Research Methods in Second Language Psycholinguistics

Edited by Jill Jegerski and Bill VanPatten

ROUTLEDGE

RESEARCH METHODS IN SECOND LANGUAGE PSYCHOLINGUISTICS

This timely volume provides up-to-date overviews of methods used in psycholinguistic research with second languages. Included are chapters on self-paced reading and listening, textual eye-tracking, visual world eye-tracking, ERPs, fMRI, translation recognition tasks, and cross-modal priming. Each contribution is authored by an expert researcher who offers experienced insight into not only the history of the method, but what is measured, how it is measured, issues in research and stimuli design, and the pros and cons of the method. These contributions are bookended by an introductory chapter on various models and issues that inform psycholinguistic inquiry into second language learning, and a final chapter that offers comments on the various methods described in addition to issues related to research design. Intended as a text to be used with advanced undergraduate and graduate students, *Research Methods in Second Language Psycholinguistics* will be useful to researchers wishing to understand more about the various methods represented and how they are used to investigate psycholinguistic processes in the second language context.

Jill Jegerski is Assistant Professor of Spanish and SLATE (Second Language Acquisition and Teacher Education) in the Department of Spanish, Italian, and Portuguese at the University of Illinois at Urbana-Champaign.

Bill VanPatten is Professor of Spanish, Second Language Studies, and Cognitive Science at Michigan State University.

SECOND LANGUAGE ACQUISITION RESEARCH SERIES

Theoretical and Methodological Issues

Susan M. Gass and Alison Mackey, Editors

Monographs on Theoretical Issues:

Schachter/Gass
Second Language Classroom Research: Issues and Opportunities (1996)

Birdsong
Second Language Acquisition and the Critical Period Hypotheses (1999)

Ohta
Second Language Acquisition Processes in the Classroom: Learning Japanese (2001)

Major
Foreign Accent: Ontogeny and Phylogeny of Second Language Phonology (2001)

VanPatten
Processing Instruction: Theory, Research, and Commentary (2003)

VanPatten/Williams/Rott/Overstreet
Form-Meaning Connections in Second Language Acquisition (2004)

Bardovi-Harlig/Hartford
Interlanguage Pragmatics: Exploring Institutional Talk (2005)

Dörnyei
The Psychology of the Language Learner: Individual Differences in Second Language Acquisition (2005)

Long
Problems in SLA (2007)

VanPatten/Williams
Theories in Second Language Acquisition (2007)

Ortega/Byrnes
The Longitudinal Study of Advanced L2 Capacities (2008)

Liceras/Zobl/Goodluck
The Role of Formal Features in Second Language Acquisition (2008)

Philp/Adams/Iwashita
Peer Interaction and Second Language Learning (2013)

Monographs on Research Methodology:

Tarone/Gass/Cohen
Research Methodology in Second Language Acquisition (1994)

Yule
Referential Communication Tasks (1997)

Gass/Mackey
Stimulated Recall Methodology in Second Language Research (2000)

Markee
Conversation Analysis (2000)

Gass/Mackey
Data Elicitation for Second and Foreign Language Research (2007)

Duff
Case Study Research in Applied Linguistics (2007)

McDonough/Trofimovich
Using Priming Methods in Second Language Research (2008)

Larson-Hall
A Guide to Doing Statistics in Second Language Research Using SPSS (2009)

Dörnyei/Taguchi
Questionnaires in Second Language Research: Construction, Administration, and Processing, 2nd Edition (2009)

Bowles
The Think-Aloud Controversy in Second Language Research (2010)

Jiang
Conducting Reaction Time Research for Second Language Studies (2011)

Barkhuizen/Benson/Chik
Narrative Inquiry in Language Teaching and Learning Research (2013)

Jegerski/VanPatten
Research Methods in Second Language Psycholinguistics (2013)

Of Related Interest:

Gass
Input, Interaction, and the Second Language Learner (1997)

Gass/Sorace/Selinker
Second Language Learning Data Analysis, Second Edition (1998)

Mackey/Gass
Second Language Research: Methodology and Design (2005)

Gass/Selinker
Second Language Acquisition: An Introductory Course, Third Edition (2008)

RESEARCH METHODS IN SECOND LANGUAGE PSYCHOLINGUISTICS

Edited by Jill Jegerski and Bill VanPatten

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

First published 2014
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Research methods in second language psycholinguistics / Edited by
Jill Jegerski, Michigan State and Bill VanPatten, University of Illinois at Urbana-Champaign.
pages cm. — (Second language acquisition research series)
Includes bibliographical references and index.
1. Second language acquisition. 2. Psycholinguistics. 3. Linguistics—Methodology. I. VanPatten, Bill., editor of compilation. II. Jegerski, Jill., editor of compilation.
P118.2.R474 2013
401′.93—dc23
2013025362

ISBN: 978-0-415-51825-3 (hbk)
ISBN: 978-0-415-51826-0 (pbk)
ISBN: 978-0-203-12343-0 (ebk)

Typeset in Bembo
by Apex CoVantage, LLC

CONTENTS

PREFACE

Although research on outcomes of second language (L2) learning (e.g., generative approaches, ultimate attainment) has made important contributions to the field of second language acquisition (SLA), what was and had been missing for some time is research on underlying processes and processing—that is, the psycholinguistics of L2 learning. The seeds for work on processing have their roots in early L2 research, but the psycholinguistics of SLA didn't really take off until the 1990s. Since then we have seen a steady increase in work related to input processing, sentence processing, output processing, lexical processing/retrieval, explicit/implicit processing, and other related areas. We have seen a steady increase in papers presented at conferences such as the Second Language Research Forum, as well as the emergence of conferences with a sole focus on the psycholinguistics of L2 learning. Journals such as *Studies in Second Language Acquisition, Language Learning,* and *Second Language Research* regularly publish papers with a psycholinguistics bent. And in doctoral programs in L2 studies, we have seen both regular courses and specialized seminars focused almost exclusively on the psycholinguistics of L2 learning. There has been, so to speak, a boom in psycholinguistic approaches to L2 learning in the last two decades.

Against this backdrop, the present volume was conceived. Its intent is to offer students of SLA an introduction to the various methods used in psycholinguistic (and neurolinguistic) research. With the exception of Chapter 1—which lays some foundational issues for the field of psycholinguistics in L2 research—and the final chapter—which assesses and comments on the various methods presented in the volume—the chapters in this book present the most current techniques and methods used to conduct psycholinguistic research in the L2 context. Each contribution is authored by a researcher or researchers with expertise in the particular method under question, who can offer experienced insight into not only the history of the method, but what is measured, how it is measured, issues in research and stimuli design, and the pros and cons of the method. Each chapter follows the same structure and ends with discussion questions and potential research projects, thus offering the student a solid introduction to the field. The book may also be of use to more experienced researchers who are looking at different ways to approach L2 learning or who wish to critique work that uses these newer methods. It is our belief that within L2 research, psycholinguistic approaches will continue to grow and the research generated by them will contribute significantly to understanding L2 learning in its many facets. The present book, then, is one attempt to provide support for these endeavors.

ACKNOWLEDGMENTS

Volumes of this sort require the talents and efforts of a number of people. To be sure, such a volume could not exist without the contributions made by the various experts who wrote the chapters. Enthusiastic about the proposal and its aims, they tackled their chapters dutifully and worked diligently to help us achieve a uniform and cohesive set of chapters that are related by structure and scope, but unique in content. They deserve the first thanks for the publication of this volume. We are of course indebted to the folks at Routledge, beginning with Ivy Ip (who has since left Routledge, but initially approached us about doing a book such as this one), and continuing with Elysse Preposi and Leah Babb-Rosenfeld, who saw the book through to the end. We also offer thanks to those "behind-the-scenes" people, who work on cover design, page setting, and all the other stuff that turns a manuscript into a book. They do important work. Finally, we thank our loved ones and those who have supported us not just during this project, but in our careers more generally. You know who you are.

This page intentionally left blank

CONTRIBUTORS

Nikos Amvrazis (Ph.D., Aristotle University of Thessaloniki) is a teacher of Greek at the School of Modern Greek Language of Aristotle University of Thessaloniki. His main research interests are second/foreign language acquisition, research methods, and syntax. He is currently focusing on the role of formal features and interfaces in SLA.

Paola (Giuli) E. Dussias (Ph.D., University of Arizona) is Professor of Spanish, Linguistics and Psychology and Associate Director, Center for Language Science, at Pennsylvania State University. The research that she and her students conduct concerns bilingual sentence processing. Together with her students, she examines whether language-specific information is largely kept independent when bilinguals compute or *parse* an initial syntactic structure for the sentences they read, or whether information from one language influences parsing decisions in the other language. Her work, using sophisticated behavioral methods such as eye-tracking and more recently event-related potentials, has been supported by grants from the National Science Foundation (NSF) and the National Institutes of Health (NIH).

Chip Gerfen (Ph.D., University of Arizona) is Professor of Spanish and Linguistics and Chair of the Department of World Languages and Cultures at American University in Washington, DC. His monograph *Phonology and Phonetics in Coatzospan Mixtec* was published in the Studies in Natural Language and Linguistic Theory series, and his research has appeared in a range of journals, including *Language, Studies in Second Language Acquisition, Bilingualism: Language and Cognition, The Mental Lexicon, Language and Cognitive Processes,* as well as a number of edited volumes. Broadly speaking, he is interested in cross-disciplinary research that bridges theoretical and experimental approaches to issues in language science.

Jill Jegerski (Ph.D., University of Illinois at Chicago) is Assistant Professor of Spanish and SLATE (Second Language Acquisition and Teacher Education) in the Department of Spanish, Italian, and Portuguese at the University of Illinois at Urbana-Champaign. Her primary research interests include non-native sentence processing, bilingual sentence processing, psycholinguistic research methods, near-nativeness in adult SLA, and heritage language instruction.

Gregory D. Keating (Ph.D., University of Illinois at Chicago) is Associate Professor of Linguistics and Second Language Acquisition in the Department of Linguistics and Asian/Middle Eastern Languages at San Diego State University. His research focuses on sentence processing in monolingual and bilingual speakers of Spanish and English.

Michael J. Leeser (Ph.D., University of Illinois at Urbana-Champaign) is Associate Professor of Spanish in the Department of Modern Languages and Linguistics at Florida State University. His research interests include input processing in second language acquisition, sentence processing in second language learners and bilinguals, and instructed SLA.

Kara Morgan-Short (Ph.D., Georgetown University) is an Associate Professor at the University of Illinois at Chicago. She holds a joint appointment in the Department of Hispanic and Italian Studies and the Department of Psychology and directs the Cognition of Second Language Acquisition laboratory. Informed by the fields of linguistics, cognitive psychology, and neuroscience, her research takes a multidimensional approach to elucidating the cognitive and neurocognitive processes underlying late-learned second language acquisition and use.

Aaron J. Newman (Ph.D., University of Oregon) is an Associate Professor in the departments of Psychology & Neuroscience, Psychiatry, Surgery, and Pediatrics at Dalhousie University in Halifax, Nova Scotia, Canada. His research interests surround neuroplasticity in the brain systems supporting language and hearing, including second language acquisition, sign language, aphasia, and cochlear implants. In his research he employs numerous techniques including fMRI, ERP, MEG, and behavioral methods.

Despina Papadopoulou (Ph.D., University of Essex) is Assistant Professor at the Department of Linguistics of the Aristotle University of Thessaloniki. Her research interests include language processing, second language acquisition, bilingualism, and language disorders.

Leah Roberts (Ph.D., University of Essex) is Professor and Leader of the Centre for Language Learning Research at the University of York. Her research focuses on second language learning and processing at the word, sentence, and discourse levels.

Gretchen Sunderman (Ph.D., The Pennsylvania State University) is Associate Professor of Spanish at Florida State University. Her research focuses on the bilingual lexicon, individual differences in lexical processing, and psycholinguistic approaches to second language acquisition.

Darren Tanner (Ph.D., University of Washington) is an Assistant Professor of Linguistics at the University of Illinois at Urbana-Champaign, where he directs the UIUC Electrophysiology and Language Processing Laboratory. His research uses behavioral and electrophysiological methods to study the cognitive processes underlying language comprehension in native speakers, L2 learners, and bilinguals.

Ianthi Tsimpli (Ph.D., University College London) is Professor of Multilingualism and Cognition at the School of Psychology and Clinical Language Sciences at the University of Reading, and Professor of Psycholinguistics and Director of the Language Development Lab at the Department of Theoretical and Applied Linguistics at Aristotle University of Thessaloniki. She teaches and researches in first and second language acquisition, bilingualism, psycholinguistics, and theoretical syntax.

Jorge Valdés Kroff (Ph.D., Pennsylvania State University) earned his Ph.D. in Hispanic Linguistics and Language Science at Pennsylvania State University. He is currently an NSF Minority Research Postdoctoral Fellow at the University of Pennsylvania. His research focuses on the cognitive and neural processes that support code-switching in comprehension and production.

Bill VanPatten (Ph.D., University of Texas) is Professor of Spanish, Second Language Studies, and Cognitive Science at Michigan State University. His research interests include second language morphosyntax, input processing, and instructed second language acquisition.

This page intentionally left blank

1

THE PSYCHOLINGUISTICS OF SLA

Bill VanPatten

What is Psycholinguistics?

Psycholinguistics, as classically defined, is concerned with a broad array of issues involving language acquisition, language comprehension and production, and the processing of language in the mind/brain (Fernández & Cairns, 2011). This emphasis on use or process stands in contrast to linguistic theory, which is largely concerned with the characterization of language itself, including internal constraints on the nature of language (e.g., Universal Grammar [UG]) and how language is represented in the mind/brain of the individual (i.e., the nature of mental representation). If we take the case of null subjects in a language like Spanish, we can see the difference between the foci of linguistic theory and psycholinguistics. Languages like Spanish have both null and overt subject pronouns, such that both (1a) and (1b) are possible and acceptable as grammatical sentences, whereas languages like English only allow overt subject pronouns in the same context (2a) and (2b).

(1)

a. Ella habla inglés.
b. *pro* habla inglés.

(2)

a. She speaks English.
b. ⋆*pro* speaks English.

Within linguistic theory, researchers are concerned with the properties of null and non-null-subject languages. They are interested in how the syntax licenses null subjects (e.g., the relationship between null subjects and verbal morphology), universal constraints that govern the interpretation of pronouns (e.g., the Overt Pronoun Constraint), as well as possible relationships between null subjects and other parts of the grammar (e.g., subject-verb inversion). In short, linguists are interested in how a null subject such as *pro* operates in the grammar and what allows it to be there to begin with. Within psycholinguistics, researchers are not interested in the properties per se, but rather, for example, how null and overt subjects are interpreted during comprehension. In the examples that follow, both (3a) and (3b) are possible, but what do speakers of Spanish take as the antecedent of each pronoun (bolded for convenience here), given that there are two possible choices?

(3)

a. Roberto llamó a Pablo después que ***pro*** regresó de México.
b. Roberto llamó a Pablo después que **él** regresó de México.

"Robert called Paul after he returned from Mexico."

There is nothing in the "linguistics" of null and overt subject pronouns that dictates to a speaker of Spanish whether Roberto or Pablo is the antecedent of either the null subject (*pro*) or the overt subject (*él*).[1] The field of psycholinguistics researches whether people have preferences for the interpretation of null and overt subject pronouns and what might be the reasons for these preferences.

We can take another example related to the lexicon. A speaker of English has the verb *die* encoded somewhere in the lexicon, and in the realm of linguistic theory researchers are concerned with the properties of that verb: its meaning (semantics), the relationship of semantics to structure (arguments such as whether the verb takes an agent, a patient, an experiencer, and so on), and how that structure projects into the syntax. In psycholinguistics, researchers might be concerned with how that verb gets accessed in real time comprehension or production; and if we are dealing with bilinguals, we might be concerned with the extent to which the two languages are activated during lexical retrieval. As a bilingual speaker of Spanish and English since my childhood, do I automatically activate both *die* and *morir* when speaking either language? At any given time, are both languages active? Does activation depend on language dominance? These are sample questions that those working within the psycholinguistics of the lexicon might ask.

In short, whereas linguistic theory is centered on representation, psycholinguistics is centered on processes and processing. But the division between representation and processes/processing is not always so neat. Clearly, when one is processing language, one has access to a mental representation of language (Gorrell, 1995; Pritchett, 1992). One of the hallmarks of theories on language comprehension is that as we understand a sentence, we automatically build syntactic structure during real time in order to comprehend that sentence. We also access the mental lexicon, which in turn is interfaced with the grammar of the language as suggested above. Let's suppose we hear or read the verb *examined*. The lexicon tells us that this verb has certain properties and that its argument structure requires two entities: an agent (the examiner) and a patient (the thing examined). Both of these must be projected onto the syntax. For this reason (4a) is good but not (4b).

(4)
 a. The lawyer examined the witness next.
 b. *The lawyer examined next.

During comprehension, as we build structure, we expect *examined* to have two entities represented in the syntax of the sentence somewhere, by a subject and an object of the verb. Thus, if we hear or read "*The lawyer examined . . .*" we project subject status onto *the lawyer* and expect an object to follow *examined.* In other words, we have projected a verb phrase with *examined* as the main verb, and within that phrase we have projected a determiner phrase governed by the verb, thus creating the expectation of an object to follow. How do we know we build this structure in real time? Because psycholinguists have investigated what happens when people comprehend sentences in which *examined* is not a main verb as in (4a) but is part of a reduced relative clause as in (5). The phrase *by the judge* takes longer to read compared with the same phrase in a sentence such as (6) (see, for example, Trueswell, Tanenhaus, & Garnsey, 1994).

(5) The lawyer examined by the judge was cited for misconduct.
(6) The evidence examined by the judge was later thrown out.

The reason it takes longer to read *by the judge* in (5) compared to (6) is that the syntactic computation is wrong if *the lawyer* is initially projected as the subject of *examined*. When an object does not follow, the reader/listener does an unconscious double take and must reanalyze the syntactic structure to fit the new phrase. This does not happen with (6) because *evidence* is not normally projected as the subject of *examined* (*examine* typically, if not always, requires an animate subject, which *evidence* cannot be). By researching what people do when they are led down a "garden path of processing" (i.e., initially computing the wrong syntax for a sentence), psycholinguistic research has shown how there is a relationship between real time comprehension and knowledge of grammar. The two go hand in hand (e.g., Gorrell, 1995; Pritchett, 1992). To be sure, there is much more to comprehending a sentence than syntactic structure. Sounds have to be perceived, and those sounds have to be computed and organized into words. During this processing, words are retrieved from the lexicon along with the information they bring to sentence building. Intonation has to be perceived and mapped onto the sentence. And so on (e.g., Carroll, 2001; Fernández & Cairns, 2011). The point being made here is that although psycholinguistics is concerned with the process of comprehension, it must, by force, relate comprehension to the language itself and how that language works. This does not mean that all psycholinguists adhere to the same theory of language. Some psycholinguists link language processing to current generative theory while others do not (e.g., Carminati, 2002; Clifton, Frazier, & Rayner, 1994; Pritchett 1992).

As mentioned earlier, psycholinguistics is often tied to language acquisition, which is sometimes referred to as *developmental psycholinguistics.* Psycholinguists are interested in stages of acquisition, developmental patterns, and how children string words together to create meaning. They are also interested in how sentence comprehension develops in the child, and how this comprehension relates to the acquisition (e.g., Bever, 1970; Slobin, 1985). Linguists are also interested in child language acquisition, but more from the viewpoint of how linguistic theory can explain developmental patterns and to what extent features of language work together to shape a grammar in the child's mind (e.g., Guasti, 2004; Lust, 2006). In other words, linguists are more interested in seeing to what extent the principles of Universal Grammar govern and constrain a child's mental representation of language.[2]

Psycholinguistics and SLA

Historically, SLA has always been linked in some way to psycholinguistics. By the nature of being acquisition, SLA is a kind of developmental psycholinguistics in the same way first language acquisition is sometimes called developmental psycholinguistics. Although it is true that a good deal of research in SLA has focused on representation and the role of things like Universal Grammar in constraining second language grammars (e.g., White, 2003), lurking in the background has been the idea that acquisition proceeds in SLA as it does in first language acquisition—via comprehension (and, of course, the interaction of comprehension with internal language-making mechanisms; see, for example, Carroll, 2001, as well as Truscott & Sharwood Smith, 2004).[3] As we have seen, comprehension is the domain of psycholinguistic research. In 1983, Evelyn Hatch published the first book to explicitly link psycholinguistics and SLA, aptly titled *Psycholinguistics: A Second Language Perspective.* Central to her book and indicative of the time period in which she wrote, is a description of the nature of second language acquisition, with a focus on how formal properties of language develop over time (e.g., morphology, phonology, syntax), the roles of input and interaction in language development, and how SLA relates to general cognition.

Since the publication of Hatch's book, psycholinguistics within SLA has focused less on language development per se and more on the real-time comprehension and production of language (e.g., Frenck-Mestre, 2005; Pienemann, 1998). Included in this focus is a move toward using investigative techniques and methods that had only been used in adult first language research, with an eye toward comparing native and second language processing of language. For example, if we return to the issue of null and overt subject pronouns from an earlier part of this chapter, the reader will recall that languages like Spanish license null subjects so that both (a) and (b) versions of the sentence (7) below are grammatical:

(7)

a. Roberto llamó a Pablo después que ***pro*** regresó de México.
b. Roberto llamó a Pablo después que **él** regresó de México.

"Robert called Paul after he returned from Mexico."

Research on Spanish native language processing shows that native speakers have a strong preference to have *pro* take the subject (what resides in [SpecIP]) of the previous clause as the antecedent. However, these same speakers do not show an antecedent preference for the overt subject pronoun *él*; about 50% of the time speakers link it to the subject of the previous clause and 50% to the nonsubject of the previous clause (e.g., Alonso-Ovalle, Fernández-Solera, Frazier, & Clifton, 2002; Keating, VanPatten, & Jegerski, 2011). Research on second language learners, even advanced learners, tends to show that the antecedent preference for both null and overt subject pronouns is the subject of the previous clause; that is, learners seemingly treat the two pronouns the same during comprehension whereas native speakers do not (Jegerski, VanPatten, & Keating, 2011; Sorace & Filiaci, 2006). At the same time, there may be some subtle properties of discourse that transfer from the L1 (English) that affect comprehension of these pronouns (Jegerski, VanPatten, & Keating, 2011). What is not in question here is representation: The second language learners studied in this research clearly have an L2 null subject language with the formal properties and constraints intact. What seems to be different is processing; how these learners link null and overt subject pronouns to antecedents during comprehension. What the realm of psycholinguistic research such as this shows is that it may be one thing to acquire the syntactic properties of language, but something else to acquire processing "interfaces" (e.g., Sorace, 2011). As we will see below, one of the major debates in psycholinguistic research in adult SLA is whether learners come to have the same processing interfaces (e.g., syntax-discourse) as adult native speakers. Also debated, as we will see, is to what extent learners rely on their first language processing mechanisms during comprehension of the second language.

Since the 1990s, we have seen a burgeoning of models and hypotheses about second language processing. These models and hypotheses demonstrate the importance that psycholinguistic research has in SLA these days, and also delineate a particular research agenda. In the sections that follow, I briefly describe some of the major psycholinguistic models that dominate the field, including the kinds of questions they ask.

The Revised Hierarchical Model

The Revised Hierarchical Model (RHM; e.g., Kroll & Tokowicz, 2005), is a model of both representation and processing. Traditionally, the RHM has been used to examine lexical storage and lexical processing, the primary hypothesis behind the model being that second language lexical items develop not through direct mapping onto the conceptual system (meaning) but with first language lexical items as mediators. That is, second language learners' lexicons do not initially develop direct links between the second language word and its meaning; during processing, the first language word "mediates" between the second language word and the conceptual system. Direct links between second language words and the conceptual system are possible and do develop, but the model proposes that such links are weaker than those that exist between the first language and the conceptual system. What is more, the model claims that it is difficult, if not impossible, to shut out the first language completely as a mediating system. It is always there, turned on, and somehow working as a "buffer" between the second language lexicon and the conceptual system.

A good deal of the research supporting the RHM involves measuring reaction time in tasks using lexical naming and/or translation (e.g., Francis, Tokowicz, & Kroll, 2003; Kroll & Stewart, 1994; Sunderman, 2002; Talamas, Kroll, & DuFour, 1999). In these studies, various types of variables are manipulated to see how naming and translation are affected. These variables include semantic versus nonsemantic grouping of words (e.g., one group of words belongs to one category such as 'fruits' while another group involves mixed categories such as fruits, vegetables, meats) and form versus meaning (e.g., words look the same, such as *hombre-hambre* "man-hunger," as opposed to being semantically related such as *hombre-mujer* "man-woman"), among others (see Kroll & Tokowicz, 2005, for a summary and discussion). Other types of research designs are used as well, but almost all measure how long it takes learners to make a decision (measured in milliseconds) about the nature of a word. Here, I will review one study to demonstrate how the research supports the RHM.

In the classic publication, Kroll and Stewart (1994), the researchers presented the results of three experiments. I will focus on Experiment 3. In this last experiment, Kroll and Stewart tested 24 fluent bilinguals whose first and dominant language was Dutch and second language was English. The task consisted of both naming a word (i.e., reading aloud the word on the computer screen) and translating from Dutch to English or vice versa. Words appeared either in category groups (e.g., weapons, vegetables, animals) or randomized with no category. Response times for both naming and translating were recorded. Also, in the end, participants were asked to recall as many words as they could in the language in which the words were presented. Kroll and Stewart's predictions were that translation times from L2 to L1 would be faster than translation times from L1 to L2, because the latter involves strong conceptual links (for the L1) and weak translation links, while the former involves strong translation links and weak(er) conceptual links (for the L2). Their results supported this prediction. What is more, they found that whether or not words appeared in semantic categories did not affect length of time to translate in the L2 to L1 condition, presumably because these translations are lexically linked, not conceptually linked. They found that the L1 to L2 translations were affected by semantic categorization versus randomization, presumably because of the strong L1-concept links.

Work within the RHM has produced a substantial body of work. Overall, the model's predictions have been supported by the research. L1 lexical items are more strongly linked to concepts than L2 items. The latter are strongly linked to L1 lexical items while L1 items are more weakly linked to L2 lexical items. Thus, there is an asymmetry in lexical representation, which plays out in various tasks.

To be sure, the RHM has been limited to lexical representation and processing, but as Kroll and Tokowicz (2005) suggest, it could be applied to other aspects of language. Because it has been limited to lexicon, the model has not taken a stand on something like the role of UG in adult SLA—an important issue in the field (e.g., see the special issue of *Studies in Second Language Acquisition,* Slabakova, 2009). If the RHM is applied to something like syntactic processing and representation, then those working within the model will have to grapple with the role of UG.

Shallow Structure Hypothesis

The Shallow Structure Hypothesis (SSH), was launched by Clahsen and Felser (2006) as part of the debate on outcome differences between native language acquisition and second language acquisition (i.e., to what extent L2 learners can become native-like). The crux of their argument was this: the processing of language can involve both structural processing (full computation of syntactic structure during comprehension) and shallow processing (partial computation of syntactic structure with greater reliance on pragmatic and lexical information so that meaning is not lost). The basic claim of the SSH is that although native speakers make use of both structural processing and shallow processing, second language learners tend to rely on shallow processing more—and in fact, Clahsen and Felser claim that L2 learners make exclusive use of shallow processing. The classic example Clahsen and Felser provide is on the use of gaps in sentence processing (see Marinis, Roberts, Felser, & Clahsen, 2005, for an example). In this kind of research, natives and non-natives were given sentences with *wh*- gaps to process as in (8). Slash marks indicate how the sentences were divided for the task of self-paced reading (see Jegerski, Chapter 2, this volume). Marinis et al. (2005) used four types of stimulus sentences, but we will only use two here for illustration.

(8) The nurse who / the doctor argued / that/ the rude patient / had angered /
1 2 3 4 5
is refusing to work late.
6

(9) The nurse who / the doctor's argument / about/ the rude patient /
1 2 3 4
had angered / is refusing to work late.
5 6

In (8), there are several gaps linked to the clause head *who,* represented by the classic *e* (for empty): "The nurse *who*$_i$ the doctor argued *e*$_i$ that the rude patient had angered *e*$_i$ is refusing to work late." In contrast, in sentence (9), there is no intermediate gap: "The nurse *who*$_i$ the doctor's argument about the rude patient had angered *e*$_i$ is refusing to work late." The researchers were interested in how long it took natives and non-natives to read Regions 3 and 5 of this type of sentence. Regions 3 and 5 contain two gaps in sentence (8) but only one gap in sentence (9). According to how gaps are processed by natives, if L2 learners are relating the gaps to the clause head *who,* they should slow down in Region 3 compared to sentences that don't have this same gap. As for Region 5, if L2 learners make use of the gaps the way natives do, then they should show shorter reading times in Region 5 compared to sentences that have the same gap but of a structurally different kind, for example: "The nurse thought the doctor argued that the rude patient had angered the staff at the hospital." What Marinis et al. (2005) found was that natives behaved as predicted, but the L2 learners did not, suggesting they were not processing gaps the same way native speakers were. They argued that learners could rely on nonsyntactic information to "fill in" for shallower processing without a loss of sentence comprehension.

Also central to the SSH is that the first language plays little to no role in sentence processing. In the Marinis et al. study just described, the researchers tested learners from the following L1 backgrounds: Chinese, Japanese, German, and Greek. They found no discernible L1 effect during processing, concluding that all four L2 groups essentially behaved the same way on the task. (The reader who is keeping track as we move along will note that the nonrole of the L1 in the SSH is different from what would [most likely] be predicted by the revised hierarchical model if the latter were applied to syntactic processing.)

The SSH has begun to generate a good deal of research on L2 processing (see, for example, the papers in VanPatten & Jegerski, 2010), with some research supporting the SSH and some not. However, the hypothesis is still relatively new, and it will be some time before the evidence significantly tilts one way or another.

Because the hypothesis is concerned strictly with processing, it is not clear how it relates to the role of UG in adult language acquisition. Indeed, Clahsen and Felser (2006) make no mention of representation or how syntax develops in the mind/brain of the learner. Their account does assume a generative grammar (hence, the reference to gaps and movement), but their focus is on processing as an end-state and not as part of the mechanism by which a linguistic system develops (the interface between input data and UG, for example).

MOGUL and Acquisition by Processing

Committed to the generative account of acquisition, the MOGUL model (Modular On-line Growth and Use of Language) was first articulated in Truscott and

Sharwood Smith (2004), and has been elaborated in subsequent publications as the authors address a variety of issues in acquisition (e.g., Sharwood Smith & Truscott, 2006; Truscott & Sharwood Smith, 2011). The model is mostly concerned with placing the acquisition of language within a coherent processing account, and makes use of not only generative constructs regarding the nature of language and (functional) categories (e.g., the model assumes that UG provides all the constraints and primitives needed for processing), but also constructs used in cognitive areas of research (e.g., memory, activation levels, competition). As one simple example of this framework, Truscott and Sharwood Smith argue that the acquisition of prodrop (null subject) languages proceeds when the processor encounters a "subjectless" sentence. If the L1 is a non-null subject language, then the resting level of [-] will be the norm. Encountering a null subject forces the processor to generate a [+] value, but its resting level is low. A low resting level would explain why in the early stages learners with a non-null subject L1 tend to overuse subject pronouns; they have yet to acquire a strong resting level for [+]. Repeated encounters with null subjects pushes the resting level of [+] higher and higher until it reaches a level that, for the L2, becomes the norm. This example illustrates how MOGUL relates processing of formal features to more psychological constructs such as resting and activation levels.

As a framework for linking acquisition to processing, MOGUL has not generated original empirical research yet. However, the model has been used to account for a variety of observed phenomena in adult SLA, including stages versus continua in the acquisition of formal elements (Sharwood Smith & Truscott, 2004), and why explicit learning seems to play little to no fundamental role in the acquisition of linguistic properties (Truscott & Sharwood Smith, 2011).

Structural Distance and Amelioration

In contradistinction to the MOGUL model, William O'Grady has argued for a processing account of SLA without recourse to UG (e.g., O'Grady, 2003, 2010). In his account, there are two major constructs that shape processing, which in turn shapes acquisition. The first is the overarching amelioration hypothesis, which states that acquisition consists of processing amelioration (O'Grady, 2010). In short, the claim is that acquisition proceeds only as processing becomes easier for the learner. This hypothesis in and of itself does not lead to any particular predictions; it is in one of O'Grady's other hypotheses that we see actual predictions about behavior: structural distance (O'Grady, 2003). Under the structural distance hypothesis, O'Grady claims that processing is more difficult for those items that are structurally more distant from their source sites. For example, *wh-* subjects are easier than *wh-* objects because the *wh-* object has to cross more syntactic nodes to land in sentence initial position. The idea here is that the processor must link or co-index the *wh-* item with its source site, and constraints on working memory during acquisition cause structural distance to be a significant factor in performing this on-line task. In one study, O'Grady, Lee, and Choo (2003) examined the interpretation of subject relative clauses versus object relative clauses in Korean as L2 (English L1) and found that indeed, learners made significantly more mistakes interpreting object relative clauses compared to subject relative clauses (see also O'Grady, 2010). This particular research relied on what are called off-line measures of comprehension. Unlike the methodologies described and discussed in this volume, O'Grady, Lee, and Choo had participants hear a sentence and then match it to one of several pictures to indicate their comprehension. The measurement taken (in this case, picture selection) occurs after the participants have heard or read a sentence. In contrast, on-line methodologies (such as that used by Marinis et al., 2005, as described in the section above on the shallow structure hypothesis) attempt to capture moment-by-moment processing, measuring reading times (or something else) at particular points within the sentence (or text). It is not clear to what extent O'Grady's account will lend itself to on-line research, and as is the case of most research, much depends on the particulars of the research questions formulated.

As in the case of MOGUL, O'Grady's account has not yet yielded a significant body of empirical research evidence, although it has been used to account for certain phenomena in SLA (e.g., scope of negation in English). As for the role of the L1, O'Grady has suggested that L1 processing routines form the starting point of how learners process input data, and that these are "gradually pushed aside" (an example of processing amelioration); however, he allows that a second language processing routine can be used instead of an L1 routine if it is "less costly" (O'Grady, 2010). It is not clear how "less costly" is to be characterized and thus operationalized, but O'Grady is still working out the details of his model. We will await further research to see how things develop as far as amelioration and structural distance are concerned.

Input Processing and Processing Instruction

The model of input processing introduced by the present author has been the focus of considerable research for almost three decades (e.g., VanPatten, 1983, 1985, 1990, 2004, 2009). As the first attempt to link processing to morphosyntactic acquisition, two main processing principles are the center of the model (there are corollaries that we will ignore here). We will take the most current versions of these principles (VanPatten, 2009, p. 51):

> The Lexical Preference Principle (LPP): If grammatical forms express a meaning that can also be encoded lexically (i.e., that grammatical marker is redundant), then learners will not initially process those grammatical forms until they have lexical forms to which they can match them.
>
> The First Noun Principle (FNP): Learners tend to process the first (pro)noun of a sentence as the subject/agent.

The LPP is meant to account for the fact that learners generally acquire lexical items before functional items, and that functional items emerge in the grammar based on the degree to which they become important for processing of meaning in sentences. The FNP was originally developed to account for the problem of acquiring nonSVO (Subject-Verb-Object) word order (including case marking, preverbal clitic object pronouns, and other items). Taken together, these principles account for a substantial number of acquisition problems.

Although this model of input processing was initially neutral on the role of UG in adult SLA, subsequent publications have argued that there is nothing incompatible with a processing model and a UG-based account, much in the same way as has been articulated for the MOGUL model (see, for example, VanPatten, 1996, as well as Rothman & VanPatten, 2013, and VanPatten & Rothman, in press).

Because the model posits the two main principles as universal, it downplays the role of the L1 in processing. In one study, Isabelli (2008) tested learners of Italian L1 early in their learning of Spanish L2. Isabelli hypothesized that because Spanish and Italian both share OVS as a possible (albeit noncanonical) word order, the Italian learners of Spanish would not show the same processing problems as English L1 speakers with sentences such as *Lo ve María* "Him sees Mary = Mary sees him." Under VanPatten's FNP, learners would misinterpret the object clitic pronoun *lo* as the subject and misinterpret the sentence. She indeed found that beginning learners of Italian did not have the same difficulty processing these sentences as English speakers. At first blush, this does look like there is L1 transfer in processing, with Italian and English speakers having different processing problems in Spanish as L2. However, another interpretation is that the FNP is still present in Italian learners of Spanish, but because they share the same lexical items for singular clitics (*lo* "him," *la* "her"), the FNP is attenuated in this one instance because of lexical transfer. Thus, processing strategies may interact with different structures in different ways depending on lexical correspondence between two languages. That the role of L1 is not so clear in input processing is corroborated by research on English and French full passives, which are structurally identical. In Ervin-Tripp (1974), L2 learners of French did not rely on knowledge of English passive to process French passives, instead reverting to the FNP to comprehend sentences in the earlier stages of SLA. In short, the L1 did not play a facilitative role.

Processability Theory

Unlike the accounts so far, the Processability Theory (PT) is a theory about the development of output processing procedures and so far has remained agnostic about input processing/computation during comprehension as well as representation (Pienemann, 1998, 2007). According to PT, there is a hierarchy of output processing procedures that emerge over time, such that the emergence of one procedure implies all procedures underneath it in the hierarchy but not necessarily those above (see Table 1.1). The relative difficulty of the procedures is to be

TABLE 1.1 The hierarchy of output processing procedures in the Processability Theory (based on Pienemann, 2007)

The procedure	*Example in learner output*
1. No procedure/Lemma	Production of a simple word
2. Category procedure	Adding past tense morpheme to a verb
3. Phrasal procedure	Matching plurality as in "two kids"
4. Simplified-S procedure/VP procedure	Moving an adverb out of the VP to front of sentence
5. S-procedure	Subject-verb agreement
6. Subordinate clause procedure	Use of subjunctive in subordinate clause triggered by information in the main clause

found in the intersection of Lexical Functional Grammar with working memory (i.e., how difficult it is to carry grammatical information across syntactic boundaries during real-time production). For example, noun-adjective agreement is easier within a noun phrase (NP; i.e., the controlling noun and the adjective are in the same basic phrase *la casa vacía* "the vacant house") but more difficult when the adjective is found in a different phrase, such as within the verb phrase (VP; *la casa quedaba vacía* "the house was vacant") and even more difficult when the adjective is in another clause (e.g., *pintaron la casa que quedaba vacía* "they painted the house that was vacant").

Processability Theory is intended to account for observed stages of development in learner output over time, focusing not on mastery but when particular surface phenomena emerge in learner speech. As stated above, PT is agnostic about how learners process input, as well as how grammar is represented in the mind/brain. It is not, however, agnostic regarding L1 transfer. Pienemann has repeatedly argued—with empirical data that support his claim—that L1 transfer is severely constrained by the hierarchy of processing procedures. What this means is that L1 features and structures can only be transferred at the point at which there is congruence with an emerged processing procedure. Thus, subject-verb agreement from Italian L1 cannot be transferred from the outset into Spanish L2. In this case, transfer would not occur until the learner has reached the S-procedure stage in the hierarchy (see Table 1.1).

Typical studies under the PT framework are different from comprehension-based psycholinguistic research. In PT studies, the norm is to collect spontaneous (or carefully elicited) speech data and then look for "emergence" of particular structures. For example, Kawaguchi (2005), studied two learners of Japanese over the course of either a two-year or three-year period of formal study (both were beginners). She collected oral data twice per semester over the course of these learners' formal study. She then analyzed the data for the emergence of various structures and found that they emerged as predicted by the PT hierarchy; that is, lemmas (simple words) appeared before those structures that involved simple category procedures, and these before phrasal procedures, and so on. Her conclusions included that L1 transfer was constrained, as were the effects of teaching: learners could not skip any stage of development due to formal instruction.

The Two—If not Three—Major Issues, Then

As might be deduced from the brief overview of the models above, there are two major issues that processing models—whether input-oriented or output-oriented—address. The first is the role of the L1. The second is the role of UG. How each model stands on these two major issues is summarized in Table 1.2. Not all models make the same predictions about L1 transfer at the level of processing, ranging from full transfer (e.g., the RHM) to minimal or constrained transfer (e.g., Processability) to no transfer (e.g., the SSH).

A third major issue in processing has been addressed only by the MOGUL model: the role of explicit learning. As mentioned above, the MOGUL model makes a strong argument against any kind of explicitness playing a major (if any) role in processing and language development, largely because it assumes the same underlying architecture for adult SLA that is assumed for child L1A. All other models remain neutral on the role of explicitness, leaving the issue largely unaddressed. The Processability Theory does claim that formal instruction cannot override the constraints inherent in its hierarchy; however, it makes no claims regarding the explicitness or implicitness of output processing.

There is a fourth issue lurking around and now made prominent by the Shallow Structure Hypothesis: outcomes. Here the issue is the extent to which L2 learners can become native-like in their processing of sentences. The SSH is clear in claiming that although shallow processing is available in L1 processing, L2 learners make use of it to the exclusion of full syntactic processing, and that this

TABLE 1.2 Six major processing models and their stance on L1 transfer and Universal Grammar

The model	*L1 plays a role in processing*	*UG constrains the underlying grammar*
Revised Hierarchical Model	Yes	Neutral
Shallow Structure Hypothesis	No	Neutral
MOGUL	Yes	Yes
Amelioration and Structural Distance	Yes	No
Input Processing	No	Yes
Processability Theory	Yes, but constrained by the processing hierarchy	Neutral

reliance inhibits the development of native-like processing that is not shallow. Other models are mute on this point, but if one looks closely, there are hints within some of these models about outcomes. For example, the Revised Hierarchical Model suggests that although direct mapping between lexical items and conceptual structure is possible in an L2, these mappings are likely to be weaker than those that are L1 mappings. This is suggestive in that L2 learners might never become native-like in a variety of domains: speed and accuracy, for example. The MOGUL approach directly addresses the issue of outcome. In one scenario, Truscott and Sharwood Smith (2004) argue that there is input competition, in that the L1 is constantly activated during the parsing and processing of input strings. This constant activation could retard or even prevent development. In another scenario, they suggest there could be output competition. In this case, almost the mirror image of input competition, the L1 and L2 lexical items compete for access during speech production. As they state, "Given the strong connections L1 items have already formed with the [processing] chain, the L2 system inevitably faces an uphill battle, frequently losing the competition and failing to fully express the L2 competence" (p. 15). This "failure" during production does not a priori suggest a problem with underlying competence, something Truscott and Sharwood Smith clearly state. If there is a problem with underlying competence, it would be traceable to input competition in their framework.

The other models—Input Processing, Structural Distance/Amelioration, Processability—do not make any claims that address eventual outcome in any discernable way.

Psycholinguistics and Neurolinguistics

An introduction of this kind would be remiss if it did not touch upon the field of neurolinguistics and its research methods, which increasingly have been used to examine language processing. The field of L2 neurolinguistics is dominated by research that examines activity within the brain (e.g., electrical impulses, blood flow) in order to determine to what extent L1 and l2 processing converge or diverge. Classic studies involve the use of ERPs—event-related (brain) potentials. ERPs refer to electrical activity in the brain, which are measured in milliseconds, and are used to determine the processing of both semantic anomalies and morphosyntactic violations. A good deal of the research in the L2 context has examined the extent to which L2 learners process semantic and morphosyntactic violations like native speakers and what the developmental path is toward such processing. The studies have mixed results. Some suggest that native-like processing is attainable by L2 learners, and for some structures and some languages, this processing can be attained relatively soon (e.g., Osterhout, McLaughlin, Pitkanen, Frenk-Mestre, & Molinaro, 2006). Others paint a different picture, claiming that some of the ERP research suggests native-like processing is elusive—or at least very difficult—for L2 learners (e.g., Mueller, 2006). Other studies shed light on these divergent conclusions looking at the context in which learning occurs. For example, Morgan-Short, Sanz, Steinhauer, and Ullman (2010), have shown that the type of ERPs obtained may depend on how something is learned. In their study, learners were exposed to a miniature artificial language under one of two conditions: explicit (focus on grammar with practice) and implicit (as in an immersion situation). The latter yielded consistent native-like processing effects while the former did not. Thus, some of the conflicting results mentioned above could be due to the training/learning conditions of the various studies or to the history/experience of the learners with the L2. What also needs to be teased out in this research is the nature of the linguistic structure under study and how it is selected (e.g., UG-derived syntactic structures vs. nonUG-derived structures). As research in other domains has shown (e.g., VanPatten, Leeser, & Keating, 2012), learner performance in on-line research may match that of native speakers for one structure and yet diverge on another.

In addition to ERPs, various imaging techniques have been used to investigate L2 processing. These techniques include fMRI (functional magnetic resonance imaging) and PET (positron emission tomography). Most of the research using brain imaging has focused on issues related to the critical period and to what extent monolinguals and bilinguals process language in the same or different parts of the brain (see, for example, Kotz, 2009). This research has examined not only lexical processing (including the processing of idioms) but also morphosyntactic processing. Kotz, for example, argues that the research suggests that the same parts of the brain are used to process syntax in both L1 and L2. However, L2 proficiency seems to play a role. At the earlier stages, it may appear that the L2 learner is engaging different parts of the brain to process language compared to the L1 native. But, as Abutalebi, Tettamantti, and Perani (2009) argue, a closer inspection shows that L2 learners are not engaging different parts of the brain compared to L1 speakers, but parts of the brain in addition to those employed by L1 speakers—presumably because of the need to engage more executive control and attention at the lower levels of ability. With time, these "extra" parts of the brain are used less and less until the L2 learner converges on a pattern of brain activity that looks similar to if not identical to that of the L1 speaker.

Conclusion

That adult SLA is now tightly linked to psycholinguistics and language processing is clear. In this chapter, I have laid out what some of the critical issues are and how certain influential models and hypotheses address L2 processing. Although representation remains a critical area of inquiry—especially the degree to which L2 representation is similar to or different from that of native speakers (and for that matter, early bilinguals)—what has emerged over the last twenty years is an increased concern for how L2 learners process language they hear and read (or speak, in the case of the Processability Theory), and the factors that affect this processing. The bilingual mind/brain is an interesting place to examine linguistics as it plays out in real time language use. And how this processing happens will provide additional insight not only to end-state (i.e., to what extent learners can become native-like) but as to how language develops in the mind/brain over time (i.e., how processing works in the creation of a linguistic system).

Discussion Questions

1) Although research on adult L2 learning can be "divided" into representation on the one hand and processing on the other, in some ways an internal grammar and the processing mechanism must work together. How so? In what ways does the moment by moment computation of sentence structure rely on a grammar, for example?
2) Some of the models and frameworks presented in this chapter acknowledge a role for Universal Grammar in acquisition while others do not or are agnostic. The same can be said for the role of the L1. What is your perspective on these roles? What would be critical data to argue for or against a particular position, either in representation or in processing?
3) Regardless of theoretical orientation, most researchers converge on the idea that learners build an implicit linguistic system. Where there is contention is on the role of explicit processing during acquisition. Consider what it takes to compute the structure of a sentence either in the early stages or later stages of acquisition (e.g., the assignment of meaning to surface lexical and morphological forms, the projection of phrases, the internal hierarchical structure of the sentence). To what degree do you see explicit processing involved?
4) Select one of the theoretical frameworks presented in this chapter and then select an empirical study representing that framework. If you are taking a course, present the study to the class. After your presentation, what do you and your classmates see as the strengths and weaknesses of the framework for answering questions about SLA? What would be a natural follow-up project to the study you presented?

Notes

1. To be sure, null subject languages obey the Overt Pronoun Constraint, which puts restrictions on the binding of overt subject pronouns. For the example at hand, the OPC is not relevant.
2. Admittedly, the distinction between what linguists and psycholinguists look at in child language acquisition has blurred over the years. See, for example, Lust (2006).
3. This does not mean that child L1 and adult SLA are identical in terms of product. Clearly, the first language influences adult SLA, and adults tend not to be native-like in a variety of domains (e.g., Sorace, 2003, 2011). The point here is that the underlying processes are identical.

References

Abutalebi, J., Tettamanti, M., & Perani, D. (2009). The bilingual brain: linguistic and non-linguistic skills. *Brain and Language, 109,* 51–54.

Alonso-Ovalle, L., Fernández-Solera, S., Frazier, L., & Clifton Jr., C. (2002). Null vs overt pronouns and the topic-focus articulation in Spanish. *Journal of Italian Linguistics, 14*(2), 151–169.

Bever, T. (1970). The cognitive basis for linguistic structures. In J. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

Carminati, M. N. (2002). *The processing of Italian subject pronouns* (Unpublished doctoral dissertation). University of Massachusetts, Amherst, USA.

Carroll, S. (2001). *Input and evidence: The raw material of second language acquisition.* Amsterdam: John Benjamins.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Clifton, C., Frazier, L., & Rayner, K. (Eds.) (1994). *Perspectives on sentence processing.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ervin-Tripp, S. M. (1974). Is second language learning like the first? *TESOL Quarterly, 8*(2), 111–174.

Fernández, E. M., & Cairns, H. (2011). *Fundamentals of psycholinguistics.* West Sussex, UK: Blackwell Publishing.

Francis, W., Tokowicz, N., & Kroll, J. (2003). Translation priming as a function of bilingual proficiency and item difficulty. *Fourth International Symposium on Bilingualism.* Tempe, AZ.

Frenck-Mestre, C. (2005). Eye-movement as a tool for studying syntactic processing in a second-language. *Second Language Research, 21,* 175–198.

Gorrell, P. (1995). *Syntax and parsing.* Cambridge, UK: Cambridge University Press.

Guasti, M. T. (2004). *Language acquisition: The growth of grammar.* Cambridge, MA: MIT Press.

Hatch, E. (1983). *Psycholinguistics: A second-language perspective.* Rowley, MA: Newbury House Publishing.

Isabelli, C. (2008). First noun principle or L1 transfer principle in SLA? *Hispania, 91*(2), 463–476.

Jegerski, J., VanPatten, B., & Keating, G. (2011). Cross-linguistic variation and the acquisition of pronominal reference in L2 Spanish. *Second Language Research, 27,* 481–501.

Kawaguchi, S. (2005). Argument structure and syntactic development in Japanese as a second language. In M. Pienemann (Ed.), *Cross-linguistic aspects of Processability Theory* (pp. 253–198). Amsterdam: John Benjamins.

Keating, G., VanPatten, B., & Jegerski, J. (2011). Who was walking on the beach? Anaphora resolution in monolingual natives and heritage speakers of Spanish. *Studies in Second Language Acquisition, 33,* 193–221.

Kotz, S. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language, 109,* 68–74.

Kroll, J., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

Kroll, J., & Tokowicz, N. (2005). Models of bilingual representation and processing: Looking back and to the future. In J. Kroll & A. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–553). New York: Oxford University Press.

Lust, B. (2006). *Child language: acquisition and growth.* Cambridge, UK: Cambridge University Press.

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition, 27,* 53–78.

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. (2010). Acquisition of gender agreement in second language learners: An event-related potential study. *Language Learning, 60*(1), 154–193.

Mueller, J. (2006). L2 in a nutshell: The investigation of second language processing in the miniature language model. *Language Learning, 56*(1), 235–270.

O'Grady, W. (2003). The radical middle: Nativism with Universal Grammar. In C. Doughty, & M. Long (Eds.), *The handbook of second language acquisition* (pp. 19–42). Oxford, UK: Blackwell.

O'Grady, W. (2010). Fundamental universals of language. *Lingua, 120*(12), 2707–2712.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition of relative clauses in Korean as a second language. *Studies in Second Language Acquisition, 25*(3), 433–448.

Osterhout, L., McLaughlin, J., Pitkanen, I., Frenk-Mestre, C., & Molinaro, N. (2006). Novice learners longitudinal designs, and event-related potentials: A paradigm for exploring the neurocognition of second-language processing. *Language Learning, 56*(1), 199–230.

Pienemann, M. (1998). *Language processing and second language development: processability theory.* Amsterdam: John Benjamins.

Pienemann, M. (2007). Processability theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 137–154). Mahwah, NJ: Lawrence Earlbaum Associates.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance.* Chicago, IL: The University of Chicago.

Rothman, J., & VanPatten, B. (2013). On multiplicity and mutual exclusivity: The case for different theories. In M. P. García Mayo, M. J. Gutierrez-Mangado, & M. Martínez Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 243–256). Amsterdam: John Benjamins.

Sharwood Smith, M., & Truscott, J. (2006). Full Transfer, Full Access: a Processing Oriented Interpretation. In S. Unsworth, T. Parodi, A. Sorace, & M. Young-Scholten, *Paths of Development in L1 and L2 acquisition* (pp. 201–206). Amsterdam: John Benjamins.

Slabakova, R. (Ed.). (2009). The Fundamental Difference Hypothesis twenty years later [Special Issue]. *Studies in Second Language Acquisition, 31*(2).

Slobin, D. (Ed.). (1985). *The crosslinguistic study of language acquisition Vol. 1: The data.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sorace, A. (2003). Near-nativeness. In M. Long & C. Doughty (Eds.), *Handbook of second-language acquisition* (pp. 130–152). Oxford, UK: Blackwell.

Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguistic Approaches to Bilingualism*, 1–33.

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research*, 339–368.

Sunderman, G. (2002). *A psycholinguistic investigation of second-language lexical processing* (Unpublished doctoral dissertation). The Pennsylvania State University, University Park, PA.

Talamas, A., Kroll, J., & DuFour, R. (1999). From form to meaning: Stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition, 2,* 45–58.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language, 33*, 285-318.

Truscott, J., & Sharwood Smith, M. (2004). Acquisition by processing: a modular perspective on language development. *Bilingualism: Language and Cognition, 7*(2), 1–20.

Truscott, J., & Sharwood Smith, M. (2011). Input, intake, and consciousness: The quest for a theoretical foundation. *Studies in Second Language Acquisition, 33*(4), 497–528.

VanPatten, B. (1983). *Processing strategies in second language acquisition* (Unpublished doctoral dissertation). University of Texas, Austin.

VanPatten, B. (1985). Communicative values and information processing in L2 acquisition. (P. Larson, E. L. Judd, D. S. Messerschmitt, & T. Scovel, Eds.) *On TESOL '84,* 89–99.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*(3), 287–301.

VanPatten, B. (1996). *Input processing and grammar instruction: Theory and research.* Norwood, NJ: Ablex.

VanPatten, B. (2004). Input processing in second language acquisition. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–31). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

VanPatten, B. (2009). Processing matters in input enhancement. In T. Piske & M. Young-Scholten (Eds.), *Input Matters* (pp. 57–61). Clevedon, UK: Multilingual Matters.

VanPatten, B., & Jegerski, J. (Eds.). (2010). *Research in second language processing and parsing.* Amsterdam, Netherlands: John Benjamins.

VanPatten, B., Leeser, M., & Keating, G. D. (2012). Missing verbal inflections as a representational problem: Evidence from self-paced reading. *Linguistic Approaches to Bilingualism, 2*(2), 109–140.

VanPatten, B., & Rothman, J. (in press). Against "rules." In A. Benati, C. Lavale & M.J. Arche.
(Eds.), *The grammar dimension in instructed second language acquisition: theory, research, and practice.* London: Continuum Press.

White, L. (2003). *Second language acquisition and Universal Grammar.* Cambridge, UK: Cambridge University Press.

2
SELF-PACED READING

Jill Jegerski

History of the Method

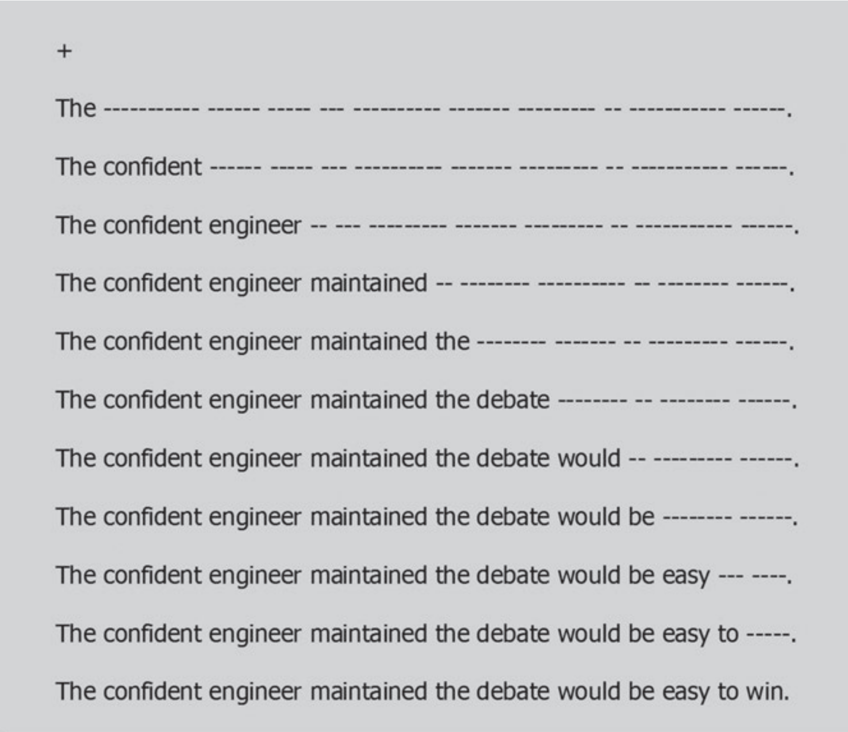
The self-paced reading (SPR) method was invented by psycholinguists in the 1970s (Aaronson & Scarborough, 1976; Mitchell & Green, 1978). SPR is so simple in design that it would be easy to assume today that it predates modern eye-tracking's appearance in reading research, but in reality the two methods in their more primitive forms appeared at around the same time, when a newfound access to computers fostered some of the most significant advances in mental chronometry since its development just over a century prior (Donders, 1868/1969, as cited in Baayen & Milin, 2010). These game-changing developments in methods for psycholinguistic research arose out of a desire in cognitive psychology to measure language comprehension processes in real time and with "tasks that are as similar as possible to normal reading" (Mitchell & Green, 1978, p. 610). Self-paced reading was the simplest way to meet these goals using modern technology and for this reason it flourished in popularity and has persisted over time—unlike many of its predecessors, including click migration (Fodor & Bever, 1965), the phoneme-monitoring task (Foss, 1970), and the sentence classification task (Forster & Olbrei, 1973). Nearly forty years after its development, SPR is still the most fundamental experimental measure employed by psycholinguists interested in processing at or above the level of the sentence. SPR was also the first on-line (i.e., real-time) method to be applied in non-native sentence processing research.

The first published investigation to apply the SPR method in the study of second language acquisition (SLA) was Juffs and Harrington (1995). The immediate theoretical motivation for the study was the debate among generative linguists in SLA as to whether observed differences between native speakers and adult second language (L2) learners were true differences in underlying grammatical competence, perhaps due to a lack of access to Universal Grammar after a critical period, or more superficial differences that arose due to the time constraints of real-time processing and were thus limited to performance. Proponents of the performance position had proposed, based on previous evidence collected via grammaticality judgments and global reaction times, that divergent behavior among non-natives could be due to processing difficulty rather than to the acquisition of a nontarget grammar (see, e.g., White & Juffs, 1998, for a more detailed discussion of the competence versus performance debate). Thus, the initial motivation for employing the SPR method in second language research was to measure linguistic performance in a way that complemented the grammaticality judgment as a measure of linguistic competence.

A decade later, when Clahsen and Felser (2006) reviewed the literature on second language processing in their development of the Shallow Structure Hypothesis, interest in psycholinguistic methods in SLA research had begun to take hold and over a dozen published studies using the self-paced reading method were available. Several of these, like Juffs and Harrington (1995), followed the generative linguistics tradition in SLA and focused on *wh*- movement (Juffs, 2005; Marinis, Roberts, Felser, & Clahsen, 2005; Williams, Möbius, & Kim, 2001) or clitics and causatives (Hoover & Dwivedi, 1998). Other researchers began to incorporate and adapt ambiguity and anomaly paradigms from the psycholinguistics tradition, such as relative clause attachment (Dussias, 2003; Felser, Roberts, Gross, & Marinis, 2003; Papadopoulou & Clahsen, 2003), subject-object ambiguity (Juffs, 1998a, 2004; Juffs & Harrington, 1996), verbal ambiguity (Juffs, 1998b), and broken agreement (Jiang, 2004). By 2009, the study of L2 processing had flourished and SPR was the single most popular on-line method among researchers at the first Conference on Second Language Processing and Parsing held at Texas Tech University, accounting for 37% of all research papers presented at the conference (see VanPatten & Jegerski, 2010, for select examples). At the same time, dozens of additional journal articles have reported self-paced reading studies of SLA, to the extent that they have become almost too numerous to be covered comprehensively and current literature reviews tend to need to be much more narrow in focus.

What is Looked at and Measured

SPR is a computerized method of recording a reading time for each designated segment (i.e., a word or phrase) of a sentence or series of sentences that is presented as an experimental stimulus. It is commonly referred to as *self-paced* and has also been called *subject-paced* because the research participant determines how long to spend reading each segment, which contrasts with fixed-pace methods like rapid serial visual presentation, or RSVP, where reading times are predetermined by the researcher. Specifically, in SPR, a button press causes the first segment of a sentence to appear together with a series of dashes masking the remainder of the stimulus, then when the participant is ready to continue a second button press reveals the next segment, then the next, and so on until the entire sentence has been read. Historically, self-paced reading has been a general term that includes several different formats. First, the display can be *cumulative,* meaning once a stimulus segment is revealed it remains visible to the participant as the next segment is revealed and the next and so on, until the entire sentence is finally displayed all together (as illustrated in Figure 2.1), or *noncumulative,* meaning only one segment is visible at a time and every time a new segment is revealed the previous one is remasked (as illustrated in Figure 2.2). Additionally, the display can be *centered,* meaning that every segment appears in the center of the display screen and overwrites the previous segment (as seen in Figure 2.3), or *linear,* meaning that segments appear in linear succession from left to right with no spatial overlap, much as they would



+

The ---------- ------ ----- --- --------- ------- -------- -- ---------- ------.

The confident ------ ----- --- --------- ------ -------- -- ---------- ------.

The confident engineer -- --- -------- ------- -------- -- ---------- ------.

The confident engineer maintained -- -------- --------- -- -------- ------.

The confident engineer maintained the -------- ------- -- --------- ------.

The confident engineer maintained the debate -------- -- -------- ------.

The confident engineer maintained the debate would -- --------- ------.

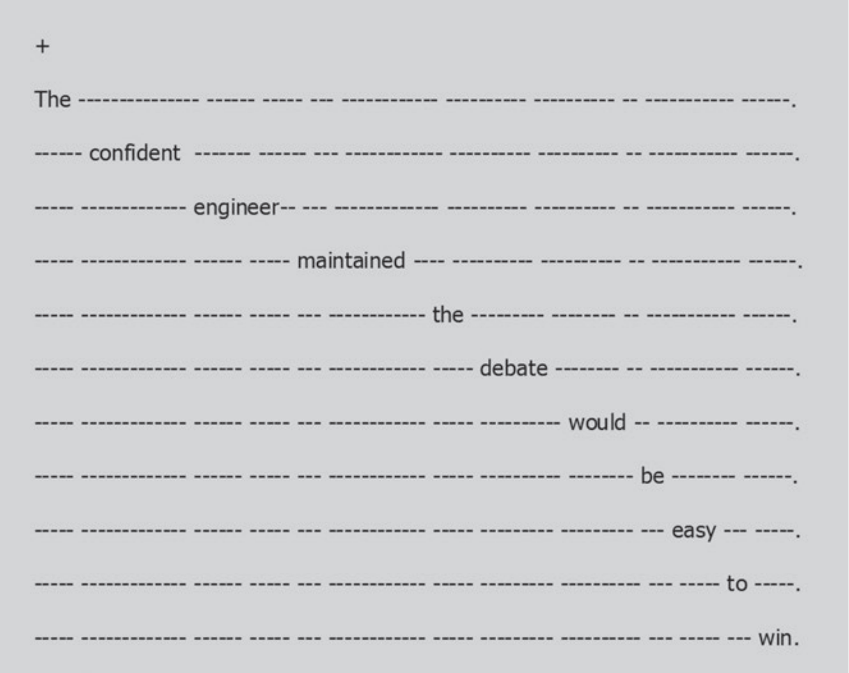
The confident engineer maintained the debate would be -------- ------.

