paved the way for the numerous studies of lexical bundles that have been conducted in the last decade by researchers from the Flagstaff school and from other research centers and educational institutions around the world. Very prestigious international journals in the applied linguistics field such as *Applied Linguistics*, *English for Specific Purposes*, *Linguistics and Education*, and the *Journal of English for Academic Purposes* among others, have published more than a dozen studies in the past decade that directly investigate lexical bundles and there are numerous chapters in edited volumes devoted to the analysis of these expressions. Even though there are a number of studies that analyzed lexical bundle use in various registers, the bulk of the research on lexical bundles focuses on the identification and analysis of these expressions in different forms of academic discourse, particularly, of academic writing. The main reason for this focus lies in the accessibility and availability of written corpora in general and of collections of texts from different academic genres in particular. In addition, the functions performed by lexical bundles in certain academic genres, such as research articles, theses, and

dissertations, for example, seem to be closely related to the communicative purposes that the writers of those registers attempt to convey at different points in their texts, making the teaching application of the use of these building blocks very desirable for researchers and practitioners in the English for academic writing field.

The purpose of the present chapter is twofold. First, the chapter goes back in time to the beginnings of the identification of formulaic expressions that employed a corpus-driven methodology, to investigate the origin of the methodological approach and the frequency thresholds that define lexical bundles. Second, the chapter provides an overview of the major studies on lexical bundles in academic writing following "the Biber et al. (1999) tradition." Thus, the chapter will first concentrate on the origins and development of lexical bundles from a chronological perspective. Later, the chapter will focus on studies of lexical bundles in academic prose highlighting the functions of bundles in discourse and their potential for the teaching of academic writing.

2. Corpus-based and corpus-driven research methods and the study of formulaic language

As previously mentioned, there have been numerous studies that focused on the identification and analysis of formulaic multi-word sequences. These studies used various approaches to identify fixed expressions. Barfield and Gyllstad (2009) stated that research on collocation has been conducted within two traditions: the frequency-based tradition and the phraseological tradition. According to these authors, the frequency-based tradition centers on the analysis of collocation based on frequency and statistics, drawing on the pioneer work conducted by Firth (1951), and later Sinclair (1987, 1991). The phraseological tradition is guided by syntactic and semantic analysis of collocation and it follows the Russian and European school of phraseological work (Aisenstadt 1979; Cowie 1981, 1998).

More recently, two distinctive research approaches to the analysis of language corpora for the identification of formulaic expressions have emerged in the frequency-based tradition. These approaches are: the corpus-based approach and the corpus-driven approach (Tognini-Bonelli 2001). The corpus-based approach to the identification of recurrent word sequences relies on expressions that have been considered formulaic in linguistic theory. This approach focuses on a group of pre-selected formulaic expressions and on the analysis of the use of those expressions in a language corpus (see for example Nattinger & De Carrico 1992; Moon 1998). The corpus-driven approach, on the other hand, is inductive. Biber (2009) provides a detailed description of the various types

of formulaic expressions that can be identified using a corpus-driven approach. This type of investigations cover, for example, studies of lexical collocations in which a corpus is used to discover the collocations of a target word. This type of studies is considered corpus-driven, even though a preliminary step is based on the analysts' selection of interesting target words for analysis. Another type of studies of multi-word combinations makes "fewer theoretical assumptions, beginning with simple words forms and using frequency distributions to identify recurrent word sequences" (Biber 2009: 276). These formulaic sequences emerge from the corpus-analysis with little pre-conceptions on the linguistic expressions that will be the target of further structural and functional analyses. This is the approach used for the identification of n-grams and lexical bundles. An n-gram is any group of 2 or more words identified in this way in a language corpus. A group of words, however, needs to meet certain characteristics to be considered a lexical bundle. In short, we can affirm that all lexical bundles are n-grams, recurrent groups of three or more words, but not all n-grams are lexical bundles. The qualities that characterize lexical bundles will be discussed in Section 4.

3. From 2-word collocations to longer recurrent expressions

Lexical bundles are defined as recurrent formulaic sequences but they are frequently not complete structural units. Although some lexical bundles are regarded as extended collocations, they differ from this type of expressions mostly in the word class of the components that make up lexical bundles. Lexical bundles are in many ways different from the two-word collocations that have been studied intuitively for decades and from other formulaic language, such as obscure idioms or pragmatic formulae. Biber and Conrad (1999: 183) explain that "words with similar meaning are often distinguished by their preferred collocations." This is the case of most lexical (or content) words (nouns, adjectives, verbs, and adverbs) that define their extended meaning in a specific set of collocates that tend to co-occur with those words. Lexical bundles, on the other hand, often incorporate many function words (articles, prepositions, pronouns, etc.) that accompany a lexical word constituting a different type of recurring word combination.

3.1 A brief account of collocations

Given that some lexical bundles have been considered extended collocations, it is important to thoroughly review this construct in order to draw reliable comparisons among these different types of formulaic expressions. Already in the 1930s, several grammarians were focusing on the use of formulas or recurrent

expressions working individually on different projects. Their work was somehow simultaneous and their findings and concepts overlapped considerably. Jespersen (1933: 18) clearly stated the difference between "formulas (or formular units) and free expressions." He explained that spoken formulas appear to be fixed and nothing can be changed, not even the intonation pattern or the insertion of pauses: formulas remain unchanged and are unchangeable. These formulas may have a meaning different from that of the words that make up the expression and according to Jespersen, memory is key in formulae production. Producing free expressions, on the other hand, involves the use of other types of mental activity as these free expressions need to be created "anew by the speaker," who needs to arrange the words in particular patterns according to the situation (Jespersen 1933: 18). At the same time, Palmer (1933) was working on developing vocabulary for teaching purposes. One of his projects was to identify and classify repeated expressions. He labeled one of these types of expressions "collocation," defined as "a succession of two or more words that must be learnt as an integral whole and not pieced together from its component parts" (Palmer 1933: Title page). Palmer's work has been an important influence in the study of fixed expressions. In addition, it is fair to acknowledge that one of the most recognized linguists in the theory of collocation is J. R. Firth. His view of collocation and collocability focused primarily on the way some lexical words prefer to get together with certain words rather than others. His work emphasized the importance of collocations in determining the extended meaning of a word or what he called "meaning by collocation" (Firth 1951: 196).

