Research Methods in Second Language Acquisition

A Practical Guide

Edited by Alison Mackey and Susan M. Gass

WILEY-BLACKWELL

Research Methods in Second Language Acquisition

Guides to Research Methods in Language and Linguistics

Series Editor: Li Wei, Birkbeck College, University of London

The science of language encompasses a truly interdisciplinary field of research, with a wide range of focuses, approaches, and objectives. While linguistics has its own traditional approaches, a variety of other intellectual disciplines have contributed methodological perspectives that enrich the field as a whole. As a result, linguistics now draws on state-of-the-art work from such fields as psychology, computer science, biology, neuroscience and cognitive science, sociology, music, philosophy, and anthropology.

The interdisciplinary nature of the field presents both challenges and opportunities to students who must understand a variety of evolving research skills and methods. The *Guides to Research Methods in Language and Linguistics* address these skills in a systematic way for advanced students and beginning researchers in language science. The books in this series focus especially on the relationships between theory, methods, and data – the understanding of which is fundamental to the successful completion of research projects and the advancement of knowledge.

Published

1. *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism*
   Edited by Li Wei and Melissa G. Moyer
2. *Research Methods in Child Language: A Practical Guide*
   Edited by Erika Hoff
3. *Research Methods in Second Language Acquisition: A Practical Guide*
   Edited by Alison Mackey and Susan M. Gass

Forthcoming

Research Methods in Clinical Linguistics and Phonetics: A Practical Guide
Edited by Nicole Müller and Martin J. Ball

Research Methods in Second Language Acquisition

A Practical Guide

Edited by Alison Mackey and Susan M. Gass

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Blackwell Publishing was acquired by John Wiley & Sons in February 2007. Blackwell's publishing program has been merged with Wiley's global Scientific, Technical, and Medical business to form Wiley-Blackwell.

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Offices
350 Main Street, Malden, MA 02148-5020, USA
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

For details of our global editorial offices, for customer services, and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data

Research methods in second language acquisition : a practical guide / edited by Alison Mackey and Susan M. Gass. – 1st ed.
p. cm. – (Guides to research methods in language and linguistics)
Includes bibliographical references and index.
ISBN 978-1-4443-3426-5 (hardcover : alk. paper)
ISBN 978-1-4443-3427-2 (pbk. : alk. paper)
1. Second language acquisition–Methodology. 2. Language and languages–Study and teaching–Methodology. I. Mackey, Alison. II. Gass, Susan M.
P118.2.R473 2012
401′.93–dc23
2011021094

A catalogue record for this book is available from the British Library.

This book is published in the following electronic formats: ePDFs 9781444347173; Wiley Online Library 9781444347340; ePub 9781444347326; Mobi 9781444347333

Set in 10/13pt Sabon by SPi Publisher Services, Pondicherry, India

1 2012

Contents

List of Contributors

Rebekha Abbuhl, California State University, Long Beach (rabbuhl@csulb.edu)
Melissa Baralt, Florida International University (missybaralt@gmail.com)
Kata Csizér, Eötvös University (weinkata@yahoo.com)
Zoltán Dörnyei, University of Nottingham (zoltan.dornyei@nottingham.ac.uk)
Patricia A. Duff, University of British Columbia (patricia.duff@ubc.ca)
Debra A. Friedman, Michigan State University (fried106@msu.edu)
Susan M. Gass, Michigan State University (gass@msu.edu)
Sylviane Granger, University of Louvain (sylviane.granger@uclouvain.be)
Tania Ionin, University of Illinois (tionin@illinois.edu)
Keiko Koda, Carnegie Mellon University (kkoda@andrew.cmu.edu)
Jenifer Larson-Hall, University of North Texas (jenifer@unt.edu)
Shawn Loewen, Michigan State University (loewens@msu.edu)
Alison Mackey, Georgetown University (mackeya@mac.com)
Kim McDonough, Northern Arizona University (kim.mcdonough@nau.edu)
Frederick L. Oswald, Rice University
Jenefer Philp, University of Auckland (philp@auckland.ac.nz)
Luke Plonsky, Michigan State University (plonskyl@msu.edu)
Charlene Polio, Michigan State University (polio@msu.edu)
Andrea Révész, Lancaster University (a.revesz@lancaster.ac.uk)
Pavel Trofimovich, Concordia University (pavel.trofimovich@concordia.ca)

1 Introduction

Alison Mackey and Susan M. Gass

Second language acquisition (SLA) research draws its research methodology and tools from a number of other fields including education, linguistics, psychology, sociology, and more. Partly for this reason, research methodology in second language studies is frequently evolving in response to developments in other fields as well as to developments in our own field. There is such a diversity of approaches to second language research methodology that a book like this one, where each chapter is authored by a person who is an experienced expert in that particular subarea, is one of the most efficient ways for research to describe and disseminate information about the method in which they have particular expertise.

Designing a research study and determining an appropriate method of investigation is a difficult task. But the task is made easier if one understands that research methods are not determined or decided upon devoid of context; research methods are dependent on the theories that they are designed to investigate. Thus, research questions are intimately tied to the methods used for determining an appropriate dataset.