The confident engineer maintained the debate would be easy --- ----.

The confident engineer maintained the debate would be easy to -----.

The confident engineer maintained the debate would be easy to win.

FIGURE 2.1 Illustration of self-paced reading, cumulative linear format with word-by-word segmentation. Each line of text shown above would be vertically centered in a separate display on a computer monitor. The participant presses a button to move through successive displays and computer software records the time between button presses, which is the primary dependent variable. A gray background is used because it is easier on the eyes than a white background.

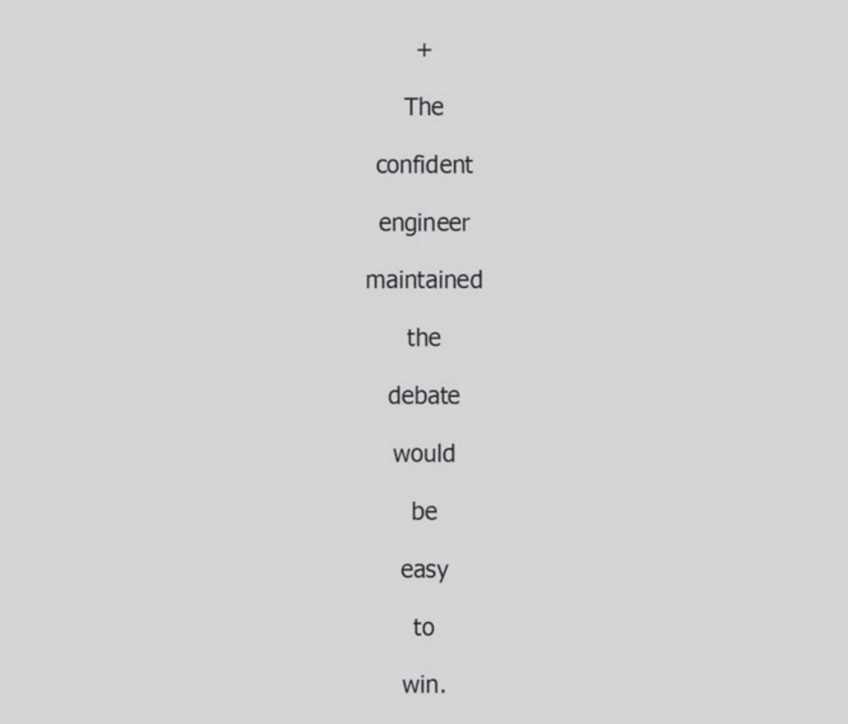


```
+

The ------------- ------ ----- --- ----------- --------- --------- -- ---------- ------.

------ confident  ------ ------ --- ----------- --------- --------- -- ---------- ------.

----- ------------ engineer-- --- ------------ --------- --------- -- ---------- ------.

----- ------------ ------ ----- maintained ---- --------- --------- -- ---------- ------.

----- ------------ ------ ----- --- ----------- the -------- -------- -- ---------- ------.

----- ------------ ------ ----- --- ----------- ----- debate -------- -- ---------- ------.

----- ------------ ------ ----- --- ----------- ----- --------- would -- ---------- ------.

----- ------------ ------ ----- --- ----------- ----- --------- -------- be -------- ------.

----- ------------ ------ ----- --- ----------- ----- -------- --------- --- easy --- -----.

----- ------------ ------ ----- --- ----------- ----- -------- --------- --- ----- to -----.

----- ------------ ------ ----- --- ----------- ----- -------- --------- --- ----- --- win.
```

FIGURE 2.2 Illustration of self-paced reading, non-cumulative linear format with word-by-word segmentation. Also known as the moving window format, this is the most common type of self-paced reading.

in normal reading. However, the cumulative display is problematic because most participants develop a reading strategy in which they reveal several segments of a stimulus at a time before reading them all at once (Ferreira & Henderson, 1990; Just, Carpenter, & Wooley, 1982) and the centered display is avoided with SPR because it is less like normal reading—though it can be necessary with some other methods, like ERPs (for further information regarding the ERP method, see Morgan-Short & Tanner, Chapter 6, this volume). For these reasons, virtually all SPR studies now elect for a noncumulative linear display, which is also referred to as the *moving window(s)*[1] technique because successive button presses cause the unmasked segment of text to proceed like a moving window across the computer screen.

The basic premise behind self-paced reading is that the eyes can be a window on cognition. Just and Carpenter (1980) proposed the eye-mind assumption, which states that the amount of time taken to read a word reflects the amount of time needed to process the word. While subsequent research has revealed that the connection between reading times and processing is in reality more complex, the basic assumption still holds in the broad sense and reading time data, as a specific class of reaction times (i.e., response times or response latencies), are interpreted with



+

The

confident

engineer

maintained

the

debate

would

be

easy

to

win.

FIGURE 2.3 Illustration of self-paced reading, centered format with word-by-word segmentation.

the goal of drawing inferences about the cognitive processing of language. Specifically, relatively longer reading times are taken as indications of processing difficulty, while faster reading times are interpreted as a sign that facilitation occurred.

Most SPR paradigms examine processing difficulties that arise during the reading of sentences that contain what could be classified as an ambiguity, an anomaly, or a distance dependency. Ambiguities arise where the grammar permits two or more distinct syntactic interpretations of a word or phrase in the sentence and observable processing strategy often occurs when the (native) parser tends towards one interpretation over the other. Such structural ambiguity can be either local, meaning it occurs temporarily during reading but is resolved within the same sentence, or global, meaning that even after the whole sentence has been read the ambiguity remains. Examples of local syntactic ambiguities include subject-object ambiguity (L1: Trueswell & Kim, 1998; L2: Juffs & Harrington, 1996) and reduced relative clause ambiguity (L1: MacDonald, 1994; L2: Juffs, 1998a). Local or temporary ambiguities are also referred to as *garden path* phenomena because such sentences are designed to initially lead the reader in the wrong direction with regard to the structure of the sentence. Garden path effects are evident in increased SPR times at or after the point in the sentence where it becomes evident to the reader that the initial interpretation was incorrect. In the example (1) below, taken from Trueswell and Kim (1998), longer reading times were observed on the embedded verb *would be* in the ambiguous version in (1a) versus in the unambiguous version in (1b).

(1) *Subject-Object Ambiguity*

a. The confident engineer maintained the debate would be easy to win.
b. The confident engineer maintained that the debate would be easy to win.

With global ambiguities, such as the attachment of ambiguous relative clauses (L1: Cuetos & Mitchell, 1988; L2: Dussias, 2003) or prepositional phrases (L1: Taraban & McClelland, 1988; L2: Pan & Felser, 2011), on the other hand, SPR effects are usually evident around the point where disambiguated versions of the stimuli become inconsistent with participants' preferred interpretation. For instance, in the example (2) below, a stimulus item from Dussias (2003), native speakers of Spanish exhibited longer reading times on the sentence-final phrase *con su esposo* "with her husband" in the forced low-attachment version in (2a) versus the forced high-attachment version in (2b), because Spanish in general tends toward high attachment for ambiguous relative clauses. Such disambiguation can be accomplished using pragmatic-contextual information, as in this example, or with grammatical dependencies like gender or number agreement, though each of these adds a layer of complexity in processing and thus has the potential to obscure reading time effects caused by the experimental manipulation, especially among non-native readers. Stimuli with local or global syntactic ambiguities are intended to present processing difficulty in the form of forced syntactic reanalysis at the point of disambiguation, but they remain grammatical in all versions.

(2) *Relative Clause Ambiguity (Disambiguated)*

a. El perro mordió al cuñado de la maestra que vivió en Chile con su esposo.

"The dog bit the brother-in-law of the (female) teacher who lived in Chile with her husband."

b. El perro mordió a la cuñada del maestro que vivió en Chile con su esposo.

"The dog bit the sister-in-law of the (male) teacher who lived in Chile with her husband."

A second class of processing phenomena targeted with SPR are anomalies, which include specific violations of grammar (i.e., error recognition or grammaticality paradigms) as well as inconsistent or noncanonical permutations of word order, semantics, discourse, and other syntactic and extrasyntactic factors that are presented in the experimental stimuli. Numerous different anomaly paradigms are available to researchers in psycholinguistics, but some examples of specific phenomena that have been examined among non-natives using SPR include gender agreement (Sagarra & Herschensohn, 2011), number agreement (Foote, 2011), tense/aspect agreement (Roberts, 2009), and case marking (Jegerski, 2012a). Stimuli with violations of the linguistic principles guiding such phenomena commonly induce longer reading times at or after the point of the violation, presumably because the parser has difficulty incorporating a word that does not fit with the existing representation of the sentence. For instance, in sentences like (3) below, a stimulus from Foote (2011), reading times on the postverbal word *de* "from" were longer in the ungrammatical stimulus condition in (3b) versus the grammatical version in (3a), as can be seen in the reading times in Table 2.1, which are also graphed in Figure 2.4 (along with hypothetical data for the other sentence regions not in Table 2.1).

(3) *Number Agreement Violation*

a. Veo que tu padre es de Texas.
b. *Veo que tu padre son de Texas.

"I see that your father is/*are from Texas."

A third type of sentence-level phenomenon that can be exploited to study processing behavior with SPR is the distance dependency. Dependency paradigms examine the computation or recognition of a syntactic relationship between two elements in the stimulus that are usually nonadjacent in the lin ear word order, which presents a particular challenge in processing. Examples include *wh-* movement (L1: Crain & Fodor, 1985; L2: Williams, Möbius, & Kim, 2001) and broken agreement (L1: Nicol, Forster, & Veres, 1997; L2: Jiang, 2004).

TABLE 2.1 Reading times in milliseconds from an SPR experiment using a grammar violation paradigm (Adapted from Foote, 2011)

Veo que tu	*padre*		*es/son*		*de . . .*	
"I see that your	father		is/are		from . . ."	
	Mean	SD	Mean	SD	Mean	SD
L1 Group						
Ungrammatical	389	66	452	94	488	109
Grammatical	414	86	409	52	403	75
Difference	–25		43		85	
L2 Group						
Ungrammatical	360	29	408	43	539	83
Grammatical	373	43	394	47	379	26
Difference	–13		14		160	

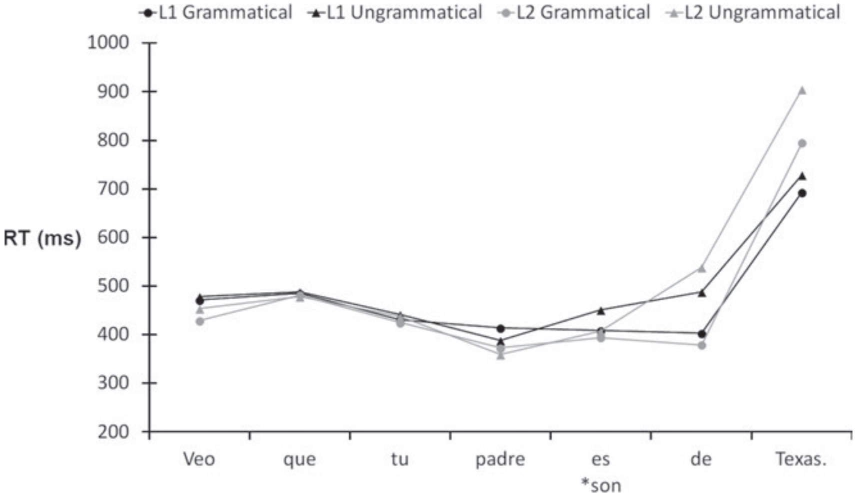


FIGURE 2.4 Example line graph of hypothetical reading time data from a SPR experiment using a grammar violation paradigm (based on a stimulus and partial data from Foote, 2011).

To illustrate, the examples of broken agreement in (4) below, taken from Jiang (2004), require the parser to compute subject-verb number agreement across several intervening words, the number feature of which can interfere with the computation. Thus, relatively longer reading times were observed on the verb in the Singular-Plural-Singular condition in (4a) versus in the Singular-Singular-Singular condition in (4b), presumably because of interference from the plural noun *cabinets* in (4a). As with syntactic ambiguities, distance dependencies can induce SPR effects without the introduction of grammar violations or other anomalies, manipulating instead the words and phrases that intervene between the two dependent elements.

(4) *Broken Agreement*

a. The key to the cabinets was kept in the main office.
b. The key to the cabinet was kept in the main office.

Over a period of nearly thirty-five years, all three types of SPR paradigms have been employed to study fundamental questions in native language sentence processing such as whether the parser considers multiple plausible analyses simultaneously or sequentially, whether all types or modules of linguistic information are immediately available or only syntax is active at first, what heuristics motivate different processing preferences, and to what extent these basic principles vary cross-linguistically, among others. Non-native sentence processing research is a relatively newer area of study that can be uniquely informative with regard to these pre-existing broad questions in psycholinguistics, and which has also begun to articulate its own research agenda within the field of second language study. SPR investigations have focused on the issue of learnability and age effects in processing, on the closely related debate as to whether divergence in adult SLA is rooted in competence or performance, and on the question of L1 transfer in processing, so far with relatively less attention dedicated to other L2 questions like mapping the developmental trajectory of non-native processing behavior. Thus, in most cases the SPR method has been employed to measure linguistic skill and knowledge for the purpose of making comparisons, either between native and non-native processing in the L2, between native processing in the L1 and non-native processing in the L2, or between the L2 processing behaviors of participant groups with different native languages.

Comparison on the basis of SPR data can be designed and interpreted from at least two different perspectives. First, because grammatical processing relies on existing knowledge of grammar that is stored in memory, the SPR method in L2 research was first viewed as complementary to previously established measures like grammaticality and acceptability judgments. From this angle, SPR data can be seen as an indirect measure of grammatical competence and are often regarded as a relatively more direct or more implicit measure of grammar than off-line judgments because the time constraints of on-line processing presumably allow less room for the application of explicit grammar rules. The most common SPR paradigms employed in this vein of research target grammar violations or anomalies and distance dependencies, both of which can be linked to the formal linguistics traditions of grammaticality judgments with relative ease. Sensitivity to an experimental manipulation of grammar, in the form of increased reading times at or near the site of a violation, is interpreted as evidence that the relevant underlying grammatical competence has been acquired. This is of course assuming that such sensitivity is also evident among a comparison group of native speakers and can therefore be reasonably expected, given that even violation-based reading time effects—which tend to be more robust and more reliable than those that occur with dependencies or ambiguities—can sometimes be inconsistent among native speakers.

Second, the SPR method can be used as a measure of performance or processing behavior itself, a perspective that is becoming dominant as the study of L2 processing expands and the body of existing published research grows. A variety of reading time effects are targeted in this line of investigation, which includes ambiguities as well as distance dependencies and anomalies. The interpretation of data can be considerably less straightforward than when SPR is employed as an indirect measure of grammatical competence, especially when the method is used to compare native and non-native processing. To illustrate, data interpretation is fairly straightforward when a group of native readers exhibits a reading time effect that is not at all evident among a group of non-native participants, as most researchers would agree that such an outcome indicates a difference between native and non-native processing. There are occasions, however, where a group of native readers shows an SPR effect that is even more pronounced among the non-native readers, meaning that the effect is sustained over more than one region of interest or it surfaces again during sentence wrap-up or while answering a poststimulus distractor question. Particularly if the SPR effect in question is presumed to signal syntactic reanalysis, a more pronounced effect among non-native readers could be interpreted as a sign of additional processing difficulty rather than native-like processing skill. Another experimental outcome that can be subject to multiple interpretations is when non-native participants display a reading time effect that occurs a region or two later than that exhibited by the native readers, or perhaps does not surface until wrap-up occurs at the last region of the stimulus. In both of these scenarios, there is some room for debate as to whether the observed differences between native and non-native processing are critical, meaning whether they represent qualitative or merely quantitative differences.

In general, reading time data from SPR experiments are more nuanced and thus tend to demand more complex interpretation than data from off-line measures like grammaticality judgments. In addition, the use of SPR to measure target L2 behavior is also by nature paradoxical, because most known L1 reading time effects are assumed to indicate some type of processing difficulty. In other words, we are investigating whether non-native readers have learned to have the same problems that native readers have while processing a given type of stimulus. In some cases, the interpretation of L2 SPR data can be relatively straightforward, but it is not always clear whether increased reading times among L2 learners reflect the target native-like processing difficulty induced by experimental manipulation of the stimuli or a different type of difficulty that has to do with the limitations of L2 processing. In the former scenario, increased SPR reading times would be interpreted as evidence of native-like processing *strategy,* whereas in the latter they would be taken as evidence of an L2-specific processing *struggle.*

Issues in the Development and Presentation of Stimuli

The creation of an SPR experiment entails the embedding of each stimulus within a trial, or experimental series of related events, which usually consists of three components: a cue, a stimulus, and a distractor. The cue phase is fairly straightforward and is the same for all trials; a "+"or similar symbol is displayed in isolation in the same screen location where the first letter of the first word in the stimulus will subsequently appear. Hence, the cue appears towards the left side of the display for a language that is read left-to-right and towards the right side for a language that is read right-to-left. The purpose of the cue is to encourage participants to direct their gaze at the location of the first word of the stimulus before it appears. Otherwise, the reading time for the first region of interest of a given stimulus may or may not include time spent initially dwelling on another screen location plus the time spent to bring the gaze to the location of the beginning of the stimulus. This would make SPR less precise as an experimental measure because, at a minimum, reading times for the sentence-initial region would be artificially but uniformly inflated, while a more likely scenario is that reading times would be affected inconsistently and therefore display unnecessarily high levels of variance.

After the cue comes the stimulus. SPR stimuli are usually one sentence in length, given that the method is most often used to measure sentence-level comprehension behavior, although discourse-level phenomena can be studied through the addition of one or more sentences that establish a context prior to the appearance of the target sentence. In either case, the development of stimuli entails the creation of a list of experimental sentences or *items,* which are directed at the research questions guiding the investigation. Within each experimental item, there are multiple versions (usually two to four) referred to as *conditions,* which correspond to the researcher's manipulation of independent linguistic variables and are thus determined by the experimental design. In order to maintain control between stimulus conditions, the corresponding regions to be directly compared in the statistical analysis should be as near to identical as possible, given the constraints of the particular experiment, as in (4) above. If the experimental manipulation of linguistic variables necessitates the use of different lexical items in different stimulus conditions (e.g., if the manipulation involves lexical specifications like verb transitivity or semantic properties), then these should be counterbalanced across stimulus conditions, minimally for objective variables like length in characters and syllables, frequency, and similarity to the equivalent L1 lexical item (including cognate status), and then verified via statistical comparisons, because such variables are known to affect reading speed. Depending on the type of stimuli and the objectives of the experiment, it may also be necessary to counterbalance the different lexical items (or phrases) to be directly compared with regard to subjective variables such as imageability, comprehensibility, concreteness, or semantic properties. As such variables are not easily measured via objective means, they are measured instead via a norming procedure, in which a group of research subjects (native speakers) are asked to rate a list of words or phrases with regard to one or more characteristics. The ratings are then compiled and analyzed and used to create counterbalanced experimental stimuli (see Sunderman, Chapter 8, this volume, for a more detailed discussion of norming procedures).

Once the experimental sentences have been created, they are broken down into regions of interest, which participants will read one-by-one and will each correspond to a separate data point in the form of a reading time in milliseconds (ms). The researcher chooses whether to use word-by-word or phrase-by-phrase segmentation, both of which are illustrated in (5) and (6) below with examples from Juffs (1998b) and Pliatsikas and Marinis (2013), respectively. The decision between the two types of segmentation is typically a compromise between the conflicting goals of maximizing the level of detail in the reading time data and maximizing the ecological validity of the experimental task. In other words, a word-by-word segmentation yields more precise data because more data points are collected per stimulus, whereas a phrase-by-phrase segmentation is closer to normal reading and may therefore eliminate some unnatural effects induced by the SPR task itself, such as a tendency towards highly incremental processing. On the other hand, data collected in a word-by-word fashion can be easily converted to phrase-by-phrase mode by summing reading times across multiple words/ regions, but once an experiment has been run in the phrase-by-phrase mode, there is no way to break the data down into word-by-word reading times without rerunning the experiment. The phrase-by-phrase mode also has the added complication of potentially influencing processing behavior through the particular grouping of words into phrases and especially the length of the phrases (Gilboy & Sopena, 1996). One way to determine the effect of stimulus segmentation on a given set of materials would be to run a pilot experiment testing two or more segmentations. Both the word-by-word and the phrase-by-phrase segmentation are common in the literature and with both modes the number of total regions per stimulus depends on the length of the stimuli, but should be exactly the same for all items in a given experiment.

(5) *Word-by-word segmentation*

Before / Mary / ate / the / pizza / arrived / from / the / local / restaurant.

(6) *Phrase-by-phrase segmentation*

The manager who / the secretary claimed / that / the new salesman / had pleased / will raise company salaries.

Within regions of interest, be they words or phrases, it is important that there be grammatical equivalency across experimental items, such that if Region 1 is a subject noun phrase (NP) in Item 1, then it should also be a subject NP for Item 2, Item 3, and all the other stimulus items. Each region of interest should also be roughly equivalent in length across different stimulus items (this is not to be confused with the control of regions across conditions of a single item, which should be identical except for the critical region).

The number of stimuli per condition is usually eight to twelve, which means that the number of sentences created for an experiment can range from 16 (2 conditions × 8 stimuli per condition) to 48 (4 conditions × 12 stimuli per condition). Because these target stimuli represent only 25 to 35% of the total experiment and even non-natives at a very high level of L2 proficiency are not commonly asked to read more than 150 to 200 sentences maximum per research session, individual SPR experiments rarely include more than 48 target stimuli. If the research questions driving a given investigation necessitate the inclusion of so many variables that the total number of target stimuli would be higher, then the variables can be broken down into separate, smaller experiments. Such simplicity has the added benefits of avoiding complexity in the statistical analyses that can render them largely uninterpretable and also keeping the required number of participants to a reasonable number, as will be explained further below.

In addition to the experimental stimuli created through the manipulation of linguistic variables, the other 65 to 75% of the sentences in an SPR experiment are not related to the research questions. There is no single consensus in the literature with regard to the ideal ratio of target versus total nontarget stimuli for an SLA experiment, but there is some evidence that a very low proportion of these might affect reading behavior during SPR and that 50% nontarget sentences is the minimum acceptable amount (Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009). Up to half of the nontarget sentences may be distractors, which are created through the manipulation of entirely different linguistic variables and may serve either to balance the target stimuli (e.g., unambiguous distractors to complement similar but ambiguous target sentences) or as target stimuli for another experiment. Whether or not the experiment includes distractors, the remaining sentences are fillers, or unrelated sentences with no specific linguistic target. All target stimuli, distractors, and fillers are created to be comparable with regard to length and other superficial characteristics so that participants cannot easily identify the target sentences. For the same reason, manipulated variables that could potentially be recognized by research participants, such as grammar violations, should also be balanced across target and nontarget stimuli. In other words, if half of the target stimuli are ungrammatical, then half of the distractors and/or fillers should also be ungrammatical, again to avoid making target stimuli stand out. Most experiments present written instructions and eight to ten additional fillers as practice items so that participants can become familiar with the SPR procedure and the distractor questions that usually follow each of the stimuli before any potentially meaningful data are recorded.

Once all of the stimuli have been finalized, the presentation lists are created. In an ideal world, there would be only one list because all participants would read all stimulus items in all conditions, but in reality this could cause any number of undesirable presentation effects, including priming and ordering effects, as well as increasing the likelihood that participants would become consciously aware of the linguistic target of the experiment. To prevent such complications, each participant reads each stimulus item only once in one of its conditions, but still reads an equal number of target stimuli in each of the conditions. In order to have each stimulus read in all of its conditions, multiple counterbalanced presentation lists are created so that one subgroup of participants reads a stimulus item in the first condition, another group reads it in the second condition, and so on. Counterbalancing here means that each participant contributes an equal number of data points to each level of a variable (e.g., by reading eight items in each of four conditions), because there may be individual differences in reading speed or other characteristics among participants. To illustrate, if there are four stimulus conditions called a, b, c, and d, and there are 32 stimulus items

TABLE 2.2 Illustration of counterbalancing the first 8 of 32 stimuli in an experiment with four conditions across four presentation lists

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Item 5*	*Item 6*	*Item 7*	*Item 8*	. . .
List I	1a	2b	3c	4d	5a	6b	7c	8d	. . .
List II	1b	2c	3d	4a	5b	6c	7d	8a	. . .
List III	1c	2d	3a	4b	5c	6d	7a	8b	. . .
List IV	1d	2a	3b	4c	5d	6a	7b	8c	. . .

numbered 1 to 32 (note that the total number of target stimuli is always a multiple of the number of conditions), then there would be four presentations lists, as illustrated in Table 2.2.

Looking at just the first eight items in Table 2.2 and ignoring for now the other 24, if the first four participants each read one of the four lists, then they would each contribute two data points in each stimulus condition and the variable levels would be counterbalanced with regard to any differences in the individual subjects. That is, that participant reading List I would yield two reading time data points in Condition a, two in Condition b, two in Condition c, and two in Condition d, so individual variations such as reading speed should not affect the outcome of the experiment. There can be individual differences among stimulus items as well, so the same type of counterbalancing applies to the stimuli. In the above example, the variable levels would be counterbalanced with regard to items as long as the number of participants is a multiple of four. If there are four participants, as above, then Item 1 contributes one data point to each condition. If there are eight participants, two would read each of the four presentation lists and Item 1 would contribute two data points to each condition, and so on. The ideal experiment would have both types of counterbalancing, both by subjects and by items. In reality, the number of items is always a multiple of the number of stimulus conditions, but at the present time it is actually quite common in the literature to see numbers of participants (within each group) that are not exact multiples of the number of stimulus conditions because planning the number of items is usually more feasible than controlling the exact number of subjects that yield usable data, especially with L2 participants. As long as the presentation lists are rotated continually through each participant group (as opposed to using List I for the first ten participants, then List II for the next ten, etc.), then the final number of data points contributed by each item to each condition will not vary by more than one.

Of course, the stimuli are not presented in numerical order and are not presented in the same order to all participants. For the ordering of stimuli within each presentation list, pseudorandomization is the preferred technique. Total randomization would be ideal as far as minimizing any effects of presentation order, especially because the order could be unique for each participant instead of only for each list, but a problem with this method is that several experimental sentences would sometimes appear one right after the other (and in theory they could even all appear together), which can lead to priming or can draw participants' attention. Unfortunately, most experimental software can be set to automatically randomize stimuli, but not to pseudorandomize with specified limitations. Thus, to prevent target stimuli from being clumped together, a limited number of randomizations are usually created and then corrected so that no two similar sentences appear in succession. Randomized lists of numbers can be generated using, for example, the RANDBETWEEN function in Excel or the online Research Randomizer (Urbaniak & Plous, 2011).

SPR studies typically involve some type of distractor task, in the form of questions that follow some or all of the stimuli, which serves in the first place to ensure that cognitive processes are engaged throughout and the participant does not end up pressing buttons without paying attention to the experimental stimuli on display. In addition, a well-designed distractor task can also prevent the participant from consciously reflecting on the primary SPR task and altering their behavior accordingly. The selection of a distractor task for an experiment should be intentional, as the type of task has been shown to affect sentence processing behavior (Aaronson & Scarborough, 1976; Havik et al., 2009; Leeser, Brandl, & Weissglass, 2011; but cf. Jackson & Bobb, 2009). Most SPR distractor tasks fall into one of two categories: acceptability judgments or comprehension questions, but given the critical distinction in SLA between acquired implicit linguistic knowledge and learned explicit information (Ellis, 2007; Krashen, 1981) and the ease with which most adult second language learners are influenced by explicit rules, it seems highly desirable in most cases to opt for meaning-based comprehension questions. After all, SPR behavior is only informative and generalizable to the extent that it reflects the same cognitive processes that are engaged during normal reading and during language comprehension in general. Furthermore, the use of metalinguistic distractor tasks can affect processing strategy, even causing certain reading time effects. For example, Leeser et al. (2011) found that intermediate L2 learners of Spanish showed on-line sensitivity to gender agreement violations during self-paced reading when the distractor task was a grammaticality judgment, but not when it was a meaningful comprehension question. Regardless of the type of distractor task, the questions or judgments tend to be binary choice items that are counterbalanced with regard to the number of correct "a" versus "b" or "acceptable" versus "unacceptable" answers.

If the researcher confirms that meaningful comprehension questions are indeed the best distractor task, but is also interested in collecting complementary data via an off-line measure like an acceptability judgment, then this task should be administered independently and not prior to the self-paced reading task, in order to avoid unnatural cross-contamination effects like priming on the reading time results. The two tasks should be constructed using different items and should be separated by additional activities, as priming is also known to affect the results of acceptability judgments (Luka & Barsalou, 2005).

An additional point of variation in the literature is with regard to how often distractor questions appear, meaning after every stimulus or randomly after only a fraction of the stimuli, such as one in four. Either method would be sufficient where the only purpose of the distractor question is that mentioned first above: to engage participants in the SPR task while simultaneously diverting their conscious attention from it. However, the comprehension questions that follow experimental stimuli are receiving increasingly more attention in L2 sentence processing research as potential loci of delayed processing effects. For instance, Roberts and Felser (2011) and Jegerski (2012b) reported and analyzed the accuracy rates and reaction times for responses to comprehension questions that appeared after every stimulus in their investigations of subject-object ambiguities. Like other researchers, they interpreted lower accuracy rates and longer reaction times as indications of processing difficulty that was delayed or spilled over from the stimulus that immediately preceded a comprehension question. Given the potential for comprehension question response and reaction time data to provide additional insight into the time-course of sentence processing, it seems that in most cases it would be advantageous to present distractor questions after every single stimulus rather than only a fraction.

Scoring, Data Analysis, and Reporting Results

Raw data collected via the self-paced reading method include reaction times in ms as well as qualitative responses for every event in the experiment that allowed input from the participant. For example, each region of the experimental sentences yields a numerical reading time plus a categorical record of the button that was pressed to advance to the next display (usually a space bar on the keyboard or a green key on a button box), so a single sentential stimulus will easily have ten or more data points associated with it, depending on how long it is and what type of segmentation was used. A distractor question also has a corresponding reaction time plus a record of which button was pressed to answer (typically one of two buttons designated on the keyboard or a button box for responding to binary choice questions). These reaction time and button press data are compiled and stored as one output file per participant by the experimental software, so there is no true scoring or coding to be conducted manually with SPR. It is usually necessary, however, to compile and sort the data before performing statistical analyses.

Data output files group data by trial (i.e., stimulus), with each data point corresponding to one row in a list or table, and all trials are listed in the order in which they were presented during the experiment. Thus, in a raw data file target items are intermingled with practice items and fillers, numerical reaction times are mixed with categorical distractor task responses, and all the different events or steps within the trial—the initial cue symbol "+" or similar, Region 1, Region 2,

TABLE 2.3 Excerpt of an unsorted SPR data output file

List	*Subject*	*Trial*	*Event*	*Response*	*RT*
List B	9121	Instructions	Instructions	GREEN KEY	29258
List B	9121	Practice 1	[SubEvent 1]	GREEN KEY	710
List B	9121	Practice 1	[SubEvent 2]	GREEN KEY	804
List B	9121	Practice 1	[SubEvent 3]	GREEN KEY	930
List B	9121	Practice 1	[SubEvent 4]	GREEN KEY	905
List B	9121	Practice 1	Practice Ques 1	A KEY	3624
List B	9121	Practice 2	[SubEvent 1]	GREEN KEY	761
List B	9121	Practice 2	[SubEvent 2]	GREEN KEY	1700
List B	9121	Practice 2	[SubEvent 3]	GREEN KEY	1357
List B	9121	Practice 2	[SubEvent 4]	GREEN KEY	1218
List B	9121	Practice 2	Practice Ques 2	B KEY	6465
List B	9121	11a	[SubEvent 1]	GREEN KEY	1175
List B	9121	11a	[SubEvent 2]	GREEN KEY	882
List B	9121	11a	[SubEvent 3]	GREEN KEY	604
List B	9121	11a	[SubEvent 4]	GREEN KEY	642
List B	9121	11a	Comp Ques	A KEY	4780
List B	9121	58filler	[SubEvent 1]	GREEN KEY	729
List B	9121	58filler	[SubEvent 2]	GREEN KEY	970
List B	9121	58filler	[SubEvent 3]	GREEN KEY	625
List B	9121	58filler	[SubEvent 4]	GREEN KEY	614
List B	9121	58filler	Comp Ques	B KEY	4007
List B	9121	25b	[SubEvent 1]	GREEN KEY	681
List B	9121	25b	[SubEvent 2]	GREEN KEY	636
List B	9121	25b	[SubEvent 3]	GREEN KEY	1103
List B	9121	25b	[SubEvent 4]	GREEN KEY	929
List B	9121	25b	Comp Ques	A KEY	4646
List B	9121	45filler	[SubEvent 1]	GREEN KEY	851
List B	9121	45filler	[SubEvent 2]	GREEN KEY	730
List B	9121	45filler	[SubEvent 3]	GREEN KEY	986
List B	9121	45filler	[SubEvent 4]	GREEN KEY	1409
List B	9121	45filler	Comp Ques	B KEY	4140

Note: This part of the data file represents only the instructions presented at the beginning of the experiment, two practice items, two experimental items, and two fillers, so a complete data file would be much longer. Each "SubEvent" is a region of interest for the stimulus.

Region 3, any subsequent stimulus regions, and the distractor question—are listed together and also in order of appearance, as seen in Table 2.3. Because each stimulus presentation list includes a different randomization of the stimuli as well as different conditions of each stimulus, the data files will initially look different for each of the presentation lists. The raw data output files are commonly in the .txt format and can easily be opened in Excel or a similar spreadsheet program, where the sorting, linking, and macros features greatly facilitate the task of preparing data for analysis. The statistical software packages SPSS and R (R Development Core

Team, 2005) both have additional features and capacity for large data sets. The initial steps in preparing data for analysis typically include compiling all of the data files into a single master file that identifies individual participants by number and specifies values for any grouping variables like nativeness or L2 proficiency level, and separating out the experimental items from the practice, distractor , and filler items, which in some cases may be analyzed separately or even included in measures of individual participant characteristics like average reading speed or overall comprehension accuracy (see Jegerski, 2012b, and Roberts & Felser, 2011, for examples of how all items have been included in posthoc measures of individual reading speed). A single master data file can be modified in order to conduct different types of statistical analyses, though it can be helpful to create a separate file or spreadsheet within a workbook for each part of the data that will undergo independent statistical analyses: one for each stimulus region with reading times, one for the distractor question with reading times, and one for the distractor question with categorical response data.

Each of these separate data sets would contain, for example, 3200 rows of data in an experiment with 40 target stimuli and 80 participants, and will be treated independently from this point forward for the purposes of statistical analyses. If the statistical analyses are to be linear mixed-effects models in R (R Development Core Team, 2005; see Baayen, Davidson, & Bates, 2008, for a motivation for the recent trend in psycholinguistics towards mixed-effects models and away from traditional parametric statistics), then there is no need to compute aggregate means or to conduct separate items analysis, and data trimming is either very minimal or entirely unnecessary (see Baayen & Milin, 2010, for further discussion of data trimming with mixed-effects models). If the more traditional ANOVAs and *t*-tests are to be conducted, then the data first need to be trimmed and converted to aggregate means, each step by both subject as well as item (see Clark, 1973, for the motivation behind items analysis in psycholinguistics). With both types of analyses, linear mixed-effects models or ANOVAs, it is possible to clean the data by removing data points from trials that ended in inaccurate distractor question responses and to transform the data, both of which are described in greater detail below.

Regardless of whether the statistical analyses are traditional parametric statistics or the newer mixed-effects models, it is currently common in both L1 and L2 SPR studies to eliminate and ignore reading time data that correspond to incorrect distractor question responses, under the assumption that inaccuracy is an indication that the participant may not have been paying attention while reading the experimental sentence. This convention is most clearly justified in the study of native language processing, where errors tend to be infrequent and the process by which readers arrive at inaccurate responses to comprehension questions is usually not of interest. It may also make sense to follow suit in research on very advanced near-natives, because their error rates are usually low as well. But particularly in those investigations where participants are not at the highest levels of proficiency, data from trials with inaccurate distractor question responses can be quite numerous (10 to 20% or more of trials) and may prove to be enlightening with regard to the mechanisms of (and obstacles to) development of L2 processing strategy, especially when analyzed independently. For example, Juffs and Harrington (1996) included data from incorrect grammaticality judgments in their report, conducting separate analyses of the data corresponding to accurate and inaccurate responses. Based on the two separate analyses, the researchers concluded that the ESL learners in their study were more likely to make an accurate grammaticality judgment for those trials in which they had dwelled longer on a critical region while reading the sentence, an interesting observation that could not have been deduced solely based on the data from accurate responses. It is also common to examine and analyze data from trials with incorrect responses in the study of aphasia, at least when such data are numerous enough to be informative (see Caplan & Waters, 2003, for an example from a self-paced listening study, or Dickey, Choy, & Thompson, 2007, for an example with visual world eye-tracking). Data from inaccurate trials have thus far been mostly ignored in SLA research, however, so at the present time there is very limited empirical evidence of how this type of data cleansing may affect experimental outcomes.

After the data from inaccurate trials are separated out and either discarded or reserved for independent analysis, the remaining steps in preparing the data for parametric tests like ANOVAs and *t*-tests are to trim the reaction time data by subject and by item and to compute aggregate means, also by subject and by item, for both numerical reaction time data and categorical response data from the distractor task. Both steps can be conducted without changing the format of the master data spreadsheet files, but it may be helpful for those new to items analysis to at least imagine the data in two different spreadsheet layouts, one by subject and one by item, in order to better understand the two types of analysis. For organization by subject, illustrated with hypothetical reading time data in Table 2.4, the data are arranged such that participants are listed down the left most column and each one corresponds to a row of data from all the different stimuli read by that subject. The stimuli are listed across the top, sorted by condition and then by

TABLE 2.4 Reading time data from SPR organized by subject

Subject	*Item 1a*	*Item 1b*	*Item 2a*	*Item 2b*	*Item 3a*	*Item 3b*	*Item 4a*	*Item 4b*
1	763			797	660			848
2		1029	883			959	823	
3	374			498	384			530
4		623	687			1102	655	

Note: This is a hypothetical partial data set from only four participants and four items in two conditions (these data are partial because of space limitations; real data would represent at least 16 stimuli and 16 participants). For ANOVAs and *t*-tests, a mean score for each subject for each stimulus condition (a, b) would be calculated on the complete data set, which in this example would yield two mean scores per row/subject. Stimulus Type or Condition would be a repeated measure or within-subjects variable because the same subject contributes to both levels of the variable.

TABLE 2.5 Reading time data from SPR organized by item

Item	*Sub 1/a*	*Sub 1/b*	*Sub 2/a*	*Sub 2/b*	*Sub 3/a*	*Sub 3/b*	*Sub 4/a*	*Sub 4/b*
1	763			1029	374			623
2		797	883			498	687	
3	660			959	384			1102
4		848	823			530	655	

Note: This is the same hypothetical partial data set from Table 2.4. For ANOVAs and *t*-tests, a mean score for each item for each stimulus condition (a, b) would be calculated on the complete data set, which in this example would yield two mean scores per row/item. Stimulus Type/Condition would be a repeated measure because the same item contributes to both levels of the variable.

item (e.g., 1a, 2a, 3a, etc., 1b, 2b, 3b, etc.). Thus, in the subjects layout, each row corresponds to a subject and each column to an item. This layout is useful for understanding both how to trim the data by subject and how to calculate aggregate means by subject. For the items layout, the same data from the subjects layout is transposed so that each row represents an item and each column represents a subject, as illustrated with hypothetical data in Table 2.5. The item layout can be helpful for visualizing how both data trimming and the computation of aggregate means produce different results when conducted by item rather than by subject. It is also interesting to note that any one of the group means (i.e., a mean of the individual means, calculated for each participant group within each stimulus condition), if calculated on untrimmed data, comes out the same by subject or by item, but the standard deviation differs. This is why data trimming is conducted separately by subject and by item and why it can result in (usually very small) differences in the group means by subject versus by item.

Data trimming is the process of cleaning reaction time data to minimize the effects of those data points which appear to have been influenced by external factors unrelated to language processing, such as minor distractions and disruptions during the SPR experiment, which can obscure real reading time effects through the addition of extraneous variance and the resulting reduction in experimental power. Trimming involves the identification and removal or replacement of extreme data points, known as outliers. Outliers are presumed to reflect measurement error rather than authentic processing behavior and their elimination is most important to maximizing the accuracy and power of parametric tests that are conducted on aggregate means, like *t*-tests and ANOVAs, and is less critical with linear mixed-effects models, although a few very extreme outliers may be removed for mixed-effects models as well. In the case of the former, the purpose of identifying outliers is to ensure that aggregate means are as accurate as possible and minimally affected by extreme values, because once the means are calculated, the range of values they represent is no longer part of the data. Given that mixed-effects models do not rely on aggregate means, the full range of values remains in the data on which the statistical tests are conducted and the presence of outliers is thus not such a concern.

There is no single accepted method for dealing with outliers in SPR data, and early L2 studies tended to leave reading time data untrimmed, but outliers can obscure reading time effects that would otherwise prove significant, so it is preferable to minimize their influence. Two common procedures for identifying outliers are the designation of an absolute cutoff and the calculation of a cutoff based on standard deviations, while two ways to mitigate the effects of outliers, once they have been identified, are deletion and substitution with a more moderate value. For the absolute cutoff method, real response times of less than 100 ms are generally not possible (Luce, 1986) and with SPR in particular reading times of less than 200 ms likely reflect unintentional button presses, so lower cutoffs usually fall within the range of those two values. Higher cutoffs vary and have been set, for instance, at 3000 ms for all participants (Roberts & Felser, 2011), 3000 ms for native readers and 4000 ms for non-native readers (Havik et al., 2009), or 6000 ms for all participants (Jackson, 2010). The absolute cutoff method has the advantage of being conducted only once on the data, uniformly across reading time data from all stimulus regions, with no need to differentiate at this point by subject and by item. A disadvantage of using absolute cutoffs is that they are uniform across participant groups and experimental conditions and thus can potentially neutralize subtle differences.

Especially where there is considerable variability between subjects or between items, it may be preferable to use the more conservative standard deviation approach to identifying outliers rather than setting absolute values. With the standard deviation method, reading times that fall greater than two to three standard deviations away from a mean, calculated either for each individual or for each participant group in each stimulus condition, are judged to be outliers. Individual calculations should be conducted where there is notable variability between subjects. A disadvantage of the standard deviation approach is that it is considerably more time consuming, especially because it is conducted twice for each stimulus region, once by subjects and once by items. Some researchers prefer to use a combination of both of the above methods, identifying the most dramatic outliers first using absolute cutoffs and then using a standard deviation method to identify additional outliers.

After the outliers have been identified using one of the above methods, they are either deleted from the data set or replaced with a more moderate value, such as the absolute value or standard deviation value used as the cutoff, or treated with a combination of both deletion and substitution. For example, extremely low reading time values of 100 to 200 ms are presumably erroneous and uninformative, and therefore can reasonably be deleted. Extremely high values, on the other hand, might reflect real processing difficulty, so it makes sense to replace them with more moderate values that are still relatively high, such as the cutoff value used to identify outliers. Spreadsheet and statistical software can greatly facilitate the identification of outliers, their deletion, and their substitution with more moderate values. In the end, by the time the reading time data are submitted to statistical tests, the trimming of outliers has rarely affected more than 5% of the total data set, although it is acceptable to go as high as 10% (Ratcliff, 1993).

Once the data have been trimmed, aggregate means can be calculated and ANOVAs and *t*-tests conducted on the means. For subjects analysis, one mean is calculated for each participant within each stimulus condition. For items analysis, one mean is calculated for each item in each stimulus condition. ANOVAs and *t*-tests, which are the most common statistical analyses conducted on SPR data in SLA research to date, both determine whether any observed differences are likely to reflect real differences in the stimuli conditions or participant groups. These tests can be performed using a commercial statistical program like SPSS or free software like R (R Development Core Team, 2005); basic *t*-tests can also be conducted in Excel using the TTEST function, which can be useful for preliminary analyses. A typical ANOVA for a simple SPR experiment is a 2 (groups) × 2 (stimulus conditions), which is run as a mixed design because group is a between-subjects factor and stimulus condition is a within-subjects factor, or repeated measure, in which the same subject contributes to both levels of the variable. In the case of an interaction, posthoc *t*-tests can be conducted to make the comparisons within each group. Statistics from these ANOVAs and *t*-tests are reported as F_1 and t_1 in order to distinguish them from those generated in the analyses by item, which are referred to as F_2 and t_2. For the analyses by item, the ANOVA for a simple experiment would again be a 2 × 2, but here both group and stimulus condition are within-subjects factors because the item is held constant across both levels of both variables. Similar sets of analyses by subject and then by item are conducted for each region of interest in the stimuli and for the distractor questions. One of the basic assumptions of parametric statistics like ANOVAs and *t*-tests is that the data are in a normal distribution, but reading time data are positively skewed and proportional data from binary choice distractor questions are negatively skewed, so the transformation of both types of data before submitting them to such analyses can also improve statistical power.

The complete results of the statistical analyses, be they parametric tests or mixed-effects models, are reported when presenting or publishing the results of an SPR experiment or series of experiments. Analyses for stimulus regions prior to the region in which the conditions differ because of the experimental manipulation do not typically yield any significant effects or interactions—barring any failure to counterbalance extraneous variables in the experimental materials—and are often not reported. In addition, descriptive statistics in the form of group means are also reported as line graphs (often with error bars) or as a table of reading times, as in Figure 2.4 and in Table 2.1, respectively. Accuracy data from distractor questions are typically displayed as separate tables or bar graphs.

An Exemplary Study

A good example of how SPR can be employed to address SLA issues can be found in Jackson (2010), a research article which provides a good level of detail on the methodology as well as the complete set of stimuli used for the experiment. This investigation was designed to address the questions of whether L1 processing strategy is transferred to L2 processing and whether L2 readers pay more attention to lexical-semantic information while processing than do L1 readers. To answer these questions, the study exploited a temporary subject-object ambiguity that can arise in German due to flexible word order, case marking that is sometimes optional, and the occurrence of lexical verbs in either verb-second or verb-final positions, depending on whether the verb tense is simple or complex, respectively. A sample of a stimulus item in its four conditions is given in (7), where slashes indicate the phrase-by-phrase segmentation used for the SPR task. Psycholinguistic research has shown that native speakers of German take longer to read the object-first version of similarly disambiguated sentences, even when a compound verb tense means that the lexical verb does not appear until the end of the sentence (Schriefers, Friederici, & Kühn, 1995). This shows not only that they prefer subject-first word order, but also that they begin to assign argument roles before encountering the lexical verb, purely on the basis of syntactic cues like case markers.

(7)

a. *Subject-first, Simple past*

Welche Ingenieurin / traf / den Chemiker / gestern Nachmittag / im Café / bevor / der Arbeitstag / anfing?
Which NOM$_{\text{/ACC}}$ engineer / met / the$_{\text{ACC}}$ chemist / yesterday afternoon / in-the café / before / the work-day / began?
"Which engineer met the chemist yesterday afternoon in the café, before the workday began?"

b. *Object-first, Simple past*

Welche Ingenieurin / traf / der Chemiker . . .
Which$_{\text{NOM/ACC}}$ engineer / met / the$_{\text{NOM}}$ chemist . . .
"Which engineer did the chemist meet . . . "

c. *Subject-first, Present perfect*

Welche Ingenieurin / hat / den Chemiker / gestern Nachmittag / getroffen . . .
Which$_{\text{NOM/ACC}}$ engineer / has / the$_{\text{ACC}}$ chemist / yesterday afternoon / met . . .
"Which engineer has met the chemist yesterday afternoon . . . "

d. *Object-first, Present perfect*

Welche Ingenieurin / hat / der Chemiker / gestern Nachmittag / getroffen . . .
Which$_{\text{NOM/ACC}}$ engineer / has / the$_{\text{NOM}}$ chemist / yesterday afternoon / met . . .
"Which engineer has the chemist met yesterday afternoon . . ."

The participants in Jackson's (2010) study were native and advanced non-native speakers of German, with 22 L1 English-L2 German, 20 L1 Dutch-L2 German, and 24 German native speakers. Dutch is similar to German with regard to the differential positioning of lexical verbs in simple versus complex verb tenses described above, while English is not, so the purpose of including the two L2 participant groups was to better address the issue of L1 transfer. Participants each read 32 target items, with eight in each of the four conditions illustrated above, along with 64 filler items in a linear, noncumulative (i.e., moving window) SPR procedure with phrase-by-phrase segmentation. The distractor task was meaning-based and asked participants to indicate whether a statement on the screen was consistent with the meaning of the stimulus sentence that preceded it.

Mixed-design ANOVAs performed on the reading times from the critical region underlined in (7) revealed that all three participant groups had similar difficulty in processing object-first word order, regardless of whether the verb tense was simple past or present perfect. Analyses of reading times for the clause-final region, however, showed that only the L1 English participants had greater difficulty with the stimuli in the present perfect conditions after reaching the lexical verb at the end of the clause. Jackson (2010) concluded that the L1 English readers may have transferred a lexical verb-driven strategy from English while processing sentences in L2 German, but that such relatively higher sensitivity to lexically-based information is not an inherent characteristic of L2 processing because the L1 Dutch readers exhibited SPR effects that were similar to those of the native German readers.

Pros and Cons in Using the Method

Pros

- SPR is inexpensive. It is probably the most economical on-line method for sentence processing research and is thus accessible to a wide range of researchers, including those conducting pilot studies and student research projects. Experiments can be built and run on a basic desktop or laptop using either free software that runs entirely script-based experiments (e.g., DMDX by Forster & Forster, 1999; Linger by Rhode, 2001; PsyScope by Cohen, MacWhinney, Flatt, & Provost, 1993) or relatively inexpensive commercial software that adds the convenience of a graphical user interface (e.g., E-Prime by Schneider, Eschmann, & Zuccolotto, 2002; SuperLab, 1992).
- SPR is highly portable. Because there is no special equipment outside the computer that runs the software and perhaps a small response device, SPR experiments can be conducted virtually anywhere.
- SPR is efficient. The researcher does not have to supervise participants as closely as with some other methods, where it can be necessary to monitor and make adjustments to equipment while the experiment is in progress. This convenience, combined with the low cost of start-up, makes it feasible

for a single researcher to run as many as six or seven subjects at a time if the laboratory facilities are sufficient.

- SPR is an exceptionally covert measure of sentence processing. Participants' conscious attention can easily be diverted away from language to a distractor task such as answering comprehension questions, which is also more familiar to them as an assessment than SPR. Additionally, participants do not need to know beforehand that the software program is recording their reading times, they are not likely to have previous assumptions regarding SPR because it is one of few methods used exclusively for psycholinguistic research, and they do not come into contact with any highly specialized technical equipment that could further invite conscious, task-specific strategy.
- SPR materials can also be relatively covert with regard to their linguistic targets. While some SPR paradigms do employ stimuli with grammatical violations that may invite explicit judgments or the activation of metalinguistic knowledge, particularly among participants formally trained in the L2, it is also possible to obtain significant reading time effects with more subtle paradigms in which all stimuli are grammatical.
- SPR can detect spillover and sentence wrap-up effects. Increased reading times on the stimulus region immediately following the site of an immediate effect or at the end of a sentence are assumed to reflect later phases of comprehension and can be indicators of processing difficulty that is either persistent or delayed, which is especially useful in L2 research.

Cons

- Text must be segmented word-by-word or phrase-by-phrase in order to generate reading time data with any level of detail. This presentation format is different from that typically seen outside of the laboratory environment and may present an additional processing load, probably because in "normal" college-level reading (as measured with eye-tracking), regressions account for about 15% of eye movements. This can raise questions of ecological validity. During debriefing in the laboratory, a few of my research participants (less than 2%) have even commented that they find the word-by-word reading mode difficult and have speculated that the SPR experiments in which they participated were tests of memory.
- Research subjects must repeatedly press a button while reading in the self-paced mode, which was historically regarded as a potentially unnatural distraction because it was markedly different from reading a book or other printed text. However, this difference between what occurs inside versus outside the laboratory is no longer as dramatic, as most participants are now accustomed to using input devices while reading because of contemporary technology like text messaging, browsing the Internet via a smart phone, or using computer software with a graphical user interface and pointing device like a mouse or touchpad.

- SPR requires that participants be relatively fluent readers. Thus, great care must be taken when using the method to study populations with relatively less reading experience in the target language, such as beginning L2 learners, heritage bilinguals, and even some native speakers with low levels of literacy. Also, the study of languages without developed and easily digitized writing systems is of course not possible with SPR.
- SPR can generate task-specific effects, especially when stimuli are segmented word-by-word. These effects include reading times that are generally slower than normal and delayed processing effects that can spill over into the next region.

Discussion and Practice

Questions

1) Give two examples of participant populations that could easily be studied using the SPR method, giving specific details regarding their L1, L2, proficiency level, and demographic characteristics. Then, give two more examples of populations for which groups it might be more desirable to select another experimental measure.
2) Define the terms *item, condition,* and *region of interest* as they relate to self-paced reading stimuli, being clear about the differences between them. Be sure to use concrete examples to support your explanations.
3) Select an example of a published research study of L2 processing using the SPR method, either from among those mentioned in this chapter or from recent SLA and psycholinguistics journals, and analyze its research design. This analysis should include a) the specific research questions guiding the study, b) the independent variables that were manipulated through the choice of participants and/or stimuli, c) the type of statistical analyses run on the data, and d) a critique of the design, with particular attention to the relationship between the research questions, the independent variables, and the statistical analyses.
4) Consider the hypothetical incomplete data set in Tables 2.4 and 2.5. Calculate the mean scores by subject and by item (you should have a total of 4 × 2 = 8 means for each table). Do there appear to be any differences in reading times between the two stimulus conditions (a, b), either by subject and/or by item? Give an example of what type of stimuli might yield this data set and which region of those experimental sentences these data might represent.
5) List the steps in preparing SPR data from one experiment for statistical analyses via ANOVAs and *t*-tests. How does the list change if the analyses are to be via linear mixed-effects models?

Research Project Option A

The off-line distractor task is an integral part of sentence comprehension research, yet not much is known about how it might influence on-line processing behavior.

One project that would serve as a good introduction to the SPR method and its potential effects on participant behavior would be to replicate a published SPR study using two different distractor task conditions: acceptability judgment and meaningful comprehension question. (If the complete set of stimuli is not part of the published article, materials, and permission to use them, can usually be obtained by writing the author.) Such an investigation could be informative with just a single L2 participant group, though ideally a comparison group of native speakers would also be included.

Research Project Option B

The majority of existing L2 sentence processing research includes participants of only one proficiency level, so our knowledge of the developmental trajectory of non-native sentence comprehension behavior is quite limited. Another good introduction to the SPR method would be to replicate a published study and include at least two participant groups with different levels of proficiency in the L2. With both this option and Research Project Option A above, the target language could be the same as in the original study, or the stimuli could be translated into another language—assuming that there is sufficient grammatical similarity between the two languages.

Note

1. In order to avoid confusion, it is worth noting that the term *moving window* is also sometimes used to refer to a gaze-contingent eye-tracking paradigm in which parafoveal vision is masked or blurred in order to examine its role in reading comprehension and in visual perception in general. In this type of eye-tracking, the "moving window" is a small, round window of clear vision that corresponds to the reader's foveal region of vision, which moves around the display screen as the reader's gaze shifts. Thus, in both eye-tracking and self-paced reading the moving window is a spatially limited view of the stimulus that is surrounded by a larger area that is masked.

Suggested Readings

Dussias, P. E., & Piñar, P. (2010). Effects of reading span and plausibility in the reanalysis of wh- gaps by Chinese–English L2 speakers. *Second Language Research, 26,* 443–472.

Jegerski, J. (2012b). The processing of temporary subject-object ambiguities in native and near-native Mexican Spanish. *Bilingualism: Language and Cognition, 15,* 721–735.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics, 32,* 299–331.

Sagarra, N., & Herschensohn, J. (2011). Proficiency and animacy effects on L2 gender agreement processes during comprehension. *Language Learning, 61,* 80–116.

References

Aaronson, D., & Scarborough, H. S. (1976). Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 56–70.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28.

Caplan, D., & Waters, G. S. (2003). On-line syntactic processing in aphasia: studies with auditory moving windows presentation. *Brain and Language, 84*(2), 222–249.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope [Computer software]. Retrieved from http://psy.ck.sissa.it/.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, & A.M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives* (pp. 94–128). Cambridge, UK: Cambridge University Press.

Cuetos, F., & Mitchell, D.C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition, 30,* 73–105.

Dickey, M. W., Choy, J. J., & Thompson, C. K. (2007). Real-time comprehension of wh-movement in aphasia: Evidence from eyetracking while listening. *Brain and Language, 100*(1), 1–22.

Donders, F. (1868/1969). On the speed of mental processes. *Acta Psychologica, 30,* 412–431. (Translated by W. G. Koster).

Dussias, P. (2003). Syntactic ambiguity resolution in second language learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition, 25,* 529–557.

Ellis, N. C. (2007). Implicit and explicit knowledge about language. In J. Cenoz & N. H. Hornberger (Eds.) *Encyclopedia of Language and Education, Second Edition, Volume 6: Knowledge about Language* (pp. 119–132). New York: Springer.

Felser, C., Roberts, L., Gross, R., & Marinis, T. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics, 24,* 453–489.

Ferreira, F., & Henderson, J. (1990). Use of verb information during syntactic parsing: Evidence from eye tracking and word by word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16,* 555–568.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior, 4,* 414–420.

Foote, R. (2011). Integrated knowledge of agreement in early and late English-Spanish bilinguals. *Applied Psycholinguistics, 21,* 187–220.

Forster, K. I., & Forster, J. C. (1999). DMDX [Computer software]. Tucson, AZ: University of Arizona/Department of Psychology.

Forster, K. I., & Olbrei, I. (1973). Semantic heuristics and syntactic analysis. *Cognition, 2,* 319–347.

Foss, D. J. (1970). Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 9,* 699–706.

Gilboy, E., & Sopena, J. M. (1996). Segmentation effects in the processing of complex NPs with relative clauses. In M. Carreiras, J. E. García-Albea, & N. Sebastián-Gallés (Eds.), *Language Processing in Spanish* (pp. 191–206). Mahwah, New Jersey: Erlbaum Associates.

Havik, E., Roberts, L., van Hout, R., Schreuder, R., & Haverkort, M. (2009). Processing subject–object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning, 59,* 73–112.

Hoover, M., & Dwivedi, V. (1998). Syntactic processing by skilled bilinguals. *Language Learning, 48,* 1–29.

Jackson, C. N. (2010). The processing of subject-object ambiguities by English and Dutch L2 learners of German. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing* (pp. 207–230). Amsterdam: John Benjamins.

Jackson, C. N., & Bobb, S. C. (2009). The processing and comprehension of *wh-* questions among second language speakers of German. *Applied Psycholinguistics, 30*(4), 603–636.

Jegerski, J. (2012a). The processing of case markers in near-native Mexican Spanish. Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY.

Jegerski, J. (2012b). The processing of temporary subject-object ambiguities in native and near-native Mexican Spanish. *Bilingualism: Language and Cognition, 15*(4), 721–735.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*(4), 603–634.

Juffs, A. (1998a). Main verb versus reduced relative clause ambiguity resolution in second language sentence processing. *Language Learning, 48,* 107–147.

Juffs, A. (1998b). Some effects of first language argument structure and syntax on second language processing. *Second Language Research, 14,* 406–424.

Juffs, A. (2004). Representation, processing, and working memory in a second language. *Transactions of the Philological Society, 102,* 199–225.

Juffs, A. (2005). The influence of first language on the processing of *wh*-movement in English as a second language. *Second Language Research, 21,* 121–151.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and object asymmetries in wh-extraction. *Studies in Second Language Acquisition, 17,* 483–516.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language processing research. *Language Learning, 46,* 286–324.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixation to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., Carpenter, P. A., & Wooley, J. D. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General, 111*(2), 228–238.

Krashen, S. (1981). *Second language acquisition and second language learning.* Oxford, UK: Pergamon.

Leeser, M., Brandl, A., & Weissglass, C. (2011). Task effects in second language sentence processing research. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching, and research: Insights from psycholinguistics* (pp. 179–198). Amsterdam: John Benjamins.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language, 52,* 436–459.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9,* 121–136.

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition, 27,* 53–78.

Mitchell, D.C., & Green, D. W. (1978). The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology, 30*(4), 609–636.

Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject-verb agreement processes in comprehension. *Journal of Memory and Language, 36,* 569–587.

Pan, H.-Y., & Felser, C. (2011). Referential context effects in L2 ambiguity resolution: Evidence from self-paced reading. *Lingua, 121,* 221–236.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition, 24,* 501–528.

Pliatsikas, C., & Marinis, T. (2013). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition, 16*(1), 167–182.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

R Development Core Team. (2005). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Rhode, D. L. T. (2001). Linger [Computer software]. Cambridge, MA: MIT TedLab. Retrieved from http://tedlab.mit.edu/~dr.

Roberts, L. (2009). The L1 influences the on-line comprehension of tense/aspect in L2 English. Paper presented at the L2 Processing and Parsing Conference, Lubbock, TX.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics, 32*(2), 299–331.

Sagarra, N., & Herschensohn, J. (2011). Proficiency and animacy effects on L2 gender agreement processes during comprehension. *Language Learning, 61,* 80–116.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). E-Prime [Computer software]. Pittsburgh, PA: Psychology Software Tools Inc.

Schriefers, H., Friederici, A.D., & Kühn, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of Memory and Language, 34,* 499–520.

SuperLab (Version 4.0.5) [Computer software]. (1992). San Pedro, CA: Cedrus Corporation.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language, 27,* 1–36.

Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language, 39,* 102–123.

Urbaniak, G. C., & Plous, S. (2011). Research Randomizer (Version 3.0) [Computer software]. Retrieved from http://www.randomizer.org/

VanPatten, B., & Jegerski, J. (Eds.) (2010). *Research in second language processing and parsing.* Amsterdam: John Benjamins.

White, L., & Juffs, A. (1998). Constraints on wh- movement in two different contexts of non-native language acquisition: Competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neil (Eds.), *The generative study of second language acquisition* (pp. 111–129). Mahwah, NJ: Lawrence Erlbaum.

Williams, J., Möbius, P., & Kim, C. (2001). Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics, 22,* 509–540.

3

SELF-PACED LISTENING

Despina Papadopoulou, Ianthi Tsimpli, and Nikos Amvrazis

History of the Method

The self-paced listening (SPL) method was originally developed by Ferreira, Henderson, Anes, Weeks, & McFarlane (1996). Ferreira et al. dubbed this technique *auditory moving window* because it is the auditory equivalent of the visual moving window paradigm (Just, Carpenter, & Woolley, 1982), which is a version of self-paced reading (see Jegerski, Chapter 2, this volume). Ferreira et al. invented SPL in order to explore whether listeners employ prosodic cues to comprehend sentences in real time. At that time, the available auditory measures—monitoring, cross-modal priming, and naming tasks—measured processing only at one specific point during sentence comprehension rather than continually as the sentence evolves. The SPL technique instead measures how much time participants take to listen to each of the words or phrases of a sentence. SPL is a rather recent psycholinguistic method compared to self-paced reading, which has been used extensively for over forty years. Still, SPL has been utilized not only in research on native adult sentence processing (Ferreira, Anes, & Horine, 1996; Ferreira, Henderson, Anes, Weeks, & McFarlane, 1996; Papangeli, 2010; Waters & Caplan, 2005) but also on sentence comprehension in monolingual (Papadopoulou, Plemenou, Marinis, & Tsimpli, 2007; Papangeli, 2010) and bilingual children (Marinis, 2007, 2008; Papangeli, 2010), as well as in populations with language disorders, such as aphasics (Waters & Caplan, 2005) and specific language impairment (SLI) children (Chondrogianni, Marinis, & Edwards, 2010; Marshall, Marinis, & van der Lely, 2007).