3.2 Extending collocations: Recurrent word combinations

Biber et al. (1999) coined the term lexical bundles for recurrent word combinations that are identified empirically rather than intuitively in a given language register. The name lexical bundles also reflects the fact that these word combinations could be among the expressions considered "compound lexical items" and that they may possess a lexical structure than ranks above the word (Sinclair 2004: 39). Biber and Conrad (1999) explained that corpus-based computational analysis provided the needed tools to empirically investigate the formulaic language phenomenon. They stated that the use of corpus-based techniques facilitates the identification of sequences of words that occur frequently across the different texts of a register. They refer to Altenberg's (1993) work as one of the first in using this type of methodologies for the identification of word combinations similar to the ones now labeled lexical bundles.

Altenberg and Eeg-Olofsson (1990) introduced their proposal for an investigation of recurrent word combinations in the London Lund Corpus, a collection

of approximately half a million words of spoken English. For the first step in their proposal, which consisted of the retrieval of this type of expressions, Altenberg and Eeg Olofsson used a procedure strongly inspired by the approach employed by Allen (1975), in his investigation of collocations in a corpus of Swedish newspaper writing. Allen's work on word combinations was the third part of a larger study (Allen, 1970) that entailed various stages, which started with a frequency dictionary of current Swedish based on newspaper language, and continued with morphological and collocational studies based on the same corpus. In spite of the fact that these studies identified frequent fixed expressions in a corpus using a data-driven methodology, the way in which the expressions were identified slightly differs from the methodology used by Biber et al. (1999) to identify lexical bundles. Allen (1975) explains that

> The starting point was provided by an alphabetical arranged concordance covering the whole word material. This concordance was further arranged with respect to the nearest word in the preceding context, the nearest word in the following context, and nearest word but one in the preceding context, the nearest word but one in the following context, etc. (Allen 1975: XXXIII)

The following step was to compare consecutive instances and eliminate those instances that did not recur. Allen stated that the sorting arrangement he used was selected due to the structure of Swedish and proved to be successful. Altenberg and Egg-Olofsson (1990) carefully described the process they used following Allen's (1975).

> The main idea has been to produce a KWIC (Key Word in Context) concordance of the entire corpus, sorted in zigzag order. This particular sorting order … offers the researcher a good view of the material for manual inspection."
> (Altenberg & Egg-Olofsoon 1990: 10)

Altenberg and Egg-Olofsson explained that this zig-zag arrangement made it easy for computer programs to retrieve all recurrent combinations that contain a certain key-word and to find all the occurrences of that word combination. Although these authors did not clearly explain if all the words in their corpus became in turn key-words, if this were the case then the recurrent word combinations identified by Allen and Altenberg and Egg-Olofsson could be considered lexical bundles if they met the pre-established high frequency and range cut-off points. It is also necessary to note that in the days when these researchers were conducting their studies, the capabilities of computers were very low, both for data storage and for data processing. Altenberg and Egg-Olofsson (1990: 8) recommended using a "mainframe computer with sufficient capacity" for processing and storing the data instead of a personal computer, which had a lot of limitations

in capacity. This limitation could result in serious data processing problems in those days, if we take into account that they were using the London Lund corpus, which consisted of half a million words from spoken texts. Another important difference in the methodology used by these authors in the identification of these recurrent word combinations is the fact that they do not place much importance in the frequency of the expressions. Even when Altenberg (1993) states that the frequency threshold he used was ten times in his corpus, there is no further explanation of why some of the examples of recurrent expression presented in his article only occurred two or three times.

Another important study that considered frequency-driven expressions identified in a corpus is the one conducted by Salem (1987) in his "Pratique des segments répétés" (the use of recurrent segments), in which he identified and analyzed lexical sequences in a corpus of political discourse made up of resolutions passed by the four major French workers' unions from 1971 to 1976, using a methodology similar to that used later by Biber et al. (1999). Salem explained that in the French linguistics community, textometric analysis consists of various methods used to reorganize word combinations based on statistical analysis of a language corpus (Lebart et al. 1998; Salem 1987). These textometric analyses use tools that divide the text into graphical forms and identify various types of textual units among which we can find repeated segments (segments répétés), which are series of consecutive graphical forms found in a corpus with frequency greater than or equal to 2 (Fleury & Zimina 2006; Lebart et al. 1998; Salem 1987).

When talking about lexical bundles, however, we need to consider Biber et al. (1999) as their real origin, at least to the extent to which these expressions are currently identified and analyzed. In their comprehensive analysis of the spoken and written grammar of English, Biber and his collaborators produced a whole chapter devoted to formulaic language and to the introduction and structural analysis of lexical bundles. Their findings will be discussed in detail in the following sections.

4. What lexical bundles are

The work just mentioned looked at formulaic expressions from different perspectives and with different purposes. It is important then to look into what Sinclair calls "the axes of patterning" (Sinclair 2004: 140). Sinclair explains that the tradition of linguistic theory has always concentrated on the paradigmatic dimension rather than its syntagmatic counterpart. Meaning seems to be better explained by

paradigmatic choice. For example, in the study of collocation introduced by Firth (1951), he illustrated the particular collocability of the word *person* by providing the frequent collocates *old* and *young*, which are exponents in this particular case of the paradigmatic dimension and, as such, are mutually exclusive. Words, however, simultaneously provide information because they have been chosen (paradigmatic dimension) and also because they are part of larger units (syntagmatic dimension). The syntagmatic relation is very important for lexical bundles, as their frequent co-occurrence is what makes these expressions salient and they are made up of words that show a strong level of syntagmatic connection at the phrase level.

Corpus-driven approaches to the identification of lexical co-occurrences produced expressions that "do not fit predefined linguistic categories" (Granger & Paquot 2008: 29). Biber et al. (1999: 990) introduced lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status." They are sequences of words that frequently recur in natural discourse, continued strings of words without any empty slots. Biber et al. (1999) emphasize the difference between three-word lexical bundles and longer expressions made up of four or more words. They explain that "Three-word bundles can be considered a kind of extended collocational association" but "on the other hand, four-word, five-word, and six-word bundles are more phrasal in nature and correspondingly less common" (Biber et al. 1999: 992). This is an important differentiation because the longer the bundle the less common it is, and bundle length also has an important influence on the type of lexical items that make up the bundle, the grammatical group the bundle aligns with, and the communicative function of its use in a particular register.

