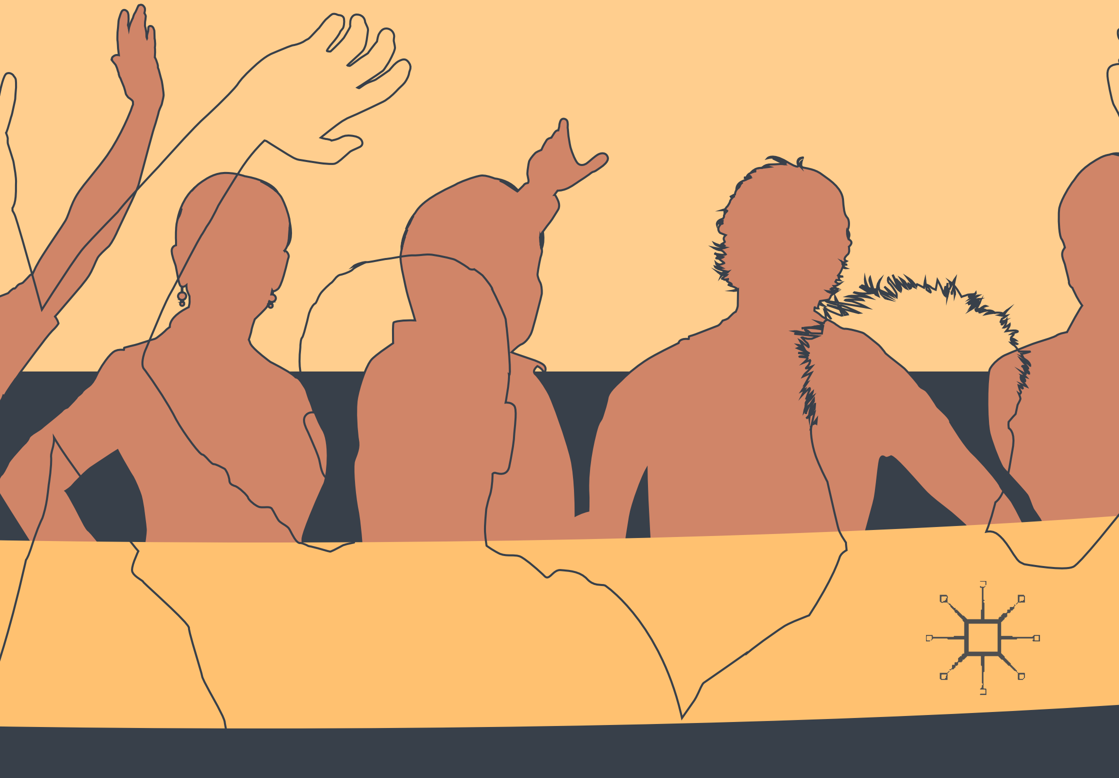
Exploring Spoken English Learner Language Using Corpora

Learner Talk

Eric Friginal, Joseph J. Lee,
Brittany Polat, and Audrey Roberson



Exploring Spoken English Learner Language Using Corpora

‘Finally, some principled empirically-based information on qualities of spoken language in context! For several decades, a promise of second language (L2) corpus linguistics has been to revolutionize ways of teaching English to speakers of other languages. But prior to this book’s publication, most L2 corpus resources have focused on genres of the written language. As a result, specialists in research and teaching of the spoken language have felt somewhat frustrated. We are intrigued by the great potential corpus tools offer since we witness the many exciting ways in which they are applied to the written language. Partly because spoken corpora are notably more difficult to generate and analyze, the infusion of corpus tools into research and teaching of the spoken language has been limited. This book goes far in alleviating such concerns since it expands the landscape of corpus studies to include several core genres of the spoken language.’

—John Murphy *Georgia State University, USA*

‘This is a long-awaited volume presenting a brief introduction to corpus linguistics and a variety of excellent corpus-based studies on spoken learner language in the university setting. The authors provide a historical overview of the research in this area, offer a range of new approaches to the analysis, introduce accessible learner corpora, and discuss pedagogical applications. The reader finds a state-of-the-art picture of research and plenty of ideas for future directions to analyze spoken learner language. I highly recommend this volume to researchers and students alike.’