In their original SPL study, Ferreira, Henderson et al. (1996) explored the following research questions:

(1) whether lexical frequency affects spoken language understanding;
(2) whether garden path effects are attested in spoken language comprehension;
(3) whether the SPL technique is sensitive to factors that affect spoken language processing.

The results revealed that low frequency words and the subsequent word were aurally processed significantly slower than were high frequency words, indicating that frequency influences not only written (Henderson & Ferreira, 1990), but also spoken language comprehension. Moreover, garden path effects were tested with the type of reduced relative clauses exemplified in (1a), which are known to induce processing effects in reading experiments (Clifton, Traxler, Mohamed, Williams, Morris, & Rayner, 2003; Ferreira & Clifton, 1986; Just & Carpenter, 1992; MacDonald, Just, & Carpenter, 1992; Rayner, Carlson, & Frazier, 1983; Rayner, Garrod, & Perfetti, 1992).

(1)
  a. The editor played the tape agreed the story was big.
  b. The editor played the tape and agreed the story was big.

Garden path effects, evidenced as longer listening times on the verb *agreed* and the noun phrase (NP), *the story,* for (1a) than for (1b), were detected in Ferreira, Henderson et al.'s study even when prosodic cues biased the participants towards the reduced relative reading. These findings were significant for two reasons; first, they extended garden path effects to spoken language processing and, second, they showed that prosody may not be employed as a cue to disambiguate sentences.

After the introduction of SPL in 1996, researchers have employed it as an on-line method to investigate subject/object ambiguities, relative clause attachment, and pronoun resolution, as well as how participants process certain grammatical phenomena, such as passive sentences, clitics, articles, and cleft sentences. More recently, SPL has become popular in research on sentence processing with children and second language learners, because it does not presuppose literacy. Moreover, a new version of the SPL task has been recently developed by Marinis (2007, 2008; Marshall, Marinis, & van der Lely, 2007), which combines the on-line auditory comprehension of sentences with picture verification and aims at making the task even more child-friendly (see also Marinis, 2008b, 2010). In SPL with picture verification, a picture is presented on the screen and then a sentence is orally presented to the participants in a word-by-word or phrase-by-phrase fashion. The picture either matches or does not match the content of the aural sentence. At the end of each sentence, participants indicate whether the picture matches what they heard or not. The rationale is that listening times on the critical segment are longer for the mismatching than the matching condition. To illustrate, consider sentences

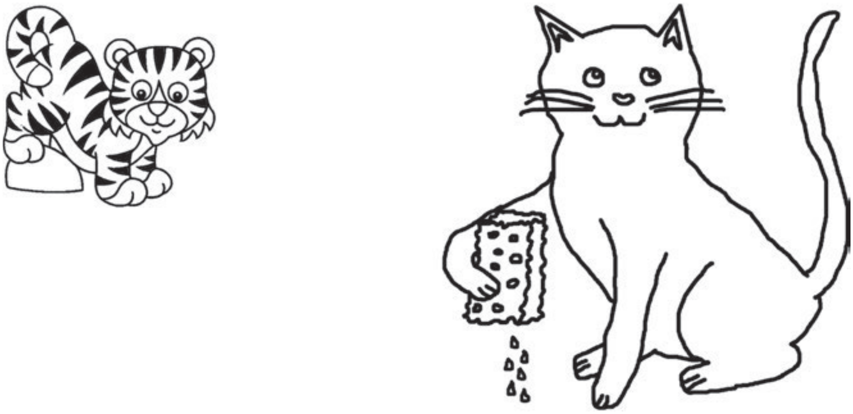


FIGURE 3.1 Illustration of a cat washing herself, which matches sentence (2b), but not sentence (2a), which has a passive interpretation. Hence, listening times on the *by*-phrase, *apó tin tíγri* ("by the tiger"), in (2a) will be longer than listening times on the PP, *me éna sfugári* ("with a sponge").

(2a) and (2b) as well as the picture that was shown to the participants (Figure 3.1) from Papangeli (2010) (slashes indicate the segments heard by the participant):

(2)

a. Dhōen pístepsa óti / pérsi to kalocéri / i γáta / plíθike / apó tin tíγri / sto Miláno.

not believed-2SG that last the summer the cat washed-pass-3SG by the tiger in Milan

"I did not believe that last summer the cat was washed by the tiger in Milan."

b. Dhōen pístepsa óti / pérsi to kalocéri / i γáta / plíθike / me éna sfugári / sto Miláno.

not believed-2SG that last the summer the cat washed-pass-3SG with a sponge in Milan

"I did not believe that last summer the cat washed herself with a sponge in Milan."

What is Looked at and Measured

SPL is implemented on a computer and records participants' listening times. More specifically, the participants listen to sentences segmented into either words or phrases as in (2) in the previous section, and the acoustic appearance of each segment is monitored by the participants' button presses. That is, the participants press a button to proceed to the next segment when they feel confident they have comprehended the current segment. Hence, the speed of stimulus presentation is controlled by the participants, which explains why the technique is called *self-paced*.

The SPL paradigm has only one possible presentation mode due to the nature of the acoustic stream—contrary to self-paced reading in which the segments can be presented in several fashions (Jegerski, Chapter 2, this volume). SPL is a moving window task, meaning that the participants do not know the length of the sentence they are going to listen to and they cannot go back and listen again to the words/phrases that they have already heard.

SPL makes the same presuppositions as reading tasks like self-paced reading and eye-tracking with respect to the interpretation of listening times for words or phrases. Namely, the assumption is that the time participants take to listen to a segment mirrors the time they need to process it, so prolonged listening times are thus a signal of difficulties that arise during language processing. That is, it is assumed that difficulties in comprehending a particular segment and integrating it with what was heard previously will be evident in longer listening times than with similar sentences for which there is no difficulty.

The SPL technique has been employed to test parsing preferences in orally presented, structurally ambiguous sentences with adults (Caplan & Waters, 2003; Ferreira, Henderson et al., 1996; Ferreira et al., 1996), children (Felser, Marinis & Clahsen, 2003; Kidd & Bavin, 2007; Papangeli & Marinis, 2010; Papangeli, 2010), and second language (L2) populations (Papangeli, 2010; Papangeli & Marinis, 2010).

The experiments conducted with monolingual adults and children have corroborated the findings from similar reading studies. For instance, Ferreira, Henderson et al. (1996) investigated the reduced relative clause ambiguity, illustrated in (1a) above and repeated for convenience in (3a), and found higher listening times on the main verb and on the subsequent segment with sentences with reduced relative clauses such as (3a), as compared to those with main clauses such as (3b). This result demonstrates the parser's preference for a main verb reading, which needs to be revised upon encountering the main verb, *agreed*.

(3)

a. The editor played the tape agreed the story was big.
b. The editor played the tape and agreed the story was big.

Kidd and Bavin (2007) tested the parsing preferences of English-speaking adults and 7- to 9-year-old children with temporary prepositional phrase (PP) attachment ambiguities like those in (4).

(4)

a. The girl cut the paper with nice colours because she was wrapping a present.
b. The girl cut the paper with sharp scissors because she was wrapping a present.
c. The girl cut some paper with nice colours because she was wrapping a present.
d. The girl cut some paper with sharp scissors because she was wrapping a present.

Both adults and children manifested a parsing preference for VP attachment—also previously attested in reading experiments—which was affected by verb subcategorization (i.e., action verbs versus light verbs), but not by the definiteness of the object (e.g., *the paper* versus *some paper*).

Felser, Marinis, and Clahsen (2003) tested relative clause attachment preferences in Anglophone adults and 6- to 7-year-old children by means of a phrase-by-phrase SPL experiment. In the examples in (5) the relative clause *who was/were feeling very tired* is ambiguous up to the verb, meaning that the relative clause could be attached to and modify either the NP1, *the nurse,* or the NP2, *the pupils.*

(5)

a. The doctor recognized the nurse of the pupils who was feeling very tired.
b. The doctor recognized the nurse of the pupils who were feeling very tired.
c. The doctor recognized the pupils with the nurse who was feeling very tired.
d. The doctor recognized the pupils with the nurse who was feeling very tired.

Their findings revealed differences between adults and children in their parsing preferences. More specifically, adults' attachment preferences were affected by the semantics of the preposition within the complex NP; namely, an NP1 preference was attested for the sentences involving the preposition *of* and an NP2 preference for the sentences incorporating the preposition *with.* On the other hand, children's preferences were affected by their listening span, in that high-span children preferred NP1 attachment and low-span children showed a preference for NP2 attachment.

In a Greek study which looked at subject/object ambiguities, Papangeli (2010) found that Greek adults and 9- to 12-year-old children—but not 6- to 9-year-old children—manifest a preference for the object reading, attested in a number of reading studies (but see Papadopoulou & Tsimpli, 2005, for opposite results in Greek).

SPL combined with picture verification was employed in a Greek study on pronoun resolution. Papadopoulou, Plemenou, Marinis, and Tsimpli (2007) tested sentences such as (6) in Greek adults and 10-year-old children in order to investigate the referential properties of null and overt pronouns:

(6) I nosokóma / voíθise / ti γramatéa / kaθós / *pro*/aftí / éγrafe / éna γráma / to vráδi.

The-FEM nurse-FEM helped the-FEM secretary-FEM while *pro*/she was writing a letter in the evening.

The null and the overt pronouns in the adverbial clause are ambiguous, as they may refer to the subject, the object, or a third referent. The participants were first presented with a picture and then they listened to sentences such as (6) segmented as illustrated above. At the end of each sentence they had to perform a sentence-picture matching task. Both groups preferred objects as referents for overt pronouns. On the other hand, null pronouns were more often interpreted as referring to the subject of the main clause by the adult group, while no preference was shown by the children.

Beyond structural ambiguity, the SPL paradigm has also been employed to investigate the on-line comprehension of grammatical structures in adults (Caplan & Waters, 2003; Papangeli, 2010), monolingual children (Booth, MacWhinney, & Harasaki, 2000; Chondrogianni, Marinis, & Edwards, 2010; Marinis, 2007; 2008; Marshall, Marinis, & van der Lely, 2007; Papangeli, 2010), L2 children (Marinis, 2007; 2008; Papangeli, 2010), and SLI children (Chondrogianni, Marinis, & Edwards, 2010).

Caplan and Waters (2003) tested the processing of subject and object cleft sentences as well as right-branching and center-embedded relative clauses in English unimpaired and aphasic participants. The task involved listening to the sentences in a phrase-by-phrase fashion and making a plausibility judgment at the end of each sentence. The findings showed that listening times reflected the syntactic complexity of the sentences; specifically, object clefts and center-embedded sentences were processed slower than subject clefts and right-branching relative clauses. Booth, MacWhinney, and Harasaki (2000) tested 8- to 12-year-old children's processing of relative clauses in a word-by-word SPL task and found that subject-object (7a) relatives were more difficult to comprehend than subject-subject (7b) relatives.

(7)

a. The monkey that followed the frog left the trees in a hurry.
b. The deer that the tiger watched entered the field from the side.

Marshall, Marinis, and van der Lely (2007) tested active and passive sentences in 6- to 10-year-old typically developing and SLI English children in a SPL task combined with picture verification. The pictures were either matched or mismatched with the sentence the participants heard. The results showed that the children were more accurate on the matched than the mismatched conditions and on the active than the passive sentences.

Chondrogianni, Marinis, and Edwards (2010) investigated the processing of articles and clitics in typically developing and SLI Greek children. Their test sentences included grammatical sentences involving articles and clitics as well as ungrammatical sentences with article and clitic omission. The ungrammatical sentences were read slower than the grammatical ones, indicating on-line sensitivity to the missing articles and clitics.

A third aim of SPL is to explore the role of prosody in language processing, and, more specifically, whether prosodic cues, such as phoneme and word duration, pause length, and suprasegmental duration and pitch facilitate sentence processing. So far, the findings are far from consistent. For example, Ferreira, Henderson et al.'s (1996) findings for reduced relative clause versus main verb ambiguities did not point to an advantage of the sentences with natural prosody, which argues against the rapid use of prosodic information during sentence comprehension (but see the discussion section of that study for possible segmentation effects). Papangeli, on the other hand, found that natural prosody eliminated Greek adults' and 9- to 12-year-old children's garden path effects in subject/object ambiguities and increased their accuracy. By contrast, prosodic cues did not have an impact on younger children, aged from 6 to 9 years, suggesting that sensitivity to prosody is prone to development.

In the last decade there has been an increasing interest in the way L2 learners process sentences on-line (see also Jegerski, Chapter 2, this volume; for a review see Clahsen & Felser, 2006). The research aims of these studies are to explore (a) whether L2 learners parse the input in the same way as native speakers and (b) whether L2 learners exhibit on-line sensitivity to ungrammatical structures which are nonetheless attested in their speech production. So far there are only a few L2 studies that have addressed these questions by means of SPL.

Marinis (2007, 2008) employed an SPL task combined with picture verification to test the processing of passive sentences as well as reflexive and nonreflexive pronouns (i.e., *him* and *himself,* respectively) in English children and Turkish children learning English as an L2. Listening times were recorded while a picture that either matched or did not match the orally presented sentence was displayed on the screen. Both the L1 and the L2 children were more accurate and faster on the match than on the mismatch conditions and on the active than on the passive sentences, which indicates similar parsing routines among both groups of children. The same technique was applied by Papangeli (2010) who investigated passives and reflexives in 6- to 9-year-old and 9- to 12-year-old Greek and Russian children learning Greek as an L2. Her findings were in line with Marinis' results, in that the processing of the structures tested was similar in both the L1 and the L2 child groups. Papangeli (2010), however, found differences between the Greek and the Russian children in that the Russian children did not show sensitivity to subject-verb agreement violations in on-line sentence comprehension.

Issues in the Development and Presentation of Stimuli

An SPL experiment consists of experimental events (or trials) that include the stimulus and an off-line task following the presentation of each sentence (to be described later in this section). The stimuli are the experimental sentences segmented into words or phrases. The development of the stimuli depends on the independent variables tested in the experiment. To illustrate this, let's consider the SPL study conducted by Felser, Marinis, and Clahsen (2003) on relative clause attachment preferences in L1 English. Relative clause attachment involves structures in which the RC is preceded by a complex NP containing two NPs and, thus, the RC can be attached either to the first or the second NP within the complex NP (see examples in (8)). In Felser et al.'s study the two independent variables were the type of preposition within the complex determiner phrase (DP), that is *of* or *with,* and disambiguation type, namely, relative clause attachment to the first or the second NP. The crossing of these two independent variables resulted in four experimental conditions, as shown below.

(8)

a. *Preposition type: of; Disambiguation type: NP1*

The doctor recognized the nurse of the pupils who was feeling very tired.

b. *Preposition type: of; Disambiguation type: NP2*

The doctor recognized the nurse of the pupils who were feeling very tired.

c. *Preposition type: with; Disambiguation type: NP1*

The doctor recognized the pupils with the nurse who were feeling very tired.

d. *Preposition type: with; Disambiguation type: NP2*

The doctor recognized the pupils with the nurse who was feeling very tired.

As illustrated in (8), the four experimental conditions of this item minimally differ; namely, the structure and the words are exactly the same in all four conditions, with the exception of the preposition in the complex NP and the number of the auxiliary verb in the relative clause. This control across stimulus conditions is crucial, because it ensures that any statistical differences between the four conditions can safely be attributed to the independent variables and not to other factors, such as word frequency, word length, or sentence length, among others, which are not relevant to the research questions. The way the experimental lists are created in terms of the number of items per condition, total number of items in the experiment, and counterbalancing procedures is similar to that employed in self-paced reading tasks and is presented thoroughly in Chapter 2 (Jegerski, Chapter 2, this volume). One difference between self-paced reading and SPL in the experimental design is that the ratio of test to distractor/filler sentences in SPL is not necessarily 1:3, as is the standard practice in self-paced reading. In fact, this ratio in SPL studies varies from 1:2 to 2:1. One reason for this is that listening experiments are usually longer than reading experiments, largely because they depend on the duration of the audio files. The extent to which this manipulation has an impact on the results cannot be addressed due to the rather small number of studies employing the SPL technique up to now. What is more, the SPL tasks testing phenomena previously investigated by means of self-paced reading have generally yielded equivalent findings, which indicates that the ratio of test and distractor/filler sentences probably does not have a significant effect on the results.

The stimuli in SPL tasks are aurally presented and the audio files need to be created in a particular manner. The experimental sentences are read by a native speaker in a soundproof recording booth, and are recorded and digitized using sound-recording software, such as SoundEdit (Dunn, 1994), Creative Media

Source Player (Creative Technology Limited), or Audacity (Mazzoni & Dannenberg, 2000), with a sampling rate of 44.1 kHz and 16-bit quantization. The sound files are stored as waveform files and are then edited by means of audio editing software, such as SoundEdit, Adobe Audition (Adobe Systems, 2003), or something comparable. In some studies, a tone is added to the waveform of the final segment in each sentence either as a means of signaling the end of the sentence, when the prosody is not natural, or as a warning that a comprehension question would follow. The presentation of each segment is controlled through button presses by the participants, as is the case in self-paced reading. Furthermore, participants are discouraged from pressing a button before listening to an audio file in its entirety (i.e., cutting off a segment midway through with a button press).

The sentences in SPL tasks are more often presented in a phrase-by-phrase fashion than in a word-by-word mode. This is because word-by-word presentation of sentences sounds unnatural. There are, however, studies that have segmented the test sentences into words rather than phrases (e.g., Booth, MacWhinney & Harasaki, 2000). The decision between the two presentation modes depends on the structures tested and the length of sentences. Another factor that may be considered is the age of the participants; word-by-word presentation is more difficult for children.

When it is crucial for the experimental aim to eliminate possible prosodic effects, certain techniques can be used which ensure flat prosodic structure. One simple technique is to have a speaker read the words of the test sentences as words in a list with flat prosody, without him/her being aware of the position of the words within the sentences. Then, the experimenter combines the experimental items by putting these words into sentences by means of audio editing software. Another way to avoid the impact of prosody on SPL tasks is through splicing; namely, words or phrases from one sentence are cut and then spliced into another sentence in order to neutralize prosodic cues across items. Prosodic changes to the experimental stimuli can also be achieved digitally by means of specially designed computer software.

When, in contrast, the sentences need to be heard with natural prosody, it is advisable to record the entire sentence. A marker—also known as *tag*—is then placed on the audio file of each sentence at the boundaries of the segments. The tags are placed at areas of low signal amplitude to make the transitions from one segment to the next smooth. When needed, the sound files are manipulated in order to avoid unintelligibility of the words (Ferreira, Henderson, Anes, Weeks, & McFarlane, 1996; Ferreira, Anes, & Horine, 1996; Waters & Caplan, 2004).

When SPL is combined with picture verification, pictures are usually presented prior to the sentence for approximately 2500 ms and remain on the screen while the participants listen to the sentence. Amvrazis (2012) is the only SPL study in which the picture was presented at the onset of the sentence. However, the findings were similar to those obtained by Papadopoulou, Plemenou, Marinis, and Tsimpli (2007), who tested the same phenomenon (pronoun resolution) and presented the pictures prior to the sentences.

SPL experiments usually incorporate an off-line task performed at the end of the sentence. The time the subjects take to respond to this task is often recorded and the accuracy scores are analyzed along with the reaction times. The off-line tasks used up to now in SPL studies include comprehension questions and plausibility/grammaticality judgments. As described previously, another type of off-line task is picture verification, in which the participants are asked to decide whether the sentence they hear matches the picture they see by pressing one of two prespecified buttons.

The results from these off-line tasks are informative, particularly when L2 learners participate in the experiment. For instance, there are studies that have indicated differences between native speakers and L2 learners as far as parsing routines are concerned but no differences between the two groups in the off-line task performed after the sentence has been processed (e.g., Papangeli, 2010). Nevertheless, in other studies the opposite pattern has been attested, namely similar processing routes for both native and L2 speakers but differences in terms of off-line responses (e.g., Amvrazis, 2012; Marinis, 2008). Obviously, these differences can be associated with the L2 learners' proficiency level and the linguistic features of the structure tested in the participants' first and second language. Still, the inclusion of an off-line task in SPL makes additional sets of data available, which can highlight interesting differences between the participant groups and the experimental conditions.

Scoring, Data Analysis, and Reporting Results

The scoring of data obtained by means of SPL experiments is to a great extent automatic due to their implementation through computer software. It is, therefore, important to have coded the data appropriately to avoid missing important information. In the example provided in the previous section from Felser, Marinis, and Clahsen (2003), the experimental list should include one column for the variable "preposition" and another one for the variable "attachment," which will be coded for each test sentence accordingly. The software includes this information in the spreadsheets of data, making data analysis straightforward. For each test sentence, there will be raw data for each sentence segment and for the off-line task performed at the end of the sentences. The raw data include reaction times in milliseconds (as reflections of how long it took to listen to something), as well as the participants' responses to the off-line task.

We will consider one example from the study by Papadopoulou, Plemenou, Marinis, and Tsimpli (2007) on pronoun resolution.

(9) O papús / milúse / δinatá / ston egonó tu / ótan / *pro* δjávaze / éna vivlío.

"The-MASC old-man / spoke-imp-3SG / loudly / to his grand-child-MASC / when / *pro* read-IMP-3SG / a book."

The experimental sentences consisted of seven segments, as illustrated in (9). For each segment, reaction times were computed, also known as *inter-response times* (Ferreira, Henderson, et al., 1996; Ferreira, Anes, et al., 1996; Waters & Caplan, 2005). Furthermore, a sentence-picture matching task at the end of each sentence was incorporated in this study, which was triggered via a question mark "?". The computer also recorded the (mis)match judgments and reaction times; that is, from the appearance of the "?" up to the subject's button press. Residual reaction times are often used in the data analysis of SPL tasks in order to avoid possible length effects of the audio files on inter-response times. Residual reaction times are calculated by subtracting the duration of the audio file from the actual reaction time.

The elimination of outliers is important to eliminate unwanted noise in the data due to fatigue, inattention, and other intervening variables. First, all data points coming from erroneous responses to the off-line task are removed. Jegerski (Chapter 2, this volume) argues that reaction times from the erroneous responses in self-paced reading may sometimes be useful for the interpretation of data, as shown in the study by Juffs and Harrington (1996). Unfortunately, there are no studies so far which employ the SPL technique and compare correct with erroneous responses, so this issue needs further investigation in future research. Moreover, the data from participants who perform at or below chance level for all conditions in the off-line task (Felser, Marinis, & Clahsen, 2003) or exhibit an accuracy rate at least 2.5 standard deviations below the group's mean (Papangeli, 2010) are also reported to be eliminated in some studies.

As noted by Jegerski (Chapter 2, this volume), there are various methods to trim the data for extreme values. In the SPL studies, a (predetermined) cutoff point is frequently used, which varies from above 2000 ms (Chondrogianni, Marinis, & Edwards, 2010; Water & Caplan, 2005) to 4000 ms (Papangeli, 2010). Additionally, reaction times which fall between two standard deviations (Chondrogianni, Marinis, & Edwards, 2010) and three standard deviations (Water & Caplan, 2005) below or above the condition mean by subject and by item are removed from further statistical analyses. In some studies, the removed reaction times are replaced with the cut-off point of the condition mean per subject and item.

The reaction times resulting from an SPL experiment are presented in means per condition and across subjects and items. When more than one group participates in the study, the means are demonstrated per participant group as well. One way to illustrate the reaction times on the sentence segments is by means of a line graph (Figure 3.2). The number of the lines depicted on the graph depends on the number of experimental conditions and each peak of the line denotes a segment. The most common statistical tests performed on the reaction time data from SPL are repeated-measures ANOVAs. The independent stimulus variables are usually the within-subject variables, while Group (e.g., native vs. L2 speakers) is the between-subject variable. When the ANOVA shows a statistically significant interaction between the two variables, *t*-tests may then be performed in order

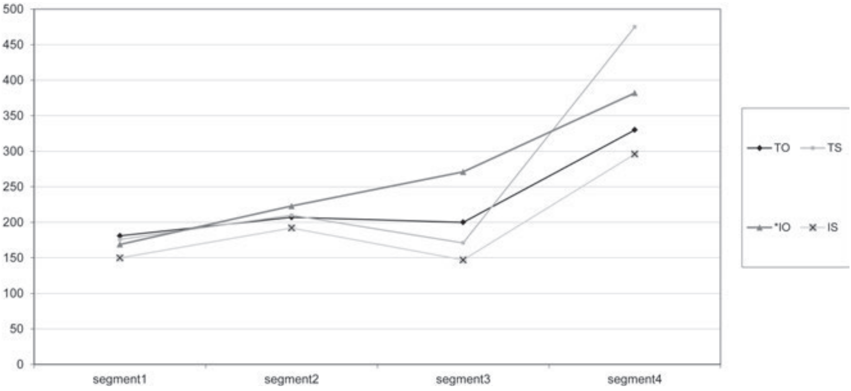


FIGURE 3.2 Listening times per segment and condition. TO: transitive verb, object reading, TS: transitive verb, subject reading, IO: intransitive verb, object reading (the * indicates that this was an ungrammatical condition), IS: intransitive verb, subject reading (taken from Papangeli, 2010, p. 60). The conditions and the items are described in detail in example (11) of the following section.

to explore the direction of this interaction. Moreover, the statistical analyses are performed on the subjects and the items to test the validity and strength of the statistical results. The standard practice is to report the subject analysis as F_1 and t_1 and the item analysis as F_2 and t_2.

An Exemplary Study

Papangeli (2010) explored two research questions relevant to SLA: (a) whether L2 learners exhibit sensitivity to ungrammaticalities in on-line comprehension; and (b) whether L2 learners' processing routes are similar to those used by native speakers. She conducted four SPL experiments; Experiments 1 and 2 tested the on-line processing of active, passive, and reflexive sentences, while Experiments 3 and 4 tested subject/object ambiguities. Here we will focus on Experiment 1, which investigated active and passive sentences by means of a SPL task combined with picture verification. We will also discuss Experiment 3, which looked at subject/object ambiguities with flat prosody through SPL. The participants were monolingual Greek adults and four child groups: 6- to 9-year-old and 9- to 12-year-old monolingual and Greek-Russian bilingual children. The L2 children had from two to six years exposure to Greek and were attending Greek school. The children also performed sections from the school-aged DVIQ test (Diagnostic Verbal IQ; Stavrakaki & Tsimpli, 2000) to assess their language skills. In Experiment 1, there were four experimental conditions, as shown below, which resulted from the crossing of two variables: (1) whether or not the picture matched the sentence and (2) active/passive voice morphology.

(10)

a. *Match—Active*

Akusa óti / prin apó δéka méres / o laγós / éδese / to fiδi / stin Práγa.
heard.1s that before ten days the.m.n rabbit.m.n tied the.n snake.n in-the Prague
"I heard that ten days ago the rabbit tied the snake in Prague."

b. *Mismatch—Active*

Akusa óti / prin apó δéka méres / to fiδi / éδese / ton laγó / stin Práγa.
heard.1s that before ten days the.n snake.n tied.3s the.m.acc rabbit.m.acc in-the Prague
"I heard that ten days ago the snake tied the rabbit in Prague."

c. *Match—Passive*

Akusa óti / prin apó δéka méres / o laγós / δéθice / apó to fiδi / stin Práγa.
heard.1s that before ten days the.m.n rabbit.m.n tied.pass.3s from the.n snake.n in-the Prague
"I heard that ten days ago the rabbit was tied by the snake in Prague."

d. *Mismatch—Passive*

Akusa óti / prin apó δéka méres / to fiδi / δéθice / apó ton laγó / stin Práγa.
heard.1s that before ten days the.n snake.n tied.pass.3s from the.m.acc. rabbit.m.acc. in-the Prague
"I heard that ten days ago the snake was tied by the rabbit in Prague."

The participants first saw a picture on a computer screen and had to press a button in order to listen to a sentence. The sentences were divided into six segments as illustrated in (10), while the critical segments were the fourth (the verb marked for passive voice), the fifth (the *by*-phrase), and the sixth (the final segment, which was an adjunct). The participants had to perform a sentence-picture matching task at the end of each sentence. The experiment included 40 critical items (10 per experimental condition), 20 fillers, and 10 practice items. The data were analyzed in terms of the participants' accuracy (i.e., whether they correctly indicated if the picture matched the sentence) and listening times on the critical regions. Moreover, residual listening times were calculated by subtracting the length of the audio file from the participant's listening time for each segment. All inaccurate responses to the sentence-picture matching task were also removed from subsequent analyses and outliers were then eliminated in two steps. First, listening times below 1000 ms and above 4000 ms on the critical regions were deleted. Secondly, the remaining listening times were further trimmed such that scores 2.5 standard deviations above or below the subject and item means were eliminated. The data from the four child groups were statistically compared with repeated measures ANOVAs, with group as the between-subjects variable and voice (active vs. passive) and matching (match vs. mismatch) as the within-subjects variables. The L2 children's accuracy and listening times were not found to differ from those of L1 children. In addition, the children's accuracy was found to correlate with language scores, which points to developmental effects on the processing of active and passive sentences. Moreover, conditions in which the sentence accurately described the picture (matched conditions) and conditions with active verbs were more accurate than mismatched conditions and passive sentences in all child groups. According to the experimenter, this data set indicates no qualitative differences between L1 and L2 children, in that both child groups are capable of rapidly using voice morphology during on-line processing.

Experiment 3 was a SPL task in which Papangeli tested the same groups of children—with the exception of the younger L2 children—parsing preferences on subject/object ambiguities (11a) and their sensitivity to ungrammaticalities involving subject-verb agreement morphology and argument structure (11b).

(11)

a. Enó majíreve / ta psária / káice/an / ston fúrno.

while cooked.imp.3s the.pl.n fish.pl.n burnt.3s/3p in-the oven
"While (s)he was cooking the fish burnt/it burnt in the oven."

b. Enó étrexe / ta psária / * káice /an / ston fúrno.

while ran.imp.3s the.pl.n fish.pl.n burnt.3s/3p in-the oven
"While (s)he was running the fish burnt/*it burnt in the oven."

In sentences such as (11b) the object reading is ungrammatical since the verb *run* is intransitive; the ungrammaticality becomes evident on the main verb *kaike* by means of subject-verb agreement morphology. There were 40 experimental items (10 per condition), 40 fillers, and 8 practice items. The sentences were segmented as shown in (11) and the critical segments were the main verb and the following segment. At the end of each sentence the participants performed a judgment task indicating whether the sentence was grammatical or ungrammatical. In this experiment the sentences were presented with flat prosody; the segments were recorded as words in a list and, subsequently, they were connected to form sentences with Adobe Audition software. The procedure followed for data trimming was the same as that used in Experiment 1. Repeated measures ANOVAs with group as the between-subjects variable and verb type (transitive vs. intransitive) and syntactic function (NP as subject or object) as the within-subjects variables were performed on the data. The L1 child groups demonstrated early sensitivity (i.e., at the main verb) to this ungrammaticality, as their listening times were longer for the ungrammatical than the grammatical conditions, while the L2 children's listening times on the ungrammatical sentences did not differ from those on the grammatical sentences. Furthermore, the older L1 child groups showed a preference for the object reading in sentences such as (11a), while the L1 younger and L2 older groups showed no preference for either reading. In addition, the children's accuracy significantly correlated with their language scores on the DVIQ test. Papangeli's findings indicate, contrary to her results in the previous experiment, that L2 parsing is different from L1 parsing in that (a) morphosyntactic information related to subject-verb agreement is not rapidly exploited in L2 sentence comprehension and (b) processing preferences for a specific interpretation, i.e., the object reading for the ambiguous sentences tested in Papangeli's study do not emerge in L2 parsing. Moreover, the divergent parsing routines employed by the 6- to 9-year-old and the 9- to 12-year-old monolingual children suggest developmental changes in parsing the input occurring between the ages of 6 and 12 years. The results from both experiments indicate that not all types of ungrammaticalities are rapidly detected in L2 on-line sentence comprehension and that native-like parsing preferences are not always evident in L2 ambiguity resolution.

Pros and Cons in Using the Method

Pros

- One of the major advantages of the SPL compared to the self-paced reading method is that SPL can be used with children and adults who are not fluent readers of the language tested. In the domain of L2 research this is a particularly welcome effect given the diversity of language backgrounds which may or may not share the same alphabetic script, as well as the heterogeneity of education and literacy backgrounds of adult second language learners.
- The SPL method enjoys the same advantages of the self-paced reading method in that it provides on-line measures of reaction times to each consecutive segment of the sentence, word, or phrase.
- SPL allows researchers to identify the hypothesized locus of a grammaticality violation, a syntactic, semantic or lexical ambiguity, or even a higher level, interface property. Depending on the phenomenon tested, the method provides results which can be interpreted both at the local level (i.e., the critical segment or segments) and also at the global level when the participant has to make a decision about the overall sentence status.
- SPL is adequate for the investigation of developmental (L1 or L2) effects on local and global sentence processing, the speed of integration and the sensitivity to local (lexical, syntactic, or morphological) factors affecting incremental processing.

Cons

- A potential problem with the SPL method has to do with the type of linguistic properties involved in the processing of the phenomenon investigated. For example, a syntax-discourse interface phenomenon, such as pronominal resolution, is largely determined at a late stage of sentence processing, although early preferences for subject or object antecedents may be found too. This is so because pronominal resolution is an interface phenomenon *par excellence;* a pronoun, null or overt, is only specified for certain grammatical features. Reference interpretation, however, is established as a combination of these features, default processing strategies (e.g., subject prominence), and tendencies associated with information-structure properties of the specific language (e.g., word-order changes signaling topic shift). In comparison, a phenomenon testing grammaticality at the local, phrasal level may produce stronger effects in the on-line processing of the critical segment and those following it. This difference in the strength of effects of discourse-related as opposed to syntax proper phenomena is relevant to self-paced reading experiments too.
- In SPL tasks that involve sentence-picture matching, a potentially controversial point has to do with the onset of the presentation of the picture and the duration of its presentation on screen. In early presentation (i.e., before the sentence is listened to), the participant can build some expectations for sentence content from the visual stimulus presented. Expectations on the content of the picture are created through the opposite route (i.e., when the participant hears the sentence and then looks at the picture to respond to the matching task).

Discussion and Practice

Questions

1) Define the terms *inter-response reaction times* and *residual reaction times* as they relate to the analysis of data obtained from SPL tasks. Clarify the difference between the two measures and give an example to better illustrate this difference.
2) Provide the characteristics of one L2 group which could be studied by means of the SPL task, and one L2 group which could be studied through the SPL task combined with picture verification. Be specific about the reasons for which one method would be more appropriate than the other for a particular L2 group.
3) Select one study that has employed the SPL technique and describe its experimental design in detail. Provide the dependent and independent variables, the experimental conditions, the number of test items per condition and the number of distractors and/or fillers. Furthermore, present the experimental procedure and describe the statistics used to analyze the data.

4) In studies on the processing of grammatical features with L2 learners that employed SPL with picture verification (e.g., Marinis, 2007, 2008a; Papangeli, 2010), the picture is displayed prior to the aural presentation of the sentence. Explain (a) how this order of presentation may affect reaction times, in particular on the critical segment, and (b) how the results would differ if the picture appeared immediately after hearing the sentence.
5) Explain why prosodic cues are removed in some studies investigating sentence processing by providing a specific example. Describe one procedure that can be used to eliminate prosodic cues.

Research Project Option A

The SPL experiments conducted with native speakers replicate the findings obtained from reading tasks. However, there are not many studies employing SPL with second language learners. One good introduction to the use of SPL would be to run an SPL task with the experimental items used in reading studies on L1 sentence processing. The research questions would be (a) to investigate whether the L2 learners exhibit the same parsing routines as the native speakers in SPL and (b) to explore possible differences between SPL and self-paced reading in terms of the processing routes applied by L2 learners.

Research Project Option B

The SPL technique combined with picture verification has not yet been used with adult L2 learners. Moreover, the issue of whether "match" responses are faster than "mismatch" ones is far from resolved even in native speakers. Replicate one of the L2 studies using SPL with adult L2 learners of the same language background in order to investigate the following research questions: (a) Do the adult L2 learners process the input the same way as the L2 children? (b) Are there significant differences between the "match" and the "mismatch" responses in terms of reaction times? (c) Does the elimination of erroneous responses affect the listening time data of the L2 participants?

Suggested Readings

Felser, C., Marinis, T., & Clahsen, H. (2003). Children's processing of ambiguous sentences: A study of relative clause attachment. *Language Acquisition, 11,* 127–163.

Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the Auditory Moving Window technique. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 324–335.

Marinis, T. (2010). Using on-line processing methods in language acquisition research. *Experimental Methods in Language Acquisition Research* (pp. 139–162). Amsterdam, Netherlands: John Benjamins.

Papangeli, A. (2010). *Language development and on-line processing in L1 and L2 children* (Unpublished doctoral dissertation). University of Reading, UK.

References

Adobe Audition [Computer Software]. (2003). San Jose, CA: Adobe Systems.

Amvrazis, N. (2012). Anaphora resolution in near-native speakers of Greek. In S. Ferre, P. Prevost, L. Tuller, & R. Zebib (Eds.), *Proceedings of the Romance Turn IV* (pp. 54–81). Newcastle, UK: Cambridge Scholars Publishing.

Booth, J., MacWhinney, B., & Harasaki, Y. (2000). Developmental differences in visual and auditory processing of complex sentences. *Child Development, 71,* 981–1003.

Caplan, D., & Waters, G. S. (2003). On-line syntactic processing in aphasia: Studies with auditory moving window presentation. *Brain and Language, 84,* 222–249.

Chondrogianni, V., Marinis, T., & Edwards, S. I. (2010). On-line processing of articles and pronouns by Greek successive bilingual children: Similar or different from children with SLI? In K. Franich, K. M. Iserman, & L. L. Keil (Eds.), *Proceedings of BUCLD 34* (pp. 78–89). Somerville, MA: Cascadilla Press.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Clifton, C., Jr., Traxler, M., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language, 49,* 317–334.

Creative Media Source Player [Computer Software]. Milpitas, CA: Creative Technology Limited.

Dunn, J. (1994). SoundEdit (Version 2) [Computer Software]. San Francisco, CA: Macromedia.

Felser, C., Marinis, T., & Clahsen, H. (2003). Children's processing of ambiguous sentences: A study of relative clause attachment. *Language Acquisition, 11,* 127–163.

Ferreira, F., Anes, M. D., & Horine, M. (1996). Exploring the use of prosody during language comprehension using the auditory moving window technique. *Journal of Psycholinguistic Research, 30,* 3–20.

Ferreira, F., & Clifton, C. E. (1986). The independence of syntactic processing. *Journal of Memory and Language, 30,* 725–745.

Ferreira, F., Henderson, J. M., Anes, M. D., Weeks, P. A., & McFarlane, D. K. (1996). Effects of lexical frequency and syntactic complexity in spoken-language comprehension: Evidence from the Auditory Moving Window technique. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22,* 324–335.

Henderson, J. M., & Ferreira, E. (1990). The effects of foveal difficulty on the perceptual span in reading. Implications for attention and eye movement control. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16,* 417–429.

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning, 46,* 283–326.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension. Individual differences in working memory. *Psychological Review, 98,* 122–149.

Just, M. A., Carpenter, P. A., & Wooley, J. D. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General, 111*(2), 228–238.

Kidd, E., & Bavin, E. L. (2007). Lexical and referential influences on on-line spoken language comprehension: A comparison of adults and primary school children. *First Language, 27,* 29–52.

MacDonald, M. C., Just, M. A., & Carpenter, P.A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology, 24,* 56–98.

Marinis, T. (2007). On-line processing of passives in L1 and L2 children. In A. Belikova, L. Meroni, & M. Umeda (Eds.), *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition* (pp. 256–276). Somerville, MA: Cascadilla Press.

Marinis, T. (2008). On-line processing of sentences involving reflexive and non-reflexive pronouns in L1 and L2 children. In A. Gavarro & M. J. Freitas (Eds.), *Language acquisition and development: proceedings of GALA 2007* (pp. 348–358). Newcastle, England: Cambridge Scholars Publishing.

Marinis, T. (2010). Using on-line processing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental Methods in Language Acquisition Research* (pp. 139–162). Amsterdam: John Benjamins.

Marshall, C. R., Marinis, T., & van der Lely, H. K. J. (2007). Passive verb morphology: The effect of phonotactics on passive comprehension in typically developing and Grammatical-SLI children. *Lingua, 117,* 1434–1447.

Mazzoni, D., & Dannenberg, R. (2000). Audacity [Open Source Computer Software]. Retrieved from http://audacity.sourceforge.net/.

Papadopoulou, D., Plemenou, L., Marinis, T., & Tsimpli, I. (2007). *Pronoun ambiguity resolution: Evidence from adult and child Greek.* Paper presented at the Child Language Seminar, Reading, UK.

Papadopoulou, D., & Tsimpli, I. M. (2005). Morphological cues in children's processing of ambiguous sentences: a study of subject/object ambiguities in Greek. In A. Brugos, M. R. Clark-Cotton, & S. Ha (Eds.), *Proceedings of the 29th Annual Boston University Conference on Language Development* (pp. 471–481). Somerville, MA: Cascadilla Press.

Papangeli, A. (2010). *Language development and on-line processing in L1 and L2 children* (Unpublished doctoral dissertation), University of Reading, UK.

Papangeli, A., & Marinis, T. (2010). Processing of structurally ambiguous sentences in Greek as L1 and L2. *Proceedings of the Annual Meeting of the Department of Linguistics, Aristotle University of Thessaloniki, 30,* Greece.

Rayner, K., Carlson, M., & Frazier, L. (1983). The interaction of syntax and semantics during sentence processing: Eye-movements in the analysis of semantically biased sentences. *Journal of Verbal Learning and Verbal Behavior, 22,* 358–374.

Rayner, K., Garrod, S. C., & Perfetti, C. A. (1992). Discourse influences in parsing are delayed. *Cognition, 45,* 109–139.

Stavrakaki, S., &. Tsimpli, I. M. (2000). Diagnostic verbal IQ test for school and preschool children: Standardization, statistical analysis, and psychometric properties. *Proceedings of the 8th conference of the Panhellenic Association of Speech and Language Therapists* (pp. 95–106). Athens: Ellinika Grammata.

Waters, G. S., & Caplan, D. (2004). Verbal working memory and on-line syntactic processing: evidence from self-paced listening. *Quarterly Journal of Experimental Psychology A, 57,* 129–163.

Waters, G. S., & Caplan, D. (2005). The relationship between age, processing speed, working memory capacity and language comprehension. *Memory, 13,* 403–413.

4

EYE-TRACKING WITH TEXT

Gregory D. Keating

History of the Method

Eye-tracking is an experimental method that consists of monitoring and recording the eye movements that a person makes while performing a task that involves complex visual cognitive processing. Given that vision—and, therefore, oculomotor behavior—is involved in a myriad of human cognitive processes, eye-tracking stands apart from some of the other methods discussed in this volume in terms of the breadth of cognitive processes it can illuminate and in its applicability outside of academic research. In addition to its use in the study of reading (the topic of this chapter) and scene perception during spoken language comprehension (see Dussias, Valdés Kroff, & Gerfen, Chapter 5, this volume), both of which are studied by psycholinguists, eye-tracking has been used in fields as diverse as Communications and Engineering. Among its many applications to industry, eye-movement monitoring has been used to evaluate the effectiveness of online advertisements and has led to the development of systems to monitor driver and pilot fatigue. Nevertheless, the primary use of eye-tracking is to study the comprehension of written text (i.e., reading).

Interest in and documentation of oculomotor behavior during reading began as early as 1879 (see Huey, 1908), but the origins of modern day eye-movement research are best traced to seminal studies that emerged in the mid-1970s (for comprehensive reviews, see Rayner 1978, 1998). Psycholinguists working in the field of psychology have used the eye-tracking method for more than 40 years to study reading comprehension in college-aged readers of English. Relative to its longstanding history in that field, eye-tracking is a newcomer to the field of second language acquisition (SLA), having emerged only in the last 15 years or so. The first published eye-tracking study that examined the reading behavior of adult second language (L2) learners was Frenck-Mestre & Pynte (1997). Their study does not always receive the merit it deserves, in part because it appeared in a psychology journal and wasn't initially known to many SLA researchers, and also because on-line L2 processing research was largely nonexistent at the time, save for a pair of self-paced reading studies (Juffs & Harrington, 1995, 1996). However, their study ushered the L2 learner into the debates between structure-based versus lexically-driven parsing models that were ongoing in native language processing research, while also tackling issues of importance to mainstream SLA research, such as the role of the first language (L1) in L2 sentence comprehension, which had previously only been examined via off-line measures. In addition, given that Frenck-Mestre and Pynte's study examined L2 syntactic processing and the potential effect of lexical information on the same, its results are relevant to current accounts of L2 processing; namely, the shallow structure hypothesis (Clahsen & Felser, 2006) described in VanPatten's introductory chapter in this volume.

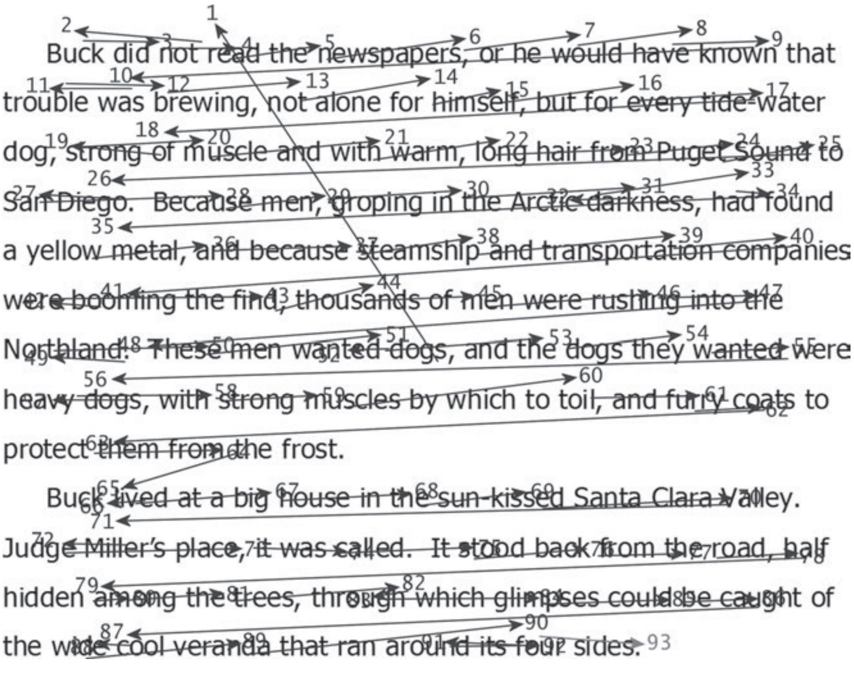
More than a decade passed before the first eye-tracking study of adult L2 learners appeared in a mainstream SLA journal in 2008 (Roberts, Gullberg, & Indefrey, 2008). Since then, published studies using the eye-tracking method to study linguistic behavior in adult L2 learners are on the uptick and by 2012 they numbered more than a dozen. The bulk of available studies come from the emerging subfield of L2 sentence processing research and all of them address, in one way or another, the question of whether native-like processing is attainable in an L2 that is learned in adulthood. Furthermore, the authors of most of these studies interpret their findings in light of the shallow structure hypothesis. Despite these broad similarities, the studies examine the processing of a wide variety of linguistic phenomena, including gender agreement (Foucart & Frenck-Mestre, 2012; Keating, 2009, 2010), *wh-* dependencies (Cunnings, Batterham, Felser, & Clahsen, 2010; Felser, Cunnings, Batterham, & Clahsen, 2012), pronoun resolution (Roberts, Gullberg, & Indefrey, 2008), binding relations (Felser & Cunnings, 2012; Felser, Sato, & Bertenshaw, 2011), attachment ambiguities (Witzel, Witzel, & Nicol, 2012), idioms (Siyanova-Chanturia, Conklin, & Schmitt, 2011), and liaison words (Tremblay, 2011). Given that eye-tracking provides a fine-grained measure of real-time language processing and is capable of detecting subtle differences between native and non-native language processing, it is likely that this method will continue to gain momentum in L2 psycholinguistic research.

What is Looked at and Measured

Before delving into how eye-movement data are used to make inferences about language behavior, it is helpful to introduce some key terms used to refer to eye movements and to review the basic characteristics of eye movements during reading, as documented in decades of research on mature, college-aged readers of English. The descriptive facts about eye movements detailed below are summarized from the many regular reviews of the eye-tracking technique published by Keith Rayner and colleagues (see, among others, Clifton, Staub, & Rayner, 2007; Rayner, 1998; Rayner, Juhasz, Pollatsek, 2005; Rayner & Pollatsek, 1989; Rayner & Sereno, 1994; Staub & Rayner, 2007).

Characteristics of Eye Movements

If you asked typical people to describe how their eyes move when they read, they'd probably say they move in a smooth, straight line from left to right—or from right to left, as the case would be for some languages—much like the sweeping movement that a highlighter makes when it is passed over the words of a sentence. In reality, the movement of the eyes across a sentence is much more turbulent. The eyes move in a series of jumps, called *saccades.* Saccades are separated by short periods during which the eyes remain relatively still, called *fixations.* Readers extract meaningful information about a text during fixations, but not during saccades. Saccades are rapid movements that are completed in 20 to 40 milliseconds (ms). The average saccade moves the eyes seven to nine characters, but saccades as short as one and as long as twenty characters are possible. In mature readers, 85 to 90% of saccades move the eyes forward in the text and the remaining 10 to 15% move them backwards. These backward, or regressive, saccades are called *regressions.* Regressions are usually the result of difficulties in comprehension or of errors in the programming of forward saccades that cause the eyes to overshoot their intended target. Figure 4.1 illustrates the typical pattern of saccadic activity observed in native readers of English. Each saccade is represented by an arrow that points in the direction of the endpoint of the saccade. The majority of the arrows point to



Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the find, thousands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

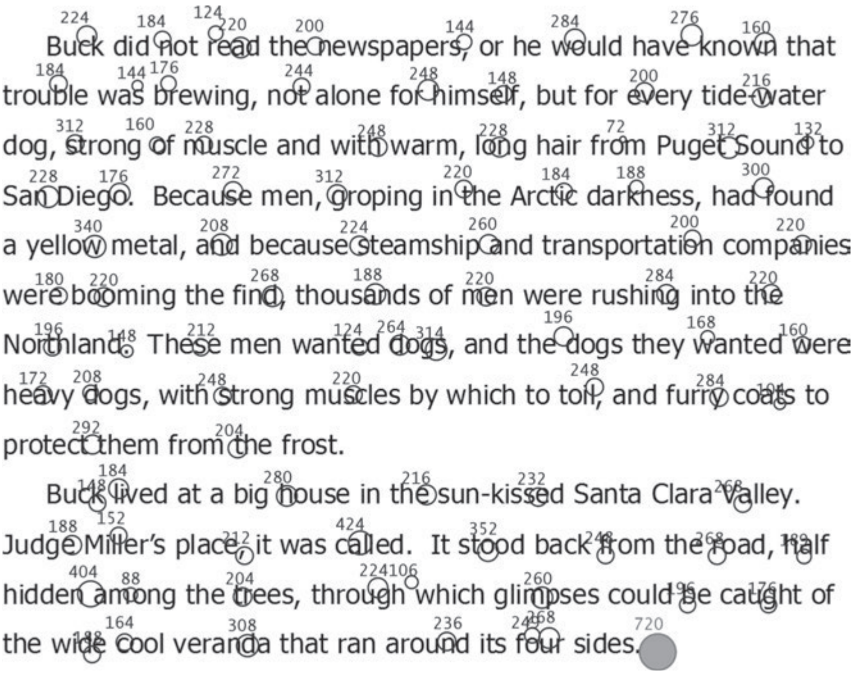
Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

FIGURE 4.1 Spatial overly of saccades made during reading. Adapted from the EyeLink Data Viewer User Manual Version 1.8.1, with permission from SR Research.

the right, indicating forward-moving saccades. Of the left-pointing arrows, most represent *return sweeps* (i.e., leftward movements necessary to move the eyes from the end of one line of text to the beginning of the next line of text), not regressions to reread words previously read.

Relative to saccades, fixations are longer and typically last one quarter of a second (200 to 250 ms), but range from about 50 to 500 ms. The location of the first fixation on a word usually falls between the beginning and the middle of the word. Figure 4.2 shows the fixations—depicted as hollow circles—that a reader made while reading a paragraph-long text. The accompanying numerals indicate the duration of each fixation (in ms).

The amount of useful information that a reader can extract from a text on a given fixation is called the *perceptual span.* Perceptual span is asymmetric in that readers (of languages that are read left to right) can perceive more information to the right of fixation than to the left. For example, the perceptual span of English readers extends up to 14 to 15 characters to the right of fixation, but only three to four characters to the left. However, although readers can make out letters that appear 14 to 15 characters downstream, they usually can't identify words that appear beyond seven to eight characters to the right of fixation due to anatomical limitations in visual acuity (i.e., the ability of the eyes to resolve detail).



Buck did not read the newspapers, or he would have known that trouble was brewing, not alone for himself, but for every tide-water dog, strong of muscle and with warm, long hair from Puget Sound to San Diego. Because men, groping in the Arctic darkness, had found a yellow metal, and because steamship and transportation companies were booming the find, thousands of men were rushing into the Northland. These men wanted dogs, and the dogs they wanted were heavy dogs, with strong muscles by which to toil, and furry coats to protect them from the frost.

Buck lived at a big house in the sun-kissed Santa Clara Valley. Judge Miller's place, it was called. It stood back from the road, half hidden among the trees, through which glimpses could be caught of the wide cool veranda that ran around its four sides.

FIGURE 4.2 Spatial overlay of fixations made during reading. Adapted from the EyeLink Data Viewer User Manual Version 1.8.1, with permission from SR Research.

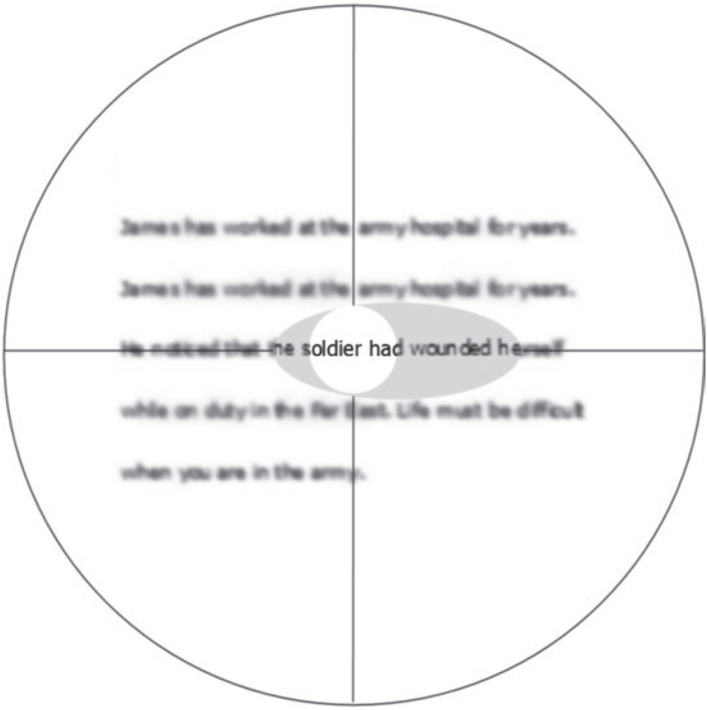

James has worked at the army hospital for years.
James has worked at the army hospital for years.
He noticed that the soldier had wounded herself
while on duty in the Far East. Life must be difficult
when you are in the army.

FIGURE 4.3 Perceptual span in normal readers of English. Adapted from Schuett, Heywood, Kentridge, and Zihl (2008). Text from Felser and Cunnings (2012).

Figure 4.3 illustrates the asymmetry in the perceptual span of adult English readers. An interesting fact about perceptual span that may need to be considered in L2 processing research is that it is influenced by attentional factors and the characteristics of writing systems. For example, the perceptual span of readers of languages read right to left, such as Hebrew, extends farther to the left of fixation than to the right (Pollatsek, Bolozky, Well, & Rayner, 1981), the opposite of English. In addition, perceptual span is smaller in native readers of languages that use character-based writing systems. Perceptual span in Chinese, which is read from left to right like English, is about one character to the left and two to three characters to the right of fixation (Inhoff & Liu, 1998). It is not yet clear what effects perceptual span might have on reading in an L2 when the direction of reading and/or the writing systems differ between the L1 and the L2, but it is reasonable to suspect that they might impact parafoveal processing (described below).

Given that perceptual span extends several characters beyond the point of fixation, it oftentimes includes not only the word currently fixated, but also the word immediately to its right (in a language that is read left to right). That is, when fixating a word, readers may also retrieve useful information about the upcoming word. Research shows that the word to the right of a fixated word is read faster when it is visible during a previous fixation compared to when it is not (Rayner, 1998). This reading time advantage is known as a *parafoveal preview benefit.* The term *parafoveal* refers to a particular range within the human field of vision and is used to refer to the portion of a text that is visible within that range. Reading researchers divide the text that is visible on a fixation into three regions. The *foveal* (central) region, indicated by the central white oval in Figure 4.3, includes the text within 1° of visual angle on either side of the fixation (about three to four characters to the left and right of fixation). A fixated word is in the foveal region. The *parafoveal* region, indicated by the grey ellipse in Figure 4.3, includes text within 5° of visual angle to either side of the fixation point. A word that appears immediately to the right of a fixated word likely lies in the parafoveal region, an area in which visual acuity declines dramatically. The *peripheral* region includes everything beyond the parafoveal region and is limited to spatial information about a text, such as where a line ends (Staub & Rayner, 2007).

The astute reader who has been keeping track of the facts about saccades, fixations, and human visual acuity has probably figured out an important fact about eye movements; namely, that a word can be perceived and processed in real time without being fixated. Words that do not receive fixations during reading are said to be *skipped.* Readers may skip a word when a programmed saccade propels the eyes too far forward in a sentence, or when a word is visible and identifiable in the parafovea. Word skipping is also influenced by factors intrinsic to words themselves. Content words are skipped less than 20% of the time, whereas function words are skipped between 60 and 80% of the time (Carpenter & Just, 1983). Word length also plays a role. Rayner and McConkie (1976) found that words two to three letters long are skipped about 75% of the time, whereas words eight letters and longer are rarely skipped. As we'll see in a later section of this chapter, the probability of word skipping is an important factor to consider in the design of an eye-tracking experiment.

The foregoing introduction to eye movements provides a solid foundation for understanding the remainder of this chapter, but it only scratches the surface of what is known about oculomotor behavior. For additional (and considerably more advanced) readings about eye movements and eye-tracking, interested readers are invited to consult a comprehensive volume on the topic, such as the *Oxford Handbook of Eye Movements* (Liversedge, Gilchrist, & Everling, 2011) or *Eye Movements: A Window on Mind and Brain* (van Gompel, Fischer, Murray, & Hill, 2007). In what follows, I discuss the measures that eye-tracking researchers use to establish links between eye movements and cognitive processes of interest to psycholinguists, such as anomaly detection, ambiguity resolution, and syntactic dependency formation.

Measures of Eye Movements

One advantage that eye-tracking is claimed to have over other on-line methods is that it more closely resembles normal reading outside of the laboratory (Clifton Jr. & Staub, 2011; for counterarguments, see Mitchell, 2004). Compared to the noncumulative moving window technique used in self-paced reading and the rapid serial visual presentation technique used in the event-related potentials paradigm (see Jegerski, Chapter 2, and Morgan-Short & Tanner, Chapter 6, this volume, respectively), in which a text is read one word or phrase at a time without the opportunity to view segments previously read, participants in eye-tracking studies view a text in its entirety and can reread words previously fixated. Eye-tracking provides two sources of data that researchers use to make inferences about cognitive processes engaged during reading: fixation times and regressions. Given that approximately 90% of reading time is spent in fixations (the remaining time spent in saccades), fixation times are the primary variable analyzed in eye-tracking studies. In line with the eye-mind assumption (Just & Carpenter, 1980), the time a reader takes to read a word is believed to reflect the time needed to process it. Words that are difficult to process will result in longer reading times and may induce regressions to words previously read. As a result, regressions constitute a secondary focus of analyses.

In addition to the purported naturalness of the method, a significant advantage that eye-tracking has over other on-line methods is its incredibly fast temporal resolution. For example, a 1000 Hz eye-tracker samples eye movements every millisecond, which makes eye-tracking one of the most sensitive measures of moment-by-moment processing currently available. The detail provided in the eye-movement record allows researchers to fractionate reading time in a *region of interest* (i.e., a particular word or series of words in a text) into different subcomponents. For example, the fixations that readers make the first time they view a region can be analyzed separately from fixations made when they reread a region, something that is not possible with other on-line techniques. To illustrate, suppose a researcher wants to examine whether non-native readers of English are sensitive to violations of subject-verb agreement in sentence pairs such as those in (1), in which version (1a) is the grammatical control sentence and version (1b) is the ungrammatical test sentence.

(1)

a. The students take classes in University Hall.
b. *The student take classes in University Hall.

Both versions of the sentence contain seven words, each of which constitutes a region of interest. The verb is the primary region of interest because it is the locus of the agreement anomaly. For this reason, the verb is called the *critical* region whereas all other regions are *noncritical* (i.e., not manipulated by the researcher to produce anomaly effects). If errors in subject-verb agreement result in processing difficulty, reading times on the verb should be longer in (1b) relative to (1a). Reading researchers would examine several measures to determine the time course of the effect of anomaly on reading. *First fixation duration* refers to the duration of the first fixation that a reader makes on a word. *Gaze duration* refers to the sum of the durations of all fixations that a reader makes on a word from first entering it from the left, until the word is exited to the right or the left. When the region of interest is more than one word, gaze duration is called *first-pass time.* In multiword regions, first-pass time is usually reported instead of—and not in addition to—first fixation duration, especially when the critical portion of the text (i.e., the anomalous or disambiguating information) may not be included in the first fixation (Clifton Jr., Staub, & Rayner, 2007).

To contextualize these measures, Figure 4.4 provides the sequence of fixations that a reader might make when reading the ungrammatical sentence (1b) above. Focusing on the critical verb *take,* the first time the reader's eye gaze enters the verb region that reader fixates it twice (fixations 3 and 4 in Figure 4.4) before exiting the verb to the left to reread *student.* In this example, first-fixation duration is 210 ms and gaze duration is 475 ms (210 + 265). If the reader had only made one fixation on the verb, first fixation duration and gaze duration would be equal. To avoid conflation of the two measures, researchers sometimes report the *single fixation duration,* which refers to the time spent on trials in which only one fixation was made on a word.

Some researchers report *regression path time,* also called *go-past time.* Regression path time begins with the first fixation on a word and includes all fixations that a reader makes until exiting the word to the right, including any time spent rereading earlier parts of the sentence if regressions were made. Extending the example depicted in Figure 4.4, this entails summing the reading times of fixations 3, 4, 5, and 6. After making the initial fixations of 210 ms and 265 ms on *take,* the reader regressed to reread *student* for 225 ms, and then made a forward saccade to reread *take* for an additional 300 ms before finally exiting the verb to the right to read *classes.* Regression path time on the verb is 1000 ms (210 + 265 + 225 + 300). The rationale behind this measure is that it reflects the time it takes the reader to process text before being ready to receive new information (Pickering, Frisson, McElree, & Traxler, 2004). One can envision how the anomalous verb in (1b) might impose the type of difficulty that prevents the reader from "going past" the verb until the difficulty is resolved.