This volume is intended as a guide for students as they design research projects. Each chapter presents some basic background to the area of research. This is a necessary feature since methodologies, as we noted above, cannot be understood in a vacuum. The book also has a pedagogical focus, with each chapter providing a practical, step-by-step guide to the method it covers, often informed by reference to studies using the method, carried out by the chapter's author. The method is discussed together with the theoretical frameworks within which it is commonly used. This how-to section takes students from beginning to end of a particular area. Finally, project ideas and resources (e.g., analytical tools when appropriate, references to more detailed discussions of a particular area), are also included, together with additional readings, and brief summaries of studies that have used the particular methodology, together with study questions that can be used as a basis for class

Research Methods in Second Language Acquisition: A Practical Guide, First Edition.
Edited by Alison Mackey and Susan M. Gass.

discussions. Summary study boxes are given to help readers grasp the main ideas of studies that have used the method in question.

The book is divided into two parts. The first is on data types, which includes representative types of the wide range of data that is commonly studied in SLA, including both newer data types, such as learner corpora, along with more traditionally studied data, such as case studies. The second part is on data coding, analysis, and replication, where we present chapters on topics like meta-analyses. We will briefly summarize the contributions and explain how they fit together. We must also remember, however, that no elicitation instrument or methodology is foolproof; all have their advantages and limitations. And, as we have stressed in our other books dealing with research methods (Mackey & Gass, 2005; Gass & Mackey, 2007), no research project should be undertaken without extensive pilot testing.

In chapter 2, "How to Use Foreign and Second Language Corpora," Sylviane Granger covers learner corpus research, which she describes as originating in the late 1980s and involving the study of computerized databases of written or spoken texts. She focuses on frequency, variation, and co-text, and describes the powerful automatic analysis that can reveal quantitative information on a wide range of language from morphemes to lexical phrases. Tania Ionin's chapter 3, "Formal Theory-Based Methodologies," focuses on methods used in formal, generative SLA research. She describes the collection of empirical data on learners' production and comprehension of the target language, which are used to draw conclusions about the underlying grammar. Methodologies she focuses on include grammaticality judgment tasks and interpretation tasks. In chapter 4, "Instructed Second Language Acquisition," Shawn Loewen and Jenefer Philp focus on an often-studied context in SLA, providing a short review of the ways in which research on instructed SLA has been done, focusing on the practicalities of carrying out each one. Their chapter focuses in general on second language (L2) classroom instruction, and does not specifically address reading and writing research or investigations of individual differences, since those topics are covered in chapters 5, 8, and 9 in this volume.

In chapter 5, "How to Design and Analyze Surveys in Second Language Acquisition Research," Zoltán Dörnyei & Kata Csizér explain how survey studies are carried out in the context of SLA research, including the required steps for designing a survey that can provide valid and reliable data. They also discuss quantitative data analysis in relation to questionnaire data, as well as how to report survey results. In chapter 6, "How to Carry Out Case Study Research," Patricia A. Duff explains the background of one of the earliest methods used to underpin the field, characterizing its focus on a small number of research participants and occasionally just one individual (a focal participant or case) and explaining how behaviors, performance, knowledge, and perspectives are examined closely and intensively, often over an extended period of time. In chapter 7 "How to Use Psycholinguistic Methodologies for Comprehension and Production," Kim McDonough and Pavel Trofimovich explain psycholinguistics as having the twin goals of understanding how people comprehend and produce language. In other words, these authors describe the methodologies used in the attempts to figure out what processes, mechanisms, or procedures underlie language use and learning. In chapter 8, "How to Research Second Language Writing," Charlene Polio classifies empirical studies of L2 writing on the basis of the ways data

are collected, coded, analyzed, and interpreted with the goal of understanding L2 learning processes. This chapter on writing is complemented by Keiko Koda's chapter 9, "How to Do Research on Second Language Reading," in which she explains that reading is a multidimensional construct involving a wide range of subskills whose acquisition depends on various learner-internal and learner-external factors. Different approaches to SLA see reading as cognitive or sociocultural and, as she argues, it is important to clarify the theoretical and methodological orientations in relation to the problem motivating the research.

The final chapter in part I, by Debra A. Friedman, focuses on "How to Collect and Analyze Qualitative Data." As she explains, the rise of theoretical and analytical frameworks such as sociocultural theory, L2 socialization, and learner identity has brought important insights to the field. She first provides her perspective on what qualitative research is and what it can contribute to the field, and then takes the reader through the process of designing and conducting a qualitative research project, including theoretical and practical aspects of qualitative methods for data collection and analysis.

In part II, we move away from a focus on data types, and instead the chapters provide input on how to analyze and code data. Complementary to Friedman's chapter is chapter 11 by Andrea Révész, "Coding Second Language Data Validly and Reliably," which brings a welcome perspective on a topic which is critical to all areas of SLA research. Coding, as Révész explains, involves organizing and classifying raw data into categories for the purpose of further analysis and interpretation. She explains the concepts of validity and reliability in relation to coding with a focus on relatively top-down, theory- and instrument-driven coding methods. Qualitative coding which emerges bottom-up from the data is the topic of the preceding chapter, by Friedman, as well as the next chapter, by Baralt. In chapter 12, "Coding Qualitative Data," Melissa Baralt focuses on how to code data using NVivo in qualitative research. NVivo is a type of software that assists researchers in managing data and in carrying out qualitative analysis. As Baralt explains, qualitative data often include text, notes, video files, audio files, photos, and/or other forms of media, and SLA researchers are increasingly using computer-assisted qualitative data analysis software to manage of all these data types, even if the researchers are not doing the kind of corpus work described earlier by Granger. Baralt provides coding examples based on NVivo software, but as she explains, the basic procedures presented in her chapter are also applicable to traditional pen-and-paper methods and other software programs.