Frequency is the ultimate quality of lexical bundles. In order to be considered a lexical bundle, a three or four-word combination has to be extremely frequent in a given register. Biber et al. (1999) established an arbitrary cut-off point of 10 times per million words (pmw) but they incorporated different benchmarks for frequency (20, 40, and 100 times pmw) to demonstrate that many expressions in both everyday conversation and academic prose, two of the registers included in their study, the recurrent expressions now labeled lexical bundles repeated much more frequently than that minimal pre-established cut-off point. Later studies (Biber et al. 2004a; Cortes 2004) used more conservative cut-off points for lexical bundle identification (20 or 40 occurrences pmw). In addition, the use of the expressions by different language users also needs to be considered in lexical bundle studies. Biber et al. (1999) called this quality range and established a range of five or more texts for lexical bundle use in order to avoid the idiosyncrasies of a single language user or a few language users in their corpus. When studying bundles of different

lengths (3-, 4-, 5, or 6+-word bundles) discriminating frequencies may become necessary as longer bundles are rare in certain registers. Cortes (2004) explained that four-word bundles are often ten times more frequent than five-word bundles. Furthermore, Biber et al. (1999) only found a few six-word bundles in their study. Specific registers, however, seem to produce longer lexical bundles (longer than six words) which perform very specific functions (Cortes 2013a). This new development in the study of lexical bundles will be discussed in Section 7. 3.

The one-million word corpus threshold established in many studies for the identification of lexical bundles is also a convention but it is linked to an important issue related to corpus size in the comparison of lexical bundles identified in small corpora and corpora of different sizes. Potentially, a combination of three or more words identified in a corpus of any size could be a lexical bundle if it recurs very frequently. Comparison of bundles yielded by small corpora and large corpora has been shown to be problematic because applying the usual normalization formula results in unreliable figures. Cortes (2002b: 72–74) showed that smaller corpora may yield many more lexical bundles than larger corpora (after normalization) and that in order to meet the cut-off point, word combinations do not need to repeat very frequently because when their frequencies are normalized any phrase that repeats a couple of times could be considered a lexical bundle.

In comparison to intuitive ways to identify recurrent word combinations, the methodology used in the identification of lexical bundles is empirical and, as previously explained, corpus-driven (Cortes 2012). Researchers approach corpora in search of lexical bundles leaving behind their intuition or perception. When a program used to identify lexical bundles finishes processing the text, it yields a list of those fixed combinations that met the pre-established cut-off points for frequency and range. These programs often start reading the first word of the first text of a corpus and move one word at a time, recording all expressions of the established word number (three-word, four-word, etc.) and identifying those that meet the cut-off points as lexical bundles (see Cortes 2012 for a detailed description of one of those programs, the Lexical Bundles Program, LBP).

The units that make up lexical bundles are orthographic word units, although sometimes these units may combine separate words, as in the case of contractions, which are extremely frequent in spoken genres. In addition, when the program that identifies lexical bundles comes across a punctuation mark, it immediately stops and starts processing the text again after the punctuation mark. As Biber et al. (2003) stated, only uninterrupted strings of words are processed as lexical bundles. Lexical bundles are then uninterrupted strings of three or more words that frequently recur in a register, identified empirically by running a computer program in a corpus of language texts.

5. What lexical bundles are not

Regarding their internal constituency, lexical bundles are very different from idiomatic expressions or other pragmatic formulas. Bundles are not often idiomatic, as the meaning of the bundle can be generally retrieved from the meaning of the words that form the expression. In expressions such as *the fact that the*, *the extent to which*, or *it is possible that* the words in these bundles retain their own meaning when forming the bundle and help create the meaning of the expression as a whole.

Structurally, most lexical bundles are not complete units, although there are some exceptions. In academic writing, for example, lexical bundles are often composed of a grammatical phrase or clause fragment with some other phrase or clause fragments embedded. They do not incorporate many lexical words and these lexical words are often special types of abstract or shell nouns (*in the context of*), copular or reporting verbs (*has been suggested that*), and qualifying adjectives (*it is important to*). In addition, lexical bundles are continuous fixed sequences in contrast to formulaic frames with internal fixed or variable slots.

As previously explained, lexical bundles are identified by frequency, which is their ultimate quality. Frequency and range cut-off points are pre-established in order to ensure that the expressions identified are really frequent. Recent studies have tried to rely on statistical measures other than pure frequency and range for the identification of frequent recurrent expressions, as in the case of Mutual Information (MI) scores (Ellis et al. 2008). MI scores compare the frequency of a word combination to the overall frequency of each of the words that constitute a fixed expression. Biber (2009: 287) explains that these scores may reflect collocational strengths for two-word collocations, particularly when these expressions are made up of two lexical words. This statistical measure does not favor combinations that incorporate high frequency words and it only shows that two words are likely to occur together. MI scores, however, can be problematic when used for lexical bundle identification and analysis because bundles are fixed expressions of three or more words and in addition, they incorporate many function words, like articles and pronouns that are extremely frequent.

6. Lexical bundles: Internal structure

Hoey (2005: 13) explains that "every word is primed to occur (or avoid) certain grammatical positions and to occur in (or avoid) certain grammatical functions; these are its colligations." Colligation is a very important concept for the study

of lexical bundles, because their fixedness could be related to the grammatical function of the words that compose the expression. In academic prose, for example, frequent lexical bundles such *as as a result of*, *on the other hand*, or *the extent to which*, include many function (also called grammatical) words like prepositions, determiners (mostly articles), and relativizers (relative pronouns) or complementizers, that surround a lexical word, generally a noun and sometimes a verb. In those lexical bundles that are made up of noun phrase fragments, many of these nouns do not behave like regular lexical words do: they are often what have been labeled in the literature shell nouns or signaling nouns (Flowerdew 2003; Schmid 2000). These nouns are abstract nouns that do not convey a lot of meaning on their own: their meaning can be retrieved from the surrounding context. Words like *context, fact, form, purpose*, and *result* have been found to be often used as shell nouns in academic writing acting like empty shells that enclose or anticipate the surrounding discourse (Aktas & Cortes 2008).