The measures discussed so far emphasize the rich detail provided in the eye-movement record and highlight the types of fine-grained analyses that can be

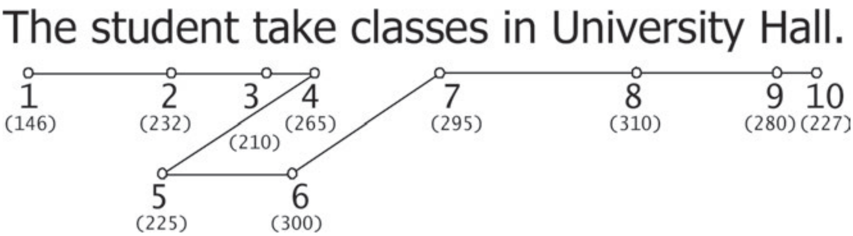


FIGURE 4.4 Sample fixation sequence for sentence (1b).

conducted, such as isolating first fixation durations from gaze durations. First fixation duration and gaze duration are thought to reflect word identification processes and one or both of these measures are usually affected by frequency, word familiarity, age of acquisition, number of meanings, and other factors (for a review of the findings, see Clifton Jr., Staub, & Rayner, 2007). However, it is important to point out that no measure of eye movement behavior has yet to be directly linked to a distinct phase of cognitive processing (Pickering, Frisson, McElree, & Traxler, 2004; Rayner & Liversedge, 2011). Nevertheless, researchers make a broad distinction between early versus late measures of eye movements. The measures discussed so far—first-fixation duration, gaze duration/first-pass time, and (sometimes) regression path (go-past) time—are deemed early measures of language processing and are believed to reflect readers' initial parses and interpretations of each word when it is first encountered. Regression path time is somewhat of a hybrid measure in that regressions out of a region to regions previously read reflect an early effect, whereas the additional reading time to overcome the difficulty is indicative of a later effect (Clifton Jr., Staub, & Rayner, 2007), the latter being a reason why some researchers consider it a late measure. If an effect is significant in regression path times but not in gaze durations/first-pass times, it is likely due to regressions to earlier parts of the sentence (Pickering, Frisson, McElree, & Traxler, 2004).

Researchers also look at a number of late measures, which are believed to be indicative of reanalysis that results from processing difficulty. *Total time* refers to the sum of the durations of all fixations made on a word, including all time spent rereading a word. Total reading time on the verb in Figure 4.4 is 775 ms (210 + 265 + 300). Some researchers also look at *second-pass time,* which is calculated in different ways. In some studies, second-pass time refers to the sum of the durations of all fixations made on a region after the region was exited and re-entered for the first time. In other studies, it refers to the sum of the durations of all fixations made on a word, excluding gaze durations/first-pass times (i.e., total time minus gaze duration/first-pass time). The latter calculation includes fixations made if a word is reread more than once. Second-pass time also goes by the more general term *rereading time.* Rereading time on the verb *take* in Figure 4.4 is 300 ms regardless of the method of calculation, given that the reader only re-entered the region once.

The foregoing discussion reviews the most common measures of fixation times used in eye-tracking studies with text. The example used to contextualize these measures focused on the reading times on one critical word, the verb *take* in (1b) compared to a verb in a control condition (1a). However, given that the eyes are usually ahead of the mind, the processing of one word can spill over onto the next. Accordingly, researchers use the aforementioned measures to analyze not only a critical region, but also the region immediately afterward, the so-called spillover region (i.e., the word *classes* in [1]).

In addition to reading times, some researchers look at the number or proportion of regressions that readers make into or out of a region. Suppose one wanted to assess the likelihood of a reader making a regression out of the critical verb in example (1) the first time it is read (i.e., during the gaze duration period). One could calculate the probability of making a *first-pass regression* in ungrammatical (1b) versus grammatical (1a) sentences by taking the number of sentences in each condition in which the reader makes a regression out of the verb and dividing that sum by the total number of sentences read in each condition. For example, suppose that participants in the hypothetical subject-verb agreement study read 24 sentences like (1), 12 grammatical versions like (1a) and 12 ungrammatical versions like (1b). If a reader made a first-pass regression out of the verb on eight of the ungrammatical sentences and three of the grammatical sentences, the probability of making a first-pass regression would be 67% (8/12) for the ungrammatical condition and 25% (3/12) for the grammatical condition. Given that processing of a critical region can spill over onto a postcritical region, it can also be informative to calculate the proportion of regressions back into a critical region (i.e., after having exited it to the right). Regressions into a region reflect delayed processing and are therefore considered a late measure, whereas first-pass regressions reflect early effects (Pickering, Frisson, McElree, & Traxler, 2004).

Issues in the Development and Presentation of Stimuli

Eye movement monitoring and recording requires specialized equipment. A typical eye-tracking system consists of two stations, one for the participant and one for the experimenter. The participant's station contains a desktop computer that displays the texts to be read, the eye-tracking device, and restraints or supports to limit head movement. Most tracking devices consist of a video-based infrared camera that tracks pupil movement. Although viewing is binocular (i.e., participants see and read the text with both eyes), reading studies typically record the movement of one eye only, usually the right eye. Placement of the infrared camera varies by model, with some models allowing multiple configurations, as shown in Figure 4.5, which highlights EyeLink models made by SR Research. Some cameras are placed on the desktop in front of the computer monitor and

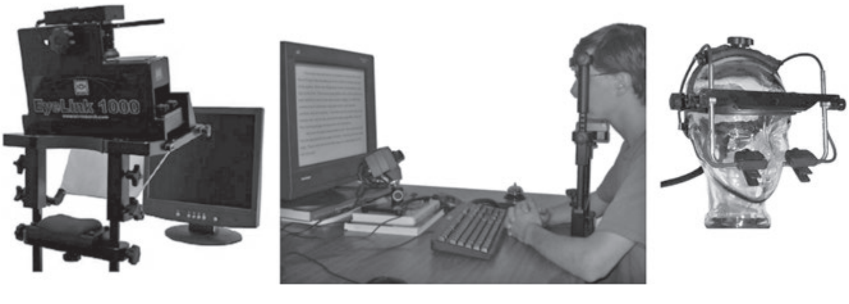


FIGURE 4.5 Tower-, desktop-, and head-mounted eye-tracking systems by SR Research.

angled up toward the participant's eyes. Others are mounted to a tower that sits above the participant's head. In other systems, miniature cameras are mounted to a lightweight headset that the participant wears during the experiment. In configurations in which the camera is stationary, it is often necessary to minimize head movements that could result in *track loss* (i.e., instances during reading in which the camera loses the pupil momentarily). Given that potentially informative reading time data are not captured when the pupil is lost during tracking, it is important to minimize track loss. This is accomplished via cushioned chin rests and head supports—and sometimes (less comfortable) bite bars—that are placed at the edge of the desktop. Nevertheless, some current eye-tracking systems can track reliably without any head stabilization.

The experimenter's station consists of a desktop computer that displays the text that the participant is currently reading. In addition, the experimenter's display contains a superimposed gaze cursor—a large dot that represents the participant's tracked eye—that mimics the precise movements of the reader's eye as it moves across the text in real time. The movement of the gaze cursor, along with other technical information provided on the screen, allows the experimenter to monitor the quality of the tracking on each and every sentence or passage read. If tracking becomes unreliable during the experimental session, the researcher can intervene at the next convenient point and make adjustments.

Despite the specialized equipment required of eye-tracking studies, the average participant is up and reading in less than 10 minutes. Setup includes adjusting the height of chairs and/or chin rests to obtain a comfortable seating position for participants and making basic camera adjustments. Once setup is complete, participants take a brief calibration test. In this test, participants are asked to follow a target with their eyes as it moves randomly to different locations on the screen, pausing momentarily at each one. The test calculates the correspondence between the reader's point of gaze on the screen and pupil position in the camera image. The calibration test is then repeated to validate adjustments made by the system to ensure the most accurate tracking possible during eye-movement recordings. Once calibration and validation are complete, the experiment proper begins.

Each trial in an eye-tracking study begins with a blank screen that contains a fixation target. To view a text, participants must look at the fixation target while simultaneously pressing a designated button on a hand-held controller. (An experiment can also be set up to allow the researcher to initiate the trial when s/he sees that the participant is fixated on the target.) The fixation target appears in the position that will be occupied by the first letter of the first word in the text to ensure that the reader's eye gaze is in the appropriate place when the text is displayed. Once a text appears on the screen, participants view it in its entirety and read it at their own pace. The text disappears when the participant presses a designated button to indicate that they have finished reading, or after a designated threshold of time has been reached. Some or all of the texts in an eye-tracking study are usually followed by a distractor task, such as answering a comprehension question. As with other methods discussed in this volume, the experimental session is preceded by a block of practice sentences to familiarize the participant with the procedure and to allow time for the participant to get accustomed to reading with the head stabilization mechanism in place.

The basic design principles that underlie the creation of critical stimuli and distractor tasks for an eye-tracking study are identical to those that one would adhere to using self-paced reading (see Jegerski, Chapter 2, this volume), except that one does not usually have to consider how to divide a text into readable segments. As in the case of self-paced reading, the following issues must be addressed in the development of stimuli for eye-tracking studies with text:

- When creating the different versions (i.e., conditions) of an experimental sentence (i.e., item), the critical and noncritical regions should be as close to identical as possible, as is the case in sample item (1) above, in which all words in version (1a) and (1b), including the critical verb, are identical, except for the preceding (noncritical) noun, which had to be different to create the agreement anomaly.
- The total number of items needed for a study usually ranges from 8 to 12 per condition, such that a study with two conditions like the subject-verb agreement study mentioned previously requires between 16 and 24 items.
- Experimental items don't usually exceed one third of the total items read in a study, meaning that two thirds to three fourths of the items should consist of distractors (items that target a specific linguistic variable that differs in some way from the one tested in the critical items) or fillers (items with no specific linguistic target).
- The different versions of each experimental item should be assigned to different presentations lists so that participants read just one version of each item (i.e., a participant in the hypothetical subject-verb agreement study should read either sentence 1a or sentence 1b, but not both, which means that at least two presentation lists are necessary).
- Once the presentation lists are created, the items within each list are pseudorandomized to ensure that items of the same condition do not appear consecutively (i.e., a reader in the subject-verb agreement study should not encounter two ungrammatical sentences in a row).
- To ensure that readers are paying attention to the meaning of the stimuli, most, if not all, critical and distractor items should be followed by a distractor task, such as a comprehension question (see Jegerski, Chapter 2, this volume, for a more in-depth discussion of distractor tasks).

In addition to the factors mentioned above, there are a couple of word-level variables that one ought to consider closely when designing the target and spillover regions of texts used in eye-tracking studies. One important factor is word length, which should be held as constant as possible between conditions. Word length is also important as it pertains to word skipping. In light of the fact that short words are likely to be skipped because they are processed parafoveally, short words may not elicit sufficient data to conduct meaningful statistical analyses. Figure 4.4 captures the likelihood of word skipping in that the preposition *in* does not receive a fixation, likely because it was read parafoveally while the participant's gaze was fixated on *classes.* Indeed, one potential advantage that the word-by-word self-paced reading paradigm has over eye-tracking is that readers are forced to fixate each region individually, which eliminates the possibility of parafoveal processing and reduces the amount of missing data that is present before trimming procedures are employed. This is not to say that eye-tracking cannot be used to examine the processing of articles, prepositions, or other categories of words that tend to be short, frequent, and processed parafoveally. However, researchers will have to be strategic in deciding how to define the critical region, perhaps by including words prior to or after the target word. In addition, some software packages, such as Charles Clifton Jr.'s PCEXPT package, contain algorithms that can calculate parafoveal preview effects from the reading times obtained from the word prior to the skipped word.

A second factor to control for is word frequency, for two reasons. First, high-frequency words are more likely to be skipped than low-frequency words. Second, low-frequency words are fixated longer than high-frequency words, even in the earliest measures of language processing such as first fixation durations and gaze durations (Rayner & Duffy, 1986; Inhoff & Rayner, 1986). Controlling for frequency across conditions will ensure that any obtained effects are not due to those trials that contained unusually high- or low-frequency words. In the context of example (1) above, one wants to be sure that a reader fixates longer on the ungrammatical verb relative to a grammatical one because of a violation of person-number agreement, not because the verb is infrequent regardless of its grammaticality.

Scoring, Data Analysis, and Reporting Results

In addition to the hardware described in the previous section, eye-tracking systems come with sophisticated software for storing collected data and for viewing it afterward. Video-based eye-tracking systems record a playable "movie" of each text a participant reads in a study. Each video displays the text that was read along with a superimposed gaze cursor that represents the participant's tracked eye. When the video is played, the gaze cursor reenacts in real time each and every saccade and fixation that the participant made while reading the text. The videos are exciting to watch and are useful for identifying unusual fixation patterns due to poor tracking or head movements. However, given that participants read up to 200 sentences in a study, and that a study may include dozens of participants per group, the videos are too time-consuming to watch and don't provide the crucial reading time data in a form that the researcher can use to easily conduct statistical analyses.

Generating a data report for a participant entails opening a participant's video-based data file in the data viewing software and selecting from an array of menus the various measures one wants to include in the report (e.g., first-fixation durations, gaze durations). The aims of the research will dictate which fixation time measures to choose and whether regressions into or out of a region will be informative. However, most researchers select a couple of "early" measures (first fixation durations, gaze durations, and regression path times) and at least one "late" measure (second-pass/rereading times and total times). Once the desired measures are selected, the software generates a report in the form of a .txt file that can be opened in Excel. The process is repeated for each participant.

The initial data reports contain more data than most researchers care to analyze. Although text is read in its entirety, it must be divided into segments for the purposes of conducting statistical analyses. The spaces between words can be used to parse a text into segments, or the researcher can define regions of interest during the programming stages (e.g., the three-word unit *in University Hall* in Figure 4.4 could be designated as one region of interest). Each region of the text appears in a separate row of the data file and the fixation time data for each region are displayed in columns to the right—a separate column for each variable included in the report. A researcher may only be interested in the fixation times of one or two regions, such as the verb and the spillover regions in example (1) above. Reporting descriptive statistics and the results of statistical analyses on noncritical regions is not as common in eye-tracking studies as in self-paced reading studies. Therefore, one step in data cleaning is to remove irrelevant rows of data from each file.

The data for each measure and region will contain some missing values (i.e., cells that lack reading times). Missing values in early measures such as first fixation duration, gaze duration/first-pass time, and regression path duration are left as missing and not replaced with another value under the assumption that readers perceived and processed the regions parafoveally. Track loss also results in missing values that are left unchanged. In contrast, missing values in second-pass/rereading times are usually replaced with zeros to reflect the fact that readers did not need to reread a region previously fixated. Furthermore, as with other on-line methods, it is common to remove data for those trials for which a participant answered the poststimulus comprehension question incorrectly.

After removing inaccurate trials, many researchers employ trimming procedures to purge the data of extreme values that could impinge on the results of the study. Although the average fixation lasts approximately 225 ms, words may receive fixations lasting less than 50 ms to more than 1000 ms. Readers are not believed to be able to extract useful information from a text on fixations lasting less than 50 ms (Inhoff & Radach, 1998). There are multiple ways to handle short fixations. If more than one fixation is made on a region, an automated procedure can merge the short fixation with a longer one according to a particular criterion (e.g., that the short fixation be within 1° of a longer one—that is, in the foveal region). However, this procedure needs to be executed before reports are generated in the data-viewing software. This procedure won't remove short fixations from regions that receive just one fixation lasting less than 50 ms. These data are usually removed and treated as missing values after the data report is generated. The precise cut-off for conducting either of the above-mentioned procedures varies from study to study, but usually ranges from less than 50 ms to less than 100 ms. In addition to short fixations, some fixations may be unusually long and may reflect processes other than those under investigation, such as fatigue or distraction. Here, too, the cut-off will vary from study to study, but many researchers remove fixations greater than 800 ms and treat them as missing values.

In addition to trimming values based on absolute cutoffs, many researchers screen eye-tracking data for outliers using the standard deviation method (separately by subjects and by items). Similar to self-paced reading studies (Jegerski, Chapter 2, this volume), values are usually deemed outliers when they are 2, 2.5, or 3 standard deviations above or below the participant's mean in a given condition. Outliers can be removed and treated as missing values, or substituted with an alternative value (e.g., the mean for the condition) so as not to lose statistical power.

Once data-handling procedures are complete, means are calculated for the analyses by subjects and by items. As with most other on-line reading techniques, ANOVAs and *t*-tests (conducted separately by subjects and by items) are the most common means of analyzing eye-tracking data. Given that eye-tracking yields multiple dependent measures, the number of analyses that a researcher runs on each region of interest is significantly larger compared to self-paced reading, which only yields one measure per segment read. For example, a researcher interested in the effect of anomaly on the critical verb in example (1) above, might run ten ANOVAs on the fixation time data alone: one ANOVA by subjects and one ANOVA by items for each of five dependent measures (first fixation duration, gaze duration, regression path time, rereading time, and total time). The same analyses would also be run on the data for the spillover region. So, at least 20 ANOVAs might be necessary before even considering regressions into or out of a region.

The most common way to display the results of eye-tracking studies is to provide mean fixation times and regression proportions (and standard deviations) in a table. Table 4.1 shows sample descriptive statistics for the hypothetical subject-verb agreement study described previously in this chapter. Included are data for a control group of native English speakers and an experimental group of ESL learners.

TABLE 4.1 Means and standard deviations for the critical verb

Measure	*Group*	*N*	*Condition*	*Mean*	*SD*
Gaze durations	Native	24	Ungrammatical	280	65
			Grammatical	222	62
	ESL	24	Ungrammatical	405	53
			Grammatical	411	50
Go-past times	Native	24	Ungrammatical	985	94
			Grammatical	642	79
	ESL	24	Ungrammatical	745	126
			Grammatical	721	101
Rereading times	Native	24	Ungrammatical	523	110
			Grammatical	292	101
	ESL	24	Ungrammatical	487	133
			Grammatical	462	117
First-pass regression proportions	Native	24	Ungrammatical	.38	.17
			Grammatical	.09	.08
	ESL	24	Ungrammatical	.19	.11
			Grammatical	.18	.10

An Exemplary Study

Research in the burgeoning subfield of L2 sentence processing has already uncovered interesting differences between native and non-native processing with the eye-tracking technique. An exemplary study is Felser and Cunnings (2012). Positioned as a test of the shallow structure hypothesis, this study addressed the claim that adult L2 learners are limited in their ability to make use of underlying structural information to interpret sentences in real-time and instead rely primarily on nonstructural cues. Specifically, this study investigated whether adult L2 learners' initial antecedent preferences for English reflexive pronouns such as *himself* and *herself* are guided by structural constraints (i.e., Binding Principle A) or by discourse-level constraints. Binding Principle A (Chomsky, 1981) captures the fact that anaphors such as *himself* and *herself* in English must be bound by the closest c-commanding antecedent. In (2), the reflexive pronoun *himself* must refer to *Paul* and cannot refer to *James.* Put differently, the only licit interpretation of this sentence is that Paul got cut.

(2) James says that Paul cut himself with a knife.

To examine whether L2 learners of English initially interpret these anaphors in accordance with Binding Principle A, Felser and Cunnings conducted two eye-tracking experiments that exploited a gender-stereotype violation diagnostic used previously by Sturt (2003) to examine the same phenomenon in native language processing. This review focuses on the first of Felser and Cunnings's two experiments. The experimental materials consisted of short texts, each of which consisted of three sentences, as in (3) below.

(3)

a. *Accessible Match, Inaccessible Match*

James has worked at the army hospital for years. He noticed that the soldier had wounded himself while on duty in the Far East. Life must be difficult when you are in the army.

b. *Accessible Match, Inaccessible Mismatch*

Helen has worked at the army hospital for years. She noticed that the soldier had wounded himself while on duty in the Far East. Life must be difficult when you are in the army.

c. *Accessible Mismatch, Inaccessible Match*

Helen has worked at the army hospital for years. She noticed that the soldier had wounded herself while on duty in the Far East. Life must be difficult when you are in the army.

d. *Accessible Mismatch, Inaccessible Match*

James has worked at the army hospital for years. He noticed that the soldier had wounded herself while on duty in the Far East. Life must be difficult when you are in the army.

The first sentence introduced a named referent (*James* or *Helen*) and established the scene for the mini-discourse. The second sentence began with a pronoun (*He* or *She*) that referred back to the main character of the first sentence, the discourse focus. The second sentence also introduced a new character using an occupational label (*the soldier, the nurse*) and a reflexive pronoun (*himself* or *herself*). Thus, each reflexive anaphor was preceded by two antecedents: a named referent (*James* or *Helen*) and an occupational NP (e.g., *the soldier*). The occupational NP was dubbed the "accessible" antecedent because it was the one allowed to bind the reflexive according to Binding Principle A, and the named referent was dubbed the "inaccessible" antecedent. The third sentence served as a wrap-up to complete the discourse.

As depicted in (3), each item appeared in four different versions in which gender congruency between the reflexive pronoun and the accessible and inaccessible antecedents was manipulated. In the "a" and "b" versions of items, the stereotypical gender of the binding-accessible antecedent *the soldier* matched the gender of the reflexive anaphor *himself.* The "a" version constituted a double-match in that the gender of the inaccessible antecedent (*James*) also matched the gender of the reflexive pronoun, whereas in the "b" versions it did not. In the "c" and "d" versions of items, the stereotypical gender of the binding-accessible antecedent *the soldier* violated the gender of the reflexive pronoun (*herself*). The "d" versions constituted a double-mismatch in that the gender of the inaccessible antecedent also mismatched that of the reflexive whereas in the "c" versions it matched. The reflexive anaphor was the primary region of interest and the two subsequent words (*while on*) constituted the spillover region.

Felser and Cunnings tested 28 native speakers of English and 25 adult ESL learners whose native language was German, a language that patterns like English with respect to the requirement of binding argument reflexives locally. Participants read 24 texts similar to those in (3), six in each of the four conditions shown. The 24 target texts were pseudorandomized among 56 filler texts and distributed across four presentation lists. Comprehension questions of the yes/no variety appeared after two thirds of all trials and never probed readers' interpretations of the reflexive pronoun. Participants' eye movements were recorded with a head-mounted tracking device. Because Felser and Cunnings were interested in the time-course of the application of binding constraints, they examined three early measures of language processing—first fixation durations, gaze durations/first-pass times, and regression path times—and one later measure: rereading time.

Repeated-measures ANOVAs conducted separately for each group—with the factors accessible antecedent (match, mismatch) and inaccessible antecedent (match, mismatch)—revealed that native English speakers' initial interpretations of English reflexives were guided by structural cues to coreference. First-fixation durations and regression path times on the reflexive anaphor were statistically faster when the reflexive matched the stereotypical gender of the binding-accessible antecedent (3a, b) compared to when it did not (3c, d). This effect also was manifest in their rereading times of the critical region.

By contrast, the early reading profile of the German ESL learners showed a pattern opposite that of the native speakers. First fixation and gaze durations on the reflexive anaphor were statistically faster when the *inaccessible* antecedent (3a, c) matched the gender of the reflexive compared to when it did not (3b, d). However, main effects of the accessible antecedent emerged in their rereading times, an indication that learners' applied Binding Principle A during real-time sentence comprehension, but did so at a later stage of processing. Additional support for delayed application of Binding Principle A was found in the German group's reading profile for the spillover region, which was virtually identical to the native speakers' reading profile for the critical anaphor. Whereas self-paced reading is sensitive enough to pick up the effect observed in the spillover region, eye-tracking is the only method sensitive enough to detect the delayed effect observed in the critical region because eye-tracking is the only on-line method that allows participants to reread text that was previously read. The observed differences between the native and non-native readers cannot be attributed to lack of knowledge of binding constraints. First, the binding of argument reflexives in the learners' L1 is constrained by Binding Principle A. Second, the non-natives performed in a manner indistinguishable from that of the native speakers on an off-line antecedent identification task designed to assess participants' knowledge of Binding Principle A. Felser and Cunnings interpret their results as reflecting qualitative differences between native and non-native processing in that native processing makes immediate use of structural constraints during initial syntactic processing, whereas L2 processing privileges nonsyntactic cues, and only later applies syntactic constraints.

Pros and Cons of Using the Method

Pros

- Eye-tracking studies closely resemble reading outside the laboratory. First, texts used in eye-tracking studies do not need to be divided into segments. This eliminates the critical guesswork in deciding how to segment a text so as to capture a potential effect without introducing task-specific confounds that might influence the results. Second, the words in the experimental texts do not disappear from view once read. Presenting texts in their entirety allows readers to engage in natural—and potentially informative—processes, such as rereading words previously read.
- Eye-tracking captures fine-grained details about moment-by-moment language processing. Compared to self-paced reading, which yields just one measure of reading time per segment, the detail captured in the eye-movement record yields multiple measures of processing behavior that are very informative with respect to the time-course of language processing.The exemplary study reviewed in this chapter illustrates just how informative time-course information can be in comparative L1-L2 processing research. What will certainly be debated is how to interpret delayed effects observed in L2 learners. Depending on the nature of the linguistic target and the theoretical stance taken on an issue, delayed effects could be taken as a quantitative difference between native and non-native processing or a qualitative one.
- Eye-tracking provides insights into how readers respond to processing difficulties. Once a reader detects the anomaly in subject-verb agreement in sentences such as "* *The student take . . .*," eye-tracking will show whether readers merely slow down on the verb only to continue reading, or whether they regress to reread the subject (presumably to check its person-number features). Different responses to anomalies could be indicative of different underlying processes, or they might reflect individual differences among readers. By contrast, self-paced reading is capable of detecting sensitivity to grammatical anomalies—usually in the form of increased reading times in the spillover region as opposed to the target region—but does not provide information about how readers respond to a difficulty in comprehension.
- Setup time is fast relative to other sophisticated on-line reading techniques. Setup time in an eye-tracking study is considerably shorter than in an ERP study, which involves precise placement of numerous electrodes on the scalp and practice in learning to read while minimizing blinking

(see Morgan-Short & Tanner, Chapter 6, this volume, for more information). On the other hand, setup time in a self-paced reading or listening experiment is probably a little faster than in an eye-tracking study.

Cons

- Eye-tracking is expensive. A high-resolution, video-based eye-tracking system appropriate for conducting eye-tracking studies with text (including hardware, software, setup, and training) costs in the neighborhood of $30,000 U.S. dollars (and some systems may cost much more). This is much less expensive than the equipment required of ERP studies, but significantly more expensive than what is needed to conduct self-paced reading studies. Furthermore, as the eye-tracking system ages it will require some maintenance and will eventually become outdated and need replacing.
- Eye-tracking is less efficient. Owing to the price of equipment, many eye-tracking labs are equipped with just one eye-tracking system. This means that many researchers can test just one participant at a time. By contrast, the software required to run self-paced reading experiments can be uploaded to multiple computers at a fraction of the cost of the original license by purchasing inexpensive run-time-only licenses (which are capable of running finalized experiments but can't be used for editing them or building new ones).
- Eye-tracking is less portable compared to other methods. Given the amount of hardware involved, eye-tracking equipment is difficult to transport safely and conveniently for offsite data collection.
- Participants sometimes need to be turned away. Quality eye-tracking is contingent on the infrared camera's ability to track the pupil. Although not a common occurrence, occasionally an otherwise eligible participant must be turned away when reliable tracking cannot be obtained. Before coming to the lab, participants should be told not to wear eye makeup. Eyeliner and mascara detract infrared light away from the pupil and are difficult to remove completely during an experimental session. Other causes of unreliable tracking are intrinsic to a participant's choice of corrective vision and cannot be changed. For example, eye glasses with small lenses whose edges intrude in the tracking area pose problems, as do hard contact lenses. Anatomical features can also influence the quality of eye-tracking. The pupil can be difficult to track in participants who blink excessively or who have occluded eye-lids. I once had to turn away a participant whose eye-lashes were so long, thick, and dark that reliable pupil tracking could not be obtained.
- Eye-tracking data takes somewhat longer to analyze. The trade-off of having multiple measures of processing behavior at one's disposal is that there is more data to analyze relative to what is obtained in a self-paced reading experiment, for example.

- Eye-tracking is not a covert measure of language processing. Participants enrolled in eye-tracking studies know their eye movements are being monitored, even if they don't know the precise aims of the research or anything about the eye-tracking technique itself. Some overly inquisitive participants will engage in unnatural reading strategies to try to determine the purpose of the study. The researcher can usually spot this during testing and make a note to discard the participant's data if the pattern is persistent.
- The amount of missing data can be higher than with other methods. In addition to the data lost from incorrect end-of-trial comprehension questions and data trimming, word skipping is an additional source of data loss.

Discussion and Practice

Questions

1) Of the fixation time measures typically reported in eye-tracking studies, which one is most analogous to the reading times obtained on a region of interest in the *noncumulative* self-paced reading paradigm? Which is most analogous to what would be obtained using the *cumulative* self-paced reading method (assuming the cumulative method was actually used)?
2) Assume a researcher wants to use the eye-tracking method to test learners' knowledge of English prepositions in sentences such as *John got <u>on</u> the northbound train* versus the odd alternative *John got <u>in</u> the northbound train*. What could the researcher do to minimize the likelihood of having little to no reading-time data on the critical preposition due to word skipping?
3) Is word skipping more of a concern in the data of native or non-native readers? Explain your answer.
4) Assume that you want to conduct an on-line processing study to determine whether non-native speakers of English are sensitive to the violations of subject-verb agreement depicted in example (1) of this chapter. How would the presentation of the stimuli, the types of measures collected, and the statistical analyses run on the measures differ between the three methods discussed so far in this volume (i.e., self-paced reading, self-paced listening, and eye-tracking)?
5) Compare and contrast the pros and cons of the three methods discussed so far in this volume. Which factors stand out as being most important to consider in choosing a method?

Research Project Option A

Much of the L2 processing research conducted to date has used the self-paced reading paradigm. Many of these studies, particularly those cited in support of the shallow structure hypothesis, show a null result for L2 learners (i.e., the L2 learners tested did not perform in a native-like way). One possibility, among others, is that self-paced reading was not sensitive enough to detect similarities between natives and non-natives. One project that would serve as a good introduction to the eye-tracking method while potentially making a relevant contribution to the field would be to replicate a self-paced reading study whose results suggest that native-like processing is unattainable.

Research Project Option B

Published studies of L2 processing using behavioral methods typically report the results of one on-line method (e.g., self-paced reading/listening, eye-tracking, or cross-modal priming). An interesting project that would make a solid contribution to the field of L2 processing would be to examine whether L2 learners' sensitivity (or lack thereof) to a particular linguistic phenomenon converges across different behavioral methods. Conducting such a study requires testing the same participants on the same linguistic phenomenon using two different techniques, such as self-paced reading and eye-tracking. It also requires double the number of stimuli so that participants do not read any version of the same sentence in both methods. As an alternative to this project, one could compare one behavioral method with one neurological one, such as ERPs or fMRI (see the chapters in this volume for relevant details about these methods).

Suggested Readings

Felser, C., Cunnings, I., Batterham, C., & Clahsen, H. (2012). The timing of island effects in nonnative sentence processing. *Studies in Second Language Acquisition, 34,* 67–98.

Keating, G. D. (2010). The effects of linear distance and working memory on the processing of gender agreement in Spanish. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing* (pp. 113–134). Amsterdam: John Benjamins.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27,* 251–272.

Tremblay, A. (2011). Learning to parse liaison-initial words: An eye-tracking study. *Bilingualism: Language and Cognition, 14,* 257–279.

References

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275–307). New York: Academic Press.

Chomsky, N. (1981). *Lectures on government and binding.* Dordrecht, Netherlands: Foris.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Clifton, C., Jr., & Staub, A. (2011). Syntactic influences on eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 895–909). Oxford, UK: Oxford University Press.

Clifton, C., Jr., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. P. G. van Gompel, M. H. Fischer, W. S. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 341–371). Oxford, UK: Elsevier.

Cunnings, I., Batterham, C., Felser, C., & Clahsen, H. (2010). Constraints on L2 learners' processing of *wh-* dependencies. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing* (pp. 87–110). Amsterdam: John Benjamins

Felser, C., & Cunnings, I. (2012). Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics, 33,* 571–603.

Felser, C., Cunnings, I., Batterham, C., & Clahsen, H. (2012). The timing of island effects in nonnative sentence processing. *Studies in Second Language Acquisition, 34*(1), 67–98.

Felser, C., Sato, M., & Bertenshaw, N. (2011). The on-line application of binding principle A in English as a second language. *Bilingualism: Language and Cognition, 12*(4), 485–502.

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66*(1), 226–248.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *The Quarterly Journal of Experimental Psychology, 50A*(1), 119–148.

Huey, E. B. (1908). *The psychology and pedagogy of reading.* New York: Macmillan.

Inhoff, A. W., & Liu, W. (1998). The perceptual span and oculomotor activity during the reading of Chinese sentences. *Journal of Experimental Psychology: Human Perception and Performance, 24*(1), 20–34.

Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 29–53). Oxford, UK: Elsevier.

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception and Psychophysics, 40*(6), 431–439.

Juffs, A., & Harrington, M. (1995). Parsing effects in L2 sentence processing: Subject and object asymmetries in *wh-* extraction. *Studies in Second Language Acquisition, 17*(4), 483–512.

Juffs, A., & Harrington, M. (1996). Garden-path sentences and error data in second language sentence processing. *Language Learning, 46*(2), 283–323.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning, 59*(3), 503–535.

Keating, G. D. (2010). The effects of linear distance and working memory on the processing of gender agreement in Spanish. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing* (pp. 113–134). Amsterdam: John Benjamins.

Liversedge, S. P., Gilchrist, I. D., & Everling, S. (Eds.) (2011). *The Oxford handbook of eye movements.* Oxford, UK: Oxford University Press.

Mitchell, D.C. (2004). On-line methods in language processing: Introduction and historical overview. In M. Carreiras & C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 15–32). New York: Psychology Press.

Pickering, M. J., Frisson, S., McElree, B., & Traxler, M. J. (2004). Eye movements and semantic composition. In M. Carreiras & C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 33–50). New York: Psychology Press.

Pollatsek, A., Bolozky, S., Well, A.D., & Rayner, K. (1981). Asymmetries in the perceptual span for Israeli readers. *Brain and Language, 14*(1), 174–180.

Rayner, K. (1978). Eye movements in reading and information processing. *Psychological Bulletin, 85*(3), 618–660.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 85*(3), 372–422.

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition, 14*(3), 191–201.

Rayner, K., Juhasz, B. J., & Pollatsek, A. (2005). Eye movements during reading. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 79–97). Oxford, UK: Blackwell.

Rayner, K., & Liversedge, S. P. (2011). Linguistic and cognitive influences on eye movements during reading. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 751–766). Oxford, UK: Oxford University Press.

Rayner, K., & McConkie, G. W. (1976). What guides a reader's eye movements? *Vision Research, 16*(8), 829–837.

Rayner, K., & Pollatsek, A. (1989). *The science of reading.* Englewood Cliffs, NJ: Prentice-Hall.

Rayner, K., & Sereno, S. C. (1994). Eye movements in reading: Psycholinguistic studies. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 57–81). San Diego, CA: Academic Press.

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition, 30*(3), 333–357.

Schuett, S., Heywood, C. A., Kentridge, R. W., & Zihl, J. (2008). The significance of visual information processing in reading: Insights from hemianopic dyslexia. *Neuropsychologia, 46*(10), 2445–2462.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research, 27*(2), 251–272.

(2007). Eye movements and on-line comprehension processes. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford, UK: Oxford University Press.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language, 48*(3), 542–562.

Tremblay, A. (2011). Learning to parse liaison-initial words: An eye-tracking study. *Bilingualism: Language and Cognition, 14*(3), 257–279.

van Gompel, R. P. G., Fischer, M. H., Murray, W. S., & Hill, R. L. (Eds.) (2007). *Eye movements: A window on mind and brain.* Oxford, UK: Elsevier.

Witzel, J., Witzel, N., & Nicol, J. (2012). Deeper than shallow: Evidence for structure-based parsing biases in second-language sentence processing. *Applied Psycholinguistics, 33*(2), 419–456.

5

VISUAL WORLD EYE-TRACKING

Paola E. Dussias, Jorge Valdés Kroff, and Chip Gerfen

Preliminary Issues

Our goal in this chapter is to describe the basic experimental design of a visual world study and to discuss the effects that researchers test for, including how these help address questions in second language processing. In part, we cover these points by illustrating two studies that have been particularly influential in the field: Allopenna, Magnuson, & Tanenhaus (1998) and Lew-Williams and Fernald (2007). Several chapters and articles have been written that extensively cover the visual world paradigm in depth (for thorough technical reviews concerning the paradigm as mainly applied to monolingual research, see; Altmann, 2011b; Huettig, Rommers, & Meyer, 2011; Tanenhaus, 2007; Tanenhaus & Trueswell, 2006). We also discuss some of the recent work in the second language (L2) literature to highlight how the method has helped researchers inform key issues in second language acquisition (SLA).

Before reviewing how to carry out a visual world study, we introduce four important core elements that will partially determine the decisions researchers make when designing visual world experiments. These decisions will depend greatly on a number of factors including but not limited to: the resources available to the researcher in terms of equipment, the population that the researcher wants to test (e.g., children versus adults), the sampling rate of the system, and the instructions to participants. All of these are likely to be dependent on the research questions that the researcher wants to address, while one relates to equipment. As discussed in Chapter 4 (Keating, this volume), an eye-tracking setup is considerably less expensive than an ERP setup, but more expensive than typical behavioral methods which require a single PC and perhaps a button box and/or a microphone and some software (e.g., self-paced reading, see Jegerski, Chapter 2, this volume). Eye-tracking systems vary in terms of the type of hardware they use and consequently in terms of the software necessary to develop an experiment and to extract and analyze data. Studies involving children (e.g., Snedeker & Trueswell, 2004), as well as those that utilize a paradigm referred to as the *looking-while-listening* paradigm (Fernald, Perfors, & Marchman, 2006), employ commercial video cameras. In one version of these studies, participants sit in front of an inclined podium, where a video camera is hidden beneath the podium. The podium has a hole in the center to allow the lens of the camera to focus on the participant's face (see Figure 5.1).

In each quadrant of the podium, there is a prop that is used by participants to perform certain actions. Participants hear a prerecorded command (e.g., "put the doll in the box) and are asked to perform the action. A second camera is placed behind the participant to record the actions and the location of the props. Using hidden cameras as a method of recording eye movements is desirable with small children because the method is not invasive, it is less expensive than commercially-available systems, and is more portable (Snedeker & Trueswell, 2004). An alternative to the basic video camera setup is to employ one of the many models of experimental eye-tracking systems, developed by a variety of companies, which come with computer eye-tracking algorithms to measure fixations. The most common are head-mounted, desk-mounted, and tower-mounted systems, although lately technological advances have been made toward the development

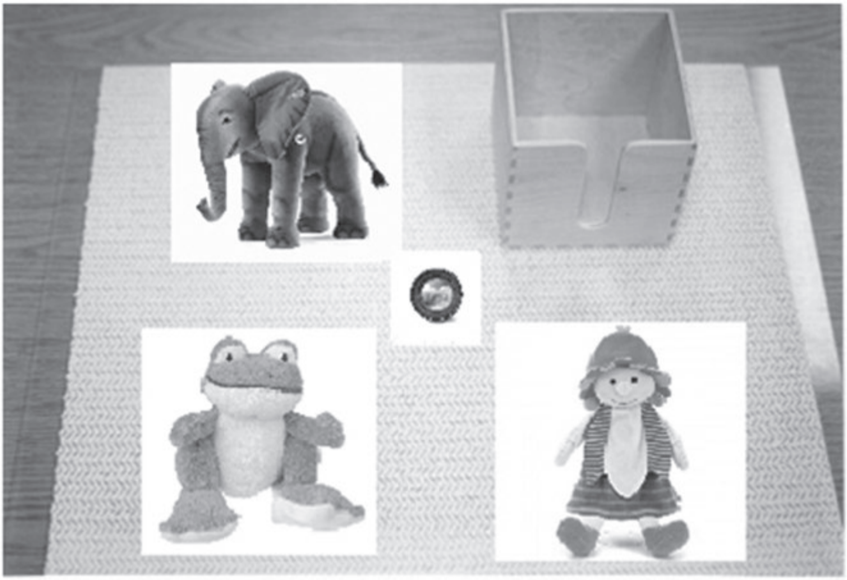


FIGURE 5.1 Sample visual scene using a hidden camera setup and real objects for an action-based experiment. Participants would hear a recorded stimulus such as "Put the doll in the box," and follow the instructions using their hands to manipulate the objects.

of eye-tracking goggles. Most language labs are likely to have an eye-tracker developed by Applied Science Laboratories, SensoMotoric Instruments (SMI), SR Research, or Tobii Technology. Head-mounted eye-trackers require that a participant wear a padded headband with miniature cameras mounted on the headband to record eye movements. This system requires more participant set-up and training time than other systems. Some eye-trackers are directly embedded into specialized computer monitors. These systems are generally more portable (although they can still be cumbersome) and require a less intense participant set-up. Eye-trackers also come as small desk-mounted or tower-mounted devices, generally set at a fixed position just below a computer monitor. These devices are easier to set up than head-mounted eye-trackers and do not produce the discomforts associated with head-mounted devices.

Cost restrictions, portability, and population of interest also influence the kind of eye-tracker used. As mentioned earlier, commercial video cameras are considerably cheaper than eye-trackers specifically designed for experimental research; however, they require intense manual data codification and extraction (explained in more detail below). In contrast, commercially available eye-trackers are generally more expensive, but come with experimental software specifically created to conduct eye-tracking research and with technical support staff who have expertise with the visual world and other research paradigms (e.g., the support group at SR Research). Commercial eye-trackers also allow experimenters to track the progress of a participant and make any necessary adjustments if the eye is not being properly tracked during the course of an experimental session.

A third issue to consider is the sampling rate of the system. Sampling rates determine the frequency with which a data point is recorded and are indicated in hertz (Hz) but are easily converted into time measurements by dividing the value into 1000 (milliseconds). For example, an eye-tracker with a sampling rate of 500 Hz records a data point every 2 milliseconds (1000/500 = 2) whereas an eye-tracker with a sampling rate of 1000 Hz records a data point every millisecond (1000/1000 = 1). In essence, there is a tradeoff between the sampling rate and the degree of movement that the participant can engage in. The higher the sampling rate, the more stable the participant's head must be. Because of this, eye-trackers with higher sampling rates are not particularly well-suited for young children. If a particular research question aims to investigate fairly subtle changes in the time-course of sentence processing, such as whether second language speakers are able to process in a native-like fashion a vowel contrast that does not exist in their native language (e.g. *bit* versus *bet* versus *beet* for Spanish learners of English), a higher sampling rate will be necessary to have sufficient data.

Finally, visual world studies differ with regard to the instructions that participants are given. These can be broadly defined as falling into two categories—action-based and passive listening (Tanenhaus & Trueswell, 2007). As the name implies, action-based instructions require participants to carry out an action related to the linguistic stimuli that they have just heard. Under this version of a

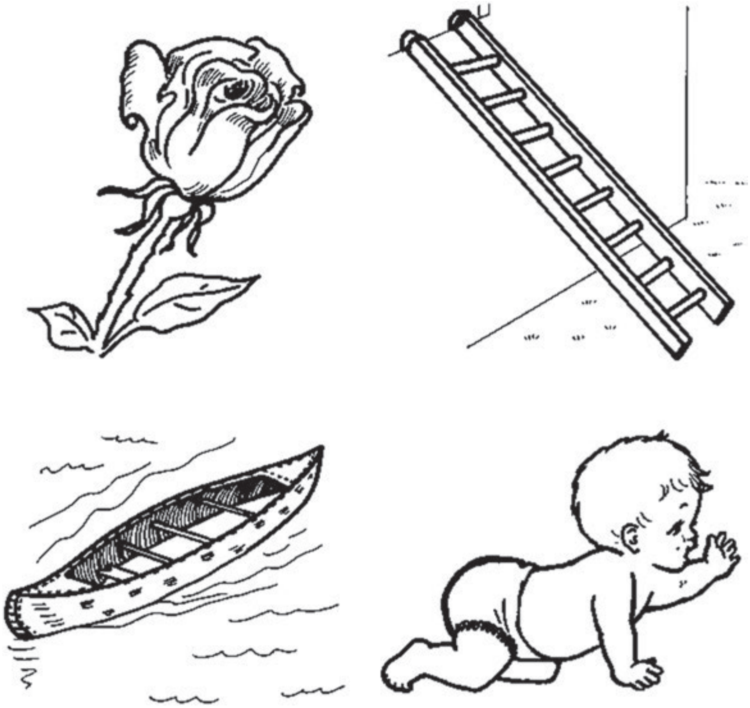


FIGURE 5.2 Sample visual scene employed in an action-based visual world experiment using a commercial eye tracking system and digital images. Participants would hear a recording such as "Click on the flower," and follow the instructions using the computer mouse.

visual world study, participants generally see a series of objects (animate or inanimate) either presented in realia (i.e., stuffed animals, plastic toys, or other types of props; Figure 5.1) or as images on a computer screen, as illustrated in Figure 5.2.

Objects are thus presented without any contextual visual information (i.e., no visual scene). In the simplest and most classic version of this design (e.g., Allopenna et al., 1998), participants hear a sentence like "*Put the doll in the box*" or "*Click on the flower*" and then carry out the action, either by physically moving the target object or by using a computer mouse while their eye movements are recorded. In contrast, in a passive listening task, objects are embedded within a contextually rich visual scene (see Figure 5.3). In one of the first studies to employ this design, Altmann & Kamide (1999) presented participants with a picture of a boy seated in a room surrounded by various objects including a cake and a toy car. Here, participants heard a sentence such as "*The boy will eat the cake.*" Instead of clicking on any named objects, participants were instructed to respond *yes* or *no* if the visual scene was compatible with what they heard. Participants responded both vocally and via a button box but, most importantly, the critical

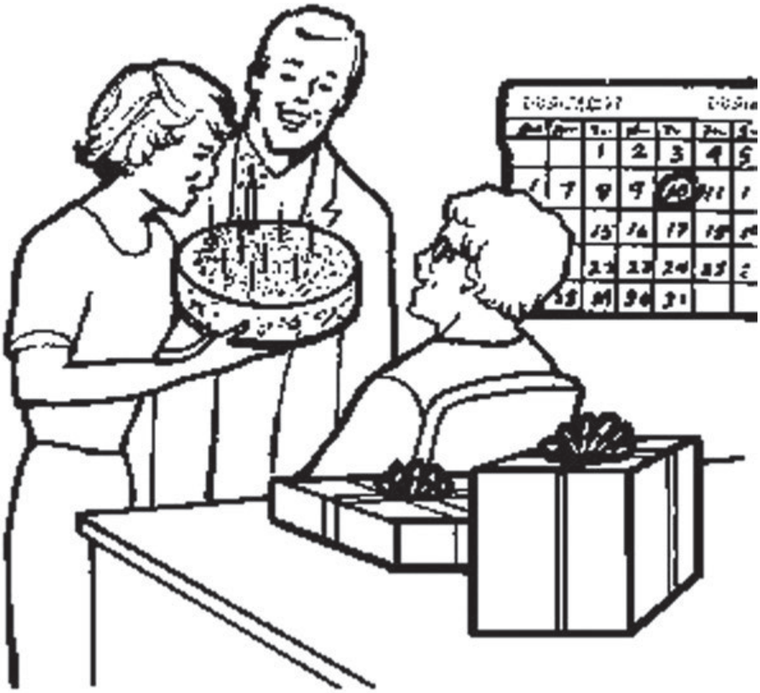


FIGURE 5.3 Sample visual scene employed in a passive listening visual world experiment. Participants would hear a recorded stimulus such as "The boy will eat the cake," and indicate whether the statement was consistent with the image via both a verbal yes/no response and a button press.

eye movements were those produced while they listened to the linguistic stimuli (i.e., before a response was made).

There are advantages and disadvantages to each approach but, broadly speaking, action-based tasks produce cleaner data (i.e., with less variation) because all participants are instructed to carry out the same task and therefore their eye movements follow a similar trajectory. During passive listening, participants are inspecting a context-rich visual scene that is closely linked to the linguistic stimuli that they are hearing. Because there are differences in the way in which individuals inspect visual scenes, data are more variable and may require more trials and/or participants in order to perform statistical analyses (Altmann, 2011b).

History of the Method

Cooper (1974) is widely cited as the first scholar advocating for the use of speech and a visual field containing objects semantically related to the speech signal to study real-time perceptual and cognitive processes. However, it was not until the publication of Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy (1995) that researchers took notice of the strong link between eye movements and comprehension. In that study, Tanenhaus et al. presented participants with one of two visual scenes, each containing four objects—for example, a towel, a towel with an apple on it, a pencil, and an empty box. Participants were asked to carry out a simple task by following an auditory instruction such as "*Put the apple on the towel in the box*." The auditory instruction is syntactically ambiguous at the point participants hear "towel" because it could be interpreted as the goal (i.e., move the apple to the empty towel) or as a modifier (i.e., move the apple that is on the towel to a to-be-named location). The ambiguity is resolved at the moment that participants hear the actual goal, "box." While listening to experimental instructions, participants' eye movements were recorded by way of a head-mounted eye-tracker. Tanenhaus et al. found that upon encountering the region of ambiguity, participants' eye movements toward the incorrect target location (e.g., the empty towel) increased. In other words, participants' eye movements reflected the local syntactic ambiguity, thereby confirming that eye movements are closely time-locked to the unfolding auditory signal. Tanenhaus et al. compared eye movements of this so-called *one-referent* scene to eye movements to a scene that included two apples, one on a towel, as in the previous scene, and one on a napkin, thus creating a *two-referent* scene. When participants were given the same instructions, looks to the empty towel (i.e., the incorrect target location) were significantly reduced and looks to the potential target location (i.e., the empty box) increased as compared to the one-referent scene. Thus, participants were more likely to interpret the towel as a modifier NP in the two-referent scene and subsequently anticipate the probable goal location. Tanenhaus et al. interpreted this finding as suggesting that participants were able to integrate contextual information (i.e., how many referents were present in a visual scene) as a relevant cue to modulate syntactic ambiguity. This at-the-time novel approach to understanding the spoken language processing of syntactic ambiguity proved highly fruitful in informing debates on syntactic modularity (Huettig et al., 2011).

The last decade has seen an impressive growth of experimental approaches using the visual world paradigm. This experimental paradigm has successfully been used to answer research questions related to virtually any area of monolingual language comprehension, including studies of subphonetic variation (e.g., Salverda, Dahan, Tanenhaus, Crosswhite, Masharov, & McDonough, 2007); phonological and phonetic processing, including effects of word frequency, cohort density, neighborhood density, lexical stress, and voice onset time (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; McMurray, Tanenhaus, & Aslin, 2002; Reinisch, Jesse, & McQueen, 2010); the influence of semantic and syntactic context on the activation of word meanings (e.g., Dahan & Tanenhaus, 2004); morphosyntactic processing (e.g., Lew-Williams & Fernald, 2007); early influence of multiple contextual cues during sentence processing (e.g., Chambers, Tanenhaus, & Magnuson, 2004; Snedeker & Trueswell, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, &

Sedivy, 1995; Trueswell, Sekerina, Hill, & Logrip, 1999); predictive processing and event representation (e.g., Altmann & Kamide, 1999; Altmann & Kamide, 2009; Kamide, Altmann, & Haywood, 2003); pragmatic inferencing (e.g., Engelhardt, Bailey, & Ferreira, 2006; Sedivy, Tanenhaus, Chambers, & Carlson, 1999); and linguistic relativity (e.g., Huettig, Chen, Bowerman, & Majid, 2010; Papafragou, Hulbert, & Trueswell, 2008). More recently, researchers have been able to extend the use of the visual world paradigm into the realm of language production and message generation to answer questions about temporal links between eye movements and speech planning (Bock, Irwin, Davidson, & Levelt, 2003; Griffin & Bock, 2000). The paradigm has also been used in research with nontraditional populations such as young children and individuals with aphasia (e.g., Trueswell et al., 1999; Snedeker & Trueswell, 2004, for children; Yee, Blumstein, & Sedivy, 2008; Thompson & Choy, 2009; Mirman, Yee, Blumstein, & Magnuson, 2011, for aphasic populations).

To our knowledge, the first study to use the visual world paradigm to investigate second language processing was Spivey and Marian (1999). The study examined the parallel activation of words in both the native and second language when participants heard a word in one language only (for related work, see Ju & Luce, 2004; Marian & Spivey, 2003a, 2003b; Weber & Cutler, 2004). L1 Russian, L2 English participants sat in front of a display that contained four real objects. The critical manipulation was the presence of an object in the visual scene which shared word onset between the two languages. For example, participants heard the Russian instruction *Poloji <u>marku</u> nije krestika* "Put the stamp below the cross", while viewing a display containing a stamp (the target), a marker (which bares phonetic similarity with *marku*—the Russian word for stamp), a keychain, and a quarter. The findings showed that participants made eye movements to the between-language phonological competitor (e.g., marker), suggesting that lexical items in both languages were activated simultaneously, despite the fact that only one language was being used during the experimental session. Furthermore, the results provided evidence that even a second language learned in adulthood can influence spoken language comprehension in the native language.

Compared to the study of native/monolingual language comprehension, relatively little work has been conducted using the visual world paradigm within the field of SLA (Blumenfeld & Marian, 2005; Dussias, Valdés Kroff, Guzzardo Tamargo, & Gerfen, 2013; Hopp, 2012; Ju & Luce, 2004; Lew-Williams & Fernald, 2007, 2010; Marian & Spivey, 2003a, 2003b; Perrotti, 2012; Spivey & Marian, 1999; Weber & Cutler, 2004, Weber & Paris, 2004). The paucity of SLA studies using the method is attributable in part to the relative infancy of the visual world paradigm relative to other experimental methods used in SLA research. However, the method's wide applicability is particularly appealing to SLA research. Specifically, the use of eye movements as an index of comprehension is a covert dependent measure with high ecological validity (i.e., it does not rely on overt responses such as button responses or linguistic judgments as the dependent measure). Its focus on spoken language comprehension (and production) with visual nonlinguistic input circumvents the need to depend on literacy as a means to test second language processing. Hence, the visual world paradigm has the potential to broaden the traditional scope of SLA research by including typically underrepresented research populations such as immigrant language learners who may not have full command of a second language in its written domain, yet successfully communicate in a second language.

What is Looked at and Measured

The visual world combines experimental designs typically employed in spoken language comprehension studies with eye-tracking, a research tool that is becoming increasingly ubiquitous at research institutions. The previous chapter introduced the different states associated with eye movements. To briefly summarize, although our experience as viewers suggests that eye movements proceed smoothly as we examine a visual scene, we are in fact engaging in a series of short, ballistic movements known as *saccades.* These saccades occur between moments of measurable stability known as *fixations.* Of course, individuals also engage in *blinks* during any given period of time. As in the case of eye-tracking studies that involve the reading of text, in visual world studies researchers are also interested in the fixations. Some researchers, particularly Altmann and colleagues, additionally argue that saccades should be measured and analyzed (Altmann 2011a, 2011b); however, this view is a minority position within the field. Because visual inspection of nonlinguistic stimuli is not as "linear" as in reading studies, in visual world studies researchers are more interested in the fixations that occur in aggregate over a timescale rather than in the different kinds of fixations that may occur over time (e.g., *first-pass fixation*, *regression*, *total fixation*; see Keating, Chapter 4, this volume).

To determine whether participants fixated on a particular object in a visual world study, *regions of interest* need to be defined amongst the different objects displayed in the visual scene. Generally speaking, in experiments that make use of experimental eye-trackers and visual scenes presented on a computer monitor, regions of interest are predetermined by way of the tools included in the experimental software. To illustrate, if a scene consists of a four-picture display of line-drawing objects (e.g., a canoe, a baby, a flower, and a ladder), the regions of interest would be defined by drawing geometric shapes around each object using the experimental software. It is important to note that these regions of interest are invisible to the participant. Because participants do not always directly fixate on the center of objects, regions of interest are drawn larger than the image itself. Researchers differ on how large they may draw the region of interest. One common practice is to draw square boxes that contain the entire image in use. Thus, for the image of the *canoe* in Figure 5.2, the region of interest would include the picture of the canoe as well as the surrounding white space. Another approach is to split the monitor into a grid and define each region as a quadrant (as in the case of a four-picture display) in the grid. Of these two stances, we lean more towards the first approach, as it allows researchers to label fixations that occur at a distance from the objects as "outside looks." These outside looks can capture very broadly the variability that may exist in the data (explained in greater detail below). Other researchers, however, are more conservative and draw the regions of interest to match the shape of the object itself (i.e., having a free-form canoe-like region drawn around the canoe).

Visual world designs that use moving video images or that make use of commercial video cameras present some unique challenges for determining where a participant is fixating. For video images, the issue is that the region of interest itself moves dynamically during the video. Some experimental software comes with tools that can define dynamic regions of interest; however, they track moving images with varying success and thus may necessitate labor-intensive check-ups on the part of the researcher to ensure that the region of interest is well-defined. For researchers who rely upon commercial video cameras, typically the camera is trained upon the participants' eyes and, therefore, the data coding relies more upon which direction the eye is looking. It is important for the participant to be seated at a fixed distance from the visual scene and for the distance between objects to be fairly large in order to increase the likelihood that an eye movement be associated with any particular region of interest. Absent any custom software, researchers must examine video recordings of these eye movements frame by frame (with a standard frame refresh rate of 33 ms) in order to manually code in which direction the eye is looking. This approach is highly labor-intensive and is best suited for large research teams in order to provide the necessary resources to code such data and to provide inter-rater reliability in the data coding.

Perhaps one of the most challenging aspects of a visual world study is understanding how the raw data extracted from experimental software or manually coded is then transformed into proportional data that is plotted as curvilinear time-course plots. Fixations at any given time and in any given region are binary and exclusive: either there is a fixation (1) or there is not (0), and if there is a fixation in Region A, then there cannot be a fixation in Region B. Thus, an individual's fixation record for any object in one experimental trial consists of a series of 1s and 0s for long stretches. Researchers differ in the exact protocol that they follow to extract eye-tracking data (largely dependent on the experimental software available to them). In Table 5.1, we illustrate a typical raw data file from our lab, generated by Data Viewer, a data-extraction software program provided as a part of the Eyelink system (SR Research).

At the moment of data extraction, a *sample report* is created containing the region of interest in which the eye has been tracked, and whether the eye is in a state of blinking or saccadic movement. If the eye is not blinking or launching a saccade, then it is counted as *in fixation*. The spreadsheet in Table 5.1 includes two main types of variables: those that identify the data and those that constitute measurements. The variables RECORDING SESSION LABEL, CONDITION,

TABLE 5.1 Sample spreadsheet of extracted eye-tracking data before conversion to proportion data

Recording session label	*Sample message*	*Cond*	*Timestamp*	*Trial label*	*Left interest area id*	*Left in blink*	*Left in saccade*	*Response*	*File*
Part01	SOUND _WORD ONSET	2	15477730	Trial: 1	1	0	0	incorrect	1
Part01	.	2	15477732	Trial: 1	1	0	0	incorrect	1
Part01	.	2	15477734	Trial: 1	1	0	0	incorrect	1
Part01	.	2	15477736	Trial: 1	1	0	0	incorrect	1
Part01	.	2	15477738	Trial: 1	1	0	0	incorrect	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Part01	.	4	15559754	Trial: 10	1	0	0	correct	1
Part01	.	4	15559756	Trial: 10	1	0	0	correct	1
Part01	.	4	15559758	Trial: 10	1	0	0	correct	1
Part01	.	4	15559760	Trial: 10	1	0	1	correct	1
Part01	.	4	15559762	Trial: 10	1	0	1	correct	1
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Part05	SOUND _WORD ONSET	1	21629970	Trial: 53	2	0	0	correct	2
Part05	.	1	21629972	Trial: 53	2	0	0	correct	2
Part05	.	1	21629974	Trial: 53	2	0	0	correct	2
Part05	.	1	21629976	Trial: 53	2	0	0	correct	2
Part05	.	1	21629978	Trial: 53	2	0	0	correct	2

TIMESTAMP, TRIAL LABEL, and FILE are identifying variables. Other variables are the measurement variables and include LEFT INTEREST AREA ID, LEFT IN BLINK, LEFT IN SACCADE, and RESPONSE. The measurement variables are composed of three columns related specifically to eye-tracking data (all of those beginning with LEFT) and one column, RESPONSE, associated with the behavioral response to a secondary task (i.e., clicking on a target item with a computer mouse). SAMPLE MESSAGE (column 2) represents the point in the recording session at which the data was extracted. During the programming of a visual world experiment, the experimenter must flag the specific moment (in milliseconds) in each sound file that marks the beginning of the critical region, which in this example was a noun onset. Thus the onset of each trial should simply state that the sample message was SOUND_WORDONSET. This column is normally included as a means to verify that the data have been extracted from the appropriate noun onset in each trial. Its inclusion in the sample report is not strictly necessary, but it is a useful step to ensure correct data extraction.

Returning to the sample data points in Table 5.1, the first five rows (labeled Part01) show data for Participant 1 (RECORDING SESSION LABEL) on Trial 1 (TRIAL LABEL) corresponding to a trial from Condition 2 (CONDITION). Turning to the specific eye-tracking columns, the participant's eye is tracked in Region 1 (LEFT INTEREST AREA ID) and is not blinking or in saccadic movement—as indicated by all 0 values in both LEFT IN BLINK and LEFT IN SACCADE. However, note that in the RESPONSE column, the trial is marked as "incorrect." This response indicates that the participant has clicked on the wrong item for this trial. Following established practice for unimpaired, native language participants, this trial would typically be excluded from the data analysis. The second set of five sample data points also comes from Participant 1 but now from Trial 10. This trial is identified as a trial representing Condition 4. Here, the participant's eye is tracked in Region 1, and the participant has identified the correct target item, thus the data will be included for data analysis. Note that the last two rows of this sample set indicate that the participant began to launch a saccade (indicated by the 1 value found in LEFT IN SACCADE). Therefore, these rows will not be counted as fixations. Finally, in the last set of five sample data points, the data come from Trial 53 from a different participant, Participant 5. This trial is another trial labeled "Condition 1." Here, the participant's eye is tracked in Region 2; none of the rows reveal any blinks or saccades and the participant has clicked on the correct item, so all of the data points will be included.

How do these binary values get converted into proportional data? Simply put, the conversion happens as an aggregate of all binary values for a given time point for each experimental condition. To illustrate, let's use a simple two-picture display, a design which results in two predefined regions of interest—a target region of interest and a distractor region of interest. According to our lab protocol, this setup further results in a third region for fixations that fall *outside* either region of interest. Let's suppose that there are 10 trials per condition and the data presented corresponds to the raw data file for Participant 1. In Condition A, at time 0 (onset of the temporal region of interest), the eye-tracking device records three fixations in the target region, four fixations in the distractor region, and three fixations falling outside either region. Then for this single time point, proportions are determined by dividing the number of actual fixations in any given region out of the total number of fixations observed. Thus, for the target region, the proportion of fixations is 0.3 (three fixations in the target region divided by 10 observed fixations); for the distractor region, the proportion of fixations is 0.4 (four fixations in the distractor region divided by 10 observed fixations); and for fixations to the outside region, the proportion of fixations is 0.3 (three fixations to neither target nor distractor regions divided by 10 observed fixations). Recall, however, that the eye may be in saccadic movement or blinking. These states would also be coded in the raw data file, which may result in less fixations for any given time point than the total number of possible fixations. To continue with our example, let's assume that in the same condition but at time 200 ms, Participant 1 had four fixations and two saccades in the target region, two fixations in the distractor region, one fixation in the outside region, and one blink on a trial. Following our lab protocol, the corresponding proportions would be calculated out of a total of seven fixations, resulting in the following proportions: 4/7 = 0.571 for the target region, 2/7 = 0.286 for the distractor region, and 1/7 = 0.143 for fixations outside of either region. Other labs may continue to calculate the proportions out of 10 *possible* fixations. Regardless of the method to determine proportions used, it is important to remain consistent throughout all calculations.