Coding in both quantitative and qualitative paradigms having been considered, chapter 13, by Jenifer Larson-Hall, focuses on "How to Run Statistical Analyses." As Larson-Hall explains, inferential statistics let the reader know whether the results that have been found can be generalized to a wider population. She provides a brief survey of how to understand and perform the most basic and frequently used inferential statistical tests in the field of SLA. Chapter 14, by Luke Plonsky and Frederick L. Oswald, "How to Do a Meta-Analysis," defines meta-analyses in both their narrow and broader senses, and focuses primarily on the practical aspects of meta-analysis more broadly conceived. Meta-analyses and research syntheses are becoming more common in the field, representing a coming of age of the field, and

also the ability to draw more general conclusions from our increasingly wide body of knowledge. In the final chapter, by Rebekha Abbuhl, "Why, When, and How to Replicate Research," we cover another crucial topic in the field and one which we believe is critically important for the future. If SLA is to continue to go from strength to strength, we need to proceed from a position of confidence in our findings. Replication will be a key part of that. This chapter by Abbuhl (and Porte [in press]) both suggest that replications, when carefully done, represent a cornerstone of our field. A recent UK grant by the Economic and Social Research Council (ESRC) to Emma Marsden (University of York, UK) and Alison Mackey (Georgetown University, US) for the project 'Instruments for Research into Second Languages' (IRIS) will support a database where research instruments can be uploaded and downloaded. This database will be fully searchable by a wide range of parameters including the first and second languages under investigation, the type of instrument, the age of the learner, and so on. The IRIS project aims to make the process of selecting and locating data collection instruments much more streamlined and efficient, which in turn will assist the process of replication in SLA research and, in the longer term, the scope and quality of meta-analyses. IRIS will also facilitate the scrutiny of instruments, so that researchers can more easily evaluate the validity, reliability, and generalizability of tools used for data collection. Replication, along with careful methodological approaches ranging from case studies to surveys to corpus-based studies, represent the past and future of SLA research. An understanding of the topics addressed in this volume is essential for the formation of a solid foundation for doing SLA research.

References

Gass, S. M., & Mackey, A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum.

Porte, G. (Ed.). (in press). *Replication studies in applied linguistics and second language acquisition*. Cambridge, England: Cambridge University Press.

Part I Data Types

2 How to Use Foreign and Second Language Learner Corpora

Sylviane Granger

A New Resource for Second Language Acquisition

Background

Learner corpus research (LCR) originated in the late 1980s within the theoretical and methodological paradigm of corpus linguistics, which studies language use on the basis of corpora, that is, computerized databases of written or spoken texts. Although still relatively young, corpus linguistics has already had a big impact on language theory and description. One of its major contributions is the light it throws on three major facets of language: frequency, variation, and co-text. First, the combined use of large amounts of natural language data and powerful automatic analysis provides unparalleled quantitative information on all types of linguistic units, from morphemes to syntactic structures through single words and lexical phrases. Second, the comparison of corpora representing different varieties of language – geographical (e.g., British English vs. South African English), temporal (nineteenth-century vs. twentieth-century), or stylistic (informal conversation vs. academic writing) – helps uncover the distinguishing features of each variety and generally enhances our appreciation of the multifaceted variation inherent in language. Third, the remarkable ease with which computers identify the immediate context of words, that is, their co-text, has demonstrated the interrelation between lexis and grammar and generally led to a better understanding of the syntagmatic aspects of language.

The idea of compiling learner corpora – computerized databases of foreign or second learner language – and applying corpus linguistic tools and methods to analyze them arose from the wish to bring to the field of second language acquisition (SLA) the same kinds of benefits that corpora were providing to the linguistic field. Several linguists with a keen interest in SLA, often because they were also language teachers, concurrently but independently started to compile and analyze large electronic

Research Methods in Second Language Acquisition: A Practical Guide, First Edition.
Edited by Alison Mackey and Susan M. Gass.

collections of second language (L2) data. Their objectives in embarking on this new type of research were theoretical, in that they wanted to gain a better understanding of the process of learning a foreign language or L2, and/or practical, in that they had a view to designing more efficient language teaching tools and methods.

LCR is at the crossroads between corpus linguistics and SLA. So far, it is mainly corpus linguists that have been active in the field. This can be seen as positive, as the first task that needed to be done was to adapt corpus linguistic techniques for learner corpus data and/or design new ones, and this required extended corpus expertise. The downside is that the grounding in SLA theory has been relatively limited to date. However, recent research shows that the LCR community wishes to situate itself firmly within the current SLA debate and, simultaneously, there is a growing - though admittedly still limited - awareness among SLA specialists of the tremendous potential of learner corpora.