—Eniko Csomay *San Diego State University, USA*

Eric Friginal
Joseph J. Lee • Brittany Polat
Audrey Roberson

Exploring Spoken English Learner Language Using Corpora

Learner Talk

palgrave
macmillan

Eric Friginal
Applied Linguistics and ESL
Georgia State University
Atlanta, Georgia, USA

Brittany Polat
Georgia State University
Atlanta, Georgia, USA

Joseph J. Lee
Ohio University
Athens, Ohio, USA

Audrey Roberson
Hobart and William Smith Colleges
Geneva, New York, USA

ISBN 978-3-319-59899-4 ISBN 978-3-319-59900-7 (eBook)
DOI 10.1007/978-3-319-59900-7

Library of Congress Control Number: 2017946322

Cover illustration: © chipstudio / Getty Images

Printed on acid-free paper

This Palgrave Macmillan imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Summary

As second language (L2) corpus studies expand into their third decade, innovations in computational technology and corpus creation have facilitated unprecedented access to authentic language in the classroom, including among non-native speakers (NNSs) of English. This book focuses on corpus-based analyses of learner oral production in university-level English or English as a Second Language (ESL) classrooms. Our analyses highlight three specialized corpora collected for the three empirical parts of this book, explored using a range of corpus approaches and methods: (1) learner talk in the English for Academic Purposes (EAP) classroom, (2) learner talk in English language experience interviews, and (3) learner talk in peer response/feedback activities. Historical and methodological perspectives in exploring spoken learner corpora, pedagogical applications, and future directions in studying learner language are discussed. A synthesis of corpus-based research of spoken learner language, list of available corpora and online databases, and an introduction to corpus linguistics and corpus tools and approaches are provided in the first two chapters of the book.

Acknowledgement

We would like to thank our mentors and colleagues at the Department of Applied Linguistics and ESL at Georgia State University (GSU), especially Gayle Nelson, John Murphy, and Sara Cushing for their guidance and critical perspectives in developing the three empirical studies presented in this book. Thanks to Mike Cullom for his valuable insights and reviews of earlier drafts of this book, our Palgrave Macmillan commissioning editors and reviewers, and the staff of the Longview Public Library, Longview, WA. We recognize Douglas Biber and Randi Reppen at Northern Arizona University; Lucy Pickering at Texas A&M, Commerce; John Swales and Rita Simpson-Vlach at University of Michigan and for their work with MICASE; Laurence Anthony at Waseda University; and the Learner Corpus Association (Founding Members: Gaëtanelle Gilquin, Sylviane Granger, Fanny Meunier and Magali Paquot at the Centre for English Corpus Linguistics, Université Catholique de Louvain) for leading the way with their corpus tools and seminal studies of learner language. Much appreciation to our GSU colleagues: Diane Belcher, Stephanie Lindemann, Scott Crossley, Youjin Kim, Ute Römer, Jack A. Hardy, Pam Pearson, Nic Subtirelu, Cassie Leymarie, and many other collaborators who have assisted in the data collection and analyses in various parts of this book. We are grateful to all our study participants and especially the students, instructors, and administrators at GSU's Intensive English Program (Cheryl Delk-Le Good, John Bunting, Debra Snell, Louise

Gobron, and Alison Camacho). Data collection in Part 3 of this book was supported by grants from the Educational Testing Service (ETS) and *Language Learning*. Finally, we dedicate this book to our families and friends, and thank them for their love and support: Mike and Beth Cullom, Donna and Ela Friginal; Chang Keun Lee and Joanne Y. Lee, and Helen, Hetty, and Jules Lee; Ali, Guinevere, Thomas, and Charlie Polat; Susan and Jim Roberson, and Michael Mills.