Even though lexical bundles are usually not complete units, they have strong grammatical correlates. In their structural classification of lexical bundles, Biber et al. (2004a: 380–381) identified three major categories:

1. Lexical bundles that incorporate verb phrase fragments: These bundles begin with a subject followed by a verb phrase fragment, or they begin with a discourse marker followed by a verb phrase fragment or start directly with a verb phrase. Expressions such as *is going to be*, *can be used to*, and *as shown in figure* are some examples of this type of bundles.
2. Lexical bundles that incorporate dependent clause fragments: In addition to the verb phrase, these bundles incorporate a dependent clause fragment. This structural correlate can be seen in expression like *if you want to* and *I want you to*, among many other bundles.
3. Lexical bundles that incorporate noun phrase and prepositional phrase fragments: These lexical bundles are phrasal in nature as opposed to the previous types, which were clausal. They are made up of noun phrases or prepositional phrases that start the bundle, followed by other noun or prepositional phrase fragments as in, for example, *the end of the*, *in the context of*, or *the way in which*.

While lexical bundles identified in everyday conversation and other spoken registers (such as university lectures) cover the three types of lexical bundles just introduced, they are majorly clausal. Lexical bundles frequently found in written academic genres, on the other hand, are mostly phrasal, made up of fragments of noun or prepositional phrases.

7. From structure to function to communicative purpose

Grouping lexical bundles according to their grammatical correlates is a useful primary step to structurally organize lexical bundles and identify tendencies in the use of bundles in specific registers. It is even more important to identify the functions lexical bundles are performing in a given register in order to analyze the saliency of this frequent expressions, particularly in written academic genres. Various researchers and research groups have been working on the design of functional taxonomies for the classification of lexical bundles. These researchers are interested in analyzing bundles in the contexts in which they frequently occur to identify the way in which they are used and the meanings they convey when used in particular registers.

7.1 Functional taxonomy development

Biber et al. (1999) introduced a group of functions for some of the bundles these authors had identified and grouped according to their structural correlates, indicating for example that the bundles they identified in the academic prose section of their corpus were used to express existence or presence (e.g. *the presence of the, the existence of the*), to identify abstract qualities (*as in the nature of the, the value of the*) or to report stance (in bundles such as *it is possible to, it is important to*), among other functions.

In her study of lexical bundles in a corpus of freshman composition, Cortes (2002a) introduced an initial taxonomy for the functional classification of lexical bundles. Her taxonomy included four categories, accompanied here by bundles that illustrate each function: (1) Location markers used to refer to physical locations: *in the middle of, the other side of, the top of the*; (2) Temporal markers used to refer to a point or period of time: at *the same time, at the end of*; (3) Text markers used to guide the reader to certain parts of the writing: *at the end of; the rest of the*, and (4) Special use bundles in expressions such as *in a way that*; *in the form of*. This classification was completed and improved in Cortes (2002b), in which more categories and subcategories were incorporated for the analysis of the lexical bundles identified in Biber et al. (1999) in both conversation and academic prose and more specifically for the bundles in her corpus of published and student writing in history and biology. This preliminary taxonomy included three major categories with some sub-categories: referential bundles (time markers, place markers, and text deixis markers), text organizers (contrast/comparison, inferential, focus, and framing), and stance markers (epistemic-certain/uncertain/probable possible, desire, ability, and obligation). A year later, Biber et al. (2003) published an improved initial taxonomy for the

lexical bundles previously identified in conversation and academic prose by Biber et al. (1999).

Biber et al. (2004a) presented a final version of their taxonomy that shares many features with the initial versions. They used this version to classify lexical bundles identified in the university lectures and textbooks section of the T2K SWAL corpus (Biber et al. 2004b). This fully developed taxonomy has been used in many studies of lexical bundles in academic prose in the literature (Ädel &Erman 2012; Biber 2006; Biber & Barbieri 2007; Chen & Baker 2011; Cortes 2004), even studies of lexical bundles in different languages like Spanish and Korean (Cortes 2008; Kim 2009; Tracy-Ventura et al. 2007), which shows that the taxonomy is flexible enough to accommodate lexical bundles that frequently occur in different registers (particularly in written academic register such as, research articles, university textbooks, research reports, reflection papers, etc.). The categories and sub-categories in this taxonomy derived from the bundles identified in academic prose include:

1. Stance expressions:
   Epistemic stance (impersonal): the fact that the, are more likely to)
   Attitudinal modality stance – obligation/directive (impersonal): it is important to; it is necessary to
   Attitudinal modality stance – ability (impersonal): can be used to; it is possible to
2. Discourse organizers:
   Topic introduction/focus: in this chapter we
   Topic elaboration/clarification: on the other hand; as well as the
3. Referential expressions:
   Identification/focus: is one of the; one of the most
   Specification of attributes – quantity: the rest of the
   Specification of attributes – tangible framing attributes: the size of the; in the form of
   Specification of attributes – intangible framing attributes: the nature of the; in terms of the
   Time/place/text reference – place: in the United States
   Time/place/text reference – time: at the same time; at the time of
   Time/place/text reference – text deixis: as shown in figure
   Time/place/text reference – multi-functional reference: at the beginning of; at the end of (Biber et al. 2004a: 384–388)

It is important to mention that even thought this taxonomy was designed following a bottom-up approach, that is, analyzing the bundles in their contexts

and holistically identifying the functions they were performing, many bundles performed more than one function across the context in which they appeared. The taxonomies just introduced matched the bundles to the function they performed most frequently in the corpora used for those studies (Biber et al. 2003, 2004; Cortes 2002a).

7.2 Other taxonomies

At the same time that Cortes (2002b) was working on her earliest version of the taxonomy for the functional classification of lexical bundles, Culpeper and Kytö (2002: 54) were working on a taxonomy for the classification of bundles in early modern English dialogues. Their corpus consisted of pseudo conversational style language extracted from trial proceedings and drama comedies. They considered this corpus a sample of availability for modern English spoken samples: trial proceedings representing authentic language and drama comedies representing constructed language. Their taxonomy contained several categories such as speech act fragments; modalizing fragments, discourse fragments, narrative fragments, and circumstantial fragments.

Hyland (2008) presented a taxonomy designed to analyze specialized academic genres such as Master's theses and dissertations. While many of the categories were based on Biber et al. (2004), "differences in the two corpora necessitated modifications" so the taxonomy designed by Hyland presented categories that more directly suited the registers in his corpus (Hyland 2008: 13). The three main categories in Hyland's taxonomy are: research oriented bundles (with subcategories such as location, procedure, quantification, description, topic); text oriented bundles (with transition signals, resultative signals, structuring signals, and framing signals as sub-categories), and participant oriented bundles (which could be sub-categorized into stance features, and engagement features). Although there is a lot of overlapping between these two taxonomies (in sub-categories and examples), Hyland brings up an important characteristic of lexical bundles: when the register under examination is very specific, the functions performed by the bundles frequently used in that register become very specific too. This specificity quality will be discussed in the next sub-section.