The Specificity of Learner Corpus Data

Learner corpus data fall within the more open-ended types of SLA data distinguished by Ellis (1994, pp. 670–672), namely natural language use data and clinical data. Natural language use data is produced by learners who use the L2 for authentic communication purposes. In principle, only this type of data should qualify as bona fide learner corpus data, since corpora are supposed to be "authentic," containing data "gathered from the genuine communications of people going about their normal business" (Sinclair, 1996). However, fully natural learner data is difficult to collect, especially in foreign language settings which give learners few opportunities to use the L2 in authentic everyday situations. Therefore, learner corpus researchers often resort to clinical data, that is, open-ended elicited data such as written compositions or oral interviews. Experimental data, such as fill-in-the-blanks exercises, which force learners to choose between a limited number of options rather than allowing them to select their own wording, clearly falls outside the learner corpus range. Admittedly, in between fully natural data and fully experimental data, there is a wide range of data types which are situated at various points on the scale of naturalness. To reflect this continuum, Nesselhauf (2004, p. 128) suggests distinguishing a category of "peripheral learner corpora," which contain more constrained data such as picture description or translation.

Uncontrolled production data has been relatively neglected in SLA studies in favor of introspection data (especially grammaticality judgment tests) and the more controlled types of production data. The reason is that naturalistic data has been found to suffer from a number of drawbacks, among them (a) the impossibility of studying some language features because of insufficient data, (b) lack of control exerted over the main variables that can influence production, and (c) difficulty in interpreting the data. While fully valid with reference to previous data collections, these three arguments lose some of their validity when applied to learner corpus data, for reasons explained below.

The first argument suggests that unconstrained data collection fails to provide enough occurrences of relatively infrequent linguistic items, therefore making it

impossible for researchers to investigate them. Larsen-Freeman and Long (1991, p. 26) provide the following example: "A researcher would have to wait a long time, for example, for subjects to produce enough gerundive complements for the researcher to be able to say anything meaningful about their acquisition." As learner corpora tend to be quite big, often over 1 million words, this oft-cited criticism loses much of its relevance. For a large number of linguistic phenomena, learner corpora provide a wealth of occurrences, to the point that researchers often cannot study the whole set and have to select a representative sample. It remains true, however, that the optimal size of a learner corpus depends on the targeted linguistic phenomenon. Articles, which are very frequent, can be investigated on the basis of a small corpus, while for lexical words – except the high-frequency ones – much larger collections are required. In addition, when assessing the size of an L2 corpus, one should consider not only the total number of words, as is customary in corpus research, but also the number of learners that produced the data. There is no direct relation between the size counted in number of words and representativeness measured in number of learners. For example, while the 80,000-word corpus used by Chen (2006) contains data produced by 10 students, a similar-sized sample from the International Corpus of Learner English (ICLE; see 'Project Ideas and Resources' below) would contain data from c. 130 different learners. The size and representativeness of learner corpora are a major asset of this new resource, which goes some way to meeting a frequent weakness of SLA studies, namely that it is often "difficult to know with any degree of certainty whether the results obtained are applicable only to the one or two learners studied, or whether they are indeed characteristic of a wide range of subjects" (Gass & Selinker, 2001, p. 31).

While previous collections of learner production data have often been criticized for their lack of rigor (e.g., Odlin, 1989, p. 151, for transfer studies, or Ellis, 1994, p. 49, for error analysis studies), this criticism cannot be leveled at learner corpus data, which, like all corpus data, is accompanied by rich ethnographic data. As rightly pointed out by Cobb (2003, p. 396), "It is a common misconception that corpus building means collecting lots of texts from the Internet and pasting them all together." However large it may be, a learner corpus will only be useful if it has been compiled on the basis of strict design criteria. Among the variables that are regularly recorded are learner variables, such as age, gender, mother-tongue background, or knowledge of other foreign languages, and task variables, such as medium, genre, topic, length, or task conditions (timing, use of reference tools, etc.). These variables are used as search criteria by researchers to compile their own tailor-made subcorpora. Admittedly, even in learner corpora that have been very carefully designed, not all variables are recorded. There is rarely any information on the teaching methods, the course material or the first language (L1) or L2 status of the teachers, all crucial factors in foreign language settings. In addition, proficiency level is often assigned on the basis of external criteria (number of years of study), an imperfect measure that has been denounced by a number of researchers (e.g., Pendar & Chapelle, 2008; Wulff & Römer, 2009). One way of overcoming this difficulty is to complement the ethnographic data with additional data obtained, for example, by submitting students to standardized questionnaires (motivation test, aptitude test, general proficiency test, vocabulary test).

The third drawback that has been pointed out in relation to unconstrained production data is that it is difficult to interpret, partly because learners may "avoid the troublesome aspects through circumlocution or some other device (Larsen-Freeman & Long, 1991, p. 26). This criticism also applies to learner corpus data. The absence of a feature in a learner corpus may not be evidence of a lack of knowledge but may result from an avoidance strategy. However, learner corpora provide a much more efficient and reliable basis for investigating avoidance, as they can be analyzed with software tools (see 'Project Ideas and Resources' below) that automatically extract the words or structures that are significantly underused by learners. Reliable measurements of L1 transfer effects are also greatly facilitated by the amount and variety of corpus data available to the researcher (see 'Data analysis' below). However, these advances do not solve all interpretation problems, and learner corpus researchers have recently started to complement learner corpus data with other data types, in particular experimental data (Gilquin, 2007). While in early LCR, learner corpus data and experimental data were seen as incompatible, researchers are now beginning to see the benefit of combining the two. See study box 2.1.