Eric Friginal
Joseph J. Lee
Brittany Polat
Audrey Roberson

Contents

About the Authors

Eric Friginal is Associate Professor of Applied Linguistics at the Department of Applied Linguistics and ESL, and Director of International Programs, College of Arts and Sciences, at Georgia State University. He specializes in (applied) corpus linguistics, sociolinguistics, cross-cultural communication, and the analysis of spoken professional discourse. His recent books include *Talking at Work: Corpus-Based Explorations of Workplace Discourse* (2016, Palgrave Macmillan), co-edited with Lucy Pickering and Shelley Staples; *Studies in Corpus-Based Sociolinguistics* and *Corpus Linguistics for English Teachers* (2017–2018, Routledge).

Joseph J. Lee is the Assistant Director of the ELIP Academic & Global Communication Program, and Director of ELIP Center for Academic Communication: Tutoring Services in the Department of Linguistics, Ohio University. His research and teaching interests include ESP/EAP, genre studies, classroom discourse, advanced academic literacies, applied corpus linguistics, and teacher education. His recent publications include research articles in *English for Specific Purposes, Journal of English for Academic Purposes*, and *Journal of Second Language Writing.*

Brittany Polat is an independent ESL researcher based in Lakeland, Florida. Her research interests include second language acquisition, pragmatics, and corpus linguistics. Her research has appeared in journals such as *Applied Linguistics*, *Journal of Pragmatics*, and *Corpus Linguistics Research*.

Audrey Roberson is Assistant Professor of Education at Hobart and William Smith Colleges in Geneva, New York, where she oversees TESOL certification in the department's Teacher Education Program, as well as directs a certificate program in TEFL. Her research interests include language teacher preparation, applied corpus linguistics, interaction in second language learning, and second language writing. She has co-authored articles in *Corpora* and in the composition journal *Across the Disciplines*.

List of Figures

List of Tables

Part I

Introduction

1

Exploring Spoken English Learner Language Using Corpora

As second language (L2) corpus studies expand into their third decade, innovations in computational technology and corpus creation have facilitated unprecedented access to authentic language in the classroom, including among non-native speakers (NNSs) of English. NNS writing across various written contexts (e.g., school essays, standardized tests/proficiency tests, and laboratory or research reports) has been studied extensively in both journal article and book formats using corpora by applied linguists including Douglas Biber, Ken Hyland, John Swales, Rod Ellis, Susan Conrad, Eli Hinkel, and Sylviane Granger, to name only a few. Despite these impressive contributions, gaps still remain in our knowledge of spoken English L2 registers, even those that are quite important for NNSs to master. Classroom learner speech and face-to-face NNS interviews, for example, have been researched both qualitatively and quantitatively, primarily by utilizing the assessment of learner performance. However, extensive corpus-based analyses of these registers are still relatively few in number. Given that these oral learner skills are essential in high-stakes situations, such as admission to graduate programs, job interviews in English-speaking settings, or proficiency tests like the TOEFL (Test of English as a Foreign Language) or IELTS (International

E. Friginal et al., *Exploring Spoken English Learner Language Using Corpora*,
DOI 10.1007/978-3-319-59900-7_1

English Language Testing System), it is certainly useful and worthwhile to further investigate oral learner language systematically, and especially with corpora as part of the research methodology.

This book focuses on corpus-based analyses of learner oral production in university-level English or English as a Second Language (ESL) classrooms in the USA. Our overarching goal here is to provide an in-depth discussion and analysis of learner spoken language, with specific pedagogical impetus and applications. Our analyses highlight three specialized corpora collected for the three analytical parts of the book, explored using a range of corpus approaches and (mixed) methods: (1) learner (and also teacher) talk in the English for Academic Purposes (EAP) classroom; (2) learner talk in English language experience interviews; and (3) learner talk in peer response/feedback activities in the classroom. Pedagogical applications are discussed in each section and future directions in studying learner talk are provided in the concluding chapter (Chap. 14). A synthesis of corpus-based research of spoken learner language, list of available corpora and online databases, and an introduction to corpus linguistics and corpus tools and approaches are discussed in this first chapter of the book.