7.3 From functions to communicative purposes and rhetorical moves in academic prose

Latest analyses of the functions of lexical bundles in written academic genres and sub-genres yielded interested developments in the relationship between bundles and rhetorical moves. Several studies conducted in the last decade focused on the identification of lexical items that could be used to specify the different

stages of a genre, trying to find words or expressions that could characterize the different rhetorical moves of specific texts. Rhetorical moves have been studied in many genres but the experimental research article has without doubt been the most exhaustively investigated (Brett 1994; Kanoksilapatham 2003; Swales 1981; Williams 1999; Yang & Allison 2003). In his seminal study of communicative functions in research article introductions, Swales (1981) introduced a four-move scheme for the analysis of these sections that initiated a long tradition of move identification research in written academic genres. In spite of being a small-scale study, Swales already identified a brief list of expressions that could be associated to each move in those sections. Kanoksilapatham (2003) also discovered expressions that frequently occurred in the move-scheme of the biochemistry research articles in her study but, because the focus was on single words or short expressions, the findings of this type of studies have been limited. Cortes (2013a) used a data-driven approach to the identification of lexical bundles in research article introduction. She started with a corpus of these sections and identified the most frequent bundles, starting with 4-word bundles and attempting to identify the longest possible bundles. In addition, the purpose of her study was to find the relationship between these expressions and the moves and steps in which they occurred (Swales 2004: 230–232). She identified expressions of up to 9-words which, in many cases, directly related to a specific move. Her findings stress the fact that lexical bundles are register-bound, as demonstrated by the fact that many bundles identified in this sub-register had never been identified before as bundles, even in studies that had used corpora made up of whole research articles. Some of the expressions identified by Cortes (2013a) were used to trigger the move, while others were used to comment on the move that had been triggered using other linguistic features. For example, one very frequent bundle identified in her study was *the purpose of the study is to*, which was used to trigger move 3 (introducing the present work) step 1 (announcing present research descriptively and/or purposively). A different example can be seen in the 5-word bundle *little is known about the*, which was always used to trigger move 2 (establishing a niche), step 1 (indicating a gap). Another finding of this study was the identification of long bundles that were complete structural units, such as whole clauses or sometimes even sentences such as *the rest of the paper is organized as follows*, a 9-word bundle that was frequently used in introductions of Business and Finance research articles. This methodology was also used by Cortes and Cotos (2012) and Cortes (2013b) in a corpus of methodology sections in research articles from 30 different disciplines, trying to categorize bundles in a four-move scheme. These authors found very similar results identifying expressions that can be used to characterize the rhetorical moves that are specific of methods sections. In these studies, the bundles many times reported experimental procedures (move 2) derived from

the specific disciplines represented in the corpus and the type of research conducted as in *participants were randomly assigned to* or *total RNA was extracted from*. Bundles such as a *randomized complete block design with* or as *the mean plus or minus* were used in move 4 (detailing statistical procedures). As shown in these examples, longer bundles incorporate a wider variety of lexical words, which helps make their communicative function very clear and reflects the purpose of the discourse, facilitating the identification of the bundle-move relationship. These findings could have a strong impact in the teaching of academic writing in courses that are genre-oriented. They emphasize the pervasiveness existing in the language of research articles and the frequent use of these expressions that have been considered clichés and have been excluded from the teaching of academic genres in the rhetorical tradition (Craswell 2004: 71). Introducing these lexical bundles together with the rhetorical moves they help communicate and having students analyze their communicative functions in the writing of their disciplines could emphasize the power these expressions have as building blocks for specific registers.

8. Conclusion

This chapter tried to present first a chronological view of the development of the study of lexical bundles, focusing later on the identification and classification of these expressions with an emphasis on studies that analyzed lexical bundle use in written academic registers. It is undeniable that the first studies that identified lexical bundles strongly highlighted frequency as the ultimate quality of these expressions. Being their defining quality, frequency is important but later studies have demonstrated that the functions lexical bundles perform make them also salient. The extensive use of the taxonomies previously described for the categorization of lexical bundles will help strengthen the categories and sub-categories in those taxonomies and create new ones when necessary. In addition, the latest developments that connect lexical bundles to communicative functions as expressed in rhetorical moves emphasize the need for studies that look at these expressions in relation to the discourse in which they are immersed. From the simple fact that these expressions are very frequent, it could be concluded that lexical bundles are easy to learn and use but many studies have shown that novice writers do not make use of lexical bundles with the same frequency or to express the same functions as expert writers do (Cortes 2004; Chen & Baker 2011). The field needs to continue studying the internal elements of lexical bundles as well as their relationship to their surrounding discourse through their semantic prosodies and preferences (Partington 2004).

In addition it is also important to assess the relationship between the use of lexical bundles and writing proficiency and writing expertise to better explore ways to initiate novice writers in the appropriate use of these expressions in academic registers.

References

Ädel, Annelie & Erman, Britt. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundle approach. *English for Specific Purposes* 31: 81–92. DOI: 10.1016/j.esp.2011.08.004

Aisenstadt, Esther. 1979. Collocability restrictions in dictionaries. In *Dictionaries and Their Users* [Exeter Linguistics Studies 4], Reinhard R.K. Hartman (ed.), 71–73. Exeter: University of Exeter.

Aktas, Rahime Nur & Cortes, Viviana. 2008. Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes* 7(1): 3–14. DOI: 10.1016/j.jeap.2008.02.002

Allen, Sture. 1970. *Frequency Dictionary of Present-Day Swedish Based on Newspaper Material,* 1: *Graphic Words.* Stockholm: Almqvist & Wiksell.

Allen, Sture. 1975. *Frequency Dictionary of Present-Day Swedish Based on Newspaper,* 3: *Collocations*. Stockholm: Almqvist & Wiksell.

Altenberg, Bengt. 1993. Recurrent word combinations in spoken English. In *Proceedings of the Fifth Nordic Association for English Studies Conference*, Julian D'Arcy (ed.), 17–27. Reykjavik: University of Iceland.

Altenberg, Bengt & Eeg-Olofsson, Mats. 1990. Phraseology in spoken English: Presentation of a project. In *Theory and Practice in Corpus Linguistics*, Jan Aarts & Willem Meijs (eds), 1–26. Amsterdam: Rodopi.