Study Box 2.1

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik, 55*(3), 273–91.

Background

Collocations combining a high-frequency verb and a noun phrase are notoriously difficult for learners of English. Previous studies have approached this problem from the perspective of competence (through elicitation tasks) or performance (by using learner corpora), but rarely have the two perspectives been combined.

Research questions

- How well do advanced French-speaking learners of English use *make*-collocations (quantitatively and qualitatively)?
- Do the learners' performance and competence differ in this respect?

Method

- Combination of error analysis, CIA (comparison of learner and native corpus data), and elicitation (fill-in exercise and acceptability judgment test).
- Corpora used: French subcorpus of ICLE and LOCNESS.
- Software tool: WordSmith Tools.

Statistical tools

- Chi-square test and distinctive collexeme analysis.

Results

The corpus study shows that French-speaking learners do not make many errors when using *make*-collocations, but they tend to underuse them and prefer those collocations that have a direct equivalent in French. In the elicited data, on the other hand, the error rate is much higher and learners' judgments are often unreliable. Both performance and competence are characterized by a high degree of L1 influence.

Learner Corpus Typology

Learner corpora have mushroomed in recent years.[1] An exhaustive description is therefore clearly beyond the scope of this chapter (for more details, see Granger, 2008). It is possible, however, to identify a number of dimensions along which they vary, such as time of collection, scope of collection, targeted language (L2), learner's mother tongue (L1), medium, and text type.

Time of collection

Cross-sectional learner corpora contain samples of learner writing or speech gathered from different categories of learners at a single point in time, while *longitudinal* learner corpora track the same learners over a particular time period. The overwhelming majority of learner corpora are cross-sectional. A few are *quasi-longitudinal*, that is, they contain data gathered at a single point in time but from learners of different proficiency levels. Very few are genuinely longitudinal, mainly because of the difficulty of collecting that kind of data in large quantities.

Scope of collection

Global learner corpora are collected on a large scale from a range of learners and used to inform SLA theory and/or generic reference and teaching tools. *Local* learner corpora are much smaller. They are collected by teachers as part of their normal teaching activities and directly used as a basis for classroom materials. Global learner corpora indirectly benefit learners who have the same profile as the students who produced the data (same mother-tongue background, same level of proficiency, etc.), where local learner corpora learners are both producers and users of the data,

which can be expected to enhance its relevance and boost learners' motivation. Local learner corpus compilation is still the exception rather than the rule but it is also one of the most promising avenues in LCR.

Targeted language (L2)

Learner corpora can be classified according to the target language they sample. At first, corpora of *L2 English* reigned supreme, but *other L2s* (Dutch, Finnish, French, German, Italian, Korean, Norwegian, Slovene, Spanish, and Swedish, to cite just a few) have progressively joined the learner corpus bandwagon.

Learner's mother tongue (L1)

Mono-L1 learner corpora contain data from learners of one and the same mother-tongue background, while *multi-L1* learner corpora cover learners from several mother-tongue backgrounds. Commercial corpora, such as the Longman Learner Corpus, which are compiled by publishing houses, tend to have a multi-L1 coverage, while academic corpora, collected by researchers in SLA and/or foreign language teaching, tend to be restricted to one mother tongue, although there are some exceptions (see 'Selection and/or Compilation of Learner Corpus' below).

Medium

While the term *written learner corpus* unambiguously refers to corpora of learner writing, the term *spoken learner corpus* may refer to lexical or (much less frequently) phonetic/prosodic transcriptions of oral production data, and may or may not have associated audio files and more recently, with the advent of *multi-media learner corpora*, video recordings. Unsurprisingly, in view of the difficulty of collecting and transcribing spoken data, written corpora dominate the learner corpus scene.

Text type

The two favorite text types represented in LCR to date are *argumentative essays* for writing and *informal interviews* for speech. This preference reflects the wish to sample the least constrained types of production data (see 'The Specificity of Learner Corpus Data' above). It also ensues from the necessity of comparing like with like. Some diversification in terms of textual genres is desirable and indeed has begun to materialize. A good example is the Indiana Business Learner Corpus, which is made up of application letters from native and non-native speakers of English studying in three different undergraduate business classes in Belgium, Finland, and the United States (Connor, Pretch, & Upton, 2002).

Main Stages in Learner Corpus Research

Table 2.1 sums up the seven main stages in LCR. Five of these stages are mandatory, whatever the focus and ultimate objective of the study, while two – data annotation and pedagogical implementation (in italics in the table) – are regular but not required features of LCR.

Choice of Methodological Approach

Any researcher embarking on a corpus project chooses one of two main methodological approaches – corpus-based or corpus-driven – according to his or her research question. The corpus-based approach consists in testing a hypothesis or rule against corpus data. It is therefore essentially a deductive approach, where the corpus does not act as the master but rather as the servant to confirm or refute a pre-existing theoretical construct. The corpus-driven approach exploits the full force of the corpus. It is an inductive approach, which progressively generalizes from the observation of data to build up the theory or rule.