The description above should give a sense of the sheer amount of data that is being processed in a visual world experiment. Because of this, calculating proportions of fixations manually is not efficient. Depending on the experimental software, it is possible to extract fixations directly onto an Excel spreadsheet (or a text file which can later be opened in Excel). With some basic knowledge of Excel, *macros* can be created to calculate proportional data. One drawback of this approach is that with very large data sets, Excel can become slow to respond, may crash frequently, or may not have enough rows to accommodate the entire data set. Under these circumstances, the most convenient way of performing these calculations is via the use of R (R Development Core Team, 2008), an open-source statistical software package. Learning to use R can require a steep learning curve, as actions are carried out through command-line prompts. Despite this, the program has many benefits, as it can be used to calculate proportions, to generate time-course plots, and to perform statistical tests on the data. With sufficient expertise, a lab assistant can write scripts that are specific to any given experimental design, which make proportion calculation and data visualization easy. An adequate description of how to create these scripts is beyond the scope of this chapter, but Table 5.2 shows a sample subset of what a proportional data file looks like using R.

TABLE 5.2 Sample spreadsheet of proportional data aggregated over conditions and binned into 20 ms time bins

Subject	*Condition*	*Time*	*Prop target*	*Prop distractor*	*Prop nothing*
1	1	20	0.215053763	0.430107527	0.35483871
1	1	40	0.204081633	0.489795918	0.306122449
1	1	60	0.21978022	0.43956044	0.340659341
1	1	80	0.210526316	0.421052632	0.368421053
1	1	100	0.222222222	0.444444444	0.333333333
...	...	...	...	...	...
24	2	920	0.942408377	0.031413613	0.02617801
24	2	940	0.949238579	0	0.050761421
24	2	960	0.95	0	0.05
24	2	980	0.95	0	0.05
24	2	1000	0.95	0	0.05

The data file in Table 5.2 consists of three columns that identify each row: participant, condition, and time. These columns are followed by the proportional values. In our sample case, there are three columns containing proportional data, one for the proportion of fixations to target items (Prop Target), one for distractor items (Prop Distractor), and one for the proportion of fixations that fell outside of either region (Prop Nothing). Conceivably, the data could be arranged where the dependent measure (i.e., the proportional values) are all contained in one column, so long as a second column identifies the region of the proportional value. Notice that the timescale (Time) is in increments of 20 ms. This sample subset was extracted from an eye-tracker with a 500 Hz sampling rate. As discussed earlier, this sampling rate produces a data point every 2 ms. To make the data files more manageable and to prevent complications in the time-course plots caused by overlapping data points, data have been aggregated into 20 ms time bins. This is accomplished by taking the mean of proportional values that fall within each 20 ms time bin. Thus, in Table 5.2, the first row of data represents the mean of the first 10 proportions per region of interest.

Whereas the original Tanenhaus et al. (1995) study calculated the proportion of total trials on which participants looked at one region versus another plotted as simple bar graphs, contemporary studies also include time-course information. In time-course plots, total proportion of fixations over trials and participants are calculated and plotted over a millisecond timescale. Figure 5.4 is a sample graph of what a typical time-course plot looks like. The sample data are taken from a pilot study where a group of L2 English speakers whose native language was Spanish was asked to listen to variable sentences that named one of two objects presented on a computer screen (e.g., a visual display that showed a picture of a hammer on the left and a mug on the right while participants heard "*The man told his daughter to photograph the hammer on the table*"). The y-axis represents the total proportion of

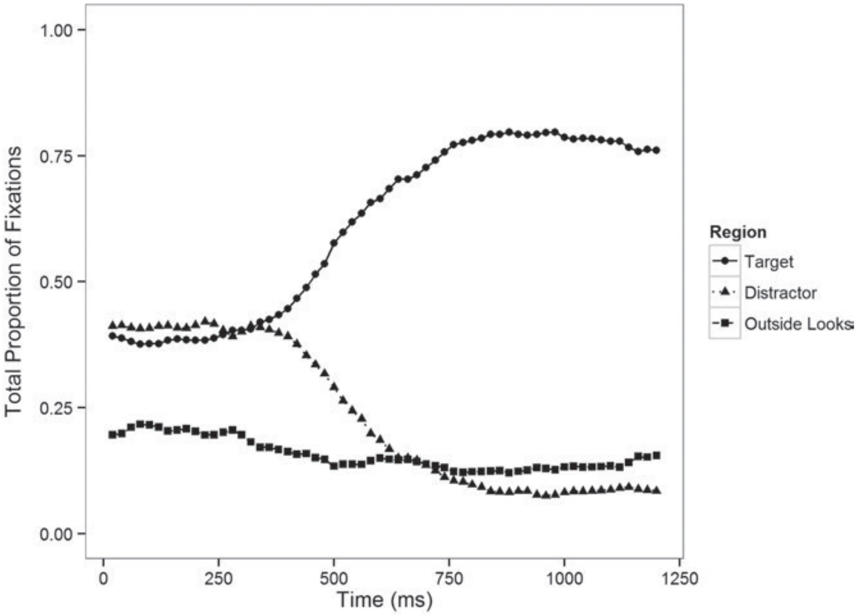


FIGURE 5.4 Sample time-course plot from a visual world experiment.

fixations; the x-axis plots time in milliseconds. Because the data are proportional, the y-axis is by definition bounded between the values of 0 and 1. Time at x = 0 typically represents the onset of a target region of interest. For this sample data, the region of interest starts at the onset of the noun pictured in the display (e.g., when participants hear "*hammer*").

This sample pilot study used a two-picture display; hence there is only one distractor. Many designs use four-picture displays; thus, a graph may show more lines than those shown in the sample graph. However, even in four-picture displays, some researchers simply aggregate distractors that are not important to the particular research question. In the sample plot, the fixations to the target noun are plotted as solid circles, fixations to the distractor noun are plotted as solid triangles, and any fixations that fell outside of either region are plotted as solid squares. From this plot, we can determine that fixations to both the target and distractor were equally prevalent until approximately 400 ms after the noun onset, where an increasing number of participants fixated on the target item. Fixations to target items continue to increase after that. One point to note from the sample plot is that fixations to target items never reach a proportion of 1, even if all participants fixate on the target item at some point during the region of interest. This somewhat counterintuitive observation results because individuals are idiosyncratic in how they will view a visual scene. Some people are faster to look at named objects than others; some individuals may be attracted to some sort of visual feature that is irrelevant to the linguistic information presented. Nevertheless, the time-course plot shows that a majority of individuals looked at the target item at some point while it was named in a majority of the trials. Although not necessarily standard in published time-course plots, it is useful to plot fixations to outside regions as a means to observe whether a majority of participants in fact fixate on any displayed objects. Here, the line that corresponds to fixations outside of the two pictures remains relatively low and stable, never going above a proportion of 0.2. This line thus represents a random factor—that is, individuals at any given point may be blinking, transitioning to another picture, or examining the visual scene in its entirety. However, if the proportion of fixations to outside regions were high and not stable, then the plot would indicate a more serious problem, such as individuals who strategically did not look at either object (explained in more detail below), a high amount of variability amongst all individuals, or even a failure to detect the eye on the part of the eye-tracking device.

Researchers using the visual world paradigm look for the presence (or absence) of *competitor* and *anticipatory* effects. Broadly, these effects are taken to reflect delayed or facilitated processing, respectively. In the illustrative pilot study introduced above, the experimental condition contained a two-picture display with phonological competitors. For example, the target item, a hammer, was paired with a picture of a hammock (instead of the picture of a mug). Both items overlap in phonology in the first syllable /hæm/. When compared to items that do not compete phonologically, such as when a hammer is displayed with a picture of a mug, studies have shown that participants take longer to identify target items (Allopenna et al., 1998). Compare the previous Figure 5.4 with Figure 5.5 (the legend is the same). Whereas Figure 5.4 shows clear divergence between target and distractor items roughly around 400 ms, in Figure 5.5, participants are showing consideration of the distractor item (i.e., the phonological competitor) much later in the time-course. Although some separation appears to happen between 400 ms and 600 ms, we do not see a reliable increase in divergence between target and distractor until after 600 ms. This later divergence is the *competitor* effect.

In contrast to the competitor effect, an effect is said to be anticipatory when eye movements to a target object are launched significantly before the presentation of the linguistic input that is predicted to initiate looks to that target. Suppose, for example, that we want to find out whether participants listening to Spanish premodifiers marked for gender (e.g., the definite article *el* and *la*) use gender information to facilitate the processing of upcoming nouns (e.g., Lew-Williams & Fernald, 2007). The experimental setup would consist of some trials in which a masculine-gendered object (micrófono/microphone$_{\text{MASC}}$) is presented alongside a feminine-gendered object (vela/candle$_{\text{FEM}}$). These are the different-gender trials. Proportion of looks to the target in different-gender trials are compared to proportion of looks to the target in same-gender trials—trials consisting of two same-gendered objects presented alongside one another (e.g., micrófono/microphone$_{\text{MASC}}$ presented next to zapato/shoe$_{\text{MASC}}$). An anticipatory effect occurs

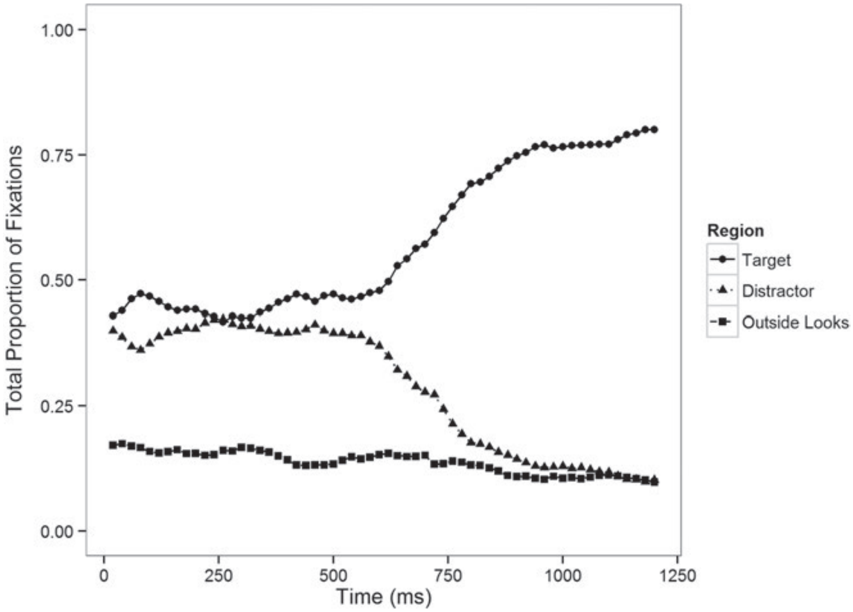
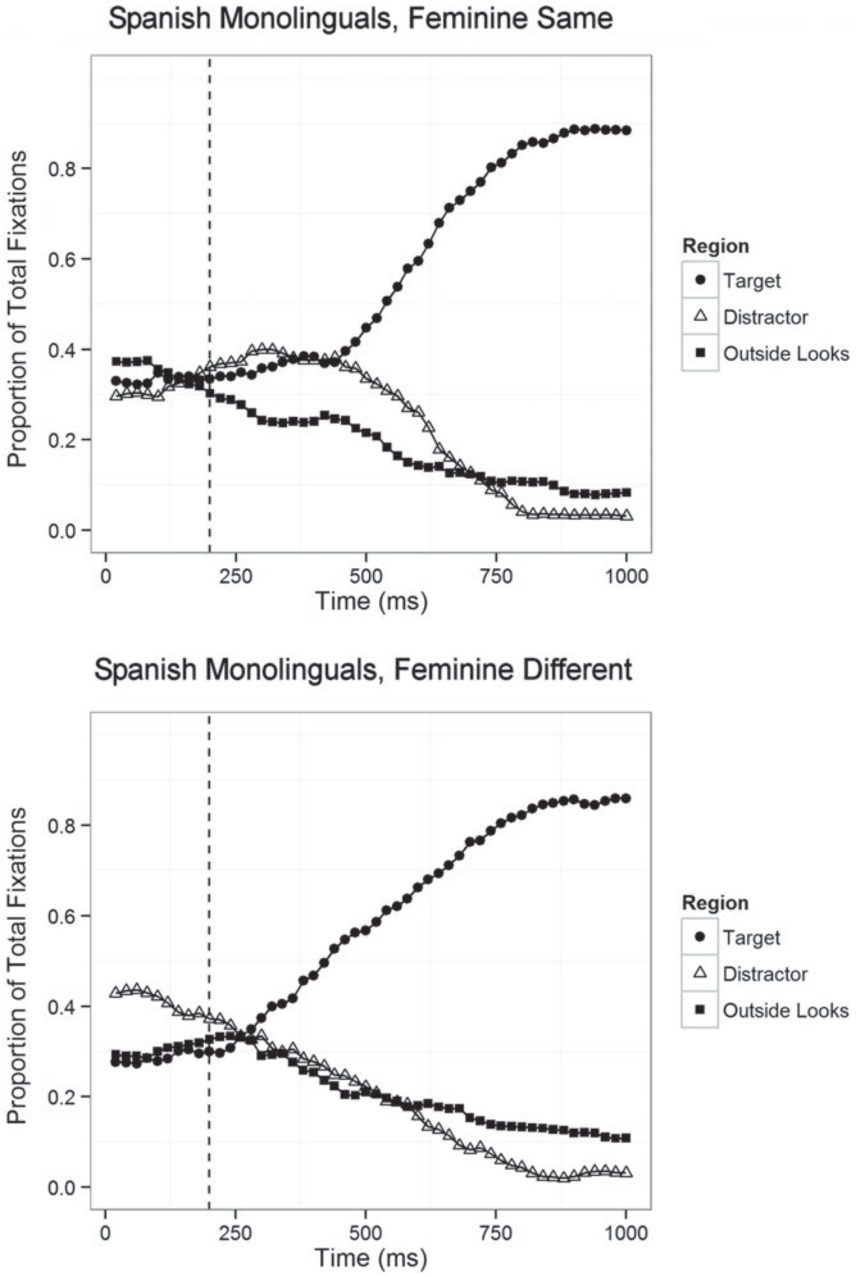


FIGURE 5.5 Sample time-course plot showing a competitor effect.

if after hearing *Encuentra la vela* "Find the candle," participants orient their eyes towards the target object *vela* more quickly on different-gender trials (i.e., when the gender information encoded in the article is informative) than on same gender trials (i.e., when participants need to wait to hear the named object before clicking on it). Figure 5.6 illustrates a time-course plot for an anticipatory effect. When comparing the panel on the top (Spanish Monolinguals, Feminine Different) to the panel on the bottom (Spanish Monolinguals, Feminine Same), we see that looks to targets on different-gender trials occur significantly before looks to targets on same-gender trials.

Although it is plausible to assume that competitor and anticipatory effects are opposite effects, in fact they are not. They are effects that can only be determined relative to a neutral baseline. For example, in the case of the Spanish grammatical gender described, if the gender information encoded in the article is not informative, the basic task is one of target word identification. This is precisely what goes on in the same-gender trials, and therefore, these trials constitute the *neutral baseline.* The effect of interest is whether the gender information present on the definite article will affect the time-course of a target relative to the neutral baseline. As described above, in the case of Spanish, it does and does so by quickening the time-course, hence the presence of an anticipatory effect.

A final potentially confusing point is the means by which researchers determine whether a competitor or anticipatory effect is present. In the Allopenna et al.


Spanish Monolinguals, Feminine Same

Spanish Monolinguals, Feminine Different

FIGURE 5.6 Sample time-course plots showing an anticipatory effect.

(1998) study, a competitor effect was determined relative to the cohort distractors that were copresent in the same visual scene. That is, the time-course of fixations to the target items was compared to the time-course of fixations to a phonological cohort, a rhyme cohort, and a nonphonological control. A competitor effect was determined by statistically comparing the distractor proportion of fixations to the proportion of fixations of the target item. In the case of the phonological cohort distractor (e.g., *beetle* for target word *beaker*), the proportion of fixations to the two items was not statistically different until roughly around 400 ms from the onset of the target noun. Nevertheless, the overall time-course for the target item was ultimately different from the time-course plot of all other distractor candidates, which also looked different from each other, meaning they all affected spoken word recognition in different ways. In contrast, Lew-Williams and Fernald (2007) compared the time-course of proportion of fixations only to target items but in separate conditions. In other words, an anticipatory effect was determined because a shift in a critical mass of looks to the target item happened faster in the different gender trials than in same gender trials. However, both trials had a similar time-course plot overall.

Visual World Studies and SLA

A few studies are beginning to emerge which use the visual world paradigm to ask longstanding questions in second language acquisition regarding the degree to which adult second language speakers recruit different types of information during real-time L2 comprehension. A recent study by Lew-Williams and Fernald (2010; see also Hopp, 2012) investigated the integration of L2-specific morphosyntactic information during on-line processing by asking whether adult L2 speakers of Spanish use grammatical gender encoded in definite articles to facilitate the processing of upcoming nouns. In a series of experiments modeled after Lew-Williams and Fernald (2007), they presented L2 learners of Spanish (L1 English) with two-picture visual scenes in which the pictured objects either matched or differed in gender. Participants heard instructions that asked them to find an object. In three experiments, they showed that when listening to sentences naming both familiar and newly-learned objects, native speakers were able to orient their eyes towards target objects more quickly on different gender trials (i.e., when the gender information in the article was informative) than on same gender trials, showing an anticipatory effect. L2 speakers of Spanish, on the other hand, waited to hear the noun to initiate a gaze shift. These findings suggested that non-native listeners were not able to integrate abstract gender information during on-line processing the way that native speakers did.

Other studies have examined transfer effects from the L1 to the L2. For example, Weber and Paris (2004) investigated whether the gender of words in the L1 exerts an effect on the recognition of words spoken in the L2. In this study, the eye movements of L1 French speakers, who had acquired German as a second language during adulthood, were monitored while they heard spoken instructions in German to click on pictures displayed on a computer screen (e.g., *Wo befindet sich die Perle* "Where is the pearl?"). The critical manipulation included target and distractor pictures in German that either shared phonological onset and grammatical gender with French words, or that shared phonological onset with French words but differed in grammatical gender. For example, in the same-gender pairs, the feminine German target *perle* "pearl" (*perle* in French, also feminine) was paired with the feminine German distractor *perücke* "wig" (*perruque* in French, also feminine). In the different-gender pairs, however, the feminine German target *kassette* "cassette" (cassette in French, also feminine) was paired with the feminine German distractor *kanone* "cannon" which was masculine in French (*canon*). Weber and Paris found that when target and competitor pairs were the same gender in German but were different gender in French (i.e., different-gender pairs), fixations to the competitor picture were reduced compared to when target and competitor pairs were the same gender in both German and French. What these results suggest is that for L1 French speakers who had acquired German as adults, the grammatical gender of French nouns modulated the processing of determiner + noun combinations in their L2, German.

The visual world paradigm has also been used to answer critical questions about the brain's ability to accommodate multiple languages, lexical activation and competition, and mechanisms of language (non)selectivity (see, for example, Blumenfeld & Marian, 2005; Canseco-Gonzalez, Brehm, Brick, Brown-Schmidt, Fischer, & Wagner, 2010; Cutler, Weber, & Otake, 2006; Ju & Luce, 2004; Marian & Spivey, 2003a, 2003b; Spivey & Marian, 1999; Weber & Cutler, 2004). These studies are not discussed here because they are not centrally related to current issues in SLA.

Issues in the Development and Presentation of Stimuli

A number of important decisions need to be made when designing visual world experiments. The first deals with the selection of the visual display. Depending on the research question, displays can vary, consisting of black-and-white line drawings or colored pictures of concrete objects displayed on a computer screen (e.g., Allopenna et al., 1998; Perrotti, 2012; Weber & Paris, 2004), arrays of objects laid out in a work space (e.g., Spivey & Marian, 1999; Snedeker & Trueswell, 2004), or line drawings or colored drawing of semirealistic scenes (Altmann & Kamide, 1999; Griffin & Bock, 2000; Arnold & Griffin, 2007).

The obvious advantage of black-and-white line drawings is that there are large repositories of pictures that have been normed for naming agreement, familiarly, and visual complexity with native-speaking children and adults as well as with some groups of L2 learners. Among these are the pictures employed in the *Boston Naming Test* (Kaplan, Goodglass, Weintraub, & Segal, 1983). The test contains 60 line drawings, graded in difficulty from easy and high frequency (e.g., *bed*) to difficult and low frequency (e.g., *abacus*), which have been normed in English and Spanish. Three other popular sources are the set normed for adults by Snodgrass and Vanderwart (1980), which contains 260 pictures normed for name agreement, image agreement, familiarity, and visual complexity, the English version of the *Peabody Picture Vocabulary Test* (Dunn & Dunn, 1997) and its Spanish equivalent, the *Test de Vocabulario en Imágenes Peabody* (TVIP; Dunn, Dunn, & Arribas, 2006). One source of line drawings that has proven to be extremely useful is the *International Picture Naming Project* (Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., . . . Bates, E., 2004; available at http://crl.ucsd.edu/˜aszekely/ipnp/). The website provides access to 520 black-and-white drawings of common objects and 275 concrete transitive and intransitive actions. A special feature of these drawings is that they have been normed in seven languages (English, German, Mexican Spanish, Italian, Bulgarian, Hungarian, and Mandarin) and with children and adults.

In the context of visual world experiments, the most significant drawback with black-and-white line drawings is that it can sometimes prove to be difficult to find pictures with the desired linguistic characteristics across a bilingual's two languages. It is because of this that many bilingualism researchers resort to other sources. Colored line drawings and photographs of real objects are the most popular. Colored line drawings are available through many sources, including IMSI MasterClips Image Collection (IMSI, 1990), and pictures of real objects abound over Google Images. One advantage of color images is that recognition of the objects is enhanced by color and texture information present in the visual display. Therefore, norming for naming agreement is greatly facilitated. A second advantage is that color images can be manually manipulated using Microsoft Paint (or a similar paint tool) to remove distracting patterns, to crop an image such that the target item is centered or to adjust the size of an image. One potential disadvantage is that visual complexity varies greatly among pictures, which can influence eye gaze.

A second important issue to consider is the position of the objects in the visual scene. Readers of left-to-right languages like English and French have a left-gaze bias (i.e., when presented with a visual scene on a computer screen, they will first direct their gaze to the upper left corner). Therefore, it is important to counterbalance the position of each item on the visual display. For example, if in one presentation list a target picture appears on the left side position and a distractor picture on the right side position of the computer screen, then a separate experimental list should have these positions reversed. It is important to note that counterbalancing of this sort results in double the number of experimental lists.

Studies have shown that eye movements generated during the recognition of a spoken word are mediated by a number of factors. It is important to be aware of these to avoid confounds in the experimental design. We have already mentioned phonological overlap as one factor (Allopenna et al., 1998). Eye movements to a display containing the picture of a *beetle* (target), a *beaker* (phonological competitor), a *speaker* (a rhyme competitor) and a *carriage* (unrelated word) generate more eye movements to the phonological competitor upon hearing the referent's name than to either of the distractor pictures in the visual display. Importantly, phonological competition is modulated by frequency effects. Dahan et al. (2001) found that when the picture of a referent such as *bench* (the target picture) was presented along with a high frequency phonological competitor (*bed*) and a low frequency phonological competitor (*bell*), a participant's early eye movements were equally likely to the target word and to the high frequency phonological competitor. The authors also found that the frequency effects during the earliest moments of lexical access could be obtained even when none of the competitors were phonologically similar to the target. Fixation latencies to targets with high-frequency names (*horse)* were shorter than those to targets with low-frequency names (*horn*). Structural similarity of the pictures is another variable. Dahan and Tanenhaus (2005) showed that when participants heard the words *snake* being spoken in the presence of a display containing a snake (target), the picture of a *rope* (visual competitor), an *umbrella,* and a sofa (two distractor objects), participants were more likely to fixate the visual competitor *rope* than either of the distractor objects.

One last issue that requires careful attention is the preparation of sound files. Because eye movements to the objects are closely time-locked to the unfolding speech signal, spoken instructions need to be carefully controlled at the critical region of interest in order to facilitate data analysis. To illustrate, the creation of sound files similar to those employed in the grammatical gender processing studies described earlier would require a number of steps. First, a speaker is asked to record several versions of a simple invariant carrier phrase (e.g., *Encuentra la/ Encuentra el* "Find the" masculine/feminine). Typically, the speaker repeats each carrier phrase in a normal declarative intonation between five to ten times. From these, measurements of the duration of each definite article are taken and an average length for each article is calculated. Subsequently, the best carrier phrase is selected and the duration of the article within each carrier phrase is matched, using Praat or other similar type of software (e.g., the duration of the definite articles *el* and *la* are matched so that each is, say, 200 ms long). Finally, the same speaker is asked to name each experimental stimulus five times. From each set of repeated tokens, the best exemplar is selected to be inserted into the appropriate carrier phrase. This procedure avoids the presence of coarticulation information in the article, which is known to influence noun recognition.

Scoring, Data Analysis, and Reporting Results

There is currently no consensus on how best to analyze eye-tracking data collected from visual world studies. Part of the issue is that generally the dependent measure in visual world studies is total proportion of fixations (although see e.g., Altmann, 2011a, for analyses done with saccadic measures). As a consequence, the dependent measure is bounded between 0 and 1, unlike other dependent measures typically used in behavioral studies. Additionally, because proportional data is plotted over time, the independent measure, time, is continuous. Altmann (2011b) describes the fundamental issues surrounding analysis:

> An entirely different class of statistical modeling needs to be carried out for analysing time-course data . . . how can one determine that any pair of curves are different from one another? How can one determine where the peak is located for any such curve (given that aggregating data for the purposes of such [time-course] plots hides the true underlying distribution of the data across subjects and trials)? And most importantly, perhaps, how can one model the dynamic changes to fixation proportions across time when successive time points are not independent of one another? (Altmann, 2011b, p. 996)

These issues are all tempered by the decisions that researchers must make on the mode of presentation of the visual scene, which further impacts how the data are analyzed. In some experiments, participants are allowed free view of the visual scene prior to the target region of interest. This protocol is in contrast to other work, which requires participants to remain on a fixation point or fixation cross until the onset of the target region of interest. Both methods have their advantages and disadvantages. Allowing free view of the visual scene represents a more ecologically valid task reflective of what participants would presumably do in nonexperimental settings. Therefore, free view presentation offers an ecological advantage over fixed visual presentation. On the other hand, free view presentation aggravates one potentially problematic issue in data analysis that is attenuated in fixed visual presentations. Specifically, because participants are idiosyncratic in the manner in which they view a visual scene prior to hearing a named object, free view presentation greatly increases the likelihood for baseline effects. Briefly, baseline effects are represented on a time-course plot by the y-intercept (or value of y at x = 0). The greater the magnitude of difference between the y-intercept of the target and any distractors, the greater the baseline effect, which subsequently represents a random effect in eye-tracking data. We illustrate an example of a time-course plot with baseline effects in Figure 5.7. Notice how the two lines representing fixations to target items and distractor items are already separated from the beginning of the time-course. In other words, at x = 0 (the onset of the critical region), the proportion of fixations to distractor items is approximately 0.477 whereas it is around 0.341 for target items. There are considerably more fixations to distractor items already present from this onset. However, this difference is not related to the experimental manipulation itself. Planned eye movements occur roughly 150 to 200 ms after their initiation. Therefore, any differences that already are present at the onset of the critical region are from planned movements occurring before the experimental manipulation. These planned movements are random effects which

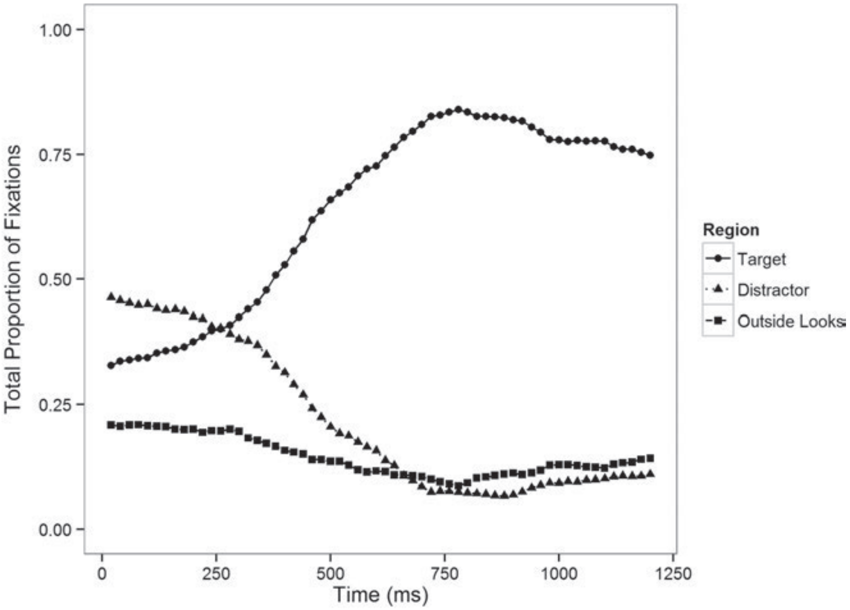


FIGURE 5.7 Sample time-course plot showing baseline effects.

are unique to the individuals and the trials. If baseline effects are very strong, then they may mask any effects driven by the experimental manipulation.

In fixed visual presentations, participants do not begin looking at the visual scene until the onset of the target region of interest; consequently, baseline effects are neutralized. In other words, all proportional data begin at 0 at the onset of the target region of interest. Although fixed visual presentations are more tenable in action-based tasks where participants are instructed in the auditory stimuli to manipulate a target item either by moving it to a new location (e.g., Tanenhaus et al., 1995) or by clicking on it with a computer mouse (e.g., Allopenna et al., 1998), the decision between a fixed visual presentation and free viewing must ultimately be made within the context of the goals of the experiment.

Researchers have implemented various approaches to analyze data which attempt to take into consideration the issue of baseline effects (see in particular special issue 59 of the Journal of Memory and Language, especially Barr, 2008, and Mirman, Dixon, & Magnuson, 2008). One such approach involves a target-contingent-based analysis (Tanenhaus, 2007). Here, researchers simply remove any trials in which participants are already on the target item at the onset of the critical region of the auditory stimulus. As in the fixed visual presentation, this technique of data trimming reduces proportional data to 0, but one disadvantage is that it can result in a high amount of data loss. In studies that employ four-picture displays, data loss may be more manageable than in experiments that employ two-picture displays, where this method of data trimming is not suitable. Another approach to data analysis involves growth curve analysis (Mirman et al., 2008) which fits nonlinear (i.e., polynomial) models to time-course data. One major advantage to this approach is that the derived models functionally describe changes over time while preserving the original data points (i.e., there is no need to aggregate the data over time bins or over trials and participants). Second, because growth curve analysis is a regression technique, models can be hierarchical and subsequently can account for random effects such as the baseline effects described above (Mirman et al., 2008; Baayen, Davidson, & Bates, 2008). However, it remains unclear how interpretable higher-order polynomial coefficients derived from the models are. Researchers would need some level of expertise in statistics and programming to create these nonlinear regressions. Another approach employed by many visual world researchers is to follow the more traditional analyses using *t*-tests and ANOVAs. Under this approach, proportional data are aggregated over time regions and subsequent analyses are carried out within each region. Researchers look for the initial time region when the proportion of fixations to target items is significantly higher than fixations to distractors. For example, researchers can conduct paired-samples *t*-tests for each condition on fixation proportions to target and distractor items in sequential 100 ms regions from 0 ms to a predetermined amount of time (typically 800 to 1000 ms). Planned eye movements generally take about 150 to 200 ms to execute. Therefore, the earliest moment in which target stimuli could affect real-time processing would be roughly 150 to 200 ms after the target onset. Then, a paired-samples *t*-test in each region would indicate an initial moment when a significantly greater proportion of fixations occurs on target items than on distractor items. Furthermore, sequential tests would reveal whether this difference is sustained throughout the rest of the time-course. In the examples presented in Figures 5.4 and 5.5, we calculated the initial time region of divergence for fixations to target items in Figure 5.4 was at Time Region 400 (i.e., the proportion of fixations to target items was significantly higher than for distractor items). In contrast, the initial time region of divergence occurred in Time Region 600 for Figure 5.5. These results lead to the interpretation that the condition in Figure 5.5 (recall that it was phonological competition) results in delayed processing (i.e., a longer time to show significant looks to the target item) when compared to trials on which there was no phonological competition. As stated previously, this is the classic competitor effect. A final approach which we have used in our lab implements a *change point* analysis (Cudeck & Klebe, 2002), described in more detail in the Exemplary Study section below.

An Exemplary Study

Dussias, Valdés Kroff, Guzzardo Tamargo, and Gerfen (2013) employed the visual world paradigm to address two questions. First, do L1 English speakers who are highly proficient in Spanish show effects of prenominal gender marking on the identification of subsequent Spanish nouns? Second, does the presence of a gender system in the L1 that overlaps significantly with the gender system of the L2 determine the extent to which grammatical gender processing in the L2 is native-like?

To investigate these questions, three groups of participants were recruited: functionally monolingual native speakers of Spanish (native controls) from the University of Granada (Spain), L1 English-L2 Spanish speakers from a large U.S. institution, and L1 Italian-L2 Spanish participants completing a year of university study in Granada. In a language history questionnaire, the native group reported having studied English or French in high school and none had spent over one month in a country where the second language was spoken. The 18 English-Spanish speakers were divided into two proficiency groups based on their performance on a standardized test of Spanish (Diploma de Español como Lengua Extranjera [Diploma of Spanish as a Foreign Language], DELE).

To assess knowledge of gender agreement in a comprehension task, participants completed a written picture identification task, which exploited the availability of nominal ellipsis in Spanish, to assess whether learners were able to select morphosyntactically appropriate nouns to complete sentence fragments. To assess gender agreement in production, participants were also administered a picture naming task in which they produced article + noun fragments to pictures displayed on a computer screen. The high mean of correct responses in these two tasks suggested that gender agreement in Spanish for these participants proved largely unproblematic. Participants knew the agreement rules in Spanish and applied them with a high degree of accuracy in a production task and a comprehension task. One remaining question was whether these same participants could access this knowledge during on-line processing of grammatical gender in Spanish.

The experiment included 112 color pictures of highly familiar concrete objects. The pictures were previously normed for naming agreement. Ten participants who did not participate in the experiment were presented with a picture and were asked to name it aloud in English. Only pictures in which there was 100% naming agreement were selected. Half of the pictures represented Spanish object names with feminine gender and half with masculine gender. Each picture served as the target on two trials and as the distractor on two additional trials. Because readers of left-to-right languages have a looking bias to direct eye gaze to the left of the screen first, the presentation side of target items was counterbalanced. To investigate whether a gender-facilitatory effect occurred when participants processed rich sentence contexts (instead of invariable sentence frames such as *Encuentra la/el* . . . "Find the . . . "), we embedded the picture names in variable sentences and distributed the target items evenly so that half appeared in the middle of the sentence (e.g., for *la espada* "the sword": *El estudiante estaba dibujando la espada que vio ayer.* "The student was drawing the sword that he saw yesterday.") and half at the end (e.g., *El niño miraba a su hermano mientras fotografiaba la espada.* "The boy watched his brother taking a picture of the sword."). To conceal the main purpose of the experiment, after listening to each sentence, participants performed a plausibility judgment task. Half of the sentences were plausible (exemplified above) and half implausible (e.g., *El señor compró la espada para la piedra.* "The man bought the sword for the rock.").

A native speaker recorded each experimental sentence between three and five times at a comfortable speaking rate in a sound attenuated chamber. The sentences were produced using standard, broad-focus intonation (i.e., no narrow focus or other emphasis was produced on any of the target noun phrases). From the master recordings, one token was selected for inclusion in the experiment. To precisely match the durational properties of the masculine and feminine articles for all of the experimental items, the article preceding the target noun in each selected sentence was hand-edited to a duration of 147 ms ± 3 ms using Praat. This duration was chosen by sampling the master recordings and calculating a mean duration of the masculine and feminine articles. In this way, the duration of the acoustic signal conveying grammatical gender prior to the onset of the target noun was identical across all items.

Data analysis entailed a change point analysis, in which we implemented a multiphase mixed-effects regression model (Cudeck & Klebe, 2002; see also Section 6.4 "Regression with breakpoints" in Baayen, 2008). The basic feature of this analysis is that any number of phases, each uniquely modeled by its own function, can be united into a more complex whole (described in more detail in Cudeck & Klebe, 2002). An advantage to this approach is that it allows researchers to estimate a point in the time course (i.e., the change point) in which there is a shift between phases. The change point describes the moment in time when one rate of change switches to a different one. In terms of a visual world experiment, researchers are interested in when participants launch their eye movements in response to the linguistic input that they hear. This observation suggests that there is a moment prior to the linguistic input that subsequently changes based on that linguistic input. To that effect, we modeled a three-phase regression model, with each phase described by a linear function, and termed these phases the preconvergence phase, the convergence phase, and the postconvergence phase. The preconvergence phase corresponds to eye movements that are not directly impacted by the critical region in the auditory stimuli; rather, they include random baseline effects due to participants' free view of the visual scene and the time dedicated to launching eye movements towards target items. The convergence phase represents the period of time whereby participants' eye movements shift towards target items. Finally, the postconvergence phase corresponds to the stage in real-time processing where participants are no longer uniformly affected by the experimental stimuli. That is, participants begin to return to a random state of free view.

Experimentally, we were interested in the first change point. The first change point between the preconvergence and convergence phase indicates the point in time when a critical mass of participants begins to shift fixations to the target item. We can then compare change points across conditions. Specifically, by conducting simple paired *t*-tests, we can determine whether one change point occurs significantly earlier (or later) than another change point. This method reveals whether an experimental condition induces a facilitatory or delayed effect when compared to a baseline condition. For the current study, we were, therefore, interested in whether the change point for different gender conditions happened significantly earlier than for same gender conditions.

As mentioned earlier, the minimum latency to plan and launch a saccade has been estimated to be approximately 200 ms (e.g., Saslow, 1967). Thus, approximately 200 ms after target onset is the earliest point at which one expects to see fixations driven by acoustic information from the target word. In line with previous findings, (Dahan, Swingley, Tanenhaus, & Magnuson, 2000; Lew-Williams & Fernald, 2007), results for the native Spanish speaker group showed evidence of the use of gender marking on articles to anticipate upcoming nouns in contexts where two pictured objects belonged to different gender classes. Analyses showed that feminine and masculine different gender trials had an earlier change point than same gender trials, indicating that Spanish monolinguals used information encoded in the article as a facilitatory cue in real-time speech. Results for the two groups of late English-Spanish learners revealed sensitivity to gender marking on Spanish articles similar to that found in native speakers, but this sensitivity was modulated by level of proficiency. The higher proficiency English-Spanish group was quicker to orient to both feminine and masculine target pictures when the article was informative (i.e., in different gender trials) than when it was not (i.e., in same gender trials). The low proficiency group did not show evidence of using grammatical gender anticipatorily, despite being highly accurate in gender assignment and gender agreement in two offline tasks. Finally, for the Italian learners, the change point in the feminine-different trials was significantly earlier than in the feminine-same gender trials. For the feminine condition, then, the results suggest that the Italian participants exploited the presence of grammatical gender on the article as a means of predicting the identity of an upcoming noun. By contrast, there was no significant difference between the change points for the same and different gender displays in the masculine article conditions, indicating that the Italian participants did not use gender as a cue in predicting the identity of a following noun when the determiner carried masculine grammatical gender.

Pros and Cons in Using the Method

Pros

- A particular advantage of the visual world paradigm is its high ecological validity. As with other eye-tracking techniques, the dependent measure in a visual world study does not require an overt behavioral response, such as a button push. Rather, visual world eye-tracking provides direct insights into the interpretation of linguistic input in real time. Specifically, given the plausible linking hypothesis, which maintains that the link between eye movements and linguistic processing is closely timed (Tannenhaus & Trueswell 2006), the visual world method allows researchers to make inferences directly, without secondary behavioral responses, via the tracking of the locus of visual fixations as the speech signal unfolds in real time.

- A second, related benefit is that the visual world paradigm allows for the examination of auditory comprehension in a controlled, laboratory setting. Much of the research on comprehension, given historical methodological constraints, has been conducted on reading. Visual world eye-tracking affords an effective means of examining auditory comprehension, providing researchers with an invaluable tool for testing the conclusions drawn from reading studies and for incorporating prosodic cues that are necessarily absent from written presentation. Arguably, for second language learners especially, for whom reading and listening skills may be vastly different, the visual world paradigm provides an important means of directly testing processing-while-listening.
- A third advantage is that very few psycholinguistic techniques allow researchers to examine both comprehension and production using a single method. It is important to note that visual world eye-tracking can also be used to test speech production, as noted above, in designs in which participants' eye movements are tracked as they speak while viewing a visual array or scene.

Cons

- Visual world eye-tracking is costly. Although the cost of eye-trackers is coming down, the technology is expensive compared, for example, to the cost of running a self-paced reading task (see Jegerski, Chapter 2, this volume), which most often requires a PC and software for constructing the experiment. Two widely used eye-trackers for language processing research are made by SR Research and by Tobii Technology.
- One challenge in eye-tracking studies involves the selection/creation of materials. Researchers must decide on whether to employ line drawings, full pictures, or even real-life objects. One obvious limitation is that the method generally works most simply with objects that are easily pictured. In second language research with the visual world paradigm, the researcher must take pains to assure not only that the target stimuli are easily pictured but also that the participants know the names for the objects, constraints which can sometimes severely limit the set of usable target stimuli.
- Another challenge in visual world studies involves programming the experiments, which can be challenging for researchers not well versed in constructing complex experiments. In a visual world study, researchers must tightly align the timing of the presentation of auditory input with the appearance of a visual display. In addition, there are a number of different experiment

builders, each with its own programming language. To cite the example of SR Research's Eyelink, which we use in our lab, researchers are faced with the choice of programming the experiment in SR Research's proprietary Experiment Builder software or with building the experiment in E-Prime and interfacing with the tracker hardware.

- Another challenge in visual world studies involves data collection and extraction. Current setups such as SR Research's Eyelink 1000 collect eye movement data at a sampling rate of 1000 Hz. For students unfamiliar with sampling issues, a 1000 Hz sampling rate means that 1000 data points will be collected per second. Thus, a single minute of eye movement data will contain 60,000 data points. With such a large quantity of data, it becomes necessary to automate data extraction, usually something that can be achieved by writing a program in Matlab or R. This adds an additional layer of work to visual world studies.
- As is discussed above, perhaps the largest current disadvantage in carrying out a visual world study involves the lack of consensus in the field regarding how to best analyze visual world eye-tracking data. Unlike reaction time data in a button press experiment or reading time on a word (or series of words) in a self-paced reading task, both of which provide researchers with a single numerical measure, visual world eye-tracking data represent proportions of fixations to a target versus a distractor or competitor item over some unit of time. In simple terms, it is not altogether clear how to best analyze the large amount of data that such designs yield, and researchers have attempted a number of approaches, ranging from binning the data and conducting successive *t*-tests at regular temporal intervals throughout a trial to more sophisticated growth curve analyses, which essentially attempt to model response patterns by fitting polynomial equations to the aggregated eye movements across time.

Discussion and Practice

Questions

1) Give two examples of types of second language learners for whom the visual world method would be a better approach than would a reading study for examining sentence processing. Be specific in discussing your answer.
2) Think of a research question (for example, the investigation of a particular issue in the processing of a particular grammatical structure) in which a reading study would be a preferable method, and explain your choice.
3) Define the following terms: a) anticipatory effect and b) competitor effect. Discuss examples of how these effects have been exploited to inform theoretical issues in language processing.

4) Discuss the research design of a published paper in which visual world eye-tracking is employed to examine issues in bilingual language processing. Make sure to explicitly address the following aspects of the article: a) the research question that motivated the study; b) how the predictor variables were selected; c) what criteria were used in selecting the participants and what data were collected on their language experience and proficiency, along with any other individual difference measures; d) how the data were presented and analyzed statistically. Finally, provide a critique of the study in which you identify at least one open question for future research that cannot be addressed by the results of the experiment(s).
5) What is meant by the claim that the visual world method has high ecological validity for examining language processing? For example, discuss why research such as Altmann and Kamide (1999) on scene processing provides a compelling example of the ecological validity of the method.

Research Project Option

A good way to begin to do research is via replication. Spivey and Marian (1999) examined late Russian-English bilinguals and showed that even in a monolingual Russian experimental context, native Russian's eye-movements reveal a competitor effect for objects whose English translation equivalents exhibit phonological overlap with Russian target items. Design and carry out a replication of this study using another language pair. If you cannot replicate the study, consider the factors that may underlie your null finding (e.g., the participants' language levels and dominance, the language pair, the stimuli, or some other factor). If you can replicate the finding, speculate on how it relates to Ju and Luce's (2004) finding that fine-grained phonetic detail in the acoustic signal is sufficient to constrain parallel activation and allow for selective lexical access in bilinguals.

Author Note

The writing of this chapter was supported in part by NSF Grant BCS-0821924 to P. E. Dussias and Chip Gerfen, by NSF Grant BCS-0955090, NSF Grant OISE-0968369, and NIH Grant 1R21HD071758-01A1 to P. E. Dussias. J. Valdés Kroff was supported by a National Science Foundation Graduate Fellowship.

Suggested Readings

Hopp, H. (2012). The on-line integration of inflection in L2 processing: Predictive processing of German Gender. *BUCLD 36: Proceedings of the 36th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151–171.

Ju, M., & Luce, P. A. (2004). Falling on Sensitive Ears—Constraints on Bilingual Lexical Activation. *Psychological Science, 15,* 314–318.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49,* 238–299.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10*(3), 281–284.

Weber, A., & Paris, G. (2004). The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening. In K. Forbus, D. Gentner, & T. Tegier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition: Evidence for continuous mapping models. *Journal of Memory and Language, 38,* 419–439.

Altmann, G. T. M. (2011a). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta Psychologica, 137,* 190–200.

Altmann, G. T. M. (2011b). The mediation of eye movements by spoken language. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye Movements* (pp. 979–1004). Oxford, UK: Oxford University Press.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73,* 247–264.

Altmann, G. T. M., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition, 111,* 55–71.

Arnold, J., & Griffin, Z. M. (2007). The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language, 56,* 521–536.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language, 59,* 457–474.

Blumenfeld, H. K., & Marian, V. (2005). Covert bilingual language activation through cognate word processing: An eye-tracking study. In *Proceedings of the twenty-seventh annual meeting of the cognitive science society* (pp. 286–291). Mahwah, NJ: Lawrence Erlbaum.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *Journal of Memory and Language, 48,* 653–685.

Canseco-Gonzalez, E., Brehm, L., Brick, C., Brown-Schmidt, S., Fischer, K., & Wagner, K. (2010). Carpet or Cárcel: The effect of age of acquisition and language mode on bilingual lexical access. *Language and Cognitive Processes, 25,* 669–705.

Chambers, C. G., Tanenhaus, M. K., & Magnuson, J. S. (2004). Actions and affordances in syntactic ambiguity resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 687–696.

Cooper, R. (1974). The control of eye-fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory and language processing. *Cognitive Psychology, 6,* 84–107.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7,* 41–63.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics, 34,* 269–284.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken word recognition: Evidence from eye movements. *Cognitive Psychology, 42,* 317–367.

Dahan, D., Swingley, D., Tanenhaus, M. K., & Magnuson, J. (2000). Linguistic gender and spoken-word recognition in French. *Journal of Memory and Language, 42,* 465–480.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Evidence from immediate effects of verb-based constraints. *Journal of Experimental Psychology: Learning, Memory & Cognition, 30,* 498–513.

Dahan, D., & Tanenhaus, M. K. (2005). Looking at the rope when looking at the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review, 12,* 453–459.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, L. M., Dunn, L. M., & Arribas, D. (2006). *PPVT-III Peabody: Test de vocabulario en imágenes.* Madrid, Spain: Tea, D. L.

Dussias, P., Valdés Kroff, J., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). When gender and looks go hand in hand: Grammatical gender processing in L2 Spanish. *Studies in Second Language Acquisition, 35,* 353–387.

Engelhardt, P. E., Bailey, K. G. D., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54,* 554–573.

Eyelink 1000 [Apparatus and software]. Mississauga, Ontario: SR Research.

Fernald, A., Perfors, A., & Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the second year. *Developmental Psychology, 42,* 98–116.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11,* 274–279.

Hopp, H. (2012). The on-line integration of inflection in L2 processing: Predictive processing of German gender. *BUCLD 36: Proceedings of the 36th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press.

Huettig, F., Chen, J., Bowerman, M., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture, 10,* 39–58.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151–171.

IMSI Master Clips Premium Image Collection [Computer software]. (1990). Novata, CA: IMSI Inc.

Ju, M., & Luce, P. A. (2004). Falling on sensitive ears: Constraints on bilingual lexical activation. *Psychological Science, 15,* 314–318.

Kamide, Y., Altmann, G. T. M., & Haywood, S. L. (2003). Prediction and thematic information in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49,* 133–156.

Kaplan, E., Goodglass, H., Weintraub, S., & Segal, O. (1983). *Boston Naming Test.* Philadelphia, PA: Lea & Febiger.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science, 18,* 193–198.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63,* 447–464.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science, 31,* 1–24.

Marian, V., & Spivey, M. (2003a). Bilingual and monolingual processing of competing lexical items. *Applied Psycholinguistics, 24,* 173–193.

Marian, V., & Spivey, M. (2003b). Competing activation in bilingual language processing: Within and between-language competition. *Bilingualism: Language and Cognition, 6,* 97–115.

McMurray, B., Tanenhaus, M. K., & Aslin, R.N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* 33–42.

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language, 59,* 475–494.

Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in Aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language, 117,* 53–68.

Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition, 108,* 155–184.

Perrotti, L. (2012). *Grammatical gender processing in L2 speakers of Spanish: Does cognate status help?* (Unpublished Undergraduate Honors Thesis). Penn State University, University Park, PA.

R [Open Source Computer Software]. (2008). Vienna, Austria: R Development Core Team. Retrieved from http://www.r-project.org/.

Reinisch, E., Jesse, A., & McQueen, J. M. (2010). Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately. *Quarterly Journal of Experimental Psychology, 63,* 772–783.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition, 105,* 466–476.

Saslow, M. G. (1967). Latency for saccadic eye movement. *Journal of the Optical Society of America, 57,* 1030–1033.

Sedivy, J. E., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental interpretation through contextual representation: Evidence from the processing of adjectives. *Cognition, 71,* 109–147.

Snedeker, J., & Trueswell, J. C. (2004). The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology, 49,* 238–299.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 147–215.

Spivey, M., & Marian, V. (1999). Crosstalk between native and second languages: Partial activation of an irrelevant lexicon. *Psychological Science, 10,* 281–284.

Szekely, A., Jacobsen, T., D'Amico, S., Devescovi, A., Andonova, E., Herron, D., . . . Bates, E., (2004). A new on-line resource for psycholinguistic studies. *Journal of Memory and Language, 51,* 247–250.

Tanenhaus, M. K. (2007). Spoken language comprehension: Insights from eye movements. In M. Gaskell (Ed.), *Oxford Handbook of Psycholinguistics* (pp. 309–326). Oxford, UK: Oxford University Press.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information during spoken language comprehension. *Science, 286,* 1632–1634.

Tanenhaus, M. K., & Trueswell, J. C. (2006). Eye movement and spoken language comprehension. In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics* (pp. 835–862). Amsterdam: Elsevier.

Thompson, C. K., & Choy, J. J. (2009). Pronominal resolution and gap filling in Agrammatic Aphasia: Evidence from eye movements. *Journal of Psycholinguistic Research, 38,* 255–283.

Trueswell, J. C., Sekerina, I., Hill, N., & Logrip, M. L. (1999). The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition, 73,* 89–134.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *Journal of Memory and Language, 50,* 1–25.

Weber, A., & Paris, G. (2004). The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening. In K. Forbus, D. Gentner, & T. Tegier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Yee, E., Blumstein, S. E., & Sedivy J. C. (2008). Lexical-semantic activation in Broca's and Wernicke's Aphasia: Evidence from eye movements, *Journal of Cognitive Neuroscience, 20,* 592–612.

6

EVENT-RELATED POTENTIALS (ERPs)

Kara Morgan-Short and Darren Tanner

History of the Method

The first reports of human brain activity recorded at the scalp were made by Hans Berger (Berger, 1929). Berger showed that by placing an electrode on the scalp and amplifying the data, a series of positive and negative voltage fluctuations could be seen. Although it took a number of years before the broader community replicated and accepted Berger's findings as reflecting brain activity, recordings of the human electroencephalogram (EEG) have since provided an exceptionally rich source of information on human cognition and, since the 1980s, on the processing and learning of language. After the initial reports of the EEG, it was several decades before EEG was seriously used to study cognition, perhaps for at least two reasons. First, the EEG in its raw form is relatively uninformative about specific neural or cognitive processes, as it reflects the summation of hundreds of (and likely many more) neural sources acting simultaneously. This makes the activity related to any specific event or process difficult to see in the raw waveform. Second, the relative expense and low power of computers in the early EEG era made higher-level quantification and analysis of EEG data infeasible. However, isolating activity relevant to specific cognitive processes is not an intractable problem. Embedded within the raw EEG are distinct signals associated with many of the ongoing processes in the brain, and some of them can be extracted using simple averaging procedures. That is, by averaging the EEG signal over numerous trials in a given experimental condition, the fluctuations in EEG activity that are time- and phase-locked to the presentation of the stimuli can be identified. The resulting waves thus reflect the brain's electrical activity associated with a specific cognitive event. This activity is known as the event-related potential, or ERP.

The 1960s saw the emergence of ERP studies of human cognitive processes, particularly with the identification of two ERP "components" that reflect processes beyond low-level sensory processing: the contingent negative variation, which reflects response preparation (the CNV; Walter, Cooper, Aldridge, McCallum, & Winter, 1964), and the P300, which is elicited by unexpected stimuli (Sutton, Tueting, Zubin, & John, 1965). However, little ERP work during this early period was specifically focused on understanding the processes underlying normal language comprehension. Instead, early ERP studies that presented linguistic stimuli generally used out-of-context word- or letter-monitoring tasks designed to study factors related to the elicitation of the CNV or P300 (e.g., Shelburne, 1972; see also Friedman, Simson, Ritter, & Rapin, 1975).

The modern era of language-related ERP research began in 1980 when Kutas and Hillyard (1980) showed that, unlike visually anomalous sentence-final words (e.g., words in large font), which elicited a large positive-going P300-like wave, semantically anomalous sentence-final words (e.g., *He spread his warm bread with socks;* anomalous word indicated by underlining) elicited a large negative-going wave between 300 and 500 ms after word presentation, with a peak around 400 ms: the N400. Since this seminal finding, the N400 has become one of the best-studied effects in the ERP literature (for example, as of the writing of this chapter, Kutas and Hillyard's original paper has been cited 2,580 times according to Google Scholar and 1,736 times according to Web of Science). A rich array of studies have investigated manipulations impacting the N400 in linguistic and nonlinguistic contexts (see Kutas & Federmeier, 2011, for a thorough review), though broadly speaking, one can generalize that the amplitude of the N400 to words covaries with numerous factors, including word frequency, orthographic neighborhood density, and semantic congruity with a preceding sentence or discourse context.

A little over a decade later, Osterhout and Holcomb (1992) reported that apparent syntactic anomalies in garden path sentences (e.g., *The broker persuaded to sell the stock. . .*) did not elicit an N400, but instead a large positive-going wave with an onset around 500 ms and maximum around 600 ms poststimulus, the P600. The P600 has subsequently been shown to be sensitive to a broad range of syntactic and morphosyntactic manipulations, including violations of agreement, case, and tense morphology, reanalysis in grammatical but difficult-to-parse sentences, and processing of filler-gap dependencies. In some cases (though not always) it is preceded by a negativity between 300 and 500 ms, which is most prominent over left anterior scalp regions (the left anterior negativity, or LAN). (See Steinhauer & Connolly, 2008, and Swaab, Ledoux, Camblin, and Boudewyn, 2012, for recent reviews of these and other language-related ERP components.)

Although there was initial promise that the N400 and P600 (and possibly the LAN) could be used to identify processes or linguistic manipulations as categorically semantic or syntactic, many counterexamples to this neat syntax/semantics distinction have recently been reported (e.g., Deutsch & Bentin, 2001; Kim & Osterhout, 2005; Kuperberg, Caplan, Sitnikova, Eddy, & Holcomb, 2006; Nakano, Saron, & Swaab, 2010; Osterhout, 1997). In light of this, the N400 and P600 have been reinterpreted by some as reflecting more general memory-based and broadly combinatorial processing streams, respectively (Kuperberg, 2007; Osterhout, Kim, & Kuperberg, 2012). It is also important to note that the N400 and P600 are not language-specific effects, although it is clear that they index processes crucial to language comprehension. Future research will surely further refine these generalizations and functional interpretations. Nonetheless, the finding that different manipulations (lexical/semantic and syntactic/combinatorial, broadly speaking) elicit qualitatively different ERP effects shows that there are at least two independent, but highly interactive, cognitive processing streams underlying the normal comprehension of language, and that the N400 and P600 components can index them.

The finding that semantic and combinatorial manipulations elicit qualitatively different brain responses set the stage for early ERP studies on second language (L2) acquisition and language processing in bilinguals. The first report of sentence processing in bilingual populations we know of was made by Ardal and colleagues (Ardal, Donald, Meuter, Muldrew, & Luce, 1990), who investigated the latency of the N400 in monolinguals, bilinguals processing their native language (L1), and bilinguals processing their L2. Weber-Fox and Neville (1996) extended bilingual research to cross-sectionally investigate age of acquisition effects in Chinese-English bilinguals processing semantic and syntactic anomalies. Another seminal study on L2 learning was reported by McLaughlin, Osterhout, and Kim (2004), who used a longitudinal design to study the acquisition of L2 vocabulary in beginning classroom learners (see below for a more complete description of this study).

Since these initial reports, ERPs have become increasingly common in the study of aspects of L2 and bilingual language processing at multiple levels, including phonological (e.g., Nenonen, Shestakova, Huotilainen, & Näätänen, 2005), lexical (e.g., Schoonbaert, Holcomb, Grainger, & Hartsuiker, 2011), and sentential. Regarding sentence processing, one of the driving questions behind many studies is whether or not L2 semantic and syntactic processing can become *native-like*—that is, whether ERP responses in L2 users can approximate those seen in native speakers, and what factors (e.g., learner proficiency, age of acquisition, or L1-L2 grammatical similarity) may modulate this native-likeness (e.g., Foucart & Frenck-Mestre, 2011; Steinhauer, White, & Drury, 2009; Tokowicz & MacWhinney, 2005; Weber-Fox & Neville, 1996). More recently, studies have begun to use ERPs to answer a broader range of questions related to language processing and language learning, including how L2 learning context impacts neural responses (e.g., Morgan-Short, Sanz, Steinhauer, & Ullman, 2010; Morgan-Short, Steinhauer, Sanz, & Ullman, 2012). Also, given findings that even monolinguals can show marked individual differences in the processing of morphosyntax and semantics, L2 research has begun to focus on how individual differences manifest themselves in the neural correlates of L2 comprehension (e.g., Tanner, Inoue, & Osterhout, 2013; Tanner, Mclaughlin, Herschensohn, & Osterhout, 2013). (See Steinhauer et al., 2009, and van Hell & Tokowicz, 2010, for recent reviews of L2 ERP research.)

What is Looked at and Measured

Scalp-recorded EEG reflects the summated postsynaptic potentials from large groups of synchronously firing cortical neurons. EEG recordings are most sensitive to cortical pyramidal neurons arranged in an open-field geometry and which are aligned perpendicularly to the scalp. ERPs then reflect the portion of the EEG activity that is both time- and phase-locked to the presentation of a particular type of stimulus. A complete account of the neurobiology of ERPs is beyond the scope of this introductory chapter, so we will focus only on the information most relevant to beginning ERP researchers; we direct the interested reader to Luck (2005) and Handy (2005) for in-depth introductions to ERPs. Throughout this chapter we will focus primarily on the processing of visually-presented stimuli (e.g., written words and sentences). However, the same basic logic of analysis and quantification extends to aural stimuli (e.g., phonemes, syllables, isolated words, and continuous speech), though issues around time-locking and waveform morphology may differ. At least with regard to ERP effects related to sentence-level semantics and morphosyntax, although the waveform morphologies elicited by stimuli in the two modalities differ, there are relatively few differences in the pattern of N400 and P600 effects elicited by auditory versus visual linguistic stimuli (Hagoort & Brown, 2000).

An example of an ERP setup is provided in Figure 6.1. Typically, an array of electrodes is affixed to the participant's head, usually with an elastic cap. A conductive gel is inserted into each of the electrodes, usually with a syringe and a blunt tip needle or the wooden end of a cotton swab, and the participant's skin is lightly abraded by gently rubbing the tip of the needle against the participant's

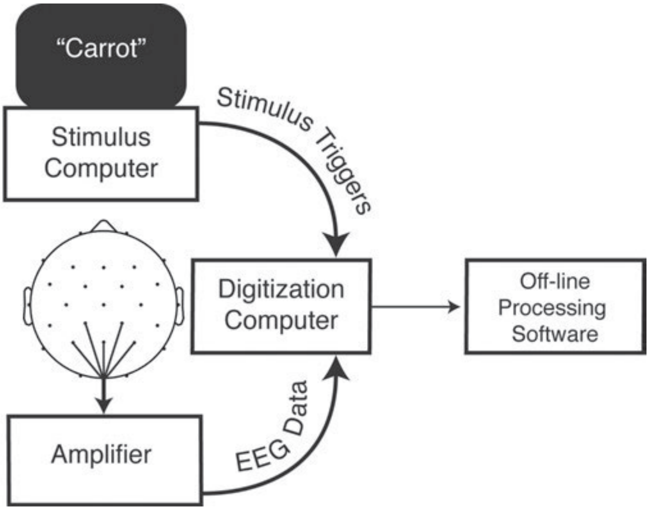


FIGURE 6.1 Schematic of a typical ERP lab setup.

scalp until a good connection is made between the skin and the electrode. The electrodes are then connected to an EEG system that amplifies and filters the raw data, which are subsequently digitized. Meanwhile, the participant views or listens to a series of stimuli presented by another computer. In a typical sentence processing experiment using visual presentation, words are presented one at a time, and generally at a fixed rate between 400 and 700 ms each. When stimuli of interest are presented (e.g., an anomalous word or its well-formed control counterpart; see discussion of issues regarding stimuli below), the stimulus computer sends a trigger code into the EEG, marking the exact millisecond when the stimulus appeared on the screen, as well as the type of stimulus it was. The EEG, along with the trigger codes, is saved on the computer for later processing.

In order to obtain the ERP signals from the EEG, a series of processing steps are carried out after the data acquisition is complete. First, additional filters can be run on the data to reduce the amount of high-frequency activity in the EEG, as most language-related ERP effects are relatively low-frequency. The EEG is then sliced up into individual epochs consisting of a time interval surrounding the trigger codes of interest (e.g., from 200 ms prestimulus to 1200 ms poststimulus). Each epoch is screened for artifacts, such as electrical activity related to eye blinks or other muscle movements, as muscle-related activity can be orders of magnitude larger than EEG activity and thus obscures the signal of interest; epochs containing artifacts are excluded from further analysis. Finally, the ERP is extracted by averaging the EEG sampled at each time point in each epoch across all trials in a particular experimental condition. This averaging procedure is carried out separately at each electrode for each participant.

The logic behind the averaging procedure is this. Brainwave activity that is consistently elicited by a particular type of stimulus (e.g., a verb that disagrees in number marking with the subject NP) will be elicited across all trials in that condition; this is the signal of interest. Background activity in the EEG not associated with processing the stimulus of interest (i.e., the "noise") will fluctuate randomly with respect to the stimulus onset. The activity that is time- and phase-locked to the onset of the stimulus will remain robust to the averaging procedure, whereas random noise will not. The signal-to-noise ratio of the resulting ERP averages thus improves as a function of the number of experimental trials in each condition: as the number of trials per experimental condition increases, the less background EEG noise will remain in the ERP waveforms.