Barfield, Andy & Gyllstad, Henrik. 2009. Researching L2 collocation knowledge and development. In *Researching Collocation in Another Language*, Andy Barfield & Henrik Gyllstad (eds), 1–21, Houndmills: Palgrave Macmillan.

Biber, Douglas. 2006. *University Language. A Corpus-based Study to Spoken and Written Registers* [Studies in Corpus Linguistics 23]. Amsterdam: John Benjamins. DOI: 10.1075/scl.23

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics* 14: 275–311. DOI: 10.1075/ijcl.14.3.08bib

Biber, Douglas & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26: 263–286. DOI: 10.1016/j.esp.2006.08.003

Biber, Douglas & Conrad, Susan. 1999. Lexical bundles in conversation and academic prose. In *Out of Corpora: Studies in Honor of Stig Johansson*, Hilde Hasselgard & Signe Oksefjell (eds), 181–189. Amsterdam: Rodopi.

Biber, Douglas, Conrad, Susan, & Cortes, Viviana. 2003. Lexical bundles in speech and writing: An initial taxonomy. In *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, Andrew Wilson, Paul Rayson & Tony McEnery (eds), 71–92. Frankfurt: Peter Lang.

Biber, Douglas, Conrad, Susan & Cortes, Viviana. 2004a. 'If you look at …': lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371–405. DOI: 10.1093/applin/25.3.371

Biber, Douglas, Conrad, Susan, Reppen, Randi, Byrd, Pat, Helt, Marie, Clark, Victoria, Cortes, Viviana, Csomay, Eniko & Urzua, Alfredo. 2004b. *Representing Language Use in the University: Analysis of the TOEFL® 2000 Spoken and Written Academic Language Corpus* [TOEFL Report MS-25]. Princeton NJ: ETS.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Brett, Paul. 1994. A genre analysis of the result sections of sociology articles. *English for Specific Purposes* 13: 47–59. DOI: 10.1016/0889-4906(94)90024-8

Butler, Christopher. 1997. Repeated word combinations in spoken and written text: Some implications for functional grammar. In *A Fund of Ideas: Recent Development in Functional Grammar,* Christopher Butler, John Connolly, Richard Gatward & Roel Vismans (eds), 60–77. Amsterdam: Institute for Functional Research into Language and Language Use.

Chen, Yu-Hua & Baker, Paul. 2011. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology* 14: 30–49.

Cortes, Viviana. 2002a. Lexical bundles in freshman composition. In *Using Corpora to Explore Linguistic Variation* [Studies in Corpus Linguistics 9], Randi Reppen, Susan Fitzmaurice & Douglas Biber (eds), 131–145. Amsterdam: John Benjamins. DOI: 10.1075/scl.9.09cor

Cortes, Viviana. 2002b. Lexical Bundles in Published and Student Academic Writing in History and Biology. Ph.D. dissertation at Northern Arizona University.

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. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3: 43–57. DOI: 10.3366/E1749503208000063

Cortes, Viviana. 2012. Lexical bundles and technology. In *The Encyclopedia of Applied Linguistics,* Carol Chapelle (ed.). Oxford: Wiley Blackwell. ⟨http://dx.doi.org/10.1002/9781405198431.wbeal0689⟩

Cortes, Viviana. 2013a. The purpose of this study is to: Connecting lexical bundles to moves in research article introductions. *Journal of English for Academic Purposes* 12: 33–43. DOI: 10.1016/j.jeap.2012.11.002

Cortes, Viviana. 2013b. The participants were randomly assigned…: Lexical bundles in research article methodology sections. Paper presented at the Conference for the American Association of Applied Linguistics, Dallas TX, March 16–19.

Cortes, Viviana & Cotos, Elena. 2012. Lexical bundles: Enhancing automated analysis of methodology sections. Paper presented at the Technology for Second Language Learning Conference, Ames IA.

Coulmas, Florian. 1981. *Conversational Routines.* Berlin: Mouton de Gruyter. DOI: 10.1093/applin/2.3.223

Cowie, Anthony. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2: 223–235.

Cowie, Anthony. 1998. *Phraseology: Theory, Analysis, and Applications*. Oxford: OUP.

Craswell, Gail. 2004. *Writing for Academic Success: A Post Graduate Guide.* London: Sage.

Culpeper, Jonathan & Kytö, Merja. 2002. Lexical bundles in early modern English dialogues: A window into the speech-related language of the past. In *Sounds, Words, Texts and Change. Selected Papers from the 11 ICHEL, Santiago de Compostela* [Current Issues in Linguistic Theory 224], Teresa Fanego, Belén Méndez-Naya & Elena Seoane (eds), 45–65. Amsterdam: John Benjamins. DOI: 10.1075/cilt.224.06cul

De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3: 59–80. DOI: 10.1075/ijcl.3.1.04dec

Ellis, Nick, Simpson-Vlack, Rita & Mynard, Carson. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42: 375–396.

Firth, John Rupert. 1951. *Modes of Meaning. Essays and Studies of The English Association* [NS 4], 118–149.

Fleury, Serge & Zimina, Maria. 2006. *MkAlign. Manuel d'utilisation*. Centre of Textometrics *CLA²T*, Paris Sorbonne University – Paris 3. ⟨http://www.translationdirectory.com/articles/article1263.htm⟩ (16 December 2013).

Flowerdew, John. 2003. Signalling nouns in discourse. *English for Specific Purposes* 22: 329–346. DOI: 10.1016/S0889-4906(02)00017-0

Granger, Sylviane & Paquot, Magali. 2008. Disentangling the phraseological web. In *Phraseology: An Interdisciplinary Perspective,* Sylvianne Granger & Fanny Meunier (eds), 27–50. Amsterdam: John Benjamins.

Hakuta, Kenji. 1974. Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning* 24: 287–297. DOI: 10.1111/j.1467-1770.1974.tb00509.x

Hoey, Michael. 2005. *Lexical Priming*. London: Routledge.

Hunston, Susan, & Francis, Gill. 1998. Verbs observed: A corpus-driven pedagogic grammar of English. *Applied Linguistics* 19*:* 45–72. DOI: 10.1093/applin/19.1.45

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

Jespersen, Otto. 1933. *Essentials of English Grammar.* London: George Allen and Unwin.

Kanoksilapatham, Budsaba. 2003. A Corpus-based Investigation of Scientific Research Articles: Linking Move Analysis with Multidimensional Analysis. Ph.D. dissertation, Georgetown University.