Most studies so far have been of the exploratory type, that is, corpus-driven. This powerful heuristic approach, which is exclusive to learner corpus data, has the advantage of not being limited by the initial hypothesis and is therefore capable of uncovering new features of interlanguage.[2] For example, automatic extraction of recurrent sequences of a particular length (two, three, or more words) has thrown invaluable light on the nature of learners' prefabricated language. The time has come, however, to exploit to the full the other approach, that is, to test SLA theoretical constructs on the basis of learner corpus data. As suggested by Myles (2005, p. 381), "it is now time that corpus linguists and SLA specialists work more closely together in order to advance both their agendas." The few studies that have used learner corpora to test an SLA hypothesis demonstrate the potential of a more SLA-informed approach. For example, Housen (2002) revisits previous morpheme studies on the basis of a longitudinal corpus of annotated oral learner data and native speaker baseline data. While the study generally confirms the general order of emergence of morphemes, it also reveals significant variation at the level of individual

Table 2.1 Main stages in learner corpus research.

1.	Choice of methodological approach
2.	Selection and/or compilation of learner corpus
3.	*Data annotation*
4.	Data extraction
5.	Data analysis
6.	Data interpretation
7.	*Pedagogical implementation*

learners and generally highlights a number of benefits that can be gained from investigations of large learner corpora (see also Rankin, 2009, on verb-second structures in advanced L2 English).

Selection and/or Compilation of Learner Corpus

As learner corpus collection is time-consuming and often difficult to undertake, it is advisable to survey the field to find out whether there might be a learner corpus that is available and suitable for the investigation. Unfortunately, although there is a wide range of learner corpora, many are not available outside the team that has compiled them. Some, however, are fully available for research purposes, among them the ICLE (Granger, Dagneaux, Meunier, & Paquot, 2009) and the French Learner Language Oral Corpora (Myles, 2005; see 'Project Ideas and Resources' below).

If no existing learner corpus fits the bill and/or a local corpus is more relevant for the planned study, there remains the possibility of compiling one's own corpus, which in today's electronic world is much less difficult than in the past. A bespoke local corpus also has the invaluable advantage of being fully controllable (Millar & Lehtinen, 2008).

If the corpus analysis stage is planned to include comparison between learner data and a control corpus of native or expert language, it is essential to identify the corpora that best suit the analysis, taking into account important variables such as geographical variety (American English, Indian English, etc.), age, and text type. Failure to ensure full comparability of the data may lead to erroneous results, as demonstrated by De Cock (2002) for prefabricated sequences in speech and by Granger and Tyson (1996) for connectors in writing.

Data Annotation

Whether the learner corpus can be used as-is in raw format or needs to be enriched with linguistic annotations very much depends on the object of study. To analyze the word *clever*, which is a single invariable lexical word which only functions as an adjective, a raw corpus is sufficient and the annotation stage can be skipped. However, this is rarely the case, and as a result, there is often much to be gained from annotating the data. To take a simple example, many words in English can belong to different word categories. This is the case with the word *order*, which can be a verb or a noun and also features in the multiword unit *in order to*. If the research project focuses only on verbs, a raw learner corpus will entail a long process of manual disambiguation to extract all and only the verbal occurrences. Fortunately, some of the most reliable and widely available corpus tools are part-of-speech (POS) taggers, which automatically assign a word category to every word in the corpus with an accuracy rate that can reach 98% (see 'Project Ideas and Resources' below). It is important to bear in mind, however, that corpus linguistic tools were developed on the basis of native corpus data, and the errors present in learner corpora may induce mistaggings (for example, the verb *lose* written *loose* may cause the tagger to assign

an adjective tag rather than the correct verb tag). It is therefore essential to start the analysis with a pilot study to check the accuracy of the tagging (see Granger, 1997, for an example of how this can be done, and Van Rooy & Schäfer, 2003, for a description of the impact of spelling errors on the accuracy rate).

There is a danger that analysts may limit themselves to the types of automatic analysis that the computer can provide. It is important to bear in mind, however, that corpus annotation software allows analysts to insert a rich variety of annotations into the text files. Although this work is largely manual and hence time-consuming, the return on investment is high as the annotations can subsequently be used as search criteria to retrieve all the occurrences in the corpus that match a particular query (see 'Data Extraction' and 'Project Ideas and Resources' below). This is the case, for example, with the CHILDES system, which was initially developed for the storage and analysis of L1 data and is particularly well suited for the annotation of spoken data (Myles & Mitchell, 2005). One type of annotation that is particularly relevant for learner corpora is error tagging. In most systems errors are coded for error type (number, gender, tense, etc.), word category (noun, verb, etc.), and/or error domain (spelling, grammar, lexis, etc.). What makes error tagging particularly useful is that the error tags are inserted into the text files and are hence presented in the full context of the text, alongside non-erroneous forms. In some studies all errors are coded (e.g., Chuang & Nesi, 2006), in others the tagging is limited to some specific categories, such as spelling errors in L2 English (Botley & Dillah, 2007) or particle errors in L2 Korean (Lee, Jang, & Seo, 2009; see Díaz-Negrillo & Fernández-Domínguez, 2006, for a survey of error-tagging systems).