Within the first 400 ms after visual presentation of a stimulus, the ERP waveform is characterized by a series of peaks and valleys sometimes referred to as *components.* Although we use the term "component" to refer to specific peaks and valleys in the averaged ERP waveform, it is important to note that the concept of an ERP component is multifaceted and may not always refer to an observable peak or valley in the waveform (see Luck, 2005, Chapter 2). Figure 6.2 depicts waveforms from a single electrode placed at the vertex of the scalp (labeled "Cz" in accordance with standard electrode placement and naming conventions;

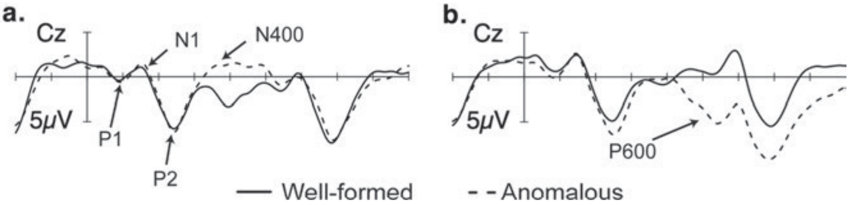


FIGURE 6.2 Depiction of N400 and P600 effects elicited by linguistic anomalies.

Jasper, 1958); positive voltage is plotted down, and each tick mark represents 100 ms of activity. The onset of the visual stimulus is marked by the vertical bar. As can be seen, the early portion of the waveform is characterized by a positive peak (the P1), followed by a negative peak (the N100, or N1), followed by another positive peak (the P200, or P2), then a broad negative deflection in the waveform (the N400 component). Here it is important to mention an important distinction: the term *N400 component* often refers to a negative-going wave around 400 ms after stimulus onset that is elicited by *any* meaningful stimulus, whereas the term *N400 effect* often refers to modulations of the amplitude of that negativity by a given experimental manipulation. Figure 6.2a depicts both an N400 component (the waveforms show a negative deflection around 300 to 400 ms) and an N400 effect elicited by a linguistic anomaly: the dashed line depicting brain responses to the anomaly shows a larger amplitude (i.e., more negative-going) N400 than the solid line depicting brain responses to the well-formed condition. Figure 6.2b depicts a P600 effect elicited by a linguistic anomaly: the brain responses to the anomalous stimuli are more positive-going than to the well-formed stimuli. Here, the divergence starts around 400 ms poststimulus, but the difference reaches its maximum around 600 ms poststimulus.

In most language-related research, it is between-condition differences, the *ERP effect,* and not necessarily the individual peaks in the ERP, that are of interest to researchers. Indeed, absolute voltage values at any time point in any single condition provide little information about whether or not a linguistic manipulation had an impact on brain responses. For example, the term *P600* generally does not refer to a well-defined peak in the waveform, but instead refers to a difference between two conditions. Because of this, researchers can examine difference waves to identify the specific impact of the experimental manipulation on participants' brain responses by computing the difference between the anomalous and well-formed conditions at each time point for each electrode. Figure 6.3 provides an example of the data from Figure 6.2a being replotted as a difference wave. Here, deviations from the x-axis reflect effects of the experimental manipulation. Again, negative voltage values are plotted up. As can be seen, relative to well-formed stimuli, brain responses to anomalous words showed a negative-going deflection between approximately 300 and 500 ms after the stimulus was presented, and the peak of the difference occurred at around 400 ms, a pattern typical of N400 effects.

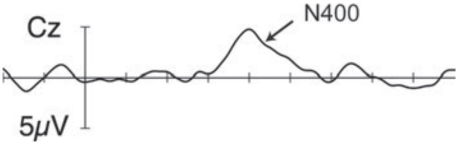


FIGURE 6.3 Data from Figure 6.2A re-plotted as a difference wave.

Thus far we have seen that ERP data can be characterized along two dimensions: effect timing (i.e., at how many milliseconds poststimulus an effect onsets or peaks) and effect polarity (i.e., whether a manipulation elicits a positive- or negative-going wave). ERPs also provide a third dimension of data, namely scalp distribution. Most ERP experiments record brain activity from a number of electrodes (from as few as three to as many as 256, with typical experiments using between 19 and 64 electrodes) placed at various locations across the scalp. This electrode array allows the investigation of both what happens and when, and also where it surfaces (e.g., over frontal, central, or posterior portions of the scalp, or over left, midline, or right hemisphere sites). For example, the N400 and LAN effects mentioned above are both negative-going waves and they occur during roughly the same time interval, approximately 300 to 500 ms poststimulus. However, the two effects can be distinguished based on characteristic scalp distributions: N400 effects are largest over central and parietal scalp sites and sometimes have a slight right-hemisphere bias, whereas LAN effects are typically most prominent over left frontal electrode sites.

An important caveat here is that the scalp topography of a given ERP effect provides no information about the location of the neural generators of the effect. A left frontal scalp distribution does not mean that the effect was generated in the left frontal cortex. In fact, it is impossible to know with 100% certainty the cortical source of an ERP effect. This is a problem known as the *inverse problem* (see Luck, 2005, Chapter 7, for a thorough discussion of issues related to source localization of ERPs). However, as in the case of the N400 and LAN effects, the difference in scalp topographies coupled with a difference in the antecedent conditions eliciting the effects (e.g., lexico-semantic manipulations and the N400 versus morphosyntactic manipulations and the LAN) allow a general inference that the two effects reflect nonidentical cognitive states or processes.

This raises a broader issue regarding ERPs, namely the types of inferences one can (or cannot) draw from a comparison of the dimensional characteristics of two ERP waveforms. For example, the timing of an effect onset in the ERP record provides a source of information on the time-course of linguistic processing, but does not necessarily reflect the exact moment at which a linguistic anomaly was detected or processed. Effect timings provide only an *upper bound* estimate on the time when the processes differed between two conditions. It may be the case that differential activity was present earlier, but it occurred in cortical or subcortical regions of the brain which ERPs are not sensitive to, that it did not reach a sufficient threshold to be detectable at the scalp, or that the earlier activity was not in sufficient phase synchrony to be robust to the ERP averaging procedure (see Bastiaansen, Mazaheri, & Jensen, 2012, regarding this final point). These caveats hold of null results in ERP research more generally, as well: failure to detect a difference in ERPs between two conditions does not unequivocally mean there was no differential neural activity. Additionally, effect polarity alone—whether a component or effect is positive- or negative-going—provides little information to the researcher. However, if two effects differ in polarity, one can infer that they reflect non-identical cognitive processes. Similarly, differing scalp topographies between two effects suggest nonidentical cortical sources of the effects, or at least differential contributions of a common set of generators. However, there are numerous issues surrounding the quantification and comparison of scalp topographies between conditions, making concrete inferences difficult (Urbach & Kutas, 2002).

Furthermore, it is important to distinguish between antecedent conditions and underlying processes. Semantic and syntactic manipulations tend to elicit N400 and P600 effects, respectively. However, this does not mean that the N400 and P600 reflect semantic and syntactic processes directly. They could instead reflect processes that are correlated with, but indeterminately removed downstream from, the actual semantic and syntactic processes themselves (Osterhout, McLaughlin, Kim, Greewald, & Inoue, 2004). Nonetheless, the multidimensional outcome measures that ERPs provide make them a tremendously useful tool for understanding neurocognitive aspects of language processing and language learning. One simply need be aware of the inferential limitations associated with the methodology (see Otten & Rugg, 2005; Rugg & Coles, 1995, for thorough discussion of these issues).

Issues in the Development and Presentation of Stimuli

Given a basic understanding of what is looked at and measured in ERP research, we now turn to some basic issues in the development and presentation of stimuli for ERP research. Although the information presented here is not sufficiently comprehensive for the purpose of fully developing an ERP experiment, we discuss critical aspects of stimuli development and presentation, and additional ERP resources are pointed out for the reader's information in the Suggested Readings section.

The first step in ERP stimuli development is to decide on the appropriate experimental paradigm and task for a given research question. Although results from ERP experiments can inform questions of L2 processing, this crucially depends on the validity and reliability of the experimental paradigm administered as the ERP data are collected. The majority of L2 ERP studies have examined aspects of L2 development with a sentence processing violation paradigm. However, other experimental paradigms, including priming and lexical decision tasks, have also proven to be useful for addressing particular research questions about L2 processing. Here we provide an overview of priming, lexical decision, and sentence processing violation paradigms in ERP research and discuss general issues related to stimuli development for these tasks. Note that we are not presenting all possible paradigms in ERP research; for example, the semantic categorization paradigm (e.g., Midgley, Holcomb, & Grainger, 2009a) is one option that will not be discussed here.

Priming paradigms can be used to investigate the organization of memory related to particular linguistic forms. In this paradigm, an initially presented stimulus—a prime such as *cat*—is expected to facilitate the response to a related stimulus—a target such as *dog*—compared to the response to a target that is preceded by an unrelated prime such as *coffee.* ERPs are then analyzed at the point of the target so as to reveal how the priming manipulation affects the processing of the target word. Priming paradigms are often paired with a lexical decision task (see below), although this is not the only type of response that can be elicited from participants. Different iterations of the priming paradigm manipulate the relationship between the prime and the target in different ways to address different research questions. For example, ERP word priming studies have addressed the organization of semantic memory in L1 and bilinguals as well as in L2 learners (e.g., Midgley, Holcomb, & Grainger, 2009b; Phillips, Segalowitz, O'Brien, & Yamasaki, 2004), whereas translation priming studies have examined the role of cross-language activation in bilinguals (Schoonbaert, Holcomb, Grainger, & Hartsuiker, 2011). ERP data collected as participants engage in other types of priming tasks, such as masked priming, syntactic priming, and cross-modal priming, has also served to advance our knowledge of the organization of memory for L1 and L2 (see Roberts, Chapter 9, this volume, for discussion of cross-modal priming).

The lexical decision task is another paradigm that can be used in conjunction with ERPs and is typically used to examine semantic memory and lexical access. In this paradigm, words and nonwords, such as *black* and *bleck,* respectively, are presented to participants who are asked to indicate whether the letter string is a word or not. ERP waveforms elicited by the word and nonword conditions are compared within subjects and allow researchers to draw conclusions about the neurocognitive processing of words from different languages or different linguistic forms (e.g., Geyer, Holcomb, Midgley, & Grainger, 2011; McKinnon, Allen, & Osterhout, 2003). Between-subject comparisons of the effect found for words versus nonwords can then be examined to explore differences between groups of learners. (See the Exemplary Study section for an example of an L2 ERP study that combines a lexical decision task with priming.)

As previously mentioned, the sentence processing violation paradigm is the most commonly used method in ERP studies of L2 processing, and is used to explore how different linguistic forms are processed during comprehension. In this paradigm, correct and violation sentences (e.g., *The winner of the big trophy <u>has/*have</u> proud parents*; from Tanner et al., 2013; target words are underlined; violation word marked with an asterisk) are presented to L2 learners who are generally asked to judge the grammaticality or acceptability of the sentences, but alternatively may be asked to respond to comprehension questions after some or all of the sentences. Researchers examine differences between the ERP waveforms elicited by (a) the target word that creates a violation in the sentence (*have* in the example above) and (b) the matched correct word (*has* in the example above).

ERP effects elicited by the sentence processing violation paradigm are generally assumed to reflect how a form is naturally processed for the purposes of comprehension, even though participants are not usually asked to respond to comprehension questions. Osterhout and colleagues have empirically examined this assumption in native speakers (Osterhout, Allen, McLaughlin, & Inoue, 2002) by comparing the ERP responses elicited during passive reading to those elicited when the participant makes an active judgment. Results from these studies showed that asking participants to judge the acceptability of stimulus sentences did not qualitatively alter the ERP processing response, although for syntactic anomalies, the P600 effect was larger when judgments were required. Therefore, it is reasonable to assume that ERP responses elicited by this paradigm are not artifacts of the task itself. The apparent validity of the sentence processing violation paradigm in regard to natural language processing and the fact that violations can be designed to probe various aspects of language, such as semantics, morphosyntax, and syntax, among others, contributes to the extensive use of the paradigm by L1 and L2 ERP researchers.

After deciding on the appropriate ERP experimental paradigm for any given research question, a sufficient number of stimuli must be developed and appropriate controls must be built into the stimuli and presentation parameters in order to ensure that any significant ERP effects can be clearly attributed to the experimental manipulation, and not to other potentially confounding factors. In order to obtain a strong within-subject signal, a typical ERP study presents 30 to 40 stimuli per condition, although this depends on the target linguistic form and the research questions. Standard controls for any psycholinguistic study must be incorporated into the development of these ERP stimuli. To the greatest extent possible, target words in different conditions should be matched for potentially confounding lexical properties, such as word category, frequency, length (orthographic and/or syllabic), and imageability, among others. When different sets of words are used for different conditions, it is advisable to report the factors that were controlled as well as statistical tests comparing these factors across conditions. Perfect control over lexical factors can be achieved in some experiments by creating a set of stimuli in which the exact same lexical items are used across experimental conditions. For example, White, Genesee, and Steinhauer (2012) examined the processing of the L2 English past tense in L1 Korean and L1 Chinese participants using the sentence sets exemplified in (1) below.

(1)

a. The teacher didn't <u>start</u> the lesson.
b. The teacher hadn't <u>started</u> the lesson.
c. The teacher didn't *<u>started</u> the lesson.
d. The teacher hadn't *<u>start</u> the lesson.

Because both *start* and *started* occur in the correct and violation conditions, any inherent differences between the two forms of the word are thus controlled across the correct and violation conditions. This type of design conforms to the Hillyard Principle, which states that researchers should "Always compare ERPs elicited by the *same physical stimuli,* varying only the psychological conditions" (Luck, 2005, p. 68, emphasis added). Although this approach is a perfect solution for controlling potentially confounding lexical factors, it can lead to the need to develop a large number of stimuli, because the different versions of each sentence must be distributed across different stimulus lists in a Latin square design, such that each participant receives only one version of the sentence.

Beyond the lexical controls necessary for any psycholinguistic method, other ERP-specific controls should be implemented. It is critically important to control the properties of the precritical material, that is, the material that precedes the critical target stimuli. This control is necessary because some later language-related ERPs may have latencies that overlap with the presentation of the subsequent word (Steinhauer & Drury, 2012). Thus, if the precritical region varies between conditions and elicits different neural responses, those differences may "bleed" into the ERPs to the target stimulus, such that differential ERP effects at the critical word may not reflect the intended experimental manipulation but rather spillover effects from uncontrolled precritical material. The possibility of ERP differences related to precritical material is most likely to result from immediately preceding words. However, precritical material that is not immediately prior to the target word can also conceivably modulate the ERP effect to the target word. In particular, material prior to the target stimuli should also be controlled such that it does not differentially influence the predictability of the target word across conditions, as predictability is known to modulate at least some language-related ERP effects, such as the N400 (Kutas & Federmeier, 2011; Laszlo & Federmeier, 2009). Note that the stimuli in (1) above meet the criteria of balancing both the critical target words and the precritical material across conditions, so any ERP differences between the violation and correct conditions (after collapsing across the base-form and past tense verbs) should represent processing of the tense violation itself.

Additional ERP-related design issues must also be considered in the development of any ERP study. First, the target word should not appear as the first or the last word of the sentence, as there are additional processes at these positions, such as processes related to sentence wrap-up that may be difficult to disentangle from the experimental effect itself (Hagoort & Brown, 1999). Ideally, at least two to three words should follow the critical word. Finally, if a response is required of participants immediately at the target word, which is common for priming and lexical decision paradigms, the hand used to make the response should be counterbalanced across participants, so that any lateralized motor potentials associated with the response are not confounded with the ERP to the stimuli. Alternately, some ERP researchers opt for a delayed response in such circumstances, so as to avoid contaminating the ERP to the target word with electrical potentials related to motor response preparation.

Because of the numerous ERP-specific issues in both the stimuli and experimental task, we provide suggested readings at the end of the chapter and recommend that novice ERP researchers consult with an experienced ERP practitioner who may be able to provide practical expertise related to experimental design.

Scoring, Data Analysis, and Reporting Results

Before ERPs can be analyzed and reported, there are several decisions that researchers make about data preprocessing. For example, there are numerous parameters related to filtering, epoching and artifact detection, and baseline correction that are meant to maximize the signal-to-noise ratio in the ERP averages, but that are beyond the scope of this chapter. Informed decisions need to be made about each processing step, as inappropriate processing can cause misleading or even spurious results. We refer the reader to Handy (see 2005), Luck (2005), and Picton et al. (2000) for more detailed information about the logic behind data processing parameters as well as what is acceptable. However, it is important to point out that after the EEG has been processed but before it is analyzed, researchers must decide whether to adopt a response-contingent analysis or to analyze data from all trials regardless of the subject's response on the trial. The use of all-trial versus response-contingent analysis varies in the L1 literature, though it is common in L2 ERP research to include all artifact-free trials. This is particularly relevant when learners are at lower levels of proficiency and do not achieve high levels of response accuracy; in these cases, adopting an all-trial analysis can be advisable because ERPs can reveal sensitivities that behavioral judgments do not (see description of McLaughlin et al., 2004, in the Exemplary Study section below).

After preprocessing the EEG and deciding whether to adopt a response-contingent or an all-trial analysis, ERP waveforms are produced by averaging the epoched EEG across items within subjects; "grand mean" group waveforms are created by averaging across individuals' ERPs. At this point, it is important to visually inspect both the grand mean waveforms and individuals' ERPs. Comparing grand mean waveforms across two or more conditions provides information about the effects of the experimental manipulation on participants' brain responses and is an important first step in understanding the data. However, one should note how well the averaged group waveform represents individuals' waveforms. Although the grand mean waveform represents the central tendency across both trials and participants, it is quite possible to observe clear ERP effects within individuals that are not represented by the group average (see Tanner et al., 2013, for a demonstration of this). The researcher should note this and consider it when interpreting group-based analyses.

Visual inspection of ERP waveforms is a necessary starting point for analyses, but robust conclusions cannot be made without proper statistical analyses. Researchers must make principled decisions about the parameters for ERP statistical analyses, as ERPs provide several possible outcome measures. Choice of the correct measure rests largely on the researcher's primary questions. For example, most ERP researchers are interested in "how much" brain response a manipulation elicited (e.g., are ERPs to anomalous words more negative-going than to control words, and are there differences across types of anomalies or groups of individuals). Although there are several options for studying waveform amplitudes, mean amplitude measures are generally the most appropriate and are the most common measure reported in language-related ERP studies. Mean amplitude is computed as the average voltage value within a time window of interest for each condition and at each electrode. Before measuring mean amplitude, researchers must determine the time window to be used for the analyses. Choosing the time window should be made in a principled manner and, ideally, the decision should be made a priori. It is fairly standard in language-related ERP research for researchers to base their initial time window on previous research (i.e., to initially assume a 300 to 500 ms time window for an N400 effect), but to adjust the time window based on visual inspection of the averaged ERP waveforms. For example, it is sometimes appropriate to use a 350 to 450 ms time window if the N400 effect appears restricted in its duration, or use a 400 to 600 ms time window if the effect appears to be delayed compared to the standard L1 time window.

Other times, the research question revolves around the latency of an ERP effect (i.e., "when" something happens). In these cases, fractional area latency (FAL) computed on difference waveforms (i.e., the point in time at which a certain fraction, e.g., 50%, of the mean area between the difference wave and x-axis has occurred) is often appropriate. Although other measures, such as peak latency (i.e., the point in a time within a time window with the highest absolute value), provide temporal information, they can be heavily impacted by even a small amount of noise in the ERP, whereas FAL is not similarly impacted. Moreover, many ERP effects of interest to language researchers (e.g., the P600) do not involve clear peaks in the waveforms, but rather differences between conditions. In such cases, peak latency is clearly not an appropriate measure. Importantly, amplitude and latency are not mutually exclusive within a study. Often a researcher is interested in both "how much" and "when," and can use both mean amplitude and FAL measures, which provide complementary information (see Newman, Tremblay, Nichols, Neville, & Ullman, 2012, for an excellent demonstration of this in L2 processing; see Handy, 2005; Luck, 2005, for thorough descriptions and evaluations of the various data quantification metrics available to ERP researchers).

A major issue that is relevant to both latency and amplitude measures is choosing the electrodes over which to conduct the analyses. In general, it is appropriate to include an array of electrodes over a broad portion of the scalp: frontal, central, and posterior electrodes, as well as electrodes over the left and right hemispheres and scalp midline. In making decisions about when and where to carry out analyses, researchers must be extremely cognizant that they are not fishing for results by only analyzing electrodes and time windows where visual inspection suggests potential effects. To the greatest extent possible, the decisions must be principled and systematic, based on a priori hypotheses and previous reports in the literature, and should be explicitly stated in the report of the study.

After having measured the data (i.e., microvolt [μV] measures for mean amplitude, or millisecond [ms] measures for latency), statistical analyses can be performed. The standard statistical approach to analyzing ERP data is to use analysis of variance (ANOVA). This will determine how the experimental manipulations impacted brain responses (e.g., if correct versus violation stimuli differed, and what the scalp topography of the relevant effects might be). ANOVAs that examine ERP effects typically include at least one repeated-measures factor, for example, Violation (violation versus control stimuli), and often also include other within-subject factors related to scalp topography, such as Hemisphere (right, left), and Anterior-Posterior (anterior, central, posterior). It is common for separate ANOVAs to be conducted for lateral and midline electrodes. The data analyzed can represent the signal either from one electrode or from a region of interest, where the signal is averaged over a group of neighboring electrodes. Depending on a study's research question, additional within- or between-subjects factors may also be included in the ANOVA. For example, if the researcher is interested in how an ERP violation effect differs for different linguistic forms (e.g., for an L2 form that is similar to an L1 form versus one that is different from an L1 form), a repeated-measures factor representing linguistic form would be included. Alternately, if the researcher is interested in how an effect differs among learners at different proficiency levels, a between-subjects factor representing proficiency group would be included in the ANOVA. As can be seen, ANOVA-based analyses of ERPs can be quite complex, as they can easily include three to five factors (or more!). Three-, four-, or five-way statistical interactions are not uncommon in ERP studies, and these can be very difficult to interpret. Because of this, simple experimental designs often yield the most interesting results in ERP studies.

Although analysis with ANOVAs is the standard method, there are several limitations inherent in this approach. One limitation of ANOVA-based analyses is that conclusions about any effects evidenced in a group of participants may not be valid at the individual level. As previously discussed, the group average is not necessarily representative of all the participants, and in fact, the average may not represent any individual's processing signature. Regression-based analyses that account for individual variation among participants can mitigate this issue to some degree (e.g., Tanner et al., 2013; Tanner et al., in 2013). Another limitation of ANOVA-based analyses is that the researcher chooses the data to include in the analysis. This choice can bias the outcomes towards significant results. This problem can be addressed through more advanced types of analyses, such as mass univariate analysis, which reduce researcher bias by systematically analyzing the complete dataset while correcting for Type 1 error rate (Groppe, Urbach, & Kutas, 2011a; Groppe, Urbach, & Kutas, 2011b). A final issue with standard analyses is that an effect in a surface ERP waveform does not necessarily correspond to one underlying cognitive process. Although ANOVA-based analyses are limited to analyzing the waveform itself, other more advanced analyses may be better at isolating the underlying components. Both time-frequency and independent component analyses may be useful in this regard (see Kappenman & Luck, 2012 for further discussion).

When reporting results from an ERP study, the ERP waveforms themselves should always be provided. Depending on the type of report or the scope of the effects investigated, waveforms from minimally one electrode most representative of the effects should be shown, though it is usually a good idea to show an array of electrodes (at least nine), depicting the distribution of effects across the scalp. Waveforms from multiple conditions may be overlaid on a single plot to allow comparison, though including more than three or four conditions on a single plot can create excessive visual clutter. At least 100 ms of prestimulus activity (the ERP's baseline) should also be depicted in the waveforms (see Osterhout et al., 2004; Steinhauer & Drury, 2012). This allows the reader to see how much noise was present in the ERPs, as well as if there were differences in the precritical region (see above). Sometimes, topographic maps can provide complementary information about the scalp distribution of the effects of interest, though these should never supplant the actual waveforms. Other forms of data visualization can also aid the reader in understanding the effects. For example, bar plots of mean amplitude measures (with standard error bars) depicting amplitudes in the various conditions can be provided, or scatterplots showing relationships between individuals' ERP effects and an individual difference measure of interest.

An Exemplary Study

McLaughlin, Osterhout, and Kim (2004) used ERPs to study how and when novice L2 learners begin to incorporate L2 vocabulary information into their on-line processing systems. To test this, they recorded ERPs while participants engaged in a French lexical decision task within an unmasked priming paradigm. The outcome of interest was any modulations to the N400 component. Stimuli consisted of prime-target pairs in three conditions, exemplified in (2), below:

(2)

a.	Semantically-related word pair:	*chien–chat*	("dog"–"cat")
b.	Semantically-unrelated word pair:	*maison–soif*	("house"–"thirst")
c.	Word-pseudoword pair:	*mot–nasier*	("word"–NONWORD)

Previous research on lexical processing using ERPs had shown that N400 amplitudes are largest to orthographically legal pseudowords (e.g., 2c), intermediate for words preceded by a semantically unrelated context (e.g., 2b), and smallest for words preceded by a semantically related context (e.g., 2a). The theoretical question of interest was how much L2 instruction was necessary before learners' brain responses were sensitive to the target words' lexical status and semantic relationship to the prime.

Participants were native English speakers enrolled in introductory French courses at a large university, and who had no previous knowledge of French. Participants were tested longitudinally, three times over the course of the first year of classroom L2 instruction, after approximately 14 hours, 63 hours, and 138 hours of instruction. For this study, instead of using a native French speaker control group, a control group of native English speakers who were not enrolled in French courses and who also had no knowledge of French was included. Control participants were also tested three times over the course of the year in order to establish that any effects seen in the learners were not a result of low-level perceptual properties of the target words themselves. Including this control group also allowed the researchers to establish whether any changes seen in brain responses across the sessions were associated with actual L2 learning, and not simply exposure effects from participating in the experiment over multiple sessions.

Each trial in the experiment consisted of a series of events: a fixation cross was presented, followed by the prime word, blank screen, target word, blank screen, and lexical decision prompt. Participants were instructed to decide whether the second word in each pair (the target) was a French word or not; ERPs were time-locked to the presentation of the target word. Results for the non-learner control group did not show any sensitivity to the words' lexical status in the behavioral decision task, and there were no differences in the amplitude of the N400 component between the conditions at any of the testing sessions. This shows that any differences between conditions in the learner group can reliably be construed as true evidence for learning.

In the learner group, the behavioral (i.e., lexical decision) results showed chance performance at the first session, indicating no ability to distinguish between French words and the pseudowords in overt judgments. This sensitivity increased only modestly by the end of the first year of L2 instruction. During the first ERP session, learners' brain responses did not show any difference between the semantically-related and unrelated prime-target pairs. However, they already showed a reliable increase in N400 amplitude for the pseudowords versus the two word conditions, despite inability to make accurate judgments about the items' word status. At this session, the magnitude of individuals' N400 effects (the difference in mean N400 amplitude between the pseudoword and word targets) was very highly correlated with the number of hours each individual had been exposed to French in the classroom prior to ERP testing ($r = 0.72$), but measures of individuals' behavioral sensitivity (d-prime scores) were fully uncorrelated with the number of hours of exposure ($r = 0.09$). This shows a systematic disconnect between brain and behavioral measures of sensitivity. By the end of the first year of instruction, N400 effects showed a pattern similar to that normally seen in native speakers: N400s were largest to pseudowords, intermediate to targets preceded by semantically-unrelated primes, and smallest to words preceded by semantically-related primes.

This study is notable for many reasons. First, it was the first longitudinal ERP study of L2 learning, not simply L2 processing. By studying learners longitudinally, McLaughlin and colleagues could see how increasing L2 competence can change individuals' responses to L2 information at the neural level, thus allowing much stronger inferences about learning and plasticity than a cross-sectional design would have. Second, this study showed that certain aspects of L2 vocabulary information are incorporated into learners' on-line processing systems very rapidly. Learners' brain responses distinguished between words and pseudowords after remarkably little instruction; however, only after several months of instruction did evidence of semantic links between words become evident in the ERP measures. Finally, and perhaps most astonishingly, this study showed that a highly sensitive measure of processing, like ERPs, can reveal evidence of learning that a behavioral measure does not. That is, ERPs showed that participants had already begun to acquire French wordform information after only a few hours of exposure, even though they could not accurately distinguish between well-formed and anomalous stimuli in a forced choice task.

Pros and Cons in Using the Method

Pros

- ERPs are multidimensional. ERPs provide an exceptionally rich source of data about processing, allowing researchers to make inferences not only about whether participants are sensitive to a given manipulation, but also about the nature of that sensitivity. For example, a reading slow-down using self-paced reading or eye-tracking lets the researcher know that an anomaly was processed, but ERPs allow further inferences about the type of processing which was engaged, or whether two different manipulations engaged processing mechanisms with different neural substrates.
- ERPs are very sensitive. Brainwave measures can sometimes give evidence of processing which goes unnoticed with behavioral measures. This aspect of ERPs can be particularly useful when studying early-stage L2 learners, for whom behavioral indices of processing might be equivocal (e.g., McLaughlin et al., 2004; Tokowicz & MacWhinney, 2005).
- ERPs effects can be observed with no secondary behavioral task. Although secondary tasks are frequently used in order to maintain participant attention (e.g., lexical decision, acceptability judgment, or comprehension question), ERP effects can be found even when the participant is specifically directed to ignore the stimuli and focus on something else (Hasting & Kotz, 2008), when participants do not consciously perceive an anomaly (Batterink & Neville,

2013), or even during sleep (Martynova, Kirjavainen, & Cheour, 2003). This is not to say that participant task has no impact on ERP responses, only that ERPs can be elicited in the absence of a secondary task.

Cons

- ERPs are expensive. Although not as expensive as fMRI research, up-front costs to build a lab are quite steep. An EEG amplifier system can cost anywhere from $25,000 to $75,000. Many EEG manufacturers require the user to additionally purchase a license for recording software, which can be several thousand dollars more. Commercial data analysis software packages can also cost several thousands of dollars; however, for users who are comfortable with a small amount of software coding, there are free toolboxes available for MATLAB (e.g., EEGLAB, ERPLAB, and FieldTrip toolboxes). These provide many high-level tools for EEG and ERP analysis, and they are rapidly becoming the standard for many labs around the word.
- ERPs are time-consuming. A recording session for a single participant in an ERP experiment can take anywhere from 1.5 to 3 hours, and often requires the presence of two research assistants for set-up. Depending on the number and type of electrodes used, simply preparing the cap on the participant's head can take over an hour for high-density recordings using traditional passive electrodes. Some modern amplifier systems have circuitry embedded within each electrode, which minimizes the need to lower scalp impedances. However, there is a trade-off between this time savings and the cost of the electrodes.
- ERPs require lots of stimuli. Because the ERP must be extracted from the EEG using an averaging procedure, a large number of stimuli must be used in each condition—many more than with behavioral measures of processing. Typical lexical or sentence processing studies usually include 30 to 40 stimuli per experimental condition, or more depending on the research question. When creating stimuli in a factorial design with counterbalanced lists (see Jegerski, Chapter 2, this volume), this can require the researcher to create several hundred (or even thousands of) stimuli for a given experiment. For certain experimental designs where lexical material must be tightly controlled (e.g., for cross-language orthographic neighborhood density), it can become infeasible to create enough stimuli to conduct an ERP experiment. In cases where enough stimuli can be created in each condition, the length of the ERP recording session puts limits on the number of conditions which can feasibly be included in a given experiment. Long recording sessions can lead to participant fatigue, which makes the data collected toward the end of the experiment noisier and less reliable. This constraint makes some multi-factorial, within-subjects designs difficult to achieve with ERP.

- The multiple dimensions of ERPs can make quantification difficult. Although multi-dimensionality is a strength of ERPs (see above), it can also be a weakness. Because ERPs can vary across trials or participants in numerous ways (latency, scalp topography, and polarity), it can be difficult to know which dimension is most impacted by the experimental manipulation. Moreover, ERP waveforms do not come "prelabeled" with components. For example, it may be difficult to know if an experimental manipulation elicited a positivity in the anomalous condition or a negativity in the control condition, or whether a negative deflection is a LAN or an N400. Also, some experimental manipulations may elicit effects which do not resemble previously reported ERP effects (e.g., an N400, LAN, or P600), making interpretation difficult. Additionally, when ERPs differ in quality across individuals (e.g., if some individuals show an N400 while others show a P600), the grand mean response will show either a combination of the two responses or a null effect. In this case, the final result does not accurately reflect the brainwave pattern in any individual in the participant, which in turn can lead to spurious conclusions.
- It is difficult to use ERPs to study language production. Because electrical signals related to muscle movement can be orders of magnitude larger than the ERP signal of interest, it is difficult to use ERPs to study brainwave fluctuations while a participant is speaking. Some studies have begun to use ERPs during isolated picture naming (e.g., Blackford, Holcomb, Grainger, & Kuperberg, 2012), although care must be taken to properly screen data for muscle artifact, and only the early portion of the data (pre-articulation) can reliably be analyzed. However, newer techniques for noise reduction, such as Independent Components Analysis, may prove to be useful in eliminating interference from muscular artifact, such that ERP correlates of more natural speech production can be analyzed in the future.

Discussion and Practice

Questions

Note: In order to respond to the second part of several discussion questions, the reader should have access to the following articles:

Rossi, S., Gugler, M. F., Friederici, A.D., & Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: Evidence from event-related potentials. *Journal of Cognitive Neuroscience, 18*(12), 2030–2048.

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience, 24*(5), 1205–1223.

1) ERP waveforms and effects can be described in terms of their characteristics along three dimensions.

   a. Identify the three dimensional characteristics used to describe ERPs.
   b. What does the term *ERP effect* generally refer to?
   c. Before reading the abstract or results of Newman et al. (2012), examine Figure 2 on page 1211. Inspect the waveforms for the native speakers and describe the characteristics of the ERP waveform elicited by the stimuli from the "acceptable" condition.
   d. For the native speakers in Figure 2, describe how the ERP waveform elicited by the stimuli for the violation condition differs from that of the acceptable condition waveform. What typical language-related ERP effect (e.g., LAN, N400, P600) is evident in the comparison of the violation and acceptable waveforms?
   e. Discuss how the ERP effect for late learners, also seen in Figure 2, is similar to or differs from the ERP effect for native speakers.

2) When designing an ERP study, both the critical target words (i.e., the stimuli to which the ERP analysis is time-locked) and the precritical material should be carefully matched or controlled between conditions.

   a. List characteristics of ERP stimuli that should be matched or controlled across conditions; for example, violation and correct (or control) conditions for target words as well as for precritical words and context.
   b. Consider the stimuli from the Newman et al. (2012) and Rossi et al. (2006) studies. For each study, discuss whether the target words and precritical material are adequately controlled.

3) Although ERP data can provide unique insight into questions about language processing, the interpretation of ERP data is constrained to particular inferences.

   a. What are the inferences that can be made based on each of the dimensional characteristics of ERP waveforms (which you identified in discussion question 1)?
   b. Note the research questions from Newman et al. (2012) and Rossi et al. (2006) and discuss what ERP characteristics would be relevant such that appropriate inferences could be made to answer the research questions.
   c. Consider whether Newman et al. (2012) and Rossi et al. (2006) arrived at appropriate conclusions based on the inferences that they made from their data. Do any conclusions represent over-interpretation of the results?

4) One peril with ERP research is the tendency to believe that collecting ERP data for any given study is more informative than collecting behavioral data or another type of on-line data. Given your own research interests, first

identify an existing L2 study that would be enhanced by the collection of ERP data. Second, identify an existing L2 study that would *not* be enhanced by the collection of ERP data. Finally, give specific reasons justifying your selection of empirical studies.

5) For discussion question 4, you identified one L2 study that could be enhanced by the collection of ERP data. Consider the changes to the stimuli and procedure that would need to be implemented in order to run that study as an ERP study.

Research Project Option A

In any scientific field of inquiry, replication of results leads to the field's ability to arrive at robust conclusions about a particular question. Replication is also an excellent way to gain new methodological skills. Two L2 ERP studies that merit replication are Tokowicz & MacWhinney (2005) and McLaughlin et al. (2004). Both studies yielded intriguing results that are highly cited in ERP literature, in part because they found significant ERP effects even in the absence of behavioral evidence of learning for L2 words (McLaughlin et al., 2004) and L2 grammar (Tokowicz & MacWhinney, 2005). The reader may choose to replicate one of these studies either by requesting the stimuli from the researchers or by creating a new set of stimuli based on the description of the stimuli in the published articles (which would then allow the reader to extend the results of the original study to a different L2).

Research Project Option B

Very little research has considered whether the task that participants engage in as ERP data is collected affects ERP results. Two studies carried out by Osterhout and colleagues (Osterhout, McKinnon, Bersick, & Corey, 1996; Osterhout et al., 2002) have shown that in L1 processing, the task, in this case reading sentences for comprehension versus judging their acceptability, does not qualitatively affect the ERP response to the stimuli. To our knowledge, no research has directly examined the effect of task on ERP results in L2 processing. Given that L2 performance depends at least partially on the type of assessment task administered (Larsen-Freeman, 1976; Sanz, 1997), it would be useful to explore whether L2 processing is also influenced by the task that participants engage in as ERP data is collected. A fruitful research project would be to replicate a selected study of L2 learners. The new study would then use a between-subjects design, where different learner groups would be assigned to different task conditions, such as directing learners to judge the grammaticality of sentences or directing learners to simply listen to sentences. A cross-group comparison would indicate whether the task itself may have influenced the results of the study. The researcher should be cognizant, though, that the validity of these results would depend on carefully matching participants across the groups on a range of variables, including L2 proficiency, type and length of L2 experience, and so on.

Suggested Readings

The following published ERP studies of L2 development can be consulted for examples of research questions, experimental design, methodology, results reporting, statistical analyses, and interpretation of results:

Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire new grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language, 66,* 226–248.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience, 24,* 933–947.

Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2013).Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition, 16,* 367–382.

White, E. J., Genesee, F., & Steinhauer, K. (2012). Brain responses before and after intensive second language learning: Proficiency based changes and first language background effects in adult learners. *PLOS ONE,* 7(12), e52318.

Although this chapter has provided a basic introduction and overview of ERP methods, it is advisable to both broaden and deepen one's knowledge of the method before investing time in conducting an ERP study. We direct the reader to the following sources that provide more detailed information about ERPs and ERP techniques, methods and standards:

Handy, T. C. (2005). *Event-related potentials: A methods handbook.* Cambridge, MA: MIT Press.

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Cambridge, MA: MIT Press.

Luck, S. J. (2012). Event-related potentials. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Volume 1: Foundations, planning, measures, and psychometrics.* Washington, DC: American Psychological Association.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., . . . Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37,* 127–152.

Rugg, M. D., & Coles, M. G. H. (Eds.). (1995). *Electrophysiology of mind.* New York: Oxford University Press.

References

Ardal, S., Donald, M. W., Meuter, R., Muldrew, S., & Luce, M. (1990). Brain responses to semantic incongruity in bilinguals. *Brain and Language, 39,* 187–205.

Bastiaansen, M., Mazaheri, A., & Jensen, O. (2012). Beyond ERPs: Oscillatory neuronal dynamics. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 31–50). New York: Oxford University Press.

Batterink, L. & Neville, H. (2013). The human brain processes syntax in the absence of conscious awareness. *Journal of Neuroscience, 33,* 8528-8533.

Berger, H. (1929). Über das Elektrenkephalogramm. *Archiv für Psychiatrie und Nervenkrankheiten, 87,* 527–570.

Blackford, T., Holcomb, P., Grainger, J., Kuperberg, G.R. (2012). A funny thing happened on the way to articulation: N400 attenuation despite behavioral interference in picture naming. *Cognition, 123,* 84–99.

Deutsch, A., & Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: Evidence from ERPs and eye movements. *Journal of Memory and Language, 45,* 200–224.

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition, 14,* 379–399.

Friedman, D., Simson, R., Ritter, W., & Rapin, I. (1975). The late positive component (P300) and information processing in sentences. *Electroencephalography and Clinical Neurophysiology, 38,* 255–262.

Geyer, A., Holcomb, P. J., Midgley, K. J., & Grainger, J. (2011). Processing words in two languages: An event-related brain potential study of proficient bilinguals. *Journal of Neurolinguistics, 24*(3), 338–351.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011a). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology, 48*(12), 1711–1725.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011b). Mass univariate analysis of event-related brain potentials/fields II: Simulation studies. *Psychophysiology, 48*(12), 1726–1737.

Hagoort, P., & Brown, C. M. (1999). Gender electrified: ERP evidence on the syntactic nature of gender processing. *Journal of Psycholinguistic Research, 28*(6), 715–728.

Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia, 38,* 1531–1549.

Handy, T. C. (2005). *Event-related potentials: A methods handbook.* Cambridge, MA: MIT Press.

Hasting, A. S., & Kotz, S. A. (2008). Speeding up syntax: On the relative timing and automaticity of local phrase structure and morphosyntactic processing as reflected in event-related brain potentials. *Journal of Cognitive Neuroscience, 20*(7), 1207-1219.

Jasper, H. H. (1958). The ten–twenty system of the International Federation. *Electroencephalography and Clinical Neurophysiology, 10,* 371–375.

Kappenman, E. S., & Luck, S., J. (2012). ERP components: The ups and downs of brainwave recordings. In S. Luck J., & E. S. Kappenman (Eds.), *The oxford handbook of ERP components* (pp. 3–30). New York: Oxford University Press.

Kim, A., & Osterhout, L. (2005). The independence of combinatory semantic processing: Evidence from event-related potentials. *Journal of Memory and Language, 52,* 205–225.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research, 1146,* 23–49.

Kuperberg, G. R., Caplan, D., Sitnikova, T., Eddy, M., & Holcomb, P. J. (2006). Neural correlates of processing syntactic, semantic, and thematic relationships in sentences. *Language and Cognitive Processes, 21,* 489–530.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62,* 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic anomaly. *Science, 207,* 203–205.

Larsen-Freeman, D. E. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning, 26*(1), 125–134.

Laszlo, S., & Federmeier, K. D. (2009). A beautiful day in the neighborhood: An event-related potential study of lexical relationships and prediction in context. *Journal of Memory and Language, 61*(3), 326–338.

Luck, S. J. (2005). *An introduction to the event-related potential technique.* Cambridge, MA: MIT Press.

Martynova, O., Kirjavainen, J. & Cheour, M. (2003) Mismatch negativity and late discriminative negativity in sleeping human newborns. *Neuroscience Letters, 340*, 75–78.

McKinnon, R., Allen, M., & Osterhout, L. (2003). Morphological decomposition involving non-productive morphemes: ERP evidence. *NeuroReport: For Rapid Communication of Neuroscience Research, 14*(6), 883–886.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience, 7,* 703–704.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009a). Language effects in second language learners and proficient bilinguals investigated with event-related potentials. *Journal of Neurolinguistics, 22*(3), 281–300.

Midgley, K. J., Holcomb, P. J., & Grainger, J. (2009b). Masked repetition and translation priming in second language learners: A window on the time-course of form and meaning activation using ERPs. *Psychophysiology, 46*(3), 551–565.

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potentials study. *Language Learning, 60,* 154–193.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of Cognitive Neuroscience, 24,* 933–947.

Nakano, H., Saron, C., & Swaab, T. Y. (2010). Speech and span: Working memory capacity impacts the use of animacy but not of world knowledge during spoken sentence comprehension. *Journal of Cognitive Neuroscience, 22,* 2886–2898.

Nenonen, S., Shestakova, A., Huotilainen, M., & Näätänen, R. (2005). Speech-sound duration processing in a second language is specific to phonetic categories. *Brain and Language, 92,* 26–32.

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M. T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience, 24*(5), 1205–1223.

Osterhout, L. (1997). On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences. *Brain and Language, 59,* 494–522.

Osterhout, L., Allen, M. D., McLaughlin, J., & Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Memory & Cognition, 30*(8), 1304–1312.

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language, 31,* 785–806.

Osterhout, L., Kim, A., & Kuperberg, G. R. (2012). The neurobiology of sentence comprehension. In M. Spivey, M. Joannisse, & K. McCrae (Eds.), *The Cambridge Handbook of Psycholinguistics* (pp. 365–389). Cambridge, UK: Cambridge University Press.

Osterhout, L., McKinnon, R., Bersick, M., & Corey, V. (1996). On the language specificity of the brain response to syntactic anomalies: Is the syntactic positive shift a member of the P300 family? *Journal of Cognitive Neuroscience, 8*(6), 507–526.

Osterhout, L., McLaughlin, J., Kim, A., Greewald, R., & Inoue, K. (2004). Sentences in the brain: Event-related potentials as real-time reflections of sentence comprehension and language learning. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs, and beyond* (pp. 271–308). New York: Psychology Press.

Otten, L. J., & Rugg, M. D. (2005). Interpreting Event-Related Brain Potentials. In T. C. Handy (Ed.), *Event-related potentials: A methods handbook* (pp. 3–16). Cambridge, MA: MIT Press.

Phillips, N. A., Segalowitz, N., O'Brien, I., & Yamasaki, N. (2004). Semantic priming in a first and second language: Evidence from reaction time variability and event-related brain potentials. *Journal of Neurolinguistics, 17*(2–3), 237–262.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Jr., . . . Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology, 37*(2), 127–152.

Rugg, M. D., & Coles, M. G. H. (1995). The ERP and cognitive psychology: Conceptual issues. In M. D. Rugg & M. G. H. Coles (Eds.), *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 27–39). Oxford, UK: Oxford University Press.

Sanz, C. (1997). Experimental tasks in SLA research: Amount of production, modality, memory, and production processes. In W.R. Glass and A.T. Pérez-Lerroux (Eds.) *Contemporary perspectives on the acquisition of Spanish* (vol. 2, pp. 41–56). Somerville, MA: Cascadilla Press.

Schoonbaert, S., Holcomb, P. J., Grainger, J., & Hartsuiker, R. J. (2011). Testing asymmetries in noncognate translation priming: Evidence from RTs and ERPs. *Psychophysiology, 48*(1), 74–81.

Shelburne, S. A. (1972). Visual evoked responses to word and nonsense syllable stimuli. *Electroencephalography and Clinical Neurophysiology, 32,* 17–25.

Steinhauer, K., & Connolly, J. F. (2008). Event-related potentials in the study of language. In B. Stemmer & H. Whitaker (Ed.), *Handbook of the cognitive neuroscience of language* (pp. 91–104). New York, NY: Elsevier.

Steinhauer, K., & Drury, J. E. (2012). On the early left-anterior negativity (ELAN) in syntax studies. *Brain and Language, 120*(2), 135–162.

Steinhauer, K., White, E. J., & Drury, J. E. (2009). Temporal dynamics of late second language acquisition: Evidence from event-related brain potentials. *Second Language Research, 25,* 13–41.

Sutton, S., Tueting, P., Zubin, J., & John, E. R. (1965). Evoked potential correlates of stimulus uncertainty. *Science, 150,* 1187–1188.

Swaab, T. Y., Ledoux, K., Camblin, C. C., & Boudewyn, M. A. (2012). Language related ERP components. In S. J. Luck & E. S. Kappenman (Eds.), *Oxford handbook of event-related potential components.* (pp. 397–440). New York, NY: Oxford University Press.

Tanner, D., Inoue, K., & Osterhout, L. (2013). Brain-based individual differences in on-line L2 sentence comprehension. *Bilingualism: Language and Cognition.* Advance online publication. doi: 10.1017/S1366728913000370

Tanner, D., Mclaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition, 16,* 367–382.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar—An event-related potential investigation. *Studies in Second Language Acquisition, 27,* 173–204.

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology, 39,* 791–808.

van Hell, J. G., & Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. *Second Language Research, 26,* 43–74.

Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C., & Winter, A. L. (1964). Contigent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. *Nature, 203*, 380–384.

Weber-Fox, C. M., & Neville, H. J. (1996). Maturational constraints on functional specializations for language processing: ERP and behavioral evidence in bilingual speakers. *Journal of Cognitive Neuroscience, 8,* 231–256.

White, E. J., Genesee, F., & Steinhauer, K. (2012). Brain responses before and after intensive second language learning: Proficiency based changes and first language background effects in adult learners. *PLOS ONE*, 7(12), e52318.

7

FUNCTIONAL MAGNETIC RESONANCE IMAGING (fMRI)

Aaron J. Newman

History of the Method

Functional magnetic resonance imaging (fMRI) celebrated its twentieth birthday in 2012 (Kwong et al., 1992; Ogawa et al., 1992). In this time, it has become the most widely-used technique for noninvasively investigating human brain activity in vivo.[1] The first publications showing noninvasive measurement of human brain activity were preceded by more than 100 years of observation and research. In 1890, the physiologists Roy and Sherrington demonstrated that brain volume increased when the peripheral nerves of animals were electrically stimulated, leading them to posit "... the existence of an automatic mechanism by which the blood-supply of any part of the cerebral tissue is varied in accordance with the activity of the chemical changes which underlie the functional action of that part" (Roy & Sherrington, 1890, p. 105). Another key observation was published by Pauling and Coryell in 1936—that hemoglobin, the molecule in our blood that transports oxygen, has different magnetic properties depending on whether it is bound to oxygen (oxyhemoglobin—Hb) or not (deoxyhemoglobin—dHb). Hb is magnetically neutral, while dHb is *paramagnetic,* meaning that it has weak magnetic properties and will be attracted to a magnetic field. Work in chemistry and physics led to the development of the nuclear magnetic resonance (NMR) technique in the 1970s, which led soon after to the MRI, which uses strong magnetic fields, for in vivo medical imaging. Ogawa, Lee, Kay, and Tank (1990) published work showing that a particular type of MRI image, called T2*-weighted, was sensitive to the amount of oxygen in the brain. Blood vessels in the brains of rats that were breathing room air appeared much darker in the MRI images than when the rats were breathing pure oxygen. This provided a potential *intrinsic contrast* mechanism—a change in MRI signal sensitive to blood oxygenation. This work led to the publication in 1992 of three papers demonstrating the use of this intrinsic contrast to measure brain activation in humans (Bandettini, Wong, Hinks, Tikofsky, & Hyde, 1992; Kwong et al., 1992; Ogawa et al., 1992). Together these three papers opened the door to a means of viewing brain activity in living humans at a spatial resolution of a few millimeters, using hardware that was widely available in most hospitals.

What is Looked at and Measured

The Nature of the fMRI Signal

The above description of fMRI highlights two important considerations regarding the technique. First of all, we are not measuring neural activity directly—fMRI is sensitive to changes in the amount of oxygen in the blood. Indeed, the fMRI signal is commonly referred to as the BOLD (blood oxygenation level dependent) signal. Increases in the BOLD signal in fMRI studies are often reported as "activation," but this is an inference based on an indirect measure. Thus, it is important to understand how such changes in blood oxygenation are related to changes in neural electrical activity. We will discuss this below. Secondly, note that Ogawa et al. (1990) found that a decrease in the Hb : dHb ratio (i.e., a net increase in dHb) was associated with a drop in signal. Given that as neural activity increases, the energy demands of the active neurons also increase, we would expect increased oxygen consumption and thus a lowered Hb : dHb ratio in active brain areas—meaning that we should see a *drop* in fMRI BOLD signal in active areas. Yet fMRI studies always report an *increase* in BOLD signal in active brain areas.

The reason for the second part of the puzzle posed above—why BOLD signal increases even though oxygen consumption should increase in active neural tissue—was solved first. Malonek and Grinvald (1996) used optical imaging, which can separately measure levels of Hb and dHB. They found that when the occipital lobes of cats were stimulated by visual input, the expected increase in dHB occurred very quickly, reflecting increased oxygen consumption. However, soon after this there was a large flood of oxygenated blood delivered to the active area, which covered a larger area than the initial increase in dHb. Malonek and Grinvald described this phenomenon as "watering the entire garden for the sake of one thirsty flower" (Malonek & Grinvald, 1996, p. 554). In other words, the increase in BOLD signal is due to the fact that neural activity leads to the delivery of much more oxygenated blood than is needed, to an area greater than that containing the neurons whose activity increased. We now know that this mechanism involves a complex cascade of biochemical events mediated by astrocytes, a type of glial cell that exist in the brain alongside neurons (Magistretti & Pellerin, 1999; Takano et al., 2006).

It is critical to keep this in mind, because it imposes certain limits on the inferences we can make from fMRI data. First, as noted above, we are not measuring neural activity, we are measuring blood oxygenation. Second, this relationship is mediated by a number of intermediate steps. Thus the timing of the fMRI BOLD response (or even the relative timing of the response between different brain areas) may not be reflective of the actual time-course of neural activity. Indeed, it should be noted that the time-course of the BOLD response is very sluggish compared to more direct measures of brain electrical activity such as EEG/ERP or MEG: in response to a transient stimulus (such as a brief presentation of a single word) the BOLD response takes approximately 2 seconds to start to rise after stimulation, 6 to 8 seconds to peak, and does not return to baseline until 12 to 20 seconds after stimulation. The time-course of the BOLD response to a brief transient event is called the *hemodynamic response function* or HRF, and is shown in Figure 7.1. Third, the magnitude of the fMRI BOLD response in different brain areas may be mediated not only by the level of neural activity, but by these intermediate physiological pathways. Thus it may be inaccurate to conclude that one brain area was more active than another in a particular condition. On the other hand, we would expect *within* a given brain area that the relative timing and magnitude of the BOLD signal between conditions would be informative.

As noted, it is also important to understand exactly how changes in neural activity mediate the BOLD response. Work by Logothetis, Pauls, Augath, Trinath, and Oeltermann (2001), combining fMRI with direct electrical recordings in the brains of monkeys, gave us our first insights into this. Logothetis

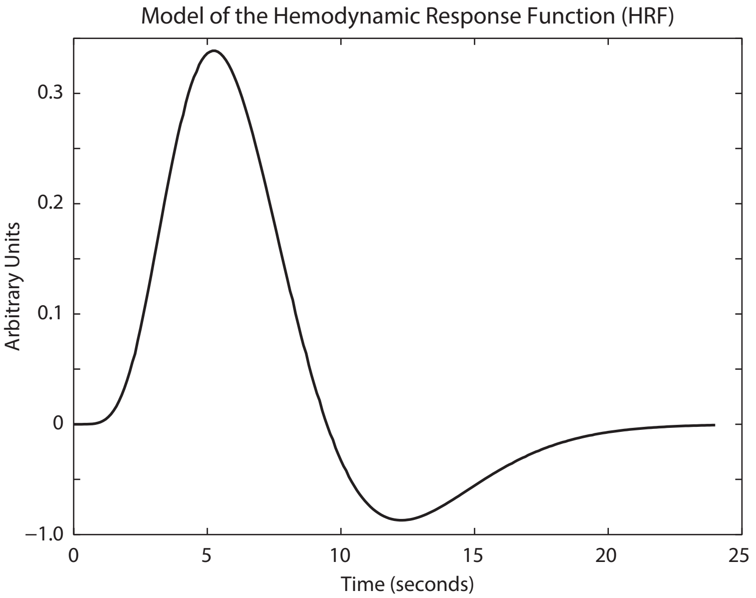


FIGURE 7.1 Model of the hemodynamic response function, based on Glover (1999).

et al. found that BOLD signal was most strongly related to local field potentials—the slow, aggregate changes in electrical potential of all the neurons within a local area of the brain—rather than to action potentials. Since local field potentials reflect the input to a brain area from other neurons, while action potentials reflect the output of the neurons in a region, this finding suggested that fMRI is more sensitive to the input that a brain area receives rather than the output of whatever processing it does. Interestingly, EEG/ERP and MEG are primarily sensitive to the same changes in brain activity, giving hope to the possibility that the combination of these different measures of brain activity would be informative.

The Output from the MRI Scanner

While different studies have different requirements, each fMRI *run* normally takes about 5 minutes (though these can range from 4 to 20 min or longer), during which time a large number of samples, or scans of the brain, are acquired. Each sample is a three-dimensional brain volume, which is composed of a series of two-dimensional slices through the brain. Each slice is commonly 3 to 4 mm thick, and is in turn composed of a two-dimensional matrix of *voxels* (volume pixels)—typically a 64 × 64 voxel matrix, with each voxel being 3 to 4 mm along each dimension. It is not uncommon for the voxels to be thicker than their within-slice dimensions, because each slice takes a certain amount of time to acquire. Thus researchers may choose to collect fewer, thicker slices in the interests of acquiring the images at a faster rate. However, most modern fMRI hardware is capable of acquiring a volume consisting of 3 × 3 × 3 mm voxels covering the whole brain every 2 seconds (the temporal sampling rate is commonly referred to as *TR* or repetition time). Thus, at the end of an fMRI scan a file is produced that has four dimensions: the three spatial dimensions plus time. Figure 7.2 shows the spatial and temporal dimensions of a sample fMRI scan.

Although researchers continually attempt to push the limits of fMRI's spatial or temporal resolution (or both), the parameters described above are common at present for principled reasons. For one, the ratio of signal to noise (commonly abbreviated as SNR) in each voxel is proportional to its volume. If we assume that the amount of noise in each measurement is fixed (determined by the scanner and the acquisition parameters), then as we increase the volume of a voxel, we should increase the ratio of signal to noise because we will capture signal from more tissue. Thus, while we would like to be as precise in our localization as possible, our sensitivity to changes in activation decreases if we push the spatial resolution too far. On the other hand, the acquisition parameters described here are by no means absolute limits. For one thing, the strength of the magnetic field of the MRI scanner in part determines the ratio of signal to noise. While 1.5 Tesla (T—a measure of magnetic field strength) scanners were the standard when fMRI was first performed, currently 3T scanners are very common, and increasing numbers of 7-and even 9T human MRI systems are being installed. These systems allow much higher spatial

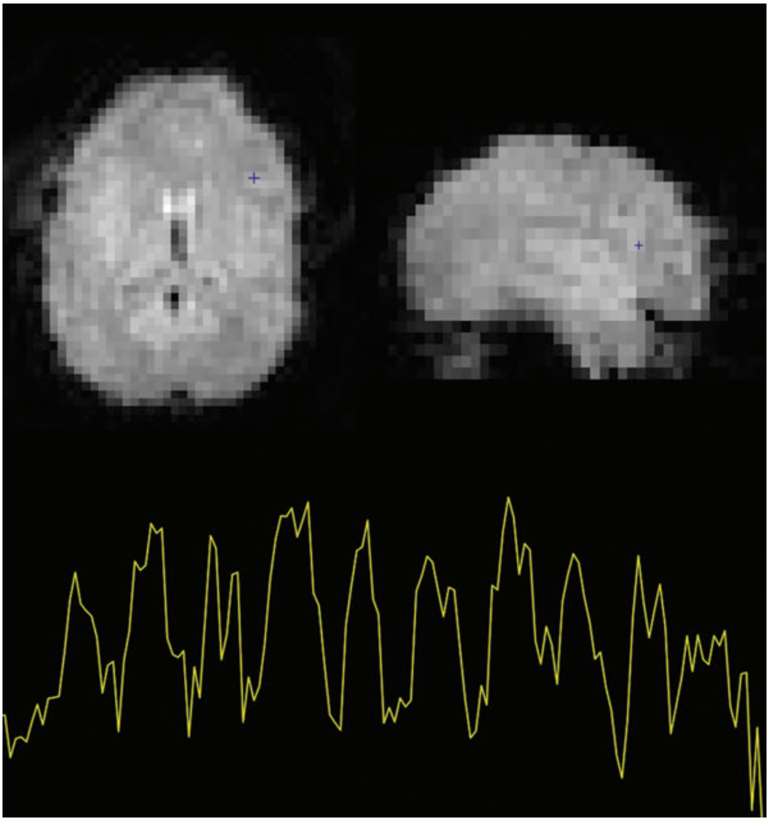


FIGURE 7.2 An example fMRI image. The top left panel shows a single axial slice from a single time point of a T2★-weighted fMRI scan; the front of the head is at the top of the image and the back of the head is at the bottom. The left side of the brain is shown on the right side of the image; this is a typical convention in medical imaging. The image is a 64 × 64 matrix of values (voxels); each voxel is 3.75 × 3.75 mm. The top right panel shows the scan in the saggital (side) view, with the front of the head to the right of the image. For reference, the blue crosshairs indicate the same point in the brain (in the left inferior frontal gyrus) in the two image views. In this study, 23 axial slices were acquired; each slice was 5 mm thick. The bottom panel shows the time course of the signal over the course of the run, in the voxel marked with the crosshairs. These data were obtained during a simple on-off block design experiment in which reading a story in English alternated with viewing consonant strings every 12–15 seconds. The selected voxel shows clear task-related modulation.

resolution, down to 1 mm or less. Such high resolution scans can yield information that is lost at lower resolution. However, higher-field MRI scanners come with greater technical challenges and costs, which may not be necessary or worthwhile to address many research questions. Many research questions concern the activity of relatively large areas of the brain and thus do not necessarily require higher spatial resolution.

In the time domain, the BOLD response is so sluggish that a 2 sec sampling rate is generally sufficient to address most questions that a researcher interested in the cognitive neuroscience of language (or most other tasks) might ask. However, since differences in the timing within a brain area between conditions may be informative, higher temporal resolution may be beneficial in some cases. Furthermore, since statistical power increases with the number of data points obtained, increasing the sampling rate for a scan of a set duration may yield more robust statistical results. However, it is important to understand that "more" spatial or temporal resolution is not always better, and the trade-offs that have to be made to get these gains in resolution may not be worth the cost.

Another important thing to keep in mind is that in addition to the fMRI scans, researchers will typically want to acquire a structural MRI scan. The low-resolution fMRI scans are not optimized to distinguish between grey matter, white matter, and other materials, so it is typically difficult to identify particular anatomical structures in an fMRI scan. In addition, as we will see below, anatomical scans are useful in mapping the activations of individuals to a common space. An anatomical MRI scan with 1 × 1 × 1 mm resolution (typically with what is referred to as T1 weighting—in contrast to the T2* weighting of fMRI images) normally takes only a few minutes to acquire.

Issues in the Development and Presentation of Stimuli

Basic Experimental Designs for fMRI

The nature of the BOLD signal, as well as the MRI hardware and the parameters used for fMRI scanning, impose some constraints on the experimental designs that can be used for fMRI. fMRI experimental designs fit broadly into two categories: block designs and event-related designs. We will describe each of these in more detail below, as well as their relative advantages and disadvantages.

Block designs for fMRI. Since the BOLD signal has such a protracted time-course (see Figure 7.1), we cannot easily distinguish the brain activity associated with two different conditions that occur closer in time than a few seconds. Early fMRI studies all used "block" designs in which a single type of stimulation or task was performed for approximately 20 to 30 sec, alternating with "off" blocks in which a control stimulation/task occurred. The alternation of conditions was necessary because fMRI is not a quantitative technique—the absolute values at each voxel are in arbitrary units. Brain activation in fMRI is defined only by the difference in BOLD signal within a voxel between different conditions. The duration of each block was chosen to optimize the signal strength. However, such block designs obviously impose certain constraints on experimental design. For example, it is impossible to conduct a block-design study in which the order of the stimuli were randomized, and the objective was to distinguish the BOLD responses to these individual stimuli. This precluded many psycholinguistic designs, such as priming (since the responses to prime and target words, typically separated by less than 1 sec, could not be distinguished) or the semantic and syntactic violation paradigms common in ERP studies of sentence processing (see Morgan-Short & Tanner, Chapter 6, this volume), which are dependent on the violations being somewhat unpredictable. Eventually, it was demonstrated that it was possible to resolve the BOLD signal to single events using different approaches, as we will see in the following section.

As an example of a simple block design experiment, let's say a researcher wanted to identify brain areas involved in processing the meanings of words. One could present a series of words to the subject (1 per second) for 30 sec, and then display a black screen for the next 30 sec, then switch back to word presentation, then the black screen again, and so on. Brain areas whose BOLD signal fluctuated on this same 30 sec cycle could be interpreted as being involved in reading words and accessing their meaning. However, the most important aspect in fMRI experimental design is thinking carefully about the two (or more) conditions being contrasted. What is different between reading words, and seeing a black screen? Certainly, in one case the subjects are receiving linguistic input and in the other they are not. However, if we "unpack" this we realize that a number of neurocognitive processes are involved.