Studies of Spoken English Learner Language

Studies of spoken learner language are often situated in the field of Second Language Acquisition (SLA), with emphasis on the documentation and assessment of learner performance. For example, Ellis and Barkhuizen's (2005) *Analyzing Learner Language* highlighted the application of discourse and conversational analysis in exploring language learning as it takes place in interaction, but also covered the use of (written) learner corpora and contrastive analysis in SLA. In many experimental research settings, spoken learner language is evaluated from a variety of angles, focusing on the acquisition of L2 pronunciation and phonology; suprasegmental features of oral production; lexis and vocabulary development; and presentation, content, coherence, and delivery. Data are primarily extracted from audio and video recordings of real-world speech, transcriptions, and performance evaluations conducted by teachers. Learner speech in the classroom has also been measured according to quality and accuracy (e.g., accuracy of response to a teacher-initiated question), frequency of participation, conversational coherence, and usage and recall. Over the years, SLA research has produced meaningful data characterizing English learner speech across a range of speech events with clearly guided pedagogical implications.

The role of conversational interaction in SLA has been extensively studied utilizing a range of methodologies, most of them in experimental research settings. As briefly reviewed in some parts of this book, L2 learners' conversational interaction studies have been motivated by a few iterations of the interaction hypothesis from, for example, seminal works by Gass (1997), Long (1983, 1996), and especially Pica et al. (1989). As discussed by Saito and Akiyama (2017), the main focus of the hypothesis involves adult SLA which is facilitated and promoted through conversational interaction with NSs and NNSs. Such settings provide many opportunities for interactants to impact various aspects of conversation and the acquisition of conversational skills and competence. This is especially effective when interlocutors work together on negotiating and solving miscommunication.

The interaction–acquisition connection in spoken L2 has often been examined using a pretest–posttest design. With this approach, researchers are able to control various features of L2 interaction as independent variables and test their impact on L2 development (Plonsky and Gass 2011). In several studies, L2 learners improved their grammatical and lexical performance when given opportunities to negotiate meaning through interaction rather than through mere exposure to simplified input (Mackey 1999). Various opportunities for learners to respond to real-world questions, ask or clarify for comprehension, and engage extensively in the conversation have proven to be beneficial in improving oral production and performance in spoken tasks. Learners' "efficacy of interaction" also increased when they had sufficient proficiency with the target structures or if they had relatively high aptitude, especially when measured through working memory (Goo 2012). Other constructs such as pedagogically elaborated feedback (Sheen 2007), interlanguage development (Ziegler 2015) and specific location (e.g., laboratory vs. classroom settings) (Gass et al. 2005) have been explored in SLA, producing conclusive information underscoring the importance of conversational interaction on the acquisition of L2 spoken discourse features.

More recent studies of learner interaction (within experimental settings) have looked at video-based conversational interaction with a more longitudinal design. Saito and Akiyama (2017), for example, analyzed L2 production by college-level Japanese English-as-a-foreign-language (EFL) learners. Learners in the experimental group were asked to participate in weekly dyadic conversation with native speakers (NSs) in the USA. The NSs were trained to provide interactional feedback (recasts) when the Japanese learners' responses had comprehensibility issues. Learners in the comparison group received "regular" EFL instruction without any interaction with NSs. Saito and Akiyama's video data showed that the experimental group developed skills related to improving many linguistic domains of language, likely in response to their NS interlocutors' interactional feedback (recasts, negotiation) during the video-based interaction. The pretest–posttest data of the students' spontaneous production showed that they made significant gains in the dimensions of comprehensibility, fluency, and lexicogrammar but not in production areas such as accentedness and pronunciation.

Clearly, recorded data from this type of experiment may be further analyzed, and the texts compiled to form a corpus of conversational interaction. The corpus approach will provide additional insights into the linguistic characteristics of NNS and NS interaction that may add supporting evidence of the importance of conversational interaction and the unique linguistic features of interlanguage speech. What are the characteristic features of L2 negotiation? How are video-based interactions similar or different from face-to-face conversation (e.g., from a corpus of study groups or classroom feedback sessions)? Questions such as these may be answered by utilizing a corpus approach, given that parameters are already aligned to facilitate successful corpus compilation.