Kim, YouJin. 2009. Korean lexical bundles in conversation and academic texts. *Corpora* 4: 135–165. DOI: 10.3366/E1749503209000288

Kjellmer, Goran. 1991. A mint of phrases. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik,* Karin Aijmer & Bengt Altenberg (eds), 111–127. London: Longman.

Lebart, Ludovic, Salem, André & Berry, Lisette. 1998. *Exploring Textual Data.* Dordrecht: Kluwer. DOI: 10.1007/978-94-017-1525-6

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach.* Oxford: Clarendon Press.

Nattinger, James & DeCarrico, Jeanette. 1992. *Lexical Phrases and Language Teaching*. Oxford: OUP.

Palmer, Harold. 1933. *Second Interim Report on English Collocations.* Tokyo: Kaitakusha.

Partington, Alan. 2004. Utterly content in each other's company: Semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9: 131–156. DOI: 10.1075/ijcl.9.1.07par

Renouf, Antoinette & Sinclair, John. 1991. Collocational frameworks in English. In *English Corpus Linguistics: Studies in Honour of Jan Svartvik,* Karin Aijmer & Bengt Altenberg (eds), 111–127. London: Longman.

Salem, André. 1987. *Pratique des segments r*épétés. Paris: Publications de L'InaLF.

Sinclair, John. 1987. *Looking Up. An Account of the Cobuild Project in Lexical Computing*. London: Collins Cobuild.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: OUP.

Sinclair, John. 2004. *Trust the Text: Language, Corpus, and Discourse.* London: Routledge.

Schmid, Hans-Jörg. 2000. English *Abstract Nouns as Conceptual Shells: From Corpus to Cognition.* Berlin: Walter de Gruyter. DOI: 10.1515/9783110808704

Swales, John. 1981. *Aspects of Article Introductions.* Birmingham: The University of Aston, Language Studies Unit.

Swales, John. 2004. *Research Genres: Exploration and Applications*. Cambridge: CUP. DOI: 10.1017/CBO9781139524827

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work* [Studies in Corpus Linguistics 6]. Amsterdam: John Benjamins. DOI: 10.1075/scl.6

Tracy-Ventura, Nicole, Cortes, Viviana & Biber, Douglas. 2007. Lexical bundles in Spanish speech and writing. In *Working with Spanish Corpora,* Giovani Parodi (ed.), 217–231. London: Continuum.

Wang Filmore, Lilly. 1979. Individual differences in second language. In *Individual Differences in Language Ability and Behavior,* Charles Fillmore, Daniel Kempler &William Wang (eds), 203–228. New York NY: Academic Press. DOI: 10.1016/B978-0-12-255950-1.50017-2

Williams, Ian. 1999. Results sections of medical research articles: analysis of rhetorical categories for pedagogical purposes. *English for Specific Purposes* 18: 347–366. DOI: 10.1016/S0889-4906(98)00003-9

Yang, Ruiying & Allison, Desmond. 2003. Research articles in applied linguistics: moving from results to conclusions. *English for Specific Purposes* 22: 365–385. DOI: 10.1016/S0889-4906(02)00026-1

Yorio, Carlos. 1980. Conventionalized language forms and the development of communicative competence. *TESOL Quarterly* 14(4): 433–442. DOI: 10.2307/3586232

Index

A

academic xi, xii, xiii, xiv, xx, 1–9, 16–17, 19, 31, 37, 49–52, 54–56, 65, 67–68, 70, 73–74, 79–81, 83–84, 91, 94–95, 100, 105–107, 110, 113, 115–119, 123–131, 138, 140–142, 153, 155–164, 166–170, 177–181, 187–193, 197–199, 204, 206–214

academic prose 2, 52, 178, 197–199, 204, 207–210

academic registers xii, 9, 55, 212–213

academic writing xii, xviii, 2, 49, 50–54, 56, 65, 67–68, 70, 73–74, 107, 113, 117–119, 125, 128, 131, 140, 166, 178, 181, 190–191, 198–199, 206–207, 212

addressee-focused, polite, and elaborated information 27, 30, 33, 34

adjectives 9, 15, 17, 23, 29, 52–53, 58, 60, 63, 69, 79–81, 83, 88–89, 200, 206

AntConc 9

appositive noun phrases 52, 58, 73

appraisal 81, 83, 89, 91

B

backchannels/backchanneling 40, 41, 47

Biber tagger 57, 130, 132

C

Call Home corpus 25, 27, 31

circumstance adverbials xx, 147, 152, 157–158, 161

circumstance adverbials and focus 161

circumstance adverbials and position 158

classroom talk 20

clausal embedding 52, 68

clausal elaboration 49, 52, 59–60, 63, 65–66, 69–70, 72

coding 83, 97, 183, 191

collocations 197–198, 200–202, 206, 213, 215

communicative markers 41

communicative purpose 3, 19, 34

complement clauses 8–9, 15, 18, 23, 52–55, 58–63, 69

complexity 37, 49–52, 56–59, 63, 65, 72–74, 139, 152

composition 28–29, 31, 33, 43, 105–107, 109, 125, 129, 142, 182, 185, 193, 195, 208

computing dimension scores 10, 32

confidence 81, 84–88, 91, 93, 141, 148, 171

context(ual), contextualization 83, 89, 99, 101

corpus design 56–57, 126, 128–129, 131, 140, 152

Corpus of Contemporary Indian English 151–154

corpus representativeness 123, 126, 129, 143

corpus-driven methodology 197–199

corpus-internal evidence 129, 141

cross-sectional 101–102, 108

curriculum 54, 79, 105–106, 113, 117–119, 195

customer service xix, 25–27, 31, 34–35, 37, 41, 43–44

cut-off points 183, 202, 204–206

D

decontextualized 100, 102, 104, 189–190

disciplinary writing 49, 57, 177

discourse xii, xiii, xiv, xviii, xix, 1–2, 6–9, 11–17, 19–20, 23, 25–26, 28–30, 33–38, 40–42, 44, 49–51, 55–56, 58–59, 67, 72, 74, 81, 85, 88, 96, 100, 127, 149, 151, 162, 177–178, 180, 190–191, 197–199, 203–204, 207, 209–210, 212, 216