Data Extraction

Corpus analysis tools, commonly referred to as concordancers, enable researchers to automatically extract a wealth of information from learner corpora. There are a number of such programs, which differ in their degree of sophistication, user friendliness, and availability (see 'Project Ideas and Resources' below). Most of them include the following functionalities.

Word list

The word list function creates lists of all the word forms in the corpus and displays these alphabetically and by frequency, together with a range of statistics (number of types, number of tokens, type/token ratio, mean sentence length, etc.).

Keyword list

The keyword function compares two previously created word lists and outputs the word forms that are statistically more frequent in one corpus than in the other. This is an extremely useful tool for LCR. Using this function, researchers can

Figure 2.1 Unsorted concordance of the verb *discuss* in the ICLE.

identify the words and word sequences (or word/error categories) that distinguish one learner sample from another, or a learner sample from a comparable native/expert corpus.

Concordancing

The concordancing option presents all the instances of a linguistic item in their immediate linguistic context. Figure 2.1 shows the concordance of the verb *discuss* in the ICLE drawn up with WordSmith Tools (see 'Project Ideas and Resources' below). Sorting the context to the left and/or to the right of the search item allows regular patterns to emerge. As shown in figure 2.2, the sorting brings out occurrences of correct uses of *discuss* (*discuss a topic, a question; discussed above*), but also several occurrences of the erroneous pattern *discuss about*. Clicking on a concordance line shows the item in its wider context. Not only words but word parts and phrases can be searched in this way. In addition, if the corpus has been annotated, it is possible to use the tags or combinations of words and tags as search strings. Any form of the verb *be* followed by a past participle or a present participle will respectively extract passive and progressive verbal forms from the corpus, while a search for forms of the verb *be* followed by the base forms of verbs will extract all morphological passive errors of the type *they were arrest* or *it must be see*. The possibilities are endless.

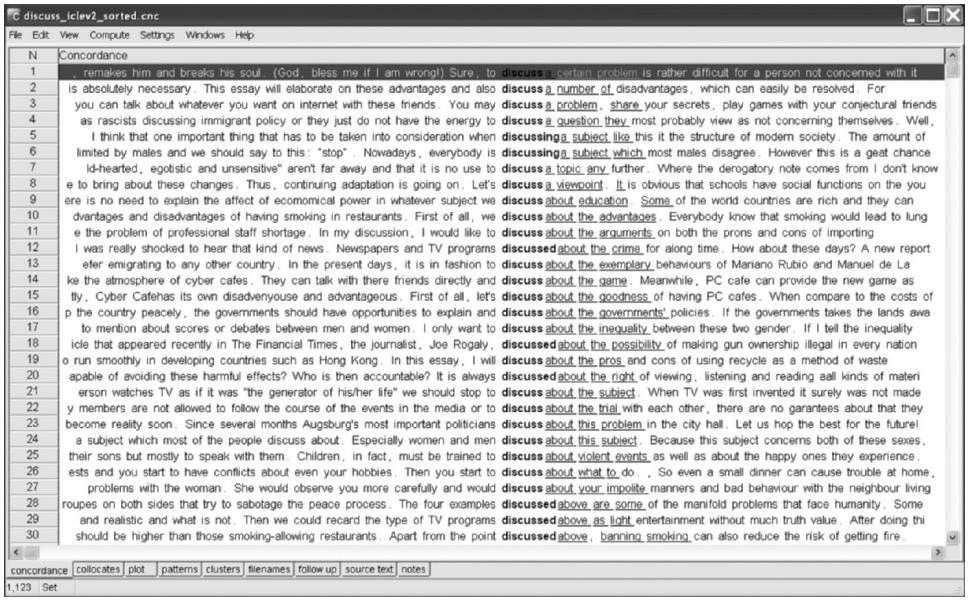

Figure 2.2 Right-sorted concordance of the verb *discuss* in the ICLE.

Distribution/range

This function allows researchers to visualize where the occurrences of the search item are situated in the corpus and hence find out whether the phenomenon under investigation is widespread in the corpus or is found in only a limited number of learner texts.

Collocates

This function retrieves the most common words to the right and/or left of the search items, that is, its collocates. Although useful in that it gives an immediate overview of the preferred company of any given word in a learner text, it cannot compete with the manual scanning of concordance lines, as it will also retrieve non-relevant items which happen to be in the near vicinity of the search item (for illustrations, see Altenberg & Granger, 2001).

Clusters

The clusters option retrieves all repeated sequences of words of a given length (two-word sequences, three-word sequences, etc.) from the concordance lines (e.g., *I will discuss, discuss about, discussed above*, from the concordance of *discuss*).

It is a powerful tool for assessing the rate and quality of prefabricated language in learner texts (De Cock, 2004).

Data Analysis

One of the hallmarks of learner-corpus-based research is the wide range of linguistic phenomena investigated. In addition to those phenomena which have featured prominently in SLA research, such as inflectional morphemes or word order, LCR has scrutinized many phenomena that have been under-researched, if not totally neglected, such as derivational morphemes, collocations, recurrent phrases, lexical richness, hedges, register, spelling, or punctuation, to name but a few.