Studies of learner comprehension and how they modify speech (e.g., in providing comprehensible input) from repetitions, emphasizing slower speech rate, and the rephrasing of utterances with more frequent and simple words have all been examined in experiments, but these may also be analyzed from a comprehensive, well-developed corpus. From simple word counts to more advanced frequencies of reformulations, various corpus methods may also allow for distributions that can be used alongside test results. Corpora will further describe the linguistic features of L2 negotiation strategies (e.g., confirmation checks, clarification requests, recasts, or information packaging). These descriptions may be used to develop testing and teaching materials, and NNSs may also be induced to notice and understand the gap between their own L2 speech system and those of other learners, NSs, and their classroom instructors.

Finally, in addition to SLA, the related sub-fields of English for Specific Purposes (ESP) and, more specifically, English for Academic Purposes (EAP) have increasingly used corpora to systematically analyze and examine spoken learner language. Spoken texts (i.e., transcriptions of oral language) are carefully designed, with additional emphasis on quantity and representation of various associated registers. The corpus approach is limited, in that phonological features (segmental and supra-segmental features of speech) may not be directly included (and assessed) in the analysis. Up to this point, transcriptions of speech have been primarily verbatim, capturing word- and sentence-level features and distributions, for the most part. Although there are attempts at more in-depth annotation of spoken texts, the process to phonologically transcribe a corpus is still in its infancy.

Exploring Spoken English Learner Language Using Corpora

Corpus-based analysis of learner language has historically focused on written rather than spoken texts. Various collections of academic written language, from popular online databases, such as the Michigan Corpus of Upper-Level Student Papers (MICUSP), the British Academic Written English (BAWE), International Corpus of Learner English (ICLE) (and many other ICLE-inspired collections), and various learner written texts from corpora including the American National Corpus (ANC) and the Santa Barbara Corpus, have been widely used to compare registers of written L2 texts. Written corpora are certainly easier and less costly to compile, especially with the internet and advanced computational techniques. Corpus-based EAP research on written genres has flourished to a greater extent in the past few years than comparable research on spoken registers (Simpson-Vlach 2013).

Pioneering efforts to also focus sufficient attention on corpus-based analysis of spoken learner language, especially in English, have been initiated in the late 1990s and early 2000s. A recognition of the importance of spoken EAP corpora paved the way for the creation of the TOEFL 2000 Spoken and Written Academic Language (T2K-SWAL) Corpus (written and spoken texts combined), compiled by Douglas Biber and his colleagues at Northern Arizona University, Georgia State University, Iowa State University, and California State University, Sacramento (Biber et al. 2004). A corpus of academic speech, the Michigan Corpus of Academic Spoken English (MICASE), developed and collected by (applied) linguists from the University of Michigan (Simpson et al. 2002) focused exclusively on speech that represents oral language in a university setting (see the MICASE section in Chap. 2 for additional description of this corpus). Simpson-Vlach (2013) noted that:

> Prior to the development of spoken language corpora, the study and teaching of spoken academic language relied heavily on some combination of written academic discourse, conversational speech, or intuition to provide models of spoken language in academic contexts. With the availability of specialized corpora of academic speech, researchers and teachers gained access to resources that permit investigations of specific questions about grammar, lexis, usage, and discourse patterns as these actually occur in spoken academic contexts. These research inquiries have begun to fill in the gaps in our knowledge about the characteristics of academic speech as a specialized language genre. Results from such investigations are of interest to both applied linguists generally as well as EAP teachers and materials writers who can use such insights to better inform their teaching and materials development. A judiciously sampled spoken academic corpus constitutes a valuable research resource and set of models characterizing the spoken language that students will encounter and need to produce in the course of their academic endeavors. (p. 453)