First, we would expect the visual system to respond differentially in the two conditions. In this design we could not distinguish between areas involved in processing meaning and those involved in low-level visual processing. A preferable design would be to use some visual control stimulus rather than a blank screen. The best control stimulus will be matched on all of the nonrelevant features of the target stimulus (in this case, words). So pictures of animals would be a poor visual control, while a series of "false fonts" (letter-like stimuli) that cover the same average area of the visual field as the real words, and have the same brightness, would be ideal. However, false fonts may not control other extraneous factors. For example, real words have a phonological realization that may be automatically activated, and the visual word form area shows stronger activation for items that follow simple vowel-consonant alternations than false fonts or even consonant strings (Dehaene & Cohen, 2011). Thus, the real words and false fonts differ in several respects besides meaningfulness. Pronounceable nonwords may be the optimal control stimulus in this study. Of course, if the pronounceable nonwords are too like real words, they may engage the semantic system *more* strongly if it is tricked into looking for a match!

Another important consideration is the task. It is generally valuable to give subjects a task to perform, rather than simply passively viewing stimuli. This way, we can be confident that subjects were awake and paying attention to the stimuli,[2] and that across subjects similar processing occurred. Of course, it is not enough to simply ask subjects to perform a task, the researcher should also collect the behavioral responses, and analyze them. This way, the fMRI data from subjects who did not perform the task (or who did not perform it properly) can be excluded from statistical analysis. Furthermore, it is possible to include behavioral performance as a covariate in fMRI analysis, since performance can modulate brain activity.

The block design described above is the simplest possible fMRI design: a basic on–off–on–off sequence, with the contrast between the on and off blocks reflecting the differences in BOLD signal between the two conditions. One can also devise more complex block designs. For example, in a study of American Sign Language (ASL), we wished to compare brain activation between three different types of ASL sentences, which varied in their syntactic and metanarrative properties (Newman, Supalla, Hauser, Newport, & Bavelier, 2010a, 2010b). For each condition (type of sentence), we had a number of video clips of someone signing the sentences (see Figure 7.3). Although we matched the lexical content of the sentences as much as possible, the sentences differed visually—some had more facial expression and body movement, were of differing duration, and included slightly different lexical items. Thus we included separate control conditions corresponding to each of the three experimental conditions,[3] by digitally overlaying three different movies of the same type played backwards; pilot testing showed that signers could not understand these. They were not ideal control stimuli in the

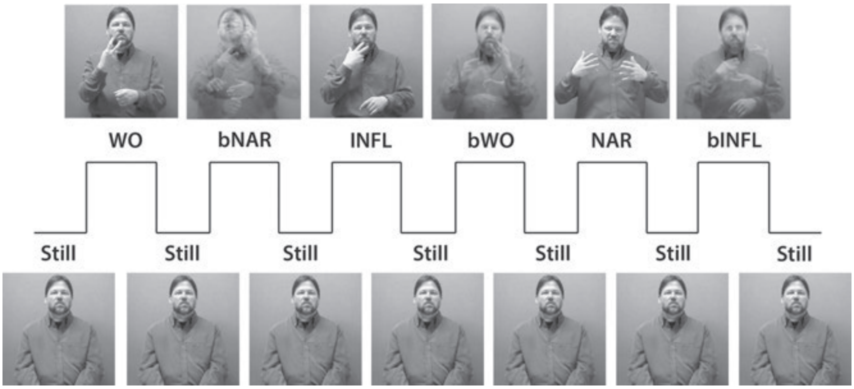


FIGURE 7.3 Examples of stimuli used in an fMRI study contrasting different linguistic devices in American Sign Language (ASL) processing. Videos of three different types of ASL sentences were presented, in separate blocks: (i) Sentences using word order (WO) to convey grammatical subject and object roles; (ii) Sentences that included complex ASL morphology (INFL); and (iii) Sentences that included narrative devices (NAR), including affective prosody and facial expressions. While lexical and semantic content were closely matched between the sentence types, each type differed in its visual-spatial properties (most noticeably in this figure, the facial expression in the NAR condition). Thus we created control videos for each sentence type by overlaying videos of three different sentences of the same type, played backward (bWO, bINFL, and bNAR). Each sentence or control block was 21 seconds in duration, and separated from each other block by a 15 second period in which a still image of the signer was displayed. These stimuli were used in the studies reported in Newman et al. (2010a, 2010b).

sense that they actually contained *more* visual/biological motion, facial expression, and such than the forward ASL sentences. However, these would not activate brain areas related to language processing, but rather other aspects of processing that we did not wish to examine anyway. Thus, the final experimental design for this study included seven different types of blocks: the three different ASL sentence types, the three corresponding control conditions, and "rest" blocks during which a still image of the signer with his hands at his sides was displayed.

These rest blocks were an important feature of the design, and should be considered for any block design. Paradoxically, a very well-controlled fMRI block design has an important shortcoming compared to one that contrasts the "on" condition with a resting baseline condition. This is due to the fact that some brain regions actually show a *decrease* in BOLD signal during tasks than during rest (Binder et al., 1999; Gusnard & Raichle, 2001), which may differ between conditions. Thus if one simply contrasts BOLD signal between two well-matched conditions, it is impossible to know whether differences in a brain area are due to differences in activation, deactivation relative to a resting baseline, or some combination of the two, as depicted in Figure 7.4. The patterns shown in Figure 7.4 are not specific to any particular experiment; however, we can illustrate them in terms of a hypothetical experiment contrasting viewing printed words with consonant strings in a study designed to activate areas associated with the recognition and understanding of words. Imagine that each of the graphs, A–E, represent the pattern of activation in a different brain area. In areas A and B the subtraction would suggest greater "activation" for words

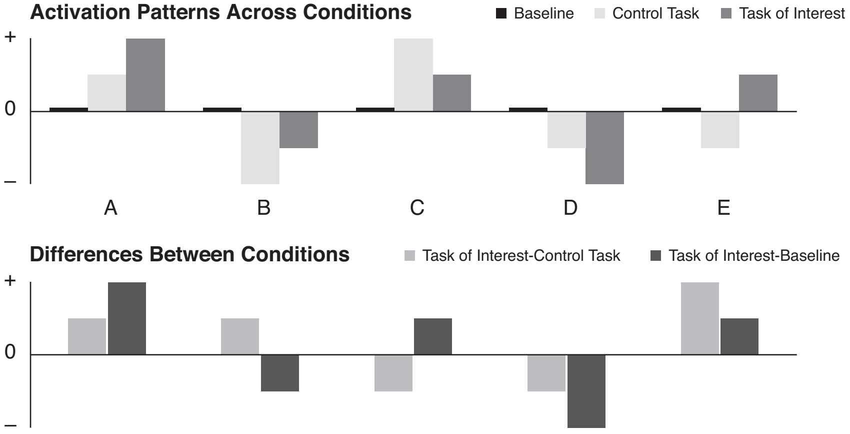


FIGURE 7.4 Issues in interpreting differences between two conditions in an fMRI experiment. These illustrations demonstrate the importance of including both a well-matched control task and a more neutral baseline condition in an fMRI study. Adapted from Gusnard and Raichle (2001).

than for consonant strings, however, in area B the BOLD signal is actually decreased relative to a neutral baseline condition (e.g., viewing a simple fixation cross), which would likely entail a rather different interpretation than the case shown in A, since areas showing the pattern in B do not show increased BOLD signal for words relative to the fixation cross. Similarly, areas C and D would appear to have reduced activation for words relative to consonant strings, but again the interpretation of these effects would differ if a baseline condition was included in the experiment. In area E, we see that a subtraction between words and consonant strings would yield double the difference in area A, even though the activation in response to words themselves, relative to baseline, is equivalent in A and B. In most cases, we would only view brain regions as truly "active" if the BOLD signal increased significantly relative to both a resting baseline *and* a well-matched control condition—so only areas A and E in this example. Thus, it is ideal to include both a resting baseline condition and a well-matched control condition.

Event-related fMRI designs. Event-related designs were developed to overcome the constraints of blocked fMRI designs. Buckner et al. (1996) first demonstrated that fMRI to single trials or events could be resolved if they were spaced far enough apart in time, a paradigm that was adapted to psycholinguistic designs such as the sentence violation paradigm (Meyer, Friederici, & von Cramon, 2000; Newman, Pancheva, Ozawa, & Neville, 2001). Later, it was shown that by "jittering" the timing between subsequent events, it was possible to present stimuli closer together in time than 12 to 20 sec. If the stimuli are expected to activate different brain areas (e.g., left vs. right visual cortex), then they can be separated by less than a second (Dale & Buckner, 1997), but if one wishes to discriminate activity between different stimuli within a single brain area, a separation of 4 sec or more may be required (Glover, 1999; Miezin, Maccotta, Ollinger, Petersen, & Buckner, 2000). The advantage of jittering the interstimulus interval (ISI) can be seen in Figure 7.5. It has been shown that the BOLD response adds approximately linearly—at least within practical limits—so that if we assume a fixed shape of the BOLD HRF within a voxel, then the BOLD response to two events that occur close together in time can be predicted by adding the HRFs that would be elicited by each individual stimulus, with the second time-shifted by the delay between the stimuli (Glover, 1999). Thus to find the BOLD response to the second stimulus, in principle we need only to subtract the expected response to the first stimulus—a process known mathematically as *deconvolution*. However, this gets more complicated, since the actual magnitude and timing of the HRF will vary somewhat from trial to trial. This makes the subtraction more complicated, and ultimately it is impossible to separate the overlapping BOLD responses to two stimuli that occur close together in time. However, if the ISI varies across trials, the problem becomes more tractable because we can use the variance in the ISI to help estimate the relative size of the response to stimuli from different conditions. Another "trick" can be used if one wishes to separate brain activation for two

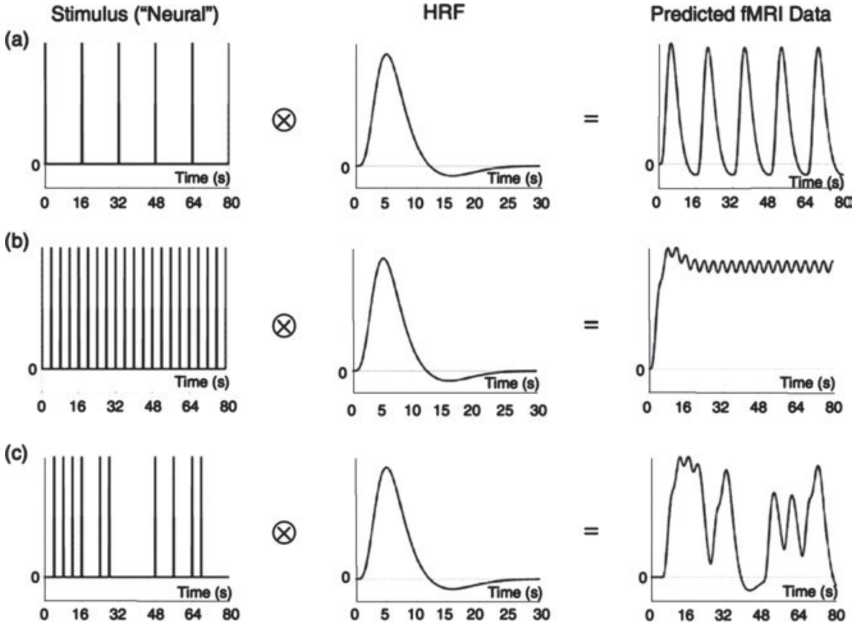


FIGURE 7.5 Different stimulus timeseries convolved with a model HRF, yielding predicted fMRI BOLD signal. The time course of the transient neural response to the stimuli would be expected to be similar to the stimulus timeseries itself. The X symbol indicates convolution, the operation by which the HRF model is multiplied by the stimulus timeseries to derive the predicted BOLD response. (a) shows a slow event-related design; events are spaced far enough apart in time that their HRFs do not overlap. This is however an inefficient use of scan time and likely very boring for the subject. (b) shows a "fast" event-related design (1 event/4 sec) in which the HRFs overlap so closely that it would be difficult to distinguish a task-related brain area from one that did not respond at all. (c) shows a jittered event-related design, using twice as many stimuli as in (a), but with variance in the strength of the BOLD signal over time due to the variance in inter-stimulus intervals. This represents the most efficient design of the three, in terms of power to recover stimulus-related BOLD fluctuation in a fixed amount of time. Reprinted from Henson (2007).

components of a trial, such as the stimulus and a subsequent response. Beyond jittering the interval between stimulus onset and a probe indicating that a response should be made, one can probe for a response on only a percentage of trials (e.g., 50%); this allows for the estimation of the response to the stimuli without contamination from the subsequent and temporally-overlapping response. A detailed discussion of this is found in Henson (2007).

Thus, the challenge in event-related fMRI designs is to make optimal use of limited scan time (i.e., by not including long "dead" periods of no stimulation while waiting for the HRF to return to baseline, which can also lead to subject boredom), while still being able to separate the BOLD responses to different trials. Ultimately, the choice between a block and an event-related design will come down to the research question and stimuli and tasks involved. If the paradigm is amenable to a block design, this should be preferred. Not only are block designs easier to implement, but they are the most efficient in terms of the strength of the contrast relative to the amount of scan time used. Because the HRFs to individual events summate over time in a block design, the overall strength of the measured BOLD response, and the relative signal change between different conditions, will tend to be larger in block than event-related designs. On the other hand, if the paradigm is such that a block design is likely to lead to adaptation or habituation effects, or the trials have a complex structure, then an event-related design would be more suitable.

The Importance of Control Conditions

As in all psycholinguistic research, in neuroimaging it is vital to control as many stimulus factors as possible. For example, the physical properties of the stimuli, such as brightness on a screen, font size, loudness of acoustic stimuli, and so forth need to be matched both across different experimental conditions, and between these and control conditions. Task selection is another important consideration. Strategy (how subjects perform a task) may vary between conditions and between subjects. It is important to give subjects explicit, detailed instructions and to debrief them after the experiment. It is also important to consider whether tasks are well-matched between the conditions that one wishes to compare. A common error in experimental design is to give subjects a task to perform during the experimental condition, but not the control condition. For example, a study investigating semantic processing might present subjects with line drawings of different animals and objects, and require subjects to make animacy judgments—after each picture is shown, a question would appear on the screen asking the subject to press one button if the depicted object is a living thing, and another button if it is not. The control condition may be well-designed from a physical perspective, for instance by using scrambled versions of the pictures (created by breaking each picture up into a grid of small squares and randomly rearranging these, thus preserving the overall luminance and color distribution; see Figure 7.6 for some examples). However, one cannot perform animacy judgments on scrambled pictures. One option is to give the subjects no task in the control condition, but then the experimental–control task contrast would include activation of the motor system, and involve differences in attention. A second option would be to simply require a button press in response to each scrambled picture. This would control motor activation;[4] however, if there are two response options in the experimental condition and only one for the control condition, there is a risk that the contrast between these conditions will include brain activity related to response selection being required in one condition and not the other. So it would be preferable to have subjects choosing responses from the same number of options in each condition. In our

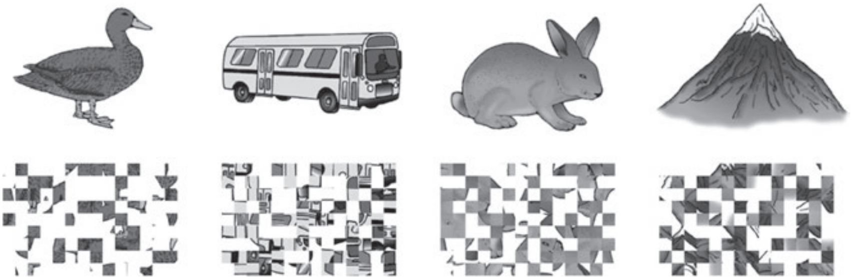


FIGURE 7.6 Example stimuli that could be used in an fMRI study in which people are asked to make animacy (living/non-living) judgments on pictures such as those shown in the top row. The bottom row shows scrambled versions of each image from the top row; these could be used as control stimuli because they are equivalent to the objet images in terms of overall luminance and color distribution. Line drawings of objects are from Rossion and Pourtois (2004).

hypothetical experiment, one could achieve this by randomly making the line color in half of the pictures (both objects and their scrambled equivalents) red, and in the other half blue. Line color would be irrelevant to performing the animacy task, but in the control task subjects would indicate by button press whether the lines were red or blue.

Another consideration is to control task difficulty as much as possible. In some cases, this is not possible, and control tasks are often less challenging than experimental tasks. In such cases, at a minimum it is important to report accuracy and reaction times for each condition and test whether statistically reliable differences were present. Ideally, if systematic differences are present, they will be included in the fMRI analysis as covariates. On the other hand, some paradigms are quite amenable to matching the difficulty of the control task to that of the experimental task, as is the case with lexical decision tasks, where the experimental stimuli are real words and the control stimuli are nonwords. For example, Vannest, Newport, Newman, and Bavelier (2011) used an event-related fMRI design incorporating a lexical decision task to investigate the processing of derivational morphology. However, even in this study, the subjects' mean reaction time for each condition was used as a covariate in the analysis (i.e., factored out) to control for differences in activation of the motor system that may have been present due to overall differences in reaction time between different experimental conditions.

A Special Consideration in SLA Research: Language Proficiency

It is well documented that many second language (L2) learners, especially those who learned later in childhood or in adulthood, show lower proficiency in their L2 than in their native language (L1) or when compared to the "average" L1 learner of the target language (Flege, Yeni-Komshian, & Liu, 1999; Johnson & Newport, 1989). Yet, many neuroimaging studies comparing L1 with L2 learners have not explicitly measured proficiency, or have made only weak attempts to control for this factor (e.g., using proxies for proficiency such as amount of time spent immersed in the language, rather than explicitly measuring proficiency using a standardized test). Thus, it is difficult to know whether any observed differences in brain activation between L1 and L2 learners are due to the effects of the age or environment of acquisition, or simply due to L2 learners having to work harder to perform the task. If the latter, one might expect to see a similar relationship between brain activation and proficiency among L1 learners and L2 learners, albeit over different ranges of proficiency.

There are several possible solutions to this problem. Matching proficiency between groups is not ideal, since in a study of "average" L2 learners one would likely need a control group of native speakers who had lower-than-average proficiency, which might be caused by mitigating factors that otherwise affect brain activity. A second approach is to titrate the task difficulty to each group, or even to each individual. This is less than ideal because brain activation may differ with the items used, and by the number of trials presented per unit of time. A preferable approach is to use a consistent task, items, and experimental parameters while ensuring through pilot testing that L2 speakers can perform the task with an acceptable degree of accuracy. Then, proficiency in the target language (the language being used for the fMRI task) is measured and included in the analysis as a covariate in an ANCOVA approach, or as a continuous variable in a general linear mixed-effects approach to fMRI analysis.

It is also important that the proficiency measure not show ceiling effects among native speakers. Many standardized language tests are designed primarily to detect significant deficits (such as aphasia or developmental disorders), and do not show a range of proficiency among people with language abilities in the normal range. In order to be able to properly regress proficiency on fMRI activation, it is critical to have variance among proficiency scores in L1 learners so that the relationship between proficiency and brain activation can be seen in native speakers. Without this, it is impossible to determine whether proficiency modulates brain activity similarly in L2 and L1 learners. This is also a problem with self-reported levels of proficiency, since virtually all native speakers will report maximum proficiency. In English, we and others (Newman, Tremblay, Nichols, Neville, & Ullman, 2012; Pakulak & Neville, 2010; Weber-Fox, Davis, & Cuadrado, 2003) have found the *Test of Adult and Adolescent Language – 3* (Hammill, Brown, Larsen, & Wiederholt, 1994) useful for these purposes, particularly as it has separate measures for vocabulary and grammar, as well as for reading, listening, speaking, and writing. The *Peabody Picture Vocabulary Test* (Dunn & Dunn, 2007) is a widely-used and standardized measure of vocabulary, with the added advantage that a standardized version is also available in French (the *Échelle de vocabulaire en images Peabody;* Dunn & Theriault-Whalen, 1993) if one happens to be studying English-French bilinguals. This list is hardly meant to be exhaustive, and certainly other standardized tests will be found for other languages. It is generally preferable to use a standardized test so that one has norms of performance on an age group comparable to the subject population. Of course, for some languages such standardized tests may not be available. In these cases it would be valuable to develop and refine a new test on groups of native speakers and L2 learners of the language to ensure that it will demonstrate a range of proficiency levels in both groups. A final point of note is to strongly discourage binarization of subjects into groups such as "high" and "low" proficiency. First, if these are based on a median split of proficiency scores, the difference between "high" and "low" will be arbitrary and sample-dependent. This is of particular concern in neuroimaging studies since the number of participants per group tends to be lower than in some other types of research. Even if there were some agreement as to what constituted "high" versus "low" on a given test, dichotomization of a continuous measure can reduce the statistical power of the analysis by as much as 30% (Cohen, 1983; Maccallum, Zhang, Preacher, & Rucker, 2002; Maxwell & Delaney, 1993). While the implementation of the statistical analysis might be somewhat simpler with dichotomized data, the additional effort involved in including continuous proficiency measures can be quite worthwhile. For details on the use of linear mixed effects modeling to analyze neuroimaging data with proficiency as a continuous predictor, see Newman et al. (2012).

Experimental Complexity and Signal-to-Noise

Language is a very complex system and words and sentences vary along a large number of dimensions. Thus, it is common in psycholinguistics to perform complex regression analyses in which the contributions of a large number of such variables are assessed (Baayen, 2008). The researcher coming to fMRI research with a background in other methods may wish to use this same fine-grained level of coding in order to determine how each variable affects brain activation. While this may be attempted, it is best to design relatively simple fMRI experiments with analyses centered primarily around a limited set of levels of a limited set of variables. This is especially true when one is first embarking on fMRI research, as one quickly learns that there are many new challenges to master before even getting to the results!

Another reason to limit the detail of stimulus coding is that while the signal-to-noise ratio of fMRI is relatively good, it is generally necessary to have repeated trials of the same stimulus category in order to find a statistically reliable effect (particularly for contrasts between closely-matched experimental and control conditions). While the number of trials required varies as a function of the size of the effect and the noise in the data, it is advisable to have *at least* 20 to 30 trials per condition in an event-related design (Huettel & McCarthy, 2001), and several blocks of at least 20 sec each in a block design. Thus, designs are necessarily limited to those for which one can find the requisite number of items that all fit a particular category. While in principle some of these limitations could be overcome by using large sample sizes (e.g., 100 subjects), the cost and logistics of doing so are typically prohibitive.

Limitations Imposed by the MRI Environment

fMRI uses a strong magnetic field that is on all the time. This field is so strong that a ferromagnetic metallic item such as a paperclip or an oxygen tank may be drawn into the center of the scanner at dangerous speeds. Metal implanted inside peoples' bodies, particularly their heads, can cause problems. While some kinds of implanted metal are not safe in an MRI scanner, since they could move or heat up and cause tissue damage, many kinds of implants are safe for conventional MRI scans (purely structural scans that might be ordered by a doctor for medical purposes). However, some of these "MR safe" metals are not compatible with fMRI scanning because of the unique demands placed on the MRI hardware by fMRI. These concerns about implanted metal are unlikely to affect study design (though they preclude studies of certain populations, such as cochlear implant users), but are vital considerations in subject recruitment and screening.

The strong magnetic field also precludes bringing most standard stimulus presentation or response collection equipment into the scanner room. Ferromagnetic material in a device could lead to it becoming a projectile hazard. Furthermore, even nonmagnetic equipment can cause distortions of the magnetic field. Some nonferromagnetic metals placed near the head may cause signal dropout in nearby brain areas due to magnetic "susceptibility" effects. In addition, electrical current flowing through wires will produce electromagnetic fields that may introduce noise into the fMRI measurements. For example, most audio headphones cannot be used in an fMRI study because electrical current flowing through wires, as well as the electronics in the speakers themselves, will distort the MRI signal. Electrical cables and other wires should not normally be run directly into the MRI room because they can act as antennae that carry electromagnetic noise into the shielded MRI room. Rather, connections must be made through a "patch panel" or radio frequency filters in the wall of the MRI room, and it is often the case that the patch panel does not have available the particular type of connector required by the researcher.

Fortunately, there are numerous "MR-compatible" products available on the market for audio, visual, and somatosensory stimulus presentation, manual response collection, and eye-tracking. Such devices typically use either carefully shielded electronics, or nonelectronic solutions such as fiber optic or air conduction. Generally the cost of these is significantly higher than non-MR-compatible alternatives. An ideal situation is that the MRI scanner is set up as a research site, with the necessary equipment in place and shared between different research groups. It is important to choose equipment based on its functionality, MR-compatibility, durability, and ease of setup. Many a research scan has failed due to broken or missing parts, or other "technical difficulties." Experiments may need to be designed around available hardware, or necessitate considerable money and time be invested to solve the technical hurdles involved in obtaining and adapting new hardware to the MRI environment.

Visual stimulation. For visual presentation a common setup involves an LCD projector located outside the MRI room, aimed through a tube in the wall (a "wave guide" designed to eliminate radio frequency interference through its combination of length and diameter) and projected onto a screen placed in the MRI bore. Other options include shielded projectors inside the MRI room, or goggles or small screens placed in the scanner bore close to the subject's eyes. Eyeglasses cannot be worn inside the MRI scanner, so with a projector and screen setup, subjects will either have to wear contact lenses, or a set of MRI-compatible glasses with a range of interchangeable corrective lenses will have to be purchased. This is an important consideration in subject recruitment and screening.

Auditory stimulation. Another hardware-related limitation is acoustic noise. Having an MRI scan has been likened to being in a machine gun nest—when scanning the scanner makes noises that can be as loud as 120 dB. Ensuring that subjects can hear acoustic stimuli may thus be a challenge. MRI-compatible headphones typically provide significant acoustic dampening (perhaps 30 to 40 dB), and noise-cancelling headphones can provide further attenuation. In many applications these solutions may be sufficient—if the stimuli are clearly audible, the background noise will be constant across conditions that the researcher wishes to compare. However, in other cases it may be necessary to present acoustic stimuli without scanner noise. One solution to this is to use "sparse" scanning (Hall et al., 1999). In this paradigm, the time between each fMRI acquisition is increased, so rather than using a typical 2 sec repetition time, one might use 3 sec or more, though the scanning (and consequent noise) still only takes 2 sec. By synchronizing the stimulus presentation software with the scanner, auditory stimuli can be presented in the silent 1 sec periods between scans. This technique takes advantage of the sluggishness of the BOLD response—the fMRI signal in response to the auditory stimulus will occur several seconds after the stimulus was actually presented.

Speech production. Speech production in the MRI scanner can also pose problems. Head motion in general is problematic for fMRI—movements of mere millimeters can lead to dramatic artifacts in the data, and speaking necessarily involves some movement of the head. Additionally, the complex changes occurring in the vocal tract during speech production change the magnetic field of the head, which is very sensitive to whether any given point in space is occupied by air or tissue, and which type of tissue. Thus, speaking during an fMRI scan can lead to artifacts in the data, which may be particularly prominent around the orbitofrontal cortex (the base of the frontal lobe) and even extending into Broca's area. However, many fMRI studies involving speech production have been published.

In general, it has been found that as long as subjects are given clear instructions to keep their heads as still as possible, and make the minimum movements necessary to speak, then the fMRI data are quite usable (Bavelier et al., 2008). Nevertheless, it is always important to examine the output from motion correction procedures during data preprocessing (see below). Another approach is to simply require subjects to think about saying the words without actually saying them. Covert production has been shown to reliably activate virtually all of the same brain areas involved in actual speech, with the exception of primary motor cortex.

Another complication is that because the scanner is so loud, it is impossible to rely on a conventional (MRI-compatible) microphone to record speech in the scanner. A low-tech approach to deal with this is to attach a long flexible tube near the subject's mouth, with the other end running out of the scanner to the ear of an experimenter sitting close by, who then writes down the words (the experimenter must wear earplugs to protect from the scanner noise). We have used this effectively in the past (Bavelier et al., 2008). A more high-tech (and consequently more expensive) approach is to use a noise-cancelling microphone built for the MRI environment.

Scoring, Data Analysis, and Reporting Results

Preprocessing of fMRI Data

Prior to statistical analysis, it is necessary to "preprocess" fMRI data, which includes a number of signal processing steps that will account for various artifacts and sources of noise in the data, leading to much more robust analyses. These include motion correction, spatial smoothing, temporal filtering, and removal of nonbrain tissues. A detailed description of fMRI preprocessing is well beyond the scope of this chapter; however, there are numerous excellent resources available, some of which are listed at the end of this chapter. Here we focus on the one step of preprocessing that is arguably the most important to understand—how to "normalize" the shape of individual brains to allow cross-subject comparisons.

Spatial normalization. Because everyone's brain is a different size and shape, we cannot simply overlay two brains and expect to have the same anatomical areas lineup. One solution to this is to trace out a particular anatomical region of interest (ROI) on the brain of each subject (e.g., the interior frontal gyrus or IFG, or Broca's area), then compute the number of active voxels per subject. This approach, however, is extremely time-consuming, and suffers from a fundamental problem: the functional organization of the brain does not necessarily respect gross anatomical landmarks. Thus a "blob" of activation in an fMRI study might cut across two such ROIs. The alternative approach is *spatial normalization,* in which the size and shape of each subject's brain is adjusted to match a standard, template brain. The problem of standardization across brains was addressed for neurosurgeons by Talairach and Tournoux (1988), whose method involved segmenting the brain into a set of rectangular volumes after first defining the horizontal plane as the line passing through the anterior and posterior commisures (two small white matter bundles connecting the two cerebral hemispheres), then scaling each volume to match their template brain. For their template they used the brain of a deceased elderly woman, which had been fixed in formalin. As MRI neuroimaging evolved, improved algorithms were developed that automatize the normalization process. *Linear* registration algorithms will shift, rotate, and adjust the size of each subject's brain to match the template. These do not match the shape of each individual gyrus or sulcus of the brain, but rather its overall size and shape. *Nonlinear* algorithms go further to account for these local differences between brains, and improve the overall quality of the matching between subject and template. An alternative method is to align major sulci with the template as a starting point for subsequent linear or nonlinear registration (Fischl, Sereno, & Dale, 1999).

In the process of developing these automated image registration algorithms, the neuroimaging community realized that most neuroimaging studies are performed on a variety of healthy young adults whose brains are still *in vivo,* rather than a single, elderly, preserved brain in a jar. Thus, a template has been developed, largely through the Montreal Neurological Institute (MNI), based on the average of several hundred healthy young brains. The Talairach and Tournoux brain served as the reference for the normalization of the initial MNI template, however the coordinate systems are not identical (because in matching the overall size and shape of the average healthy brain to that of the original template brain, not all internal structures were in the same relative positions). This is important to remember, because some software still uses the Talairach transformation and coordinates, but looking up the same three-dimensional coordinates in the two template brains may result in different anatomical labels.

Having a standard coordinate system is valuable because it gives researchers a common language in which to refer to localization of brain activation, just as latitude and longitude allow anyone with a GPS system to arrive at the same geographic location. However, there is some danger in taking these coordinates too seriously. Normalization algorithms are not perfect and cannot guarantee that their results will be entirely consistent across the brains of individuals. In addition, there may be systematic differences between different normalization algorithms. Thus, while location coordinates are typically reported to the nearest millimeter, one should not take differences of a few millimeters between studies too seriously. Furthermore, even if the gross sulcal and gyral anatomy were perfectly aligned across individuals, this would not guarantee perfect alignment of these people's *functional* anatomy. The cellular microstructure (cytoarchitectonics) of cell types, densities, and connectivity patterns across the layers of the cerebral cortex varies across the brain, and the relationship between gross anatomical landmarks and the underlying cytoarchitectonics is not consistent across subjects (Amunts et al., 1999; Rademacher et al., 2001). Work is underway to develop a probabilistic map of the cortex indicating the likelihood with which a particular voxel is in a particular cytoarchitectonic region (Eickhoff et al., 2007).

Statistical Analysis

Statistical analysis of MRI typically follows a multi-level approach. At the first level, the data from each individual run are analyzed, to determine the size of the effects of each independent variable. This is often done using a multiple regression approach, in which the expected height of the BOLD response is specified for each volume (time point) in the data, and the time-course of each voxel in the image is tested to determine how well it correlates with this model. This is shown for a simple block design in Figure 7.7. At the second level of the analysis, one may combine runs within each subject. This simplifies group-level analysis by reducing

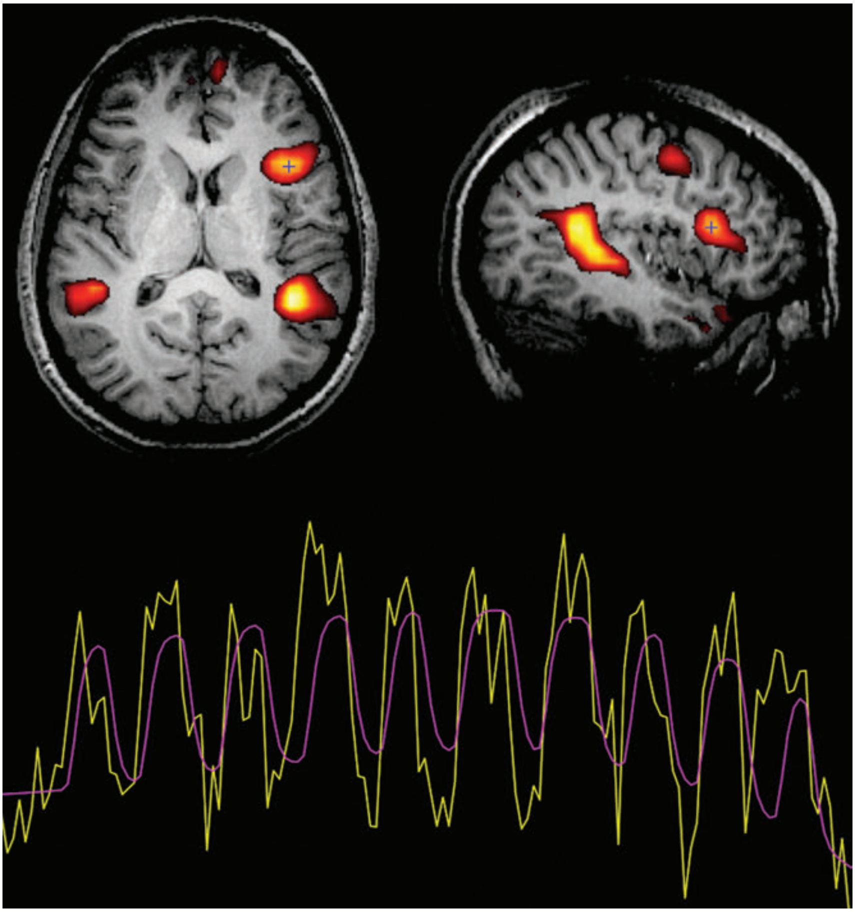


FIGURE 7.7 Sample fMRI timecourse from a block-design study (yellow), with the best-fit linear regression model superimposed (purple). The model was a square wave having a value of 0 during "off" blocks and 1 during "on" blocks, convolved with the HRF shown in Figure 7.1.

the data contributed by each subject, although the investigators should consider whether they wish to have the within-subject variance between runs included in the group level analysis. At the highest level of analysis, data from all subjects in the study are included in a group-level (or even between-group) analysis. These analyses typically follow the structure of standard mixed-effects ANOVAs,[5] with subjects treated as random effects and the experimentally-manipulated independent variables as fixed effects. The output of each of these analyses is a *statistical parametric map* (abbreviated as SPM), a 3D volume with a statistical value at each voxel (e.g., regression coefficients, z, F, or p values).[6]

Thus, the general structuring of fMRI statistical analysis is not too different from what the investigator may be familiar with from other research methods. One of the biggest differences between fMRI and other domains is that the analyses are *massively univariate*—rather than performing one or a handful of statistical tests, as one might with reaction time data, we are testing the model at every voxel in the brain. Given a typical fMRI dataset of perhaps 30 slices, each with 64 × 64 voxels per slice, we would perform 122,880 statistical tests. Besides being computationally intensive, this large number of multiple comparisons guarantees that a certain proportion of the voxels will have p values that exceed a conventional threshold of significance, such as 0.01. Indeed, $p < 0.01$ on a dataset this size would be expected to yield 1,228 voxels above threshold, purely by chance. A common solution for the multiple comparison problem in other domains is the Bonferroni correction, in which one divides the desired p value by the number of tests being performed. Thus, in this case, to achieve a nominal p threshold of 0.01 across the entire image, we would need to accept only voxels whose p value was less than 8.13×10^{-8}! This is, however, overly conservative. First, it is common to include only voxels within the brain in the analysis. Furthermore, the data from adjacent voxels are expected to be correlated, both because of the inherent smoothness of the BOLD response, and the spatial smoothing applied during preprocessing. The Bonferroni correction, however, assumes that each statistical test is independent of every other one. Alternative methods of correction have been developed, including the false discovery rate (or FDR), which adjusts the threshold for significance based on the distribution of statistical values in the particular dataset (Genovese, Lazar, & Nichols, 2002). Another approach first thresholds the statistical map at a specified p value (e.g., < 0.01), then applies a correction for cluster size by estimating the probability of finding a cluster of adjacent, suprathreshold voxels by chance given the inherent smoothness of the dataset (Worsley, 2001).

Functional Connectivity Analysis: Moving Beyond Neo-Phrenology

Neuroimaging research has been accused of being a "neo-phrenological" science, in which the ultimate goal is simply to associate particular brain areas with particular cognitive functions. Although the voxel-by-voxel, and cluster-by-cluster, approach can be argued to provide much more than this, it does have limitations. In response to these, additional approaches to analysis have been developed.

One is *functional connectivity,* which essentially looks at correlations between different brain areas and how these are modulated by experimental manipulations. Functional connectivity should not be confused with structural connectivity; two brain regions may be functionally connected, in that they show correlated BOLD signals, without having direct white matter connections between them. Their interaction may, for example, be mediated through an intermediate brain region, or they may both receive correlated input from a third brain area (e.g., two cortical areas might both be modulated by thalamic input). Functional connectivity studies can be useful in characterizing how different parts of a network interact under different task or stimulus conditions. For example, Dodel et al. (2005) examined how correlations between a network of 41 brain regions changed in bilinguals between word and sentence production tasks, as well as between English and French versions of the tasks. Dodel et al. found stronger functional connectivity during sentence production between a number of regions, particularly between the left IFG (i.e., Broca's Area) and numerous other areas. The language in which tasks were performed also modulated functional connectivity within the network, with all differences showing stronger correlations between brain regions when subjects were performing the task in their L2, English. Thus, even though certain brain areas, such as the left IFG, were active in multiple tasks, the manner in which this brain region interacted with other areas was modulated by the task conditions.

Reporting Results

Poldrack et al. (2008) have published guidelines for reporting an fMRI study, which will serve as an excellent starting point in this regard. The complexity of fMRI data lend themselves strongly to visual representations of the data, and these should be used to augment the text. Given the high costs of fMRI scanning, it is rarely the case that an experiment is repeated several times in the same laboratory, and in recent years there has been an increasing awareness of the value of meta-analyses of neuroimaging data (Kober & Wager, 2010; Turkeltaub, Eden, Jones, & Zeffiro, 2002). Reporting data thoroughly and systematically thus not only conveys the complex information to readers effectively, but contributes to the ongoing body of knowledge in the field and future discoveries by facilitating meta-analyses.

An Exemplary Study

As an example of an fMRI study of bilinguals, we will examine Saur et al. (2009). The goal of this study was to compare the brain areas involved in sentence processing among highly proficient early and late bilinguals. Saur et al. included three groups of French-German bilinguals (12 subjects/group): those who had learned both languages before the age of 3 (2L1); native French speakers who learned German in mid-childhood (mean age 16 years) (L2G); and native German speakers who learned French in mid-childhood (mean 14 years) (L2F). All subjects were given standardized L2 proficiency tests in both French and German. As expected, both late bilingual groups were more proficient in their L1 (both groups scoring around 98% correct) but both were still quite proficient in their L2 (around 85% correct in both groups). The 2L1 group was more proficient in German (95%) than French (85%), probably because they lived in Germany; nevertheless, their French proficiency was on par with the L2F group. As noted above, having these proficiency scores is valuable for interpreting the fMRI data because we can be assured that all subjects were reasonably fluent in both languages, and in particular the L2F and 2L1 groups were very closely matched on proficiency while varying in age of acquisition.

The stimuli were French and German sentences. Half of these were grammatical, and half ungrammatical. Grammaticality was crossed with subject-verb word order; half of each sentence type was SV and the other half was VS, as illustrated in (1) and (2).

(1) *SV Order*

Peter(S) kommt(V) spät von der Arbeit. (German)
Natalie(S) travaille(V) á Paris ce soir. (French)

(2) *VS Order*

Wann kommt(V) Peter(S) von der Arbeit? (German)
Où travaille(V) Natalie(S) ce soir? (French)

This manipulation was of interest because in French the VS order is more marked (involving greater grammatical complexity, e.g., in the example when it is achieved through a movement operation) than in German, where it is more common for the subject to follow the verb. Thus, it was expected that while late bilinguals would show stronger brain activation during grammatical processing in their L2 than their L1, this effect was predicted to be more pronounced for the VS sentences in the native French/late German speakers. Based on previous work, Saur et al. predicted this pattern of results in the left IFG (i.e., Broca's Area), as well as overall greater activation within areas associated with language processing for the L2 than the L1 in late bilinguals.

Saur et al. used an event-related fMRI design. The study consisted of eight scanning runs per subject, each roughly 5 min long. A given run contained only French or German sentences, with the order of runs pseudorandomized such that no more than two runs in the same language occurred in sequence. Within each run, 10 sentences of each category (grammatical/ungrammatical × SV/VS) were presented, with the order of sentences pseudorandomized. The sentences were presented aurally through MR-compatible headphones, and subjects made grammaticality judgments after each sentence by pressing a button on an MR-compatible response pad (held in their left hand, to prevent activation of left motor cortex that would occur with right-handed manual responses, potentially contaminating left-lateralized language processing activity). The sentences were of variable length (1.4 to 3.1 sec), with jittered ISIs (2.9 to 4.6 sec) and eight longer "null" events. This timing allowed for accurate estimation of the HRF to the individual stimuli without the need for very long ISIs, as discussed above. Because the researchers were interested in grammatical processing, a grammaticality judgment task was an appropriate way to ensure that subjects were processing the grammatical content of the sentences. This, combined with a desire to examine brain activation only for grammatical sentences in some contrasts, necessitated an event-related rather than a block design.

The results supported Saur et al.'s predictions. As seen in Figure 7.8, in both late L2 groups there was greater activation in the left IFG and inferior temporal lobe for their L2 than their L1 (panels identified with red outlines). These occurred in the context of other activations that did not appear to be common across the two groups' L2s. However, it is difficult to "eyeball" what is common to the two groups. Thus, Saur et al. (2009) employed a *conjunction* analysis. These are common in fMRI studies, and refer to the practice of identifying regions that are commonly activated across two different contrasts. In this case, Saur et al. used conjunction to identify brain regions that showed both significantly greater activation

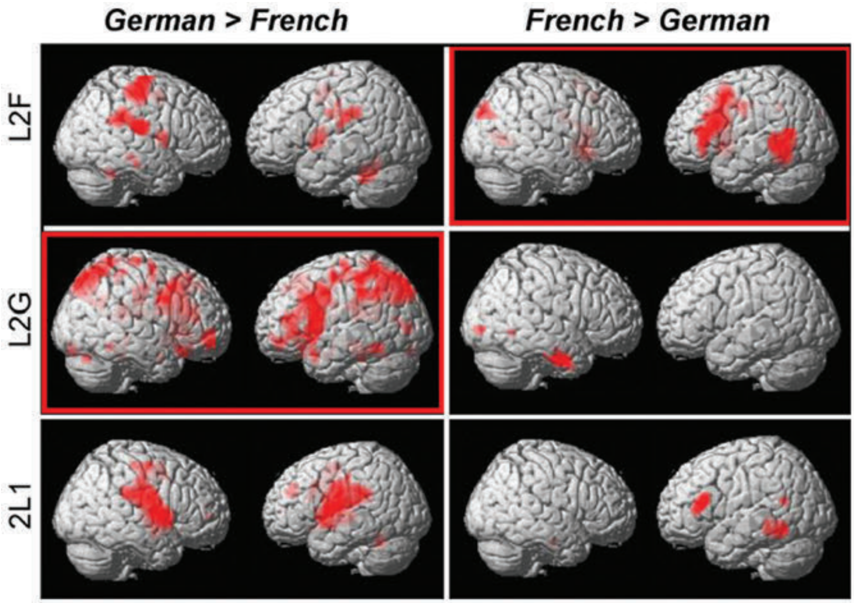


FIGURE 7.8 Statistical parametric maps showing fMRI activation for each language condition, in each group, from Saur et al. (2009). Areas in red survived statistical thresholding of $t > 4.53$, or $p < 0.05$, corrected for multiple comparisons. The images surrounded by red frames are the contrasts showing greater activation in each group's later-learned L2 relative to their earlier-learned L1.

for German than French in the L2G group, and significantly greater activation for French than German in the L2F group. As seen in Figure 7.9, this revealed greater activation for L2 than L1 in the left IFG (pars opercularis and pars triangularis of the IFG were identified as separate clusters within the IFG) and inferior temporal lobe, as suggested by our visual inspection, as well as the right caudate nucleus (which was not visible in the lateral views of the brain shown in Figure 7.8). It should be noted that while L2 > L1 activation maps shown in Figure 7.8 were generated for each group using correction for multiple comparisons (cluster size correction), for the conjunction analysis a more liberal threshold of $p < 0.001$, uncorrected for multiple tests, was used. This more liberal threshold was likely used to help increase the overlap among areas. One may debate whether such uncorrected results should be reported, but given that in this case the conjunction was between maps derived from two independent groups of subjects, it is reasonable to expect that variability in brain organization combined with a degree of inaccuracy in spatial normalization might result in functionally similar areas being slightly offset from each other.

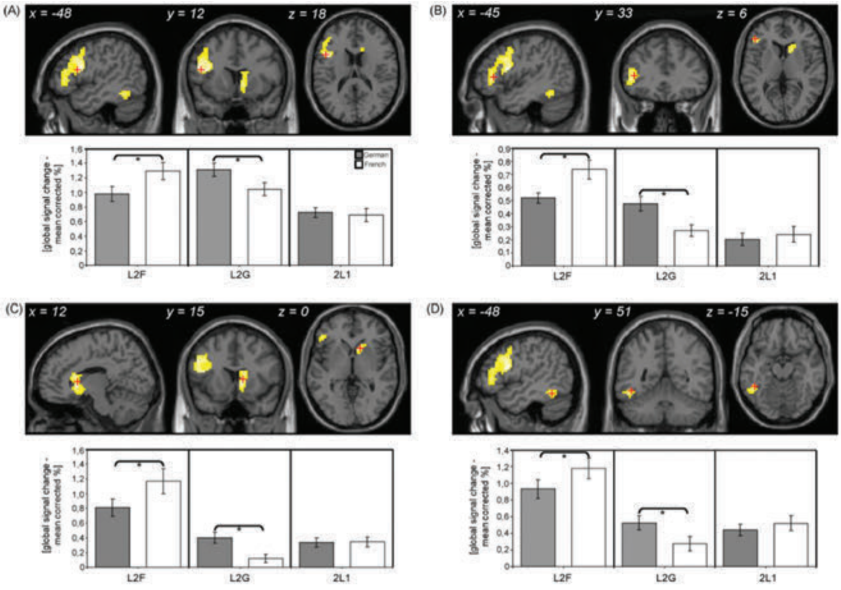


FIGURE 7.9 Conjunction analysis from Saur et al. (2009), showing brain regions that were significantly more activated during L2 than L1 processing in both late-learning L2 groups (L2F-native German speakers with French as their L2, and L2G-native French speakers with German as the L2). Each panel shows activation patterns across groups and languages within a specific region of interest (ROI): (A) left IFG, pars opercularis; (B) pars triangularis; (C) head of the right caudate nucleus of the basal ganglia; and (D) left inferior temporal gyrus. The centre of each ROI is marked in each set of brain images with a red crosshair.

Pros and Cons of Using the Method

Pros

Functional MRI has many strengths, including widespread availability, relative ease of use, good signal-to-noise ratio, the often intuitive nature of data interpretation (brain area *X* is involved in task *Y*), and ultimately the power of the technique to observe the living brain at work.

The biggest strength of fMRI is that it provides a level of neuroanatomical precision that is almost impossible to achieve with any other noninvasive technique. Perhaps its closest competitor in this regard is MEG, and MEG does have some advantages. For example, MEG is silent, and has fewer contraindications concerning implanted metal and devices (though there are some). MEG also offers far superior temporal resolution, since it is measuring the magnetic correlates of brain electrical activity directly. However, while MEG source localization algorithms can yield high accuracy, their accuracy varies with neural location, and their ability to resolve or differentiate between very close regions is more limited. MEG is also typically similar in cost to fMRI, and overall, at the time of this writing, the software tools for MEG analysis are not as polished and easy to use as those for fMRI. This can potentially add complexity and time to the research cycle, unless the necessary technical and analytical support is in place. Thus, when the research question centers around a question of functional neuroanatomy, fMRI is the ideal tool.

Another advantage of fMRI is that, in spite of the limitations imposed by the scanning environment, the method is quite flexible with regard to the types of stimuli that can be used. One can use visual and/or auditory stimuli (indeed, Braille reading has even been studied, using embossed paper; Sadato et al., 1996), including printed text, auditory stimuli, and movies. Not all methods for studying language processing have this degree of flexibility.

Cons

Despite the many advantages of fMRI, accessibility and ease of use can make it just as easy to do badly-designed research as high quality work. Additionally, many analysis software packages are so well-engineered that a novice with little understanding can move data through the software and generate images of "something" that are ultimately flawed. It is thus incumbent on researchers to learn as much as possible about the technique and analysis methods, rather than a "plug and pray" approach. New researchers should avail themselves of the help of colleagues, and the many excellent courses that are offered each year around the world on fMRI techniques and data analysis with specific software packages (these can be found by viewing the web sites of the Organization for Human Brain Mapping and analysis package developers such as SPM and FSL).

It should also be noted that even if one has a scanner that is capable of fMRI, there are many other technical hurdles to surmount before the first fMRI study can be run. These include determining the best pulse sequence (basically the program that the scanner runs to acquire data of a certain type), the appropriate parameters to use (the Methods sections of papers using a similar scanner can be used as a starting point, but one should consult with experts in this regard), how data can be moved off the scanner, and of course all the issues with stimulus presentation and response collection noted above. If one is lucky enough to conduct research at a facility where others are already doing fMRI work, then one will benefit from these people's experiences. However, if starting from scratch, the researcher should expect a relatively slow and painful process to get the operation up and running. It is generally advisable that, if sufficient funding is available, one should opt for "turnkey" fMRI-compatible systems rather than attempting to re-invent the wheel and build things from scratch. Even with systems built for fMRI research, there will be plenty of technical complications, and so anything that the researcher can do to simplify the process and benefit from the experience of others should be used to full advantage.

Another drawback to fMRI is that it is very expensive, typically costing hundreds of dollars per hour and requiring about an hour or more per subject for data collection. Thus, in the larger perspective, one must ask if fMRI really gives the best "bang for the buck" compared with lower-cost techniques. Blind "fishing expeditions" are unlikely to generate impactful or even informative data. Researchers should design their experiments around solid neuroanatomical hypotheses. Another consequence of the expense of fMRI scanning is that extensive pilot testing of the experimental paradigms is important, both before ever taking them to the MRI scanner, and then with the MRI, prior to real data collection. It is important to ensure that expected behavioral effects can be obtained in the fMRI paradigm.

Discussion and Practice

Questions

1) Which would be the best experimental design for the following psycholinguistic studies? In framing your answer, consider the relative merits of block as opposed to event-related fMRI designs.
   a. A semantic priming task in which a series of individual words are presented, some of which are semantically related to the preceding word in the series.
   b. Contrasting naming vs. performing animacy judgments on pictures of objects.
2) Identify the problems that a researcher might face, and possible solutions, in attempting to study the neural bases of categorical perception of phonemes in which a series of sounds are presented that vary continuously along some phonetic continuum (e.g., voice onset time varies between *pa, ba,* and *ga*) and participants are required to identify what syllable was heard.
3) fMRI is a powerful neuroimaging technique. However, there are a number of limitations in terms of both what we can and can't measure, and how we

interpret the data. Discuss these issues, and identify experimental questions and/or designs for which ERP might be a preferable technique.

4) Why can't BOLD fMRI signals be used to directly compare the intensity of neural activity between two different regions of an individual's brain?
5) Korbinian Brodmann (1909) published a map of the human cerebral cortex based on cytoarchitecture (the types of cells and their arrangement in different regions of the cortex). Today, the locations of fMRI activations are often reported in terms of Brodmann's areas, which is typically done based on either visual comparison with Brodmann's original drawings or with the Talairach atlas. Explain why this approach has serious limitations and how these can be (and are being) overcome.

Research Project Option

Because fMRI scanning tends to cost hundreds of dollars per hour, and experiments can be technically challenging to set up, it is generally difficult to collect data just to "play with." However, most of the analysis software packages listed in the previous section also make sample data available on their websites, along with tutorials guiding one through the design of the studies and analysis of the data. These are a highly recommended first step in designing and analyzing fMRI studies. Further, it is relatively easy to use a tool such as FEAT in the FSL software package to model the expected BOLD response of a stimulus time series. Thus one can design a stimulus sequence for a proposed fMRI design (e.g., derived from one of the discussion questions above) and use FEAT to visualize the expected fMRI response, or compare such responses between different candidate experimental designs.

Notes

1. Searches were conducted on PubMed on February 14, 2012, using the search terms fMRI, EEG, ERP, MEG, and positron emission tomography (since "PET" has multiple meanings including domestic animals and a type of plastic), restricting the publication date to the year 2011. The annual number of peer-reviewed, scientific publications mentioning fMRI in 2011 was 24,757, exceeding the combined number of publications referring to EEG (4939)/ERP (3911), MEG (470), or PET (4705).
2. A surprising number of people can fall asleep during an fMRI scan, in spite of the noise! The repetitive nature of the noise, combined with the often boring nature of some fMRI tasks, seem to contribute to this.
3. Our initial plan was to play the sentences backwards for the control condition. Speech processing studies often use reversed speech as a control condition, since it contains the same basic spectral-temporal properties as forward speech but is incomprehensible. However, in our pilot testing using backwards ASL movies, we found that signers could actually understand most of the sentences, and indeed some did not even realize that the sentences were reversed. Those subjects commented that the signing seemed "strange," as if the signer did not know the language very well, but were still understandable. Interestingly, the pilot fMRI data collected with these stimuli showed greater activation for reversed than forward ASL sentences in language-related brain regions such as Broca's area.

This example highlights the importance of thoroughly pilot-testing stimuli and paradigms, and debriefing subjects afterwards, before running a large number of subjects.

4. Note that since motor representations in the brain are lateralized, it is important to have subjects use the same hand for responses in conditions being compared. Otherwise, there will be right motor cortex activation for the condition that required left hand responses, and left motor cortex activation for the condition requiring right hand responses.
5. Technically, many packages now use linear mixed effects modeling rather than ANOVA, but the conceptual structuring of the analysis is similar.
6. One of the most widely-used analysis packages is also called SPM, but the output of any software package is an SPM.

Suggested Readings

A number of excellent books on the fMRI technique have been published. These will provide more extended coverage of the topics described here, as well as many others beyond the scope of this chapter. These books include:

Friston, K. J., Ashburner, J. T., Kiebel, S. J., Nichols, T. E., & Penny, W. D. (2006). *Statistical Parametric Mapping: The Analysis of Functional Brain Images.* London: Academic Press.

Huettel, S. A., Song, A. W., & McCarthy, G. (2009). *Functional Magnetic Resonance Imaging* (2nd ed.). Sunderland, MA: Sinauer Associates, Inc.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis.* New York: Cambridge University Press.

In addition, there are numerous free and commercial software packages available for data preprocessing and analysis. Most of these have websites with extensive documentation and tutorials. While the software packages differ somewhat in the algorithms they offer, most packages are quite suitable for most experimental designs. The websites of the following packages are readily available through any Web search engine:

Analysis of Functional Neuroimages (AFNI)
FMRIB's Software Library (FSL)
FreeSurfer
Statistical Parametric Mapping (SPM)
Brain Voyager (a commercial package)

References

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B., & Zilles, K. (1999). Broca's region revisited: Cytoarchitecture and intersubject variability. *The Journal of Comparative Neurology, 412*(2), 319–341.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics.* Cambridge, UK: Cambridge University Press.

Bandettini, P. A., Wong, E. C., Hinks, R. S., Tikofsky, R. S., & Hyde, J. S. (1992). Time course EPI of human brain function during task activation. *Magnetic Resonance in Medicine, 25*(2), 390–397.

Bavelier, D., Newman, A. J., Mukherjee, M., Hauser, P., Kemeny, S., Braun, A., & Boutla, M. (2008). Encoding, rehearsal, and recall in signers and speakers: shared network but differential engagement. *Cerebral Cortex, 18*(10), 2263–2274.

Binder, J. R., Frost, J., Hammeke, T., Bellgowan, P., Rao, S. M., & Cox, R. W. (1999). Conceptual processing during the conscious resting state: A functional MRI study. *Journal of Cognitive Neuroscience, 11*(1), 80–95.

Brodmann, K. (1909). In L. Garey (Ed.), *Brodmann's Localisation in the cerebral cortex*. River Edge, NJ: Imperial College Press.

Buckner, R. L., Bandettini, P. A., O'Craven, K. M., Savoy, R. L., Petersen, S. E., Raichle, M. E., & Rosen, B. R. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. *Proceedings of the National Academy of Sciences, 93*(25), 14878–14883.

Cohen, J. (1983). The Cost of Dichotomization. *Applied Psychological Measurement,* 7(3), 249–253.

Dale, A.M., & Buckner, R. L. (1997). Selective averaging of rapidly presented individual trials using fMRI. *Human Brain Mapping, 5*(5), 329–340.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences, 15*(6), 254–262.

Dodel, S., Golestani, N., Pallier, C., Elkouby, V., Le Bihan, D., & Poline, J.-B. (2005). Condition-dependent functional connectivity: syntax networks in bilinguals. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 360*(1457), 921–935.

Dunn, D. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test, Fourth Edition, manual.* Minneapolis, MN: NCS Pearson, Inc.

Dunn, L. M., & Theriault-Whalen, C. M. (1993). *Echel le de vocabulaire en images Peabody.* Toronto, ON: Psycan.

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M.-H., Evans, A. C., Zilles, K., & Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage, 36*(3), 511–521.

Fischl, B., Sereno, M. I., & Dale, A.M. (1999). Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage, 9* (2), 195–207.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age Constraints on Second-Language Acquisition. *Journal of Memory and Language, 41*(1), 78–104.

Genovese, C. R., Lazar, N. A., & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage, 1*(4), 870–878.

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage, 9*(4), 416–429.

Gusnard, D., & Raichle, M. (2001). Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews: Neuroscience, 2*(October), 685–694.

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., . . . Bowtell, R. W. (1999). "Sparse" temporal sampling in auditory fMRI. *Human Brain Mapping,* 7(3), 213–223.

Hammill, D. D., Brown, V. L., Larsen, S. C., & Wiederholt, J. L. (1994). *Test of Adolescent and Adult Language, third edition (TOAL-3)* (3rd ed.). Austin, TX: Pro-Ed.

Henson, R. (2007). Efficient Experimental Design for fMRI. In K. J. Friston (Ed.), *Statistical parametric mapping* (pp. 193–210). London: Elsevier.

Huettel, S. A., & McCarthy, G. (2001). The effects of single-trial averaging upon the spatial extent of fMRI activation. *Neuroreport, 12*(11), 2411–2416.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology, 21*(1), 60–99.

Kober, H., & Wager, T. D. (2010). Meta-analysis of neuroimaging data. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(April), 293–300.

Kwong, K. K., Belliveau, J. W., Chesler, D. A., Goldberg, I. E., Weisskoff, R. M., Poncelet, B. P., . . . Turner, R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences, 89*(12), 5675–5679.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *Nature, 412*(6843), 150–157.

Maccallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the Practice of Dichotomization of Quantitative Variables. *Psychological Methods,* 7(1), 19–40.

Magistretti, P. J., & Pellerin, L. (1999). Cellular mechanisms of brain energy metabolism and their relevance to functional brain imaging. *Transactions of the Royal Society of London. Series B, Biological Sciences, 354*(1387), 1155–1163.

Malonek, D., & Grinvald, A. (1996). Interactions Between Electrical Activity and Cortical Microcirculation Revealed by Imaging Spectroscopy: Implications for Functional Brain Mapping. *Science, 272*(5261), 551–554.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate Median Splits and Spurious Statistical Significance. *Psychological Bulletin, 113*(1), 181–190.

Meyer, M., Friederici, A.D., & von Cramon, D.Y. (2000). Neurocognition of auditory sentence comprehension: event related fMRI reveals sensitivity to syntactic violations and task demands. *Cognitive Brain Research, 9,* 19–33.

Miezin, F. M., Maccotta, L, Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage, 11*(6 Pt 1), 735–759.

Newman, A. J., Pancheva, R., Ozawa, K., & Neville, H. J. (2001). An event-related fMRI study of syntactic and semantic violations. *Journal of Psycholinguistic Research, 30*(3), 339–364.

Newman, A. J., Supalla, T., Hauser, P. C., Newport, E., & Bavelier, D. (2010a). Dissociating neural subsystems for grammar by contrasting word order and inflection. *Proceedings of the National Academy of Sciences, 107*(16), 7539.

Newman, A. J., Supalla, T., Hauser, P. C., Newport, E. L., & Bavelier, D. (2010b). Prosodic and narrative processing in American Sign Language: an fMRI study. *NeuroImage, 52*(2), 669–676.

Newman, A. J., Tremblay, A., Nichols, E. S., Neville, H. J., & Ullman, M.T. (2012). The influence of language proficiency on lexical semantic processing in native and late learners of English. *Journal of Cognitive Neuroscience, 24*(5), 1205–1223.

Ogawa, S, Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences, 87*(24), 9868–9872.

Ogawa, S., Tank, D. W., Menon, R., Ellermann, J. M., Kim, S. G., Merkle, H., & Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences, 89*(13), 5951–5955.

Pakulak, E., & Neville, H. J. (2010). Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. *Journal of Cognitive Neuroscience, 22*(12), 2728–2744.

Pauling, L., & Coryell, C. D. (1936). The magnetic properties and structure of hemoglobin, oxyhemoglobin and carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences, 22*(4), 210–216.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage, 40*(2), 409–414.

Rademacher, J., Morosan, P., Schormann, T., Schleicher, A., Werner, C., Freund, H., & Zilles, K. (2001). Probabilistic mapping and volume measurement of human primary auditory cortex. *Neuroimage, 13*(4), 669–683.

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception, 33*(2), 217–236.

Roy, C., & Sherrington, C. (1890). On the regulation of the blood-supply of the brain. *Journal of Physiology, 11*(1–2), 85–108.

Sadato, N., Pascual-Leone, A., Grafman, J., Ibanez, V., Deiber, M. P., Dold, G., & Hallett, M. (1996). Activation of the primary visual cortex by Braille reading in blind subjects. *Nature, 380*(6574), 526–528.

Saur, D., Baumgaertner, A., Moehring, A., Büchel, C., Bonnesen, M., Rose, M., . . . Meisel, J. M. (2009). Word order processing in the bilingual brain. *Neuropsychologia, 47*(1), 158–168.

Takano, T., Tian, G.-F., Peng, W., Lou, N., Libionka, W., Han, X., & Nedergaard, M. (2006). Astrocyte-mediated control of cerebral blood flow. *Nature Neuroscience, 9*(2), 260–267.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain: 3-dimensional proportional system: an approach to cerebral imaging.* New York: Thieme Medical Publishers.

Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-Analysis of the Functional Neuroanatomy of Single-Word Reading: Method and Validation. *NeuroImage, 16*(3), 765–780.

Vannest, J., Newport, E. L., Newman, A. J., & Bavelier, D. (2011). Interplay between morphology and frequency in lexical access: The case of the base frequency effect. *Brain Research, 1373,* 144–159.

Weber-Fox, C., Davis, L., & Cuadrado, E. (2003). Event-related brain potential markers of high-language proficiency in adults. *Brain and Language, 85*(2), 231–244.

Worsley, K. J. (2001). Statistical analysis of activation images. In P. Jezzard, P.M. Matthews, & S. M. Smith (Eds.), *Functional magnetic resonance imaging: an introduction to the methods* (pp. 251–270). New York: Oxford University Press.

8

TRANSLATION RECOGNITION TASKS

Gretchen Sunderman

History of the Method

Over the years, there has been a wealth of psycholinguistic research investigating the lexicon, or an individual's mental dictionary. Some researchers study the mental lexicon in monolinguals, or individuals who speak only one language. Other psycholinguists study bilinguals, or individuals who speak more than one language. Studying bilinguals allows researchers to ask questions about how words from two languages are organized and retrieved from the lexicon. But within this area of research, the term *bilingual* is sometimes applied to individuals who are highly proficient in two languages that they learned simultaneously from birth, as well as to individuals who are learning a language in the university setting. It is obvious that these are very different types of bilinguals. In fact, the terms *second language (L2) learner, beginning bilingual,* and *advanced bilingual* better describe those individuals learning a language later in life in an academic setting. Yet, on some level, wherever the individual falls on the language skill spectrum, that individual is no longer a monolingual. Thus, research investigating the bilingual lexicon and bilingual memory is concerned with individuals who have words from more than one language in their minds, from beginning stages of language learning through proficient, highly-skilled use. Scholars interested in second language acquisition often focus on developmental aspects of the bilingual lexicon investigating beginning bilinguals (i.e., L2 learners). However, they also focus on individuals with more advanced language skills as well, as a point of comparison in terms of lexical development. As a result, throughout this chapter, various types of individuals with differing language abilities will be discussed. The terms *L2 learner* and *beginning bilingual* will often be used interchangeably, whereas the term *highly-proficient bilingual* will be used for those with more advanced skill in a second language.

An important psycholinguistic task that has been regularly used with L2 learners and highly proficient bilinguals is the translation recognition task. The translation recognition task is perhaps not the first task one might think of when considering research on the lexicon. Many might think of the lexical decision task, a task that requires participants to judge a string of letters and decide whether or not it is a word. The lexical decision task is frequently used in both monolingual and bilingual studies of word recognition. However, it is not without its problems. De Groot (2011) has argued that the task is quite unnatural and worse, may not truly tap into word recognition processes. What then is a better task to investigate the bilingual lexicon? The answer is a task that requires participants to process words in a more natural and meaningful way. This task is the translation recognition task. The task itself is quite straightforward: present two words sequentially and ask participants to judge whether the two words mean the same thing, or are translation equivalent of each other. This powerful and versatile task has become popular in research on the bilingual lexicon, and most importantly for our purposes, with second language learners.

The translation recognition task is a relatively new task in bilingualism and second language acquisition research, emerging in the early 1990s (de Groot, 1992). Historically, the task was developed for both theoretical and practical reasons. The theoretical reason was related to the retrieval process of words from the lexicon, or selecting the proper word to speak. Translation is a popular task used by psycholinguists (e.g., Kroll & Stewart, 1994; Potter, So, Von Eckard, & Feldman, 1984). One can simply show words to participants and have those participants produce the word in the other language. Often performance on translation tasks is then compared to performance on picture or word naming tasks to investigate the nature of the underlying links in the lexicon (see VanPatten, Chapter 1, this volume for a description of Kroll & Stewart, 1994). However, translation requires production, actually speaking the word in either the first or the second language. Producing words in a second language can be quite difficult for second language learners. De Groot (1992) reasoned on a theoretical level that a translation recognition task, which does not require overt production, would help isolate any effects found in the processing stages prior to production. In other words, a translation recognition task would circumvent the translation-retrieval process, and would provide a tool to examine the preproduction stages of translation. By using a translation recognition task instead of a productive translation task, researchers could get a glimpse of what participants were activating in their minds and intending to produce.

On practical grounds, de Groot (1992) proposed that a task like translation recognition would be better suited for beginning bilinguals with lower levels of L2 proficiency. L2 production is notoriously difficult, especially for beginning language learners. A researcher investigating production either has to use a few very simple words that the participants are likely to know (which limits the generalizability of the study) or has to deal with missing data and unhappy participants who become frustrated when they cannot produce a word. Importantly, the lack of production of a word does not necessarily imply that the word is completely unknown; it could simply mean that the word has a weak representation in bilingual memory. The solution to the problem proposed by de Groot (1992) was to create a task that mimics the translation process, but does not require overt translation. Thus, the translation recognition task was created.

In order to ensure that translation recognition indeed was sensitive to the same types of experimental manipulations as translation production, de Groot and Comijs (1995) compared the two tasks directly looking specifically at a series of lexical features that are known to affect processing. In their study, de Groot and Comijs investigated word imageability (e.g., Paivio, 1986), context availability (Schwanenflugal, Harnishfeger, & Stowe, 1988), definition accuracy (de Groot, Dannenburg, & van Hell, 1994), word familiarity (Noble, 1953), log word frequency, length, and cognate status as potential variables that could affect translation recognition in L1 Dutch-L2 English bilinguals. To give an example of the first few variables, consider the words *tree* and *truth*. Word imageability can be understood as how easy it is to form a mental image of the word. Context availability is how easy it is to imagine a context where the word could be used and definition accuracy is how easy it is to define a particular word. At this point you probably have a fairly strong image of a *tree* active in your mind, and a much less active image of *truth*. These three variables are highly correlated to each other and with the variable of word concreteness, or how perceivable a word is. In other words, these variables all represent more semantic-type factors. To understand the other variables, consider the words *piano* and *ubiquitous.* Word familiarity is a subjective measure of how well-known a word is to an individual. If we asked a set of individuals to rate which word was more familiar to them, *piano* would be far more familiar than *ubiquitous.* The log word frequency is an objective measure of the occurrence of a given word in a specific corpora and length is simply the number of letters of a word. And finally, cognate status is a measure of the similarity (both orthographic and phonological) between two words. The word *piano* is more frequent than the word *ubiquitous* and is shorter in length, thus making it easier and faster to process than *ubiquitous.* In addition, *piano* is a cognate in Dutch, of *piano,* thus the retrieval of this word will also be faster relative to noncognate words.

Norming studies are incredibly important in lexical level investigations. The score a given word received related to the above factors (except log word frequency and length) was determined by de Groot and Comijs (1995) using a norming study in which participants rated the words on a 7-point scale with respect to the particular variable. For example, each participant was asked to judge how easy it was to think of an image for the following words: *umbrella, lunch, justice, hunger* (see de Groot et al., 1994, for a complete description of the norming procedures). Overall, de Groot and Comijs found that both translation recognition and translation production respond similarly to the above described manipulations and therefore translation recognition could be considered an alternative to standard translation production. However, one of the most interesting findings of their study was that participants seemed to exhibit a certain type of bias in their processing; participants were faster to accept a pair as translation equivalents if they looked perceptually similar (or were cognates). In other words, the participants were sensitive to form similarity across language. This processing bias was then turned on its head and used against participants in subsequent translation recognition studies in that field. Instead of having participants accept translation pairs, cleverly constructed nontranslation pairs were developed to potentially interfere with or slow down processing. The logic being that where there was interference, there must have been a memory trace. Thus a new tool was revealed to investigate the structure of bilingual memory.

Translation recognition has been used to investigate many questions about the developing lexicon of second language learners. Some of these questions include: How do L2 learners access meaning in the L2? Can L2 learners directly access meaning (a notion called conceptual mediation) or do they instead rely on their L1 to access meaning? What is the role of the L1 translation equivalent during lexical processing? How do the lexical and conceptual links change as an individual becomes more proficient in the L2? These questions are directly related to the predictions of the well-known developmental model of the bilingual lexicon, the Revised Hierarchical Model (Kroll & Stewart, 1994). The model predicts that in early second language learning, conceptual mediation may be difficult, and a learner will instead be more reliant on lexical or form relations between the two languages. However, as proficiency increases, the interlanguage connections change and shift from lexical processing to conceptual processing. These questions have been tested using the translation recognition task in quite a few studies (e.g., Altarriba & Mathis, 1997; Ferré, Sánchez-Casas, & Guasch, 2006; Guasch, Sánchez-Casas, Ferré, & García-Albea, 2008; Laxén & Lavaur, 2010; Quasem & Foote, 2010; Sunderman & Kroll, 2006; Sunderman & Priya, 2012; Talamas, Kroll, & Dufour, 1999) with different populations of bilinguals and language learners, with different experimental manipulations, and with different results. In the next section, a description of the various experimental manipulations that have been used to investigate the developing bilingual lexicon is presented.

What is Looked at and Measured

A translation recognition task uses behavioral measures of reaction time (RT) in milliseconds and accuracy to investigate processing. Simply put, the task measures how quickly and accurately an individual is able to look at two words (presented one after the other) and decide if those two words mean the same thing, or are translation equivalents. The underlying assumption in using measures of RT and accuracy is that longer RTs and greater error rates indicate a processing difficulty. On the one hand, if a participant takes longer and makes more mistakes in a certain condition of the experiment compared to another condition, then we assume this increased processing time signals *interference.* On the other hand, if a participant is faster and makes fewer mistakes in a certain condition of the experiment compared to another condition, then we assume this decreased processing time is signaling a *facilitation* of some sort.

Central to the notions of interference and facilitation is the notion of *activation.* An extensive body of research indicates that second language learners and bilinguals activate information about words in both languages in parallel, regardless of their intention to function within one language alone (See de Groot, 2011, for a review of lexical processing in bilinguals). In other words, it is not possible to completely "turn off" one language when processing in another language; the two languages are always active, or "turned on," to some extent. Given the activation of two languages in one mind, researchers can use experimental tasks to attempt to tap into that activation. For instance, if there is an assumption that L2 learners of Spanish automatically activate the translation equivalent *dog* when they see the word *perro,* then an experiment using a translation recognition task and the notion of interference can examine that assumption. Imagine a learner sees the word *perro.* If that learner needs the L1 crutch to access the meaning of that word, the learner will retrieve the word *dog.* Thus, *dog* will be highly activated in the learner's mind. If the learner then sees a word that looks like or sounds like *dog,* such as *dot,* the two similar words then may start competing with each other for activation. This competition gives rise to interference, and the overall processing is slowed. The learner may take a bit longer to respond, and perhaps may even make a mistake and accept that *perro* and *dot* are translation equivalents. We infer from the increased RTs and error rates that the interference from the competitor is directly related to the activation of the translation equivalent. If *dog* had not been highly activated, then seeing the word *dot* should not have caused interference.

The actual mechanics of the translation recognition task are quite straightforward. The task is completed on a computer and requires participants to simply press a "yes" or "no" button to register their response. Prior to the presentation of the two words, a fixation point (+) is presented at the center of the screen. When participants are ready, they press a key on a button box to begin. The first word appears for a set period of time (400 ms for example) followed by a brief blank screen (100 ms) and then the second word appears. The second word either remains on the screen until the participants press a response button indicating whether or not the two words are translation equivalents, or it times out after a set period of time if the participant does not respond quickly enough. The second word is then replaced by a fixation point. The process repeats until all trials are completed.

There are two sets of trials in this experiment, the "yes" trials where a participant accepts the translation pair and the "no" trials. The critical trials in a translation recognition task are those that require participants to say "no" and reject the pair. A set of distracters are created that are either orthographically, phonologically, morphologically, or semantically related to one of the words in the translation pair, depending on the question being investigated in the study. The logic of the task is to see if the distracters affect participants' abilities to reject the translation pair. If it takes a participant more time to reject a translation pair or if the participant is committing more errors compared to an unrelated control pair, then, as previously described, the form or meaning similarity that was present in the distracter is inferred to be the cause of the interference.

In a series of examples below, the logic of this task and the notion of interference across several different dimensions will be illustrated. Specifically, I will describe orthographic, phonological, morphological, and semantic manipulations that have been used in previous research with the translation recognition task. The purpose of this description is twofold. First, it will provide concrete examples of materials that have been used in various studies that can be easily understood in terms of interference. And second, the descriptions serve to illustrate the versatility of the translation recognition task to address a wide range of theoretical questions, simply by altering the materials used in the study. To that end, the section that follows immediately after the current section will provide detailed information on how to develop and create those materials for a translation recognition task.

Orthographic Manipulations

One type of experimental manipulation used in a translation recognition experiment is based on orthography, or the written word. In the examples in (1) below, taken from Talamas et al. (1999), English-dominant language learners who differed in their level of proficiency in Spanish were presented with "no" trials where the second word of the translation pair was orthographically related to the translation equivalent *soap-jabón*.

(1) *Orthographic manipulation of the second word of the translation pair*

a. Related: *soap-jamón* (ham)
b. Unrelated: *soap-limón* (lemon)

What is the logic of an orthographic manipulation? This manipulation will allow a researcher to investigate whether the translation equivalent is differentially active for learners with a range of proficiency in the L2. Indeed, Talamas et al. (1999) found that this type of related trial produced different results for the more and less proficient learners. For less proficient learners, there was significant interference for orthographic form-related pairs. In other words, when the less proficient learners had to say "no" to a pair like *soap-jamón,* the similarity between *jabón* and *jamón* caused them to be slower and more inaccurate compared to the unrelated control trials which did not share the orthographic similarity. The more proficient learners were not tricked by the form similarity. The logic is that interference suggests activation. Given the nature of the interference, we can infer that the less proficient learners were activating the translation equivalent to a greater extent than the more proficient learners (the more proficient learners were relying on semantic information which will be discussed in a subsequent section). The overall pattern of results provides support for the hypothesis that early in second language learning, lexical form relations between L2 and L1 provide the basis of interlanguage connection, a claim of the Revised Hierarchical Model (Kroll & Stewart, 1994).

The Talamas et al. (1999) study used an orthographic manipulation related to the translation equivalent. Another approach is to manipulate the orthographic similarity to the word itself. In the examples (2) below, taken from Sunderman and Kroll (2006), English-dominant bilinguals who differed in their level of proficiency in Spanish were presented with "no" trials where the second word of the translation pair was orthographically related to the first word of the correct translation pair, in these examples *cara-face.*

(2) *Orthographic manipulation of the first word of the translation pair*

a. Related: *cara-card*
b. Unrelated: *cara-lake*

What is the logic of this type of orthographic manipulation? This manipulation tests the notion that L2 learners cannot shut off their L1 when processing in the L2. The prediction was that in the context of a translation recognition task, distractor words that share lexical features with the first presented word (e.g., *cara-card*), were expected to become highly activated and would thus require additional time to correctly reject compared to unrelated controls. These predictions were based on the Bilingual Interactive Activation (BIA) model (Dijkstra & Van Heuven, 1998), a well-known word recognition model. Indeed, Sunderman and Kroll found that both more and less proficient learners were as sensitive to this type of orthographic manipulation. In other words, all participants were slower and more inaccurate in rejecting pairs that looked similar. This finding suggests that one is not able to shut off the dominant first language (L1) during word recognition. However, we will see in the Exemplary Study section that this inability to shut off the L1 may be dependent on the context of learning.

In both of the previous examples, the orthography of the distracters was hypothesized to cause interference for differing theoretical reasons, one based on the translation equivalent activation and the other on L1 competitor activation. However, orthographic distracters are only one type of manipulation that can be used in a translation recognition task. In example (2) above, the interference the participants suffered could have partially been explained by phonology. The items in their study such as *cara-card* shared not only perceptual features (i.e., orthography), but they also shared phonology. In other words, they sound alike. Given that

Spanish and English share the same alphabet, it is difficult to disentangle the relative contribution of orthography and phonology. However, Sunderman & Priya (2012) isolated the variable of phonology in a study where they asked whether highly proficient bilinguals were sensitive to phonological information when performing a translation recognition task.

Phonological Manipulations

In the examples in (3) and (4) below, taken from Sunderman and Priya (2012), highly proficient Hindi-English bilinguals performed a translation recognition task in Hindi (which uses the Devanagari script) and English (which uses the Latin script). The critical "no" trials were phonological distracters; the two languages do not share script and it would be impossible to have orthographic distracters. For each word pair such as *बिल्ली-cat* (बिल्ली pronounced as [bili]), two related distracters were created. The distracters were either phonologically related to the second item of the pair (i.e., the translation equivalent) or phonologically related to the first item of the pair. For example, in the first condition in (3) the distracter is the word *cap* which sounds like the translation equivalent in English *cat*. In the second condition in (4) the distracter is the word *Billy* which sounds like the Hindi word *बिल्ली*.

(3) *Phonological Manipulation: Translation Neighbors*

a. Related: *बिल्ली-cap*
b. Unrelated: *बिल्ली-Dan*

(4) *Phonological Manipulation: Neighbors*

a. Related: *बिल्ली-Billy*
b. Unrelated: *बिल्ली-Entry*

These phonological manipulations are parallel to the two orthographic manipulations described above in examples (1) and (2), but at the level of phonology. The phonological manipulation that was used in (3) investigated whether different-script bilinguals activate and utilize translation equivalents in a similar manner to same-script bilinguals. The manipulation in (4) examined whether phonology could be turned off in the other language. With respect to the translation neighbor manipulation in (3), the highly proficient different-script bilinguals suffered interference from the related distracters. In other words, just like the less proficient learners in several other studies (Sunderman & Kroll, 2006; Talamas et al., 1999), these types of distracters that sounded like the translation equivalents were difficult to reject for the Hindi-English participants. This suggests that the phonology of the translation equivalent was active for these bilinguals. Given that these individuals were highly-proficient in Hindi and English, this was odd to see; this type

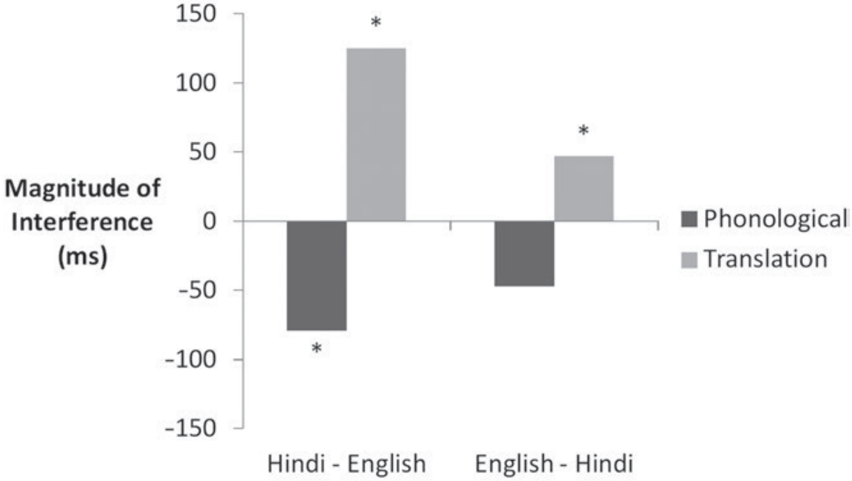


FIGURE 8.1 Degree of phonological facilitation and translation interference (in ms) in Hindi-English and English-Hindi directions. * $p < 0.01$

of interference in the task is typically only found with less proficient bilinguals. Thus, the finding suggests that factors such as phonology and script may influence the translation recognition process as well, and that different script bilinguals utilize the translation equivalent in a very different manner. With respect to the neighbor manipulation in (4), the Hindi-English bilinguals were faster to reject the phonologically related distracters. In other words, when the target word, in either direction (Hindi-English or English-Hindi) was presented and it sounded like the first word presented, the participants were facilitated in rejecting that pair. The presence of the phonological distracter helped them to quickly discard the item as being a possible translation equivalent. This finding indicates that script and phonology were playing important cues in the translation recognition task, in ways we do not fully understand at this point. Figure 8.1 above illustrates both the translation interference and facilitation that the participants suffered in this translation recognition task in both the Hindi-English and English-Hindi direction of the task. The magnitude of interference is the difference between the related and unrelated trials. Interference is plotted above the horizontal axis (increased RT) and facilitation is plotted below the horizontal axis (decreased RTs).

Morphological Manipulations

While the translation recognition task itself is relatively straightforward, the types of manipulations with the materials can become quite complex based on the theoretical question involved. Qasem and Foote (2010) asked whether individuals' activation patterns in the lexicon could be based on morphology, and not solely orthography. This idea was based on the predictions of the morphological decomposition model (Frost, Forster, & Deutsch, 1997), which suggests that the lexical activation in certain languages like Arabic is based on morphological similarity rather than orthographic similarity. In the example in (5) below, taken from Qasem and Foote (2010), native Arabic speakers who differed in their level of proficiency in English were presented with "no" trials where the second word of the translation pair was orthographically or morphologically related to the translation equivalent *shoulder-katif.*

(5) *Shoulder-*تك ف *(katif)*

a. Orthographically related: هك ف *(kafh)* "cave"
b. Morphologically related: تكات ف *(takaatuf)* "unity"

In (5a) the word *kahf* is orthographically related to the word *katif* in that it shares all but one root consonant, *t.* In Arabic, words contain root morphemes and vowel infixes to form words. For example, the root *k-t-b* with the addition of various combinations of vowels can create the following words: ***k***_a_***t***_a_***b***_a_ "wrote," ***kutib*** "has been written," and _a_***ktub*** "am writing" (O'Grady, Achibald, Aranoff, & Rees-Miller, 2010). Often the words are semantically related, as in the case above, but it is also possible to have the same root morphemes, but have a word that is *not* semantically related. In (5b) the morphologically related *katif* and *takaatuf* are both derived from the *k-t-f* consonantal root, but they are semantically *unrelated* (*shoulder* and *unity*). When presented with these two types of distracters, the participants in the study suffered interference in both the orthographically and morphologically related conditions, regardless of proficiency. Moreover, these Arabic native speakers suffered even more interference in the morphologically related condition. Thus, the pattern of the data suggests that the speakers of Arabic were accessing the lexicon based on morphological features over orthographic features. The morphological distracters caused more interference because morphological information was highly activated.

Semantic Manipulations

One area that has been particularly productive with the translation recognition task is related to the question of conceptual mediation, or whether L2 learners are able to directly access the meanings of L2 words directly without having to translate or rely on the L1. This claim has been tested using a translation recognition task and a semantic manipulation.

In the examples (6) below, taken from Talamas et al. (1999), English-dominant language learners who differed in their level of proficiency in Spanish were presented with "no" trials where the second word of the translation pair was semantically related to the translation equivalent pair *soap-jabón.*

(6) *Semantic manipulation*

a. Related: *soap-baña (bathe)*
b. Unrelated: *soap-cerca (close)*

Talamas et al. found that this type of manipulation produced different results for the more and less proficient learners. For the more proficient learners, there was significant interference for the semantically-related pairs. In other words, when the more proficient learners had to say "no" to a pair like *soap-baña,* the semantic similarity between *soap* and *baña* caused them to be slower and more inaccurate compared to the unrelated control trials which did not share conceptual overlap. The less proficient learners were not as affected by the semantic manipulation, but recall that they were tricked by orthographic translation similarity. Given the nature of the interference, we can infer that the more proficient learners were conceptually mediating. The overall pattern of results of the Talamas et al. study provides support for the hypothesis that early in second language learning, only more proficient learners can directly access concepts without relying on the L1.

However, others (e.g., Altarriba & Mathis, 1997; Sunderman & Kroll, 2006; Quasem & Foote, 2010), using a similar semantic manipulation and a translation recognition task, have found the opposite results, lending support to the notion that second language learners do indeed have access to conceptual information early in the acquisition process. While it is beyond the scope of this chapter to address these discrepant results in the literature, what is important to focus on is how scholars continue to explore the notion of semantic relatedness and test it via translation recognition tasks. Some scholars (Ferré, Sánchez-Casas, & Gausch, 2006; Gausch, Sánchez-Casas, Ferré, & García-Alba, 2008) investigated the strength of the semantic relationship, basically asking whether learners were more sensitive to very closely related semantic relations (e.g., *donkey-horse*) than weaker semantic relations (*donkey-bear*), and indeed they were. Others (Laxén & Lavaur, 2010) have investigated the effects of translation ambiguity (Tokowicz & Kroll, 2007, Tokowicz, Kroll, de Groot, & van Hell, 2002). Translation ambiguity arises when some words have more than one translation between languages, thus causing a mapping problem. Laxén and Lavaur specifically looked at how the translation equivalent of words in French that have one translation (*arbre-tree*) versus two semantically *unrelated* translations (*argent-money* and *silver*) or two semantically *related* translations (*bateau-boat* and *ship*) are accepted in a translation recognition task. The hypothesis was that it would be easier to accept unambiguous pairs compared to ambiguous pairs, and indeed it was, at least for highly skilled French-English bilinguals in this study. The bottom line with semantic manipulations, or any of the other manipulations discussed in this section, is that through clever manipulations with the stimuli, a researcher is able to investigate very specific features of the architecture of the lexicon. In the next section, I will address how to design materials for a translation recognition task.

Issues in the Development and Presentation of Stimuli

The creation of stimuli for a translation recognition task can be quite challenging. However, there are certain steps and procedures that will help mitigate any potential problems that can arise when developing stimuli for an experiment. These steps often include referencing established norms in the literature or conducting a norming study to select appropriate stimuli. Because the translation recognition task is so versatile and can be used to test predictions based on orthography, phonology, morphology, and semantics, just to name a few, it is difficult to describe the exact manner in which to develop stimuli. The stimuli will always be dependent on the manipulations of the experiment and the theoretical question at hand. However, if one has a sound understanding of how to create materials within the context of a specific study, this knowledge can then be transferred and applied to other manipulations, languages, and theoretical questions. Thus, in the following section, I will use the materials and conditions from Sunderman and Kroll (2006) to illustrate the various steps that take place in developing stimuli.

Choosing Correct Translation Pairs

Although a translation recognition task requires creating distracters that can potentially interfere with processing, before one can create distracters, the correct translation pairs must be created. In our study (Sunderman & Kroll, 2006), we selected 48 correct translation pairs. These pairs were not randomly selected; we first began with a much larger set of potential translation pairs. We knew that we would be testing L2 learners with low levels of L2 proficiency, so we had to select pairs that we were fairly certain the learners would know. Often using language textbooks or one's experience as a language instructor can help one initially choose a set of translation pairs that should be known to the lowest level of proficiency of the participants in the study. However, to ensure that the translation pairs were at an appropriate level of proficiency, we also needed to conduct a norming study. A norming study in this case entailed presenting a group of participants (who were then not eligible to participate in the study) with the correct translation pairs to see if they indeed knew those pairs. For example, we used a simple paper and pencil task and asked participants who were at an equivalent or lower proficiency level than was used in the study to write down the correct translation equivalent of our potential items. Those items with low levels of accuracy were then discarded from the pool of correct translation pairs. In the end, we settled on 48 correct translation pairs. These correct translation pairs were never actually seen in the study; they were simply the starting point for creating the distracters. The critical trials in a translation recognition task are those that require participants to say "no" and reject the pair. A set of distracters must be created that are related in some way to one of the words in the translation pair, depending on the question being investigated in the study.

Creating the Distracters by Condition

In our study (Sunderman & Kroll, 2006), we were interested in investigating how two different types of form similarity and one type of semantic similarity affected lexical processing. In addition, we were also interested in investigating whether grammatical class information (i.e., whether something was a noun or a verb) would affect the translation recognition process. Translation equivalents must necessarily share the same grammatical class, so we were interested in seeing if participants could use grammatical class as a cue when judging translation pairs. Thus, for each correct translation pair, we needed to make six different distracters, two for each type of form or semantic similarity, one with the same grammatical class (+) and one with a different (–) grammatical class. We created 48 tables like Table 8.1 for each correct translation pair, thus giving us 288 related distracters. Since we only conducted the experiment in the L2-L1 direction, we only had to create this one set of distracters. However, if we had included the L1-L2 direction, we would have had to create another table with separate distracters. Simply translating the distracters will not maintain the form similarity across language.

To illustrate, consider the correct translation pair *cara-face* in Table 8.1. There are two different distracters (which differed by whether or not they were the same as the target word in terms of grammatical class) for each condition: orthographically related to the first word of the pair, orthographically related to the second word of the pair, or semantically related. The words *card* and *care* are both orthographically related to the word *cara*. The words *fact* and *fast* are both orthographically related to the word *face.* In the third condition, the distracters *head* and *pretty* were related in meaning to the translation pair. But how did we generate those items? How did we determine orthographic similarity? Or grammatical class? Or semantic relatedness? The answer is through a series of lexical databases, calculations, and norming studies.

For orthographic similarity, we operationalized it as the onset of the word, typically the first two to three letters of the word. We then generated items that were either the same as (i.e., matched) the grammatical class of the word pair or did not match. Of course, given that some nouns can also be verbs, such as the word *face* (e.g., to face the music), words were always used according to their most

TABLE 8.1 Illustration of materials used in each condition for the pair CARA-FACE

Grammatical Class	*Form conditions*		*Meaning condition*
	Orthographically related to 1st word	*Orthographically related to 2nd word*	*Semantically related*
+	Card	Fact	Head
–	Care	Fast	Pretty

frequent grammatical sense as determined by the frequency norms of Francis and Kucera (1982). Basically, this first step requires thinking of words and consulting dictionaries and frequency norms. The CELEX frequency norm (Baayen, Piepenbrock, & Gulikers, 1995) is another to consult.

Because we generated two distracters for each form-related condition that were either matched or mismatched , we wanted to make sure that the only difference between those words was their grammatical class and not something else like orthography. In other words, we wanted to make sure the distracters in the matched grammatical class were not somehow more orthographically similar than the distracters in the unmatched grammatical class condition. Therefore, we also computed an orthographic similarity measure described by van Orden (1987) on all of the distracters. This calculation takes into account identical letters in the same position, adjacent letters, and similar first and last letters and assigns a value to each word based on the similarity to another word. A simple *t*-test between the orthographic similarity measures between the unmatched and matched grammatical distracters can reveal if there are differences. The analysis of orthographic similarity produced no significant differences as a function of grammatical class. This procedure allowed us to be confident that grammatical class was not confounded with orthography. However, had differences arisen, the stimuli would have had to have been modified until no differences existed.

In terms of semantic similarity, words for the meaning-related condition were first selected from the Edinburgh Associative Thesaurus (EAT), an interactive online associative thesaurus that generates common word associations to a given word (Kiss, Armstrong, Milroy, & Piper, 1973). We also obtained an independent measure of semantic similarity. Each of the 48 target words was paired with its semantically related distracter and presented in English to native English-speaking participants who rated them for semantic similarity. Semantic similarity is understood as the strength of a relationship between two ideas or concepts. Participants were instructed to look at the two words and decide the similarity of the two meanings of the words. They were told to use a seven-point scale in which 1 meant *very different* and 7 meant *very similar.* These similarity scores were then used in later data analysis to help understand the patterns in the data.

What is important to take away from this description of the generation of the related stimuli in the Sunderman and Kroll study is not necessarily the particulars of the study, but the type of information that must be gathered about the stimuli used. If one wants to match on orthography, then there are likely databases or algorithms to determine orthographic similarity. The same goes for phonology, morphology, and semantics associations. If no such database exists or does not quite capture the relationship in question, then one needs to conduct a norming study to have a sense of the characteristics of the stimuli. In the end, in constructing the related distracters in a translation recognition experiment, one is simply attempting to make sure to the best of their ability that the stimuli are related along the one and only one dimension they are trying to investigate in that particular condition. And, if there is any hint of differences (because it is very difficult to achieve perfection in any stimuli creation endeavor), then having an independent rating or measure of that difference is important as it will allow the researcher to use that information in the statistical analyses. Some other valuable resources to consult when creating stimuli would be:

(1) The English Lexicon Project (http://elexicon.wustl.edu). This is a free database that contains lexical characteristics, along with reaction time and accuracy measures from two different experiments (visual lexical decision and naming), for 40,481 words and 40,481 nonwords (Balota et al. 2007).
(2) The MRC Psycholinguistic database (www.psych.rl.ac.uk*)*. This is a powerful database containing words with up to 26 linguistic and psycholinguistic characteristics for each word (Wilson, 1988).
(3) WordNet (http://wordnet.princeton.edu). This is an online database evaluating lexical-conceptual and semantic relationships between words (Princeton University, 2010).

Creating the Unrelated Controls by Condition

At this point, a set of unrelated distracters must also be created. For any given distracter, the unrelated distracter has to be identical to the related distracters in terms of length and frequency, but not alike in terms of the variable in question. For example, for the related distracter *card* generated above, we created an unrelated distracter that was the same length, frequency, and grammatical class as *card,* but did not look like *card* (i.e., did not have orthographic similarity since that is what we were interested in evaluating). The word *card* is a noun, has a mean word frequency of 61 per million in English (Francis & Kucera, 1982) and is four letters long. Upon searching, we find that *lake* is also a four-letter noun with a frequency of 61. For that one distracter, we now have its unrelated control. Again, the logic is that if it takes a participant longer to reject the translation pair *cara-card* compared to the translation pair *cara-lake,* then we can infer that the orthographic similarity to the word *cara* is responsible for the interference. How do we know that? The series of matching procedures between each related and unrelated trial ensures that the only difference between the words *card* and *lake* is the orthographic similarity of *card* and *cara.* All 288 related distracters need an unrelated distracter matched on frequency, length, and grammatical class *within* each condition, not across all conditions. This would be virtually impossible. To ensure that the two types of distracters are matched in terms of frequency and length, simple *t*-tests are conducted between the related and unrelated distracters within each condition. However, since the matching of the related and unrelated distracters is done on a pair-by-pair basis within each condition, this has implications for data analysis which I will discuss later.

Design and Fillers

At this point in the design there were 576 distracters: 288 related and 288 unrelated. In other words, each of the 48 correct translation pairs now had 12 distracters associated with it. In designing the experiment, it is important that no participant see the same item twice. This would result in repetition priming. Thus, a Latin square design was used to distribute the different stimuli into six different lists, or versions, of the materials (See Jegerski, Chapter 2, this volume, for a similar procedure for counterbalancing stimuli). Each version of the materials ultimately had 96 critical items (48 related and 48 unrelated). All of these trials represent the "no" trials where the participants are rejecting the translation pair. Thus, 92 trials where the participants say "yes" must also be created. Since a balanced experiment will have participants responding "yes" and "no" an equal number of times, a set of filler trials had to be created. At first glance it may seem like any old fillers will do, but it is important to match the filler trials in terms of frequency and length with the correct trials, even though the correct trials are not seen. For example, if the critical trials are based on a correct pair like *cara-face,* then the filler trials should be matched as best as possible to the frequency and length of that pair. The logic of this matching procedure is to avoid a strategy adoption by the participants. If all of the filler trials were either more or less difficult than all of the critical trials, the participants may become aware of this and adopt a different strategy with the critical trials.

Procedure

Because this task only requires participants to read words on a screen and push a button, participants can be tested individually or in a group setting in a computer lab. Any typical experimental software for presenting stimuli can be used. As described earlier, each trial typically begins with a fixation point presented at the center of the screen. The participants begin the trial by pressing either a key on a button box connected to the computer or a key on the keyboard. The first word replaces the fixation point for a set period of time followed by a brief blank screen and then the second word typically appears in the same position. The presentation of the first word can vary from 240 ms to 500 ms. The second word remains on the screen until the participant presses either the "yes" or the "no" button or a set time has elapsed (500 ms for example). Participants are instructed to make their responses as quickly and accurately as possible, and, if they are unsure, to guess. The two measures in this task are reaction time (RT) and accuracy. The experimental software will record the RT to the nearest millisecond from the onset of the presentation of the second word. In other words, when the second word appears on the screen, the RT measurement begins, and once a key press or stroke is registered, the RT is registered. The experimental software also allows coding of what the correct response on any given trial should be. Thus, the moment the participant presses a key, the software will register whether the response was accurate or inaccurate. Any experimental software will also allow randomization of the presentation order of the word pairs so as to avoid a presentation order effect. Typically a few practice trials are given prior to the start of the experiment.

Scoring, Data Analysis, and Reporting Results

The most common method for analyzing data from a translation recognition experiment is a repeated measures ANOVA on the participant and item means. However, in order to be able to conduct this type of analysis, a series of steps must take place to sort and clean up the data appropriately.

First, any experimental software will produce a data file for each participant. The data files can be in a text or spreadsheet format. This file will contain at a minimum the participant number, the order of presentation, the trial number, the condition, the RT, and the accuracy of the trial. This is the critical information one needs to retrieve from the data file. There will likely be additional information in the data file, but it is often unnecessary for the purposes of data analysis. Therefore, it is useful to import the experimental data into a spreadsheet (if it is not already in that format) for every participant and then delete the information in the various columns that are not needed. In other words, keep the critical information (participant number, order of presentation, conditions, RTs, accuracy, etc.). Never alter the original data file produced by the experimental software, so as to always have a clean starting spot if a step is missed along the way.

Once every participant has a dedicated spreadsheet with the critical information from the experiment, one large spreadsheet will be created that essentially stacks the individual participant files one on top of the other. It is important that every individual spreadsheet be identical so that they can be stacked seamlessly. Of course, this will entail listing the participant number as many times as there are stimuli. For example, in the Sunderman & Kroll study described throughout the chapter, there were 192 trials in the experiment. The overall spreadsheet will have the following columns: *Participant, Item number, Condition, RT,* and *Accuracy.* The item numbers should range between 1 and 192 with a condition code and the corresponding RT and accuracy scores. Simply repeat the participant number in all of the cells in the participant column, 192 times. This is essentially linking every item with the participant number, which will be important when the RTs and Accuracy scores are aggregated over both participant and item. At this point, and prior to calculating means for conditions for participants and items, RTs that are outliers need to be removed. First, convert the excel file into an SPSS file. Typically, RTs that are faster than 300 ms or slower than 3000 ms are treated as outliers and removed from the analyses. In SPSS the RECODE feature can be used to complete this task. Data trimming is done in this way because it is typically thought that extremely fast scores reflect anticipatory processes, whereas extremely slow scores are due to lapses in attention or other processing strategies and, therefore, do not reflect the processes of interest (Ratcliff, 1993).

In the translation recognition task, participants either responded "yes" or "no." The response times for the "yes" and "no" trials are analyzed separately. The "yes" trials were the filler trials while the "no" trials were the critical manipulations. Moreover, only correct responses on trials are included in the RT analyses; the data from all critical trials will be included in the accuracy analysis though. Use the FILTER / SELECT CASES feature in SPSS to select the RTs for the correct trials separately for the "yes" and "no" trials for data analysis. Among the critical "no" trials, there were several conditions. It is necessary to find the mean and standard deviation per condition for each participant. In order to do this, use the AGGREGATE feature (break on Participant) in SPSS to calculate the mean and standard deviation for each participant by condition. Based on these values, RTs that are 2.5 standard deviations above or below the participant's mean RT for each of the overall conditions are excluded from the analyses and treated as outliers. The RECODE feature will be useful in eliminating outliers. The mean accuracy for "yes" and "no" trials by condition for each participant can then be calculated again, thus producing the subject analysis file. This process can be repeated for determining the item RT means by using the AGGREGATE feature except with item as the break variable instead of participants. The mean accuracy for "yes" and "no" trials for each item will be your item analysis file. Finally, using the MERGE feature in SPSS, any participant information such as proficiency level, age of acquisition, etcetera, can be included in the data files for analysis.

At this point, it is possible to analyze the data. Typically the descriptive statistics are presented in a table format, with the mean RTs and accuracy for each condition across participants. For example, recall in Sunderman & Kroll (2006), there were three different conditions, two form related and one semantically related. The distracters were either related or unrelated and were either matched on grammatical class or mismatched. In our study we were interested in L2 proficiency, so subjects were both more proficient and less proficient learners. Thus, the overall descriptive table contained the mean RT and accuracy for each of these factors (see Table 8.2).

With one glance, you can see how the less and more proficient learners in each condition performed on the related and unrelated items when they were matched or mismatched on grammatical class. In fact, to facilitate interpretation of the descriptive data, interference scores are calculated and listed. The magnitude of interference can be calculated for each type of distracter as the difference between the related and the unrelated trials. The difference scores give an indication of the processing cost associated with the form or meaning similarity.

Because of the way in which the stimuli were created, separate analyses are typically performed for each critical distracter type. Recall that the item-by-item matching of the related and unrelated distracters was only done within each condition, not across each condition. Typically, a mixed ANOVA is conducted. In

TABLE 8.2 Mean RTs (ms) and percent accuracy for translation recognition

	Proficiency							
	Less				*More*			
	Grammatical class							
	+		–		+		–	
	RT	Acc	RT	Acc	RT	Acc	RT	Acc
Orthographically Related to 1st word								
Related	1,039	77%	1,016	88%	935	87%	902	93%
Unrelated	995	87%	1,012	89%	888	95%	897	95%
Interference	44	10%	4	1%	47	8%	5	2%
Orthographically Related to 2nd word								
Related	1,027	85%	1,017	89%	902	91%	883	95%
Unrelated	941	89%	1,016	90%	901	95%	894	95%
Interference	86	4%	1	1%	1	4%	11	0%
Semantically related								
Related	1,066	73%	1,077	80%	965	83%	955	89%
Unrelated	979	88%	989	87%	879	95%	879	95%
Interference	87	15%	88	7%	86	12%	76	6%

Note: Interference is computed as the difference between related and unrelated conditions.

Sunderman & Kroll (2006) for example, we used participant means as random factors, and had one between-group factor (proficiency: less or more), and two within-group factors (related or unrelated and matched or unmatched grammatical class). The item analysis uses item means as random factors, with two between-group factors (related or unrelated and matched or unmatched grammatical class and one within-group factors (proficiency: less or more).

Reporting and interpreting the results of the analyses are fairly straightforward using this task. Increased RTs and lower accuracy scores indicate interference. The nature of the interference is based on the underlying activation in the participants' minds. The specific type of interference is related to the type of distracter, whether it was orthographic, phonologic, or semantic in nature. Often in translation recognition tasks, there are interesting interactions in the data. Less proficient learners seem to be affected by different types of form similarity, thus you may have an interaction between proficiency and relatedness, for example. Standard post-hoc tests can be performed to help interpret the interactions. Additional questions based on the properties of the stimuli can also be explored. For example, recall the norming study on semantic similarity described earlier. Based on the strength of the semantic similarity of the items, one can go back into the data and attach those values to the specific items, and then analyze whether the effects differ along the dimension of semantic similarity. In this way, the additional stimuli information allows for a more nuanced analysis of the data. In order to more fully understand and expand on the information presented here so far, it would be useful to consult the articles listed in the Suggested Readings section.

An Exemplary Study

The translation recognition task can be used to look at very specific aspects of bilingual memory, namely language activation in terms of orthography, phonology, morphology, and semantics. But perhaps scholars in SLA research might be more interested in the effects of proficiency on language activation. Or perhaps they are more interested in the role age of acquisition plays in underlying memory representation, or other individual differences. The translation recognition task can be used in a more global sense to ask questions not only about language activation, but also many other questions that may be of interest to SLA researchers. In this section, I review a study that uses a translation recognition task, but asks questions about the context of learning. Specifically, the study investigates whether cross-language interference (so typically found in the translation recognition task) can disappear in a study abroad context. It is important to show the versatility of the task to address a wide range of issues.

Linck, Kroll, and Sunderman (2009) had two groups of participants perform a translation recognition task. In fact, it was the same translation recognition task used in Sunderman and Kroll (2006) with the same materials, conditions, and so on. There was one group of learners, the *immersed group* that included 25 undergraduate students from an American university who were enrolled in a study abroad program. The *classroom group* included 20 undergraduate students from the same American university, but who enrolled in intermediate-level Spanish language courses and had no immersion experience. All participants were native speakers of English. The two groups were matched on three different proficiency measures and two cognitive ability measures that have been found to be related to on-line language processing, including working memory capacity (e.g., Miyake & Friedman, 1998) and the Simon task (e.g., Linck, Hoshino, & Kroll, 2008). In other words, the main difference between the two sets of learners was the immersion setting. One of the main questions of the study was: does an immersion setting allow a learner to turn off the L1 in a way that is not possible in the classroom?

Given that Sunderman and Kroll (2006) reported lexical interference (i.e., *cara-card*-type interference) for language learners of all proficiency levels, we hypothesized that the immersion setting would decrease the activation of the dominant L1 and perhaps attenuate the lexical interference effect. In other words, perhaps those learners in the immersion setting would not suffer interference from an orthographically related distracter since their L1 was sufficiently suppressed. Indeed, that is exactly what we found. In the translation recognition task the immersed learners showed no sensitivity to lexical form distracters, whereas the group of classroom learners with no immersion experience did, as can be seen in Figure 8.2.

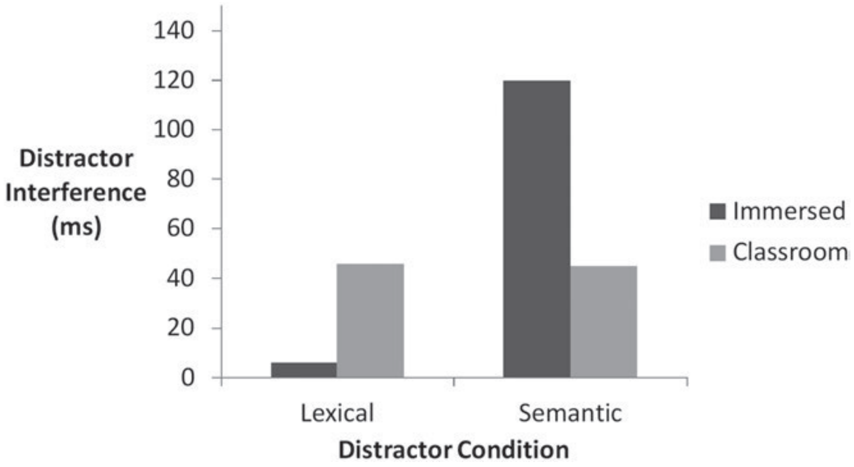


FIGURE 8.2 Degree of lexical and semantic interference (in ms) for immersed and classroom learners. Interference is calculated as mean related RT minus mean unrelated RT. (Adapted from Linck et al., 2009).

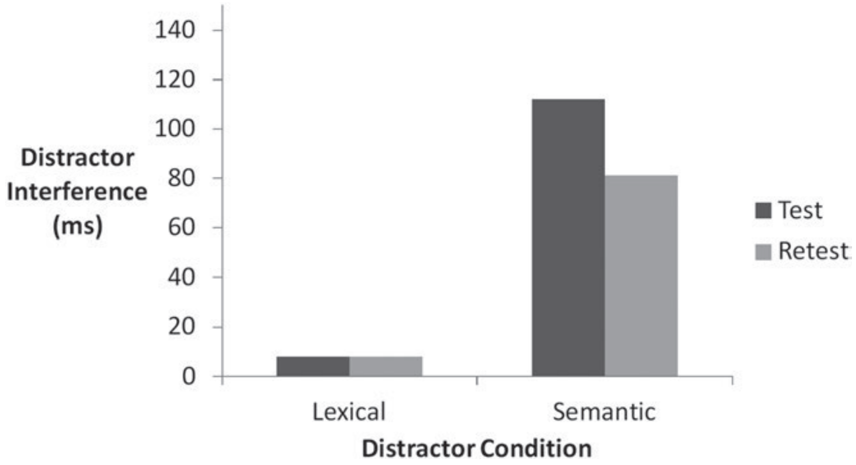


FIGURE 8.3 Degree of lexical and semantic interference (in ms) for the 14 immersed learners retested six months after returning home. Interference is calculated as mean related RT minus mean unrelated RT. (Adapted from Linck et al., 2009).

This result suggests that the L1 may have been inhibited within the immersion context. In contrast, semantic interference (i.e., *cara-head*-type interference) was far greater in the immersion setting as compared to the classroom setting.

In fact, even after retesting a subset of immersed learners on the same task six months after returning home, participants remained insensitive to the lexical form distracters, but remained sensitive to semantic distracters, as can be seen in Figure 8.3.

We argued that the L1 inhibition within the immersion environment may have allowed stronger lexical-conceptual links to develop, making the learners more resistant to L1 lexical competition during translation recognition, even upon returning home to the L1-dominant environment.

This study is an example of how the translation recognition task is being used to answer questions about language inhibition. In fact, although the focus in the previous section was on the development of specific materials in a range of empirical studies that used translation recognition tasks, those studies also investigated more global issues such as the effects of script (e.g., Sunderman & Kroll, 2006; Quasem & Foote, 2010), age of acquisition (e.g., Ferré, Sánchez-Casas, & Gausch, 2006), and proficiency (e.g., Ferré, Sánchez-Casas, & Gausch, 2006; Gausch, Sánchez-Casas, Ferré & García-Alba, 2008; Sunderman & Kroll, 2006; Quasem & Foote, 2010), thus highlighting the strength of this task. In the end, the translation recognition task is an incredibly useful tool for scholars interested in examining the psycholinguistic processes that support lexical processing in language learners and other bilinguals.

Pros and Cons in Using the Method

Pros

- Collecting data is both economical and efficient. Many free software programs (e.g., DMDX by Forster & Forster, 1999; PsyScope by Cohen, MacWhinney, Flatt, & Provost, 1993) or inexpensive software programs (e.g., E-Prime by Schneider, Eschmann, & Zuccolotto, 2002; SuperLab by Cedrus Corporation, 1992) are available to program the experiments on laptops or desktop computers. Additional equipment is not necessarily required, although a response pad or button box can be used. Once the experiment is programmed, data can be collected either in an individual setting or in a computer laboratory with multiple participants at one time.
- The data are very straightforward and immediately accessible upon completion of the task. The experimental software will produce a data file indicating the response times and accuracy on every trial. There is no additional coding of data or listening to recordings (which is often the case in production-based experiments). Moreover, the nature of the data is transparent. Increased reaction times and lower accuracy indicate processing difficulties or interference.
- Translation recognition tasks are useful for investigating developmental aspects of the lexicon since they can be used with second language learners with low levels of L2 proficiency, unlike production tasks like translation or picture naming. In other words, translation recognition tasks tap into the translation processes without requiring overt production, something that can be challenging for learners. In fact, participants have often reported that the translation

recognition task is fast and painless (which is quite the opposite of what they say about picture naming).

- The translation recognition task is versatile, in that is can be used to explore a variety of specific questions related to the underlying lexical representation in the lexicon (e.g., its orthography, morphology, semantics, etc.) or more global questions about the interaction between the architecture of the lexicon and factors such as age of acquisition, proficiency, and context of learning. The fact that the translation recognition task allows us to investigate these questions in individuals with lower levels of proficiency is of particular interest to those working in the field of second language acquisition.

Cons

- Perhaps the most salient drawback is related to the creation of the stimuli. The creation of the stimuli is often a painstaking process requiring close and careful attention to numerous lexical features such as length, frequency, cognate status, and so on. Multiple norming studies are often required to assess levels of similarity in the materials (whether orthographic, semantic, etc.) This process is necessary, but can be time-consuming to do properly.
- Although the translation recognition task mimics the underlying process of translation, it does not capture the actual production process. Thus, some may be more interested in the overt production of language as opposed to the preproduction phase.
- Translation recognition necessarily entails the activation of both languages. Those interested in investigating more covert activation may prefer using methodologies like masked-priming instead.

Discussion and Practice

Questions

1) What is the logic of the distracters in the translation recognition task? What is the logic of the unrelated distracter? What issues do you need to consider when developing distracters for a translation recognition experiment?
2) Create a set of related and unrelated distracters based on one of the manipulations described in this chapter for three different correct translation pairs. What difficulties did you encounter? How long might it take you to create stimuli for an entire experiment?
3) Select one of the lexical databases described in the "Issues in the Development and Presentation of Stimuli" section of this chapter and explore the lexical stimuli characteristics available in the resources. List five factors that you find intriguing and describe potential experimental manipulations based on those factors.

TABLE 8.3 Illustration of materials used in each condition for the pair CARA-FACE and its translation

	Form conditions		*Meaning condition*
Grammatical class	*Orthographically related to 1st word*	*Orthographically related to 2nd word*	*Semantically related*
+	Card / **tarjeta**	Fact / **hecho**	Head /**cabeza**
–	Care/ **cuida**	Fast/ **rápido**	Pretty / **bonita**

4) What are some individual differences that could affect performance on a translation recognition task? How would one go about testing those differences?
5) Consider the stimuli presented in Table 8.3 above modified from Sunderman and Kroll (2006). If one wanted to use the materials from this task but conduct the experiment in the L1-L2 direction, would it be acceptable to translate the materials? Why or why not? The translated materials are in **bold** below.

Research Project Option A

Laxén and Lavaur (2010) conducted a translation recognition task with highly skilled French-English bilinguals, investigating the notion of translation ambiguity. Translation ambiguity arises when some words have more than one translation between languages, thus causing a mapping problem. How does the right word get mapped on to the right meaning? This study was not developmental in nature. In other words, it did not focus on learners from differing proficiency levels. One potential project would be to replicate this study using learners with varying proficiency levels in order to assess whether translation ambiguity has more or less of an effect on learners. As an additional point of interest, this study used the translation recognition task in a slightly different way and does not create "no" distracters, but instead is interested in the "yes" conditions. It is yet another way the translation recognition task can be used.

Research Project Option B

Conducting a comparative norming study would be a useful research project. One could select a set of words and gather ratings on those words from one of the lexical databases described in the "Issues in the Development and Presentation of Stimuli" section of this chapter, and then also gather original ratings on the same items from second language learners or bilinguals. How would the two sets of ratings compare? Would they be correlated? In what ways might the two sets of ratings diverge? For example, are high-frequency words based on corpus samples also high-frequency words for beginning language learners? What implications does this have for research with second language learners?

Suggested Readings

Ferré, P., Sánchez-Casas, R., & Guasch, M. (2006). Can a horse be a donkey? Semantic and form interference effects in translation recognition in early and late proficient and non-proficient Spanish-Catalan bilinguals. *Language Learning, 56,* 571–608.

Sunderman, G., & Kroll, J.F. (2006). First language activation during second language lexical processing: An investigation of lexical form meaning and grammatical class. *Studies in Second Language Acquisition, 28,* 387–422.

Talamas, A., Kroll, J. F., & Dufour, R. (1999). Form related errors in second language learning: A preliminary stage in the acquisition of L2 vocabulary. *Bilingualism: Language and Cognition, 2,* 45–58.

Qasem, M., & Foote, R. (2010) Cross language lexical activation: A Test of the Revised Hierarchical and Morphological Decomposition Models in Arabic-English Bilinguals. *Studies in Second Language Acquisition, 32,* 111–140.

References

Altarriba, J., & Mathis, K. M. (1997). Conceptual and lexical development in second language acquisition. *Journal of Memory and Language, 36,* 550–568.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39,* 445–459.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database [CD-ROM]. University of Pennsylvania, Philadelphia, PA: Linguistic Data Consortium.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments, & Computers, 25,* 257–271.

de Groot, A.M. B. (1992). Determinants of word translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1001–1018.

de Groot, A.M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals: An Introduction.* New York: Taylor and Francis.

de Groot, A.M. B., & Comijs, H. (1995). Translation recognition and translation production: Comparing a new and an old tool in the study of bilingualism. *Language Learning, 45,* 467–510.

de Groot, A.M. B., Dannenburg, L., & van Hell, J. G. (1994). Forward and backward word translation by bilinguals. *Journal of Memory and Language, 33,* 600–629.

Dijkstra, A., & Van Heuven, W. J. B. (1998). The BIA model and bilingual word recognition. In J. Grainger & A. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 189–225). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ferré, P., Sánchez-Casas, R., & Guasch, M. (2006). Can a horse be a donkey? Semantic and form interference effects in translation recognition in early and late proficient and nonproficient Spanish-Catalan bilinguals. *Language Learning, 56,* 571–608.

Forster, K. I., & Forster, J. C. (1999). DMDX [Computer software]. Tucson, AZ: University of Arizona/Department of Psychology.

Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage.* Providence, RI: Brown University Press.

Frost, R., Forster, K. I., & Deutsch, A. (1997). What can we learn from the morphology of Hebrew: A masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 23*, 829-856.

Guasch, M., Sánchez-Casas, R., Ferré, P., & García-Albea, J. E. (2008). Translation performance of beginning, intermediate and proficient Spanish–Catalan bilinguals: Effects of form and semantic relations. *The Mental Lexicon, 3*, 289–308.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitken, R. W. Baileu, & N. Hamilton-Smith (Eds.), *The Computer and Literary Studies.* Edinburgh, UK: University Press.

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language, 33,* 149–174.

Laxén, J., & Lavaur, J. (2010). The role of semantics in translation recognition: effects of number of translations, dominance of translations and semantic relatedness of multiple translations. *Bilingualism: Language and Cognition, 13,* 157–183.

Linck, J. A., Hoshino, N., & Kroll, J. F. (2008). Cross-language lexical processes and inhibitory control. *Mental Lexicon, 3,* 349–374.

Linck, J., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second language learning. *Psychological Science, 20,* 1507–1515.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahway, NJ: Lawrence Erlbaum Associates.

Noble, C. (1953). The meaning-familiarity relationship. *Psychological Review, 60,* 89–98.

O'Grady, W., Archibald, J. Aronoff, M., & Rees-Miller, J. (2010). *Contemporary Linguistics.* (6th Ed.). New York: Bedford/St. Martin's Press.

Paivio, A. (1986). *Mental representations: A dual coding approach.* New York: Oxford University Press.

Potter, M. C., So, K.-F., Von Eckhardt, B., & Feldman, L. B. (1984). Lexical and conceptual representation in beginning and more proficient bilinguals. *Journal of Verbal Learning and Verbal Behavior, 23,* 23–38.

Princeton University. (2010). *WordNet: A lexical database for English* [Online Database]. Retrieved from http://wordnet.princeton.edu

Qasem, M., & Foote, R. (2010) Cross language lexical activation: A Test of the Revised Hierarchical and Morphological Decomposition Models in Arabic-English Bilinguals. *Studies in Second Language Acquisition, 32,* 111–140.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). E-Prime [Computer software]. Pittsburgh, PA: Psychology Software Tools Inc.

Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language, 27,* 499–520.

Sunderman, G., & Kroll, J. F. (2006). First language activation during second language lexical processing: An investigation of lexical form meaning and grammatical class. *Studies in Second Language Acquisition, 28,* 387–422.

Sunderman, G. L., & Priya, K. (2012). Translation recognition in highly proficient Hindi–English bilinguals: The influence of different scripts but connectable phonologies. *Language and Cognitive Processes, 27,* 1265–1285.

SuperLab (Version 4.0.5) [Computer software]. (1992). San Pedro, CA: Cedrus Corporation.

Talamas, A., Kroll, J. F., & Dufour, R. (1999). Form related errors in second language learning: A preliminary stage in the acquisition of L2 vocabulary. *Bilingualism: Language and Cognition, 2,* 45–58.

Tokowicz, N., & Kroll, J. F. (2007). Number of meanings and concreteness: Consequences of ambiguity within and across languages. *Language and Cognitive Processes, 22,* 727–779.

Tokowicz, N., Kroll, J. F., de Groot, A.M. B., & van Hell, J. G. (2002). Number of translation norms for Dutch–English translation pairs: A new tool for examining language production. *Behavior Research Methods, Instruments and Computers, 34,* 435–451.

van Orden, G. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory and Cognition, 15,* 181–198.

Wilson, M. D. (1988) The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavioural Research Methods, Instruments, and Computers, 20*, 6–11.

9

CROSS-MODAL PRIMING WITH SENTENCES

Leah Roberts

History of the Method

The benefit of using on-line methods to investigate sentence comprehension is that they can tap into the moment-by-moment automatic processing of the linguistic input. The best of these methods are those in which there is as little interference from the task as possible, that is, tasks which do not require from participants conscious problem solving or metalinguistic responses. Cross-modal priming (CMP) is arguably just such a technique. The method relies on *priming;* evidence for which comes from the observation that the processing of visual stimuli (a *target probe)* is faster if the probe immediately follows the presentation of an identical or semantically related word or *prime* (e.g., Meyer, Schvaneveldt, & Ruddy, 1975; Neely, 1991). The facilitated processing of the target probe is caused by the fact that its level of activation in the mind has been raised by the prior processing of the prime. CMP is a dual task: participants listen to a sentence for comprehension, during which the target probe—a word or a picture—appears on a computer screen at some critical point and the participant must make a response (lexical, semantic) to this probe as quickly as possible. The required response may be to name the picture or word, or to make a binary decision about it, for instance in *lexical decision* where targets are words or nonwords (e.g., *Is this a real word?*), or in semantic categorization tasks (e.g., *Is this alive or not alive?*). One of the major advantages of this technique is that the experimenter can choose where to position the critical target probe in the sentence, and can chart the rising and falling activation levels of a lexical item over the course of a sentence. In other words, since response times (RTs) between target and nontarget (*unrelated* or *control*) probes are compared, and given that faster RTs reflect greater activation levels of the critical stimulus, the researcher can examine what linguistic items are more or less activated in the comprehender's mind at specific points during the processing of an uninterrupted input stream. In the field of sentence comprehension, this task has been used to address theoretical questions relating to the organization of the mental lexicon and lexical processing in sentence contexts (e.g., Swinney, 1979; Swinney, Onifer, Prather, & Hirshkowitz, 1979) and to the establishment of syntactic dependencies during real-time sentence comprehension (e.g., Nicol, 1993; Roberts, Marinis, Felser, & Clahsen, 2007).

The first use of the CMP method is found in the work of David Swinney and colleagues (e.g., Swinney, 1979; Swinney, Onifer, Prather, & Hirshkowitz, 1979), who devised the task in response to the need to examine the nature of *semantic* priming in more detail. Semantic priming is evidenced in studies where there is facilitated processing of a semantically related probe word (e.g., TULIP, if preceded by the prime *poppy*[1]), which is assumed to be caused by the activation spreading from the mental representation of the prime (*poppy*) to (semantically, conceptually, and phonologically) related nodes in the mental lexicon. Although semantic priming had been assumed to be a major component of the comprehension process, it was as yet unclear how it might interact with other semantic/syntactic processes to produce the observed effects. Up until this point, semantic priming had been investigated almost wholly using unimodal (visual) lexical decision tasks. Swinney and colleagues asked whether it would be obtained in cross-modality contexts, an issue which speaks to the question of whether semantic priming is a central or a modality-specific process. The researchers also wished to determine whether the lexical and semantic processes at work during real-time sentence comprehension could be measured by use of the cross-modal task. Priming effects were indeed observed when sentences were presented aurally with visual probes. The effect was strongest when the presentation of the visual target for lexical decision immediately followed a semantically related word and the facilitation effect persisted up to three syllables downstream of the prime. The authors also found that the priming effect lasted across clause boundaries. This was a rather surprising finding, given a major assumption of sentence processing theory at the time was that once a clausal unit is analyzed, it is removed from working memory (Caplan, 1972).

As well as being used successfully in the study of lexical and grammatical on-line processing with adults, the CMP method with pictures (*cross-modal picture priming,* CMPP), requiring no reading or metalinguistic judgments, is highly useful for preliterate children (Love & Swinney, 1997; McKee, Nicol, & McDaniel, 1993; Roberts, et. al., 2007) and has even been used with children with Specific Language Impairment (SLI) (Marinis & Van der Lely 2007).[2] As regards second language learners, CMP with sentences has not been employed as extensively as other on-line techniques. However, it is hoped that the discussion of some of the studies in this chapter will demonstrate that the method has a lot of potential to address interesting theoretical questions in second language acquisition (SLA) research, both in the area of the mental lexicon (see e.g., Brien & Sabourin, 2012) and of sentence processing (see e.g., Felser & Roberts, 2007), and with beginning or (as yet) nonliterate second language (L2) learners as well as those who are more advanced.

What is Looked at and Measured

Lexical Processing in Sentences

Many researchers have used the CMP task