Some of the analyses have focused exclusively on misuse and have led to the revival of error analysis in the form of *computer-aided error analysis* (CEA; Dagneaux, Denness, & Granger, 1998). The majority of studies, however, have tended to focus on other linguistic features that distinguish learner language from native language, many of which manifest themselves in over- and underuse rather than misuse. This approach, referred to as *contrastive interlanguage analysis* (CIA; see Granger, 1996, and Gilquin, 2000/1), consists in comparing not two different languages, as was the case with contrastive analysis, but two varieties of one and the same language: either two learner varieties (L2 vs. L2) or one learner variety and one native (or expert) variety (L2 vs. L1). Comparing two or more learner varieties is a good method for assessing the influence of the many variables that play a part in SLA, such as task effects, the learner's mother tongue, or level of proficiency. Comparing learner and native varieties makes it possible to uncover typical features of interlanguage, not only errors, but also instances of under- and over-representation of words, phrases, and structures. Two recent studies that illustrate the L2 vs. L2 approach are those by Ädel (2008), who compares the use of involvement markers in timed vs. untimed learner essays, and Groom (2009), who investigates the use of collocations and recurrent sequences by learners who spent less than one month in an English L1 environment compared to those who spent at least one calendar year. Although the number of exclusively L2-focused studies is growing, studies which involve a native or expert control corpus in addition to the L2 data tend to be more popular. Two recent studies representing this approach are those by Luzón (2009), who investigates the use of a rhetorical strategy, namely use of the first person pronoun *we*, by Spanish English as a foreign language (EFL) students and expert writers, and Callies (2008), who compares the use of raising constructions in a native English corpus and two EFL corpora (Polish vs. German learners). See study boxes 2.2 and 2.3.

Both CEA and CIA have been criticized for being guilty of the "comparative fallacy" (Bley-Vroman, 1983), that is, for comparing learner language to a native or expert norm and thus failing to analyze interlanguage in its own right. Several arguments can be offered in defense of the learner corpus position (Granger, 2009). First and foremost, it is important not to confuse theory and method. As rightly pointed out by Tenfjord, Hagen, & Johansen (2006, pp. 93, 102) in relation to CEA, comparisons of interlanguage and native language are methodological aids, which

Study Box 2.2

Cobb, T. (2003) Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *Canadian Modern Language Review/ Revue canadienne des langues vivantes, 59*(3), 393–423.

Background

While the beginning stages of acquisition are well covered in SLA studies, intermediate-advanced interlanguage remains relatively uncharted. One of the main reasons for this neglect is lack of data.

Research question

- Is there a common pattern of interlanguage development across relatively distinct populations of advanced learners?

Method

- Replication of three European studies using comparable learner corpus data collected in Quebec: vocabulary frequency, use of prefabricated sequences, and features of reader/writer visibility in learner argumentative essay writing.
- Corpora used: Quebec teaching English as a second language (TESL) corpus (advanced) and English as a second language (ESL) corpus (intermediate).
- Software tools: WordSmith Tools and VocabProfile.

Statistical tools

Chi-square and t-test.

Results

The study suggests that advanced learners work through identifiable acquisition sequences that are systematic and more or less universal. It also provides avenues for pedagogical implementation of the results.

"can, in principle, service any theory." Errors in CEA are not to be confused with properties of the interlanguage. Rather they are "analytical concepts imposed upon the texts … in order to procure systematic data that any valid theory of SLA should be able to account for." Second, as mentioned above, CIA does not need to include an L1 norm. It is perfectly possible to focus exclusively on L2 data and analyze it in its own right either cross-sectionally or longitudinally. Third, it would seem reasonable to suggest that the comparative fallacy is in fact also present in many non-corpus-based SLA studies but in a hidden, undercover way. For example, all the

Study Box 2.3

Díez-Bedmar, M. B., & Papp, S. (2008). The use of the English article system by Chinese and Spanish learners. In G. Gilquin, S. Papp, & B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 147–175). Amsterdam and New York: Rodopi.

Background

Studies of article use in L1 and L2 acquisition have brought out a number of difficulties of a grammatical and/or pragmatic nature. The acquisition of articles has proved especially difficult for L2 learners who have no article system in their L1.

Research questions

- Will Chinese learners exhibit more non-native features in their use of articles than Spanish learners?
- Will difficulties be both grammatical and pragmatic for Chinese learners and exclusively pragmatic for Spanish learners?

Method

- Integrated contrastive model: comparison of article use in three L1 corpora (English, Chinese, and Spanish) and two L2 corpora (Chinese and Spanish learners). Obligatory context analysis based on annotated data.
- Corpora: compiled by the authors (the L2 Chinese corpus is part of ICLE).
- Software tool: WordSmith Tools.

Statistical tools

Chi-square and z-test.

Results

The results generally bear out the initial hypothesis that Chinese learners experience more difficulty than Spanish learners. The study provides a clear picture of the similarities and differences between the two L2 groups, among others the different hierarchies of accuracy of the definite (*the*), indefinite (*a*), and zero (0) articles.

studies that compare learners of different proficiency levels are in fact based on an underlying L1 norm. The same can be said of SLA studies reporting the results of grammaticality judgment tests. In LCR, the norm, rather than being implicit and intuition-based, is explicit and corpus-based (Mukherjee, 2005). Finally, L1–L2 comparisons are extremely powerful heuristic techniques which help bring to light