Both MICASE and T2K-SWAL include L2 speech, especially from learner presentations and study groups, but these corpora of spoken academic texts focus more on spoken language in academia in general than upon an in-depth learner oral production. L2 speech is tangentially represented and can be extracted, but may still be limited when it comes to fully illustrating a learner-centered speech event in US universities. The advantage in using MICASE and T2K-SWAL is that both corpora include a wide range of speech events from classroom lectures (primarily on teacher-led lectures and discussions), laboratory sessions, tutorials, advising sessions, research interviews, dissertation defenses, public colloquia, meetings, and academic service encounters. As Simpson-Vlach (2013) argued, these spoken academic corpora are valuable collections of previously unavailable data that constitute an important resource for EAP and corpus practitioners. Nevertheless, within the larger world of corpus-based research, SLA, and ESL in the classroom, these seminal corpora are still relatively limited as far as how comprehensively they represent L2 speech.

There have been encouraging and important additions to MICASE and T2K-SWAL, with specialized collections targeting very specific groups of learners and sub-registers (e.g., interviews, computer-mediated communication, and peer response). It appears that the trend is to continue exploring learner talk through very specialized corpora and register-centered analysis. For example, Oral Proficiency Interviews (OPIs), which are widely used to measure speaking ability in a second or foreign language, are also now being explored using data from, for example, The Michigan English Language Assessment Battery (MELAB) speaking assessment (which is an OPI used for academic and professional purposes around the world). A study by Staples et al. (2017) shows that the MELAB has similarities with conversation in its use of stance and is closely aligned with academic registers and nurse–patient interactions in the use of language for informational exchange.

Overall, texts in these corpora, especially those collected in the classroom, are still comparatively restricted in number of speakers and total number of words, but more qualitative evidence may be utilized from accompanying audio/video files and researcher data (e.g., teacher observation reports, test results, student papers/reflections). Triangulating corpus-based distributions with results from qualitative data sources may produce meaningful results and relevant pedagogical implications. In this book, Parts II (learner talk in the classroom), III (learner talk in English language experience interviews), and IV (learner talk in peer response/feedback activities) all utilize specialized corpora that highlight, more than other collections of learner language, L2 speech in use within a very specific language teaching and learning contexts. The numbers, overall, are still low and could be beneficially increased in future related studies, but we present a clear model of corpus-based analysis (including semantic and psychosocial analytical constructs), with results that are descriptive of the register and potentially useful in aiding L2 spoken pedagogy.

Corpus Linguistics: A Brief Introduction

Corpus linguistics, primarily a *research approach* in the study of spoken and written texts, has evolved over a few decades to support empirical investigations of naturally occurring language-in-use. From (macro) collections of millions of texts to very specialized (micro) corpora, the corpus approach has been instrumental in providing in-depth descriptions of the linguistic characteristics of spoken and written discourse. Biber et al. (2010) emphasize that corpus linguistics is not, in itself, a model of language but a methodological approach that can be characterized as follows:

- It is empirical, analyzing the actual patterns of use in natural texts
- It utilizes a large and principled collection of natural texts, known as a *corpus* (pl. *corpora*), as the basis for analysis
- It makes extensive use of computers for analysis, employing both automatic and interactive techniques
- It relies on the combination of quantitative and qualitative analytical techniques.

Corpus-based researchers argue that language use is systematic and can be extensively described using empirical, quantitative, and frequency-based methods (Biber 1988). Corpora and corpus-based research provide extensive numerical data, but these will then have to be functionally interpreted meaningfully and accurately. Biber, as cited in Friginal

(2013), notes that quantitative patterns discovered through corpus analysis should always be subsequently interpreted in functional terms. Clearly, these patterns of linguistic variation exist because they reflect underlying functional differences. With corpus data, then, descriptions of written and oral production of L2 learners in the classroom may have greater generalizability and validity, producing a range of supporting evidence that could be further examined in research settings. Results and interpretations of these findings may be used to inform pedagogy—the creation of learning and teaching materials and L2 teaching lessons utilizing corpus tools.

What Is a *Corpus*?