discourse particles 8, 15, 23, 29–30, 36–38, 40, 42

doubt raisers 80, 83

E

English for Specific Purposes 54, 193–195, 198, 213–216

evaluative language 79, 96

F

face (threats/needs) 84

face-to-face conversation 32, 35, 43, 44

final position 157–159, 170

fixed expressions 198–199, 201–202, 206, 215

formulaic language xx, 197, 199–201, 203

formulas 83, 91, 197, 200–201, 206

frames xix, 79, 82, 89–92, 198, 206

frequency xii, 10, 17, 33–35, 37, 41, 53–54, 59, 64, 67, 73–74, 82–84, 88, 90, 111–115, 125–126, 128, 131–132, 141, 157–158, 162, 166, 170, 187, 190, 198–200, 202–206, 212–213

frequency-driven, frequency-based expressions 203

functional dimensions 27

functional taxonomy 208

G

gender 26, 28, 32, 53, 80–82

grammatical xix, xx, 2, 7, 9, 51–53, 55, 58, 73, 84, 99, 111, 117, 148, 151–152, 155, 171, 177, 180, 183–184, 188, 198, 204, 206–208

I
idioms 198, 200, 215
Indian English xx, 147, 149–162, 166, 169–174
initial position 159
informational discourse 20, 49
instructional xix, 3, 37, 99–100, 102, 104, 113, 117, 119, 123–124, 126, 132, 138, 142, 190
instructional language 37
internal variation xx, 115, 147, 152, 170–171
interpretation xiii, xiv, 17, 27, 65, 67, 82, 89, 94–96, 104, 119, 134
Involved and simplified narrative 27, 30, 33, 34
involvement 27, 99, 100, 105, 109, 112, 113, 119

K
KWIC (key word in context) 202

L
learner corpus 99–102, 105, 107, 109–110, 112–113, 115, 117, 119–120
lexical xiii, xix, xx, 20, 26, 28, 36–37, 55, 58, 82–83, 88, 111, 123–126, 129, 131–132, 134, 136, 138, 140–143, 148, 151, 171, 183–184, 197–213, 215
lexical bundles xiii, xx, 197–213
lexical Bundles Program (LBP) 205
lexical co-occurrence 197
lexical distributions 131, 134, 136, 140–141, 143
lexical diversity 123, 126
lexical variability xix, xx, 123–126, 129, 136, 138, 140, 142
longitudinal 101–102, 108–109, 114–117
Longman Grammar of Spoken and Written English xviii, xix, 51, 197

M
managed information flow 27, 30, 40, 41
markers of adequacy 83, 87–89
markers of excellence 83, 87–89
medial position 157–159, 170
mitigation strategies xix, 79, 81, 91
modals xix, 9, 16, 23, 29–30, 33–34, 79, 81–89
moves (see rhetorical moves) 2, 189, 210–212
Multidimensional analysis 1, 6, 25, 27

N
negative xix, 10–11, 15, 17, 23, 30, 33–38, 79, 82–83, 88–95, 103, 192
negative information xix, 79, 82–83, 90–94
nominal pre-modifiers 52, 54, 58, 64, 68
normalization 205

O
occluded genre 79, 81
Outsourced Call Centers 44

P
paradigmatic dimension 89, 203, 204
paraphrase 178–185, 187, 189–191
participant language use 1
pedagogy, pedagogical xi, 99, 100, 103, 109–111, 179
phrasal compression 49, 52, 59, 63, 67, 71–73
phrasal embedding 52
phraseological tradition 199
plagiarism 178–180, 184–185, 187, 189–190, 194
planned, procedural talk 27, 30, 36–37
positive xix, 10–12, 15–16, 23, 30, 33–36, 38, 40–42, 69, 79, 83, 86–95, 103, 190
prediction/predictor 84–86, 91, 139
prepositional phrases 9, 15, 23, 52–54, 58, 63–64, 73, 207
process-based information 37
proficiency xix, 101, 105, 107–108, 112–114, 119–120, 178, 180–181, 192, 213

Q
qualitative xiii, xiv, xviii, xix, 56, 57, 59, 69, 73, 79, 81, 83, 96, 114, 116, 149, 183
quantitative xiii, xiv, xviii, xix, 5, 50, 56–57, 59, 65, 68–73, 79, 81, 86, 91, 149, 153, 183
quasi-longitudinal 102, 108, 116

R
recommendation letters xix, 79, 81, 82, 90, 95, 96
register features 26, 56, 73
register variation xii, xviii, 50–51, 149, 172
relative clauses 9, 15, 18, 23, 52–53, 58, 65–70
reliability 58, 80, 123, 126, 129, 138, 140–142
research articles xiv, xix, 49, 51–52, 54–58, 70–73, 125, 197–198, 209, 211–212
restrictive circumstance adverbials xx, 152
rhetorical moves 2, 189, 210–212
rhetorical structure 55, 182

S
segments répétés 203
semantic prosody 88
shell nouns 206–207
situational characteristics 3–5
spoken corpora xi, 25, 37
spoken academic corpus 1
stance xix, 1, 9, 11–15, 18–20, 23, 27–28, 42, 55, 62, 84, 113, 117, 208–210
student presentations xviii, 1–4, 6–8, 10, 16–17
student writing xix, 2, 55, 125, 166, 178, 192, 208
subject-auxiliary inversion 155–158, 166, 169–170
summary xiii, 19, 30, 43, 50, 64, 67–68, 91–92, 150, 179–183, 185–191

Switchboard corpus 25, 27, 32
syntagmatic dimension 204

T
target domains xx, 123, 127, 141
teacher involvement 99–100, 105, 109, 112, 119
teacher presentations xviii, 1, 3–4, 8, 10, 13
telephone interactions 25–26, 37, 40, 43, 45
telephone-based corpora 25, 27, 41
textometric analysis 203
textual borrowing xx, 177–181, 183, 185, 187, 188, 189, 191, 192

U
university setting xviii, 1, 2, 7–12

W
WH-questions xx, 8, 23, 147, 152, 155–158, 166, 169, 170
word combinations 197–198, 200–203, 205
word list reliability 123
word list stability 140
word lists xx, 82, 123–126, 128–129, 132–134, 136, 138–143
writing xii, xiv, xviii, xix, xx, 2–3, 6, 8, 12–13, 49–52, 54–57, 64–65, 67–68, 70, 73–74, 94–95, 97, 105–108, 110, 112–115, 117–119, 125, 128, 131, 140, 156–157, 165–166, 168, 177–182, 184–186, 189–196, 198–199, 202, 206–208, 212–214