> " … a corpus is a large and principled collection of natural texts." (Biber et al. 1998, p. 12)

> "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research." (Sinclair 2005)

> "… a corpus is a collection of (1) machine readable (2) authentic texts (including transcripts of spoken data) which is (3) sampled to be (4) representative of a particular language or language variety." (McEnery et al. 2006, p. 5)

> "Corpora may encode language produced in any mode—for example, there are corpora of spoken language and there are corpora of written language. In addition, some video corpora record paralinguistic features such as gesture (Knight et al. 2009) and corpora of sign language have been constructed (Johnston and Schembri 2006; Crashborn 2008)." (McEnery and Hardie 2012, p. 3)

> " … is a collection of spoken or written texts to be used for linguistic analysis and based on a specific set of design criteria influenced by its purpose and scope." (Weisser 2016, p. 13)

From the definitions above, a *corpus* (Latin, “body,” *corpora*, plural) can be briefly defined as a **systematically designed electronic collection of naturally occurring *texts***. The word text, as used in corpus-based research, is not limited to describing language that was initially written. Hence, a text can also be a transcription of spoken language. Even in the age of computers, the transcription of speech is still quite labor-intensive. Capturing various features of spoken language (e.g., dysfluent markers, repeats and reformulations, overlaps and backchannels, and many others) may require extensive hand coding and annotation. Although there have been recent advancements in dictation tools and “speech to text” technology (similar to the technology used in subtitles and close captioning on live television), the transcription of spoken data, especially by teachers and student researchers, is still primarily conducted manually.

A corpus is, by definition, computerized, stored electronically, and searchable by computer programs. Corpora and corpus approaches in the study of speech patterns may offer relevant options to search for a wide variety of data on vocabulary use, commonly used markers, and potential errors as they occur in transcripts. The advantage of creating spoken corpora is that they can be designed with a purpose. Researchers compile corpora and search for existing constructs or speech patterns which are identified as relevant and measurable. A corpus provides the opportunity to measure tendencies and distributions across registers and genres of speech. For example, if a lexicographer is interested in the use of oral respect markers (e.g., use of *sir* or *ma'am*, use of titles—*Dr. Williams, Atty. Johnson*) in task-based interaction by a particular group of people, he or she may construct a corpus of naturally occurring speech from speakers of the target group. If the corpus is **representative** of that group, the researcher can find the distributions of these respect markers and describe the tendencies of those patterns (Friginal and Hardy 2014).

An important distinction among corpora is the number of groups (e.g., native vs. non-native speakers, advanced L2 vs. beginning level learners) and types of language production they are designed to represent. Corpora can, therefore, be constructed to reflect the language used by very large groups of people or learners, or researchers may focus on a particular type of language user or classroom situation. Most large-scale corpora (i.e., **general** corpora) such as those representing national varieties of English (e.g., British English from the British National Corpus or BNC) contain millions of words and texts representing a range of spoken and written registers. In the early 1980s, a corpus of 1 million words was considered large (e.g., seminal corpora such as Brown and LOB corpora both had 1 million total words). In comparison, today, there are corpora of hundreds of millions of words. The size of the corpus does not necessarily make it a general (or reference) corpus. It is, instead, the inclusion and distribution of multiple registers and groups of speakers and writers that does. Note that while the Brown and LOB included many registers of English, they crucially lacked spoken language. If the goal of a corpus is to attempt to represent the language as a whole, it must also necessarily include samples of texts transcribed from speech. The BNC's latest edition is made up of nearly 97 million orthographic words, but only about 10 percent of this corpus is from spoken data, primarily because of the enormous time and manpower needed to record and transcribe naturally occurring speech. A variety of forms of written language, such as books, newspapers, and advertisements were included in the BNC to give the sample breadth across genres. The BNC's spoken texts include multiple types of speaking from education, business, public life, and leisure from three geographical regions in Great Britain (2.64% of the spoken texts came from speakers of unknown location) (Friginal and Hardy 2014).

Another popular general corpus is the Corpus of Contemporary American English (COCA). COCA is a database of more than 450 million words and is readily searchable online (http://corpus.byu.edu/coca). Mark Davies of Brigham Young University designed and developed COCA as well as his other collections including COHA (Corpus of Historical American English) and the 1.9-billion-word GloWbE (Corpus of Web-Based Global English). These freely available corpora are great resources for register-based research in contemporary and historical American English, and in the case of GloWbE, varieties of English collected from the global internet. However, spoken registers are also still not well represented in these collections. For example, COCA separates groups of texts "representing" spoken data, but these are limited to television interview transcripts (e.g., interviews from talk shows like the *Oprah Show*) and news reports. Clearly, the pattern here is that recording and transcribing speech samples may not be comprehensively represented, even in large-scale and highly regarded general collections.

For the most part, classroom-based research data may come from a limited number of sources whose context is as important to describe as the larger language domain itself. Data that have been collected in this more focused, individualized setting may allow the researcher to more clearly understand the discourse domain and target group (or groups) of speakers and writers. In corpus linguistics, this dataset is referred to as **specialized corpus**. Specialized spoken corpora like MICASE and T2K-SWAL are large enough to provide opportunities for statistical computations of significance, but are still relatively small in overall size, especially with their total number of words, text files, and registers.

Specialized spoken corpora collected from classrooms provide teachers and researchers the ability to control for many more variables to study and include in the analysis. These are designed to represent a particular register (e.g., lecture vs. small group discussion), domain, or variety of the language. This is useful especially when moving from the analysis of results to the discussion of 'generalizing' towards a bigger population, after further analysis. Overall, this is a question of scope. What is being investigated? What spoken texts are included? What are teacher and learner backgrounds? These are interesting questions, but they may be very difficult to answer as it would be problematic to collect a spoken corpus that includes an equal **representation** of all classroom talk from multiple geographic areas, groups of learners, and classroom tasks. Not only would such a corpus be difficult to collect, but also if all relevant variables are not represented in the corpus, the researcher would be unable to make valid generalizations based on his results to the population as a whole. Instead, a narrowing of scope may be necessary to ask a realistic and specific set of questions (Friginal and Hardy 2014). The classroom-based and learner interview corpora we analyze in this book are very specialized and could still be further redesigned and developed to include other settings and groups of learners and teachers. Interview questions, language activities (in the classroom and peer response activities), and other learner demographics may be added to fully represent classroom talk in US universities.

A Brief Historical Overview of Corpus Linguistics

The following is a brief historical overview of corpus linguistics adapted and synthesized from Friginal and Hardy's *Corpus-Based Sociolinguistics: A Guide for Students* (Routledge, 2014) and Biber, Reppen, and Friginal's 'Research in Corpus Linguistics' from the *Oxford Handbook of Applied Linguistics* (Oxford University Press, 2010):

The focus on collecting naturally occurring texts has been essential in corpus linguistics and recognized as an important methodological approach. Some may think that corpus-based research emerged only in the 1980s and 1990s, along with developments in desktop computing technology (Biber et al. 1998). In fact, the standard practice in language research up until the 1950s was to base language descriptions on analyses of collections of natural texts from those collected by ethnographers and field linguists. Many of these collected text samples have been used to describe the structure of languages and produce dictionaries. Dictionaries have been primarily based on the analysis of word use in natural utterances taken from interviews with speakers representing a particular dialect region. For example, the *Oxford English Dictionary*, which was published in 1928, was based on around 5,000,000 citations from natural texts (totaling approximately 50 million words), compiled by over 2,000 volunteers for more than a 70-year period. Samuel Johnson's *Dictionary of the English Language*, published in 1755, was developed from a collection of 150,000 natural sentences written on slips of papers to illustrate the natural usage of words (Biber et al. 2010).

Pre-electronic corpora of texts such as newspaper writing, short stories, and academic essays were collected to study vocabulary use empirically and also to inform grammar studies and grammar teaching in English. Influential grammar books used actual sentences taken from novels and newspapers to show various structures of formal, grammatically correct sentences and syntactic items such as verb phrases and clauses. In the 1960s and 1970s, most research in linguistics moved to what Biber (1988) referred to as *intuition-based methods* (i.e., intuition vs. empirical analysis in research), which maintained that language was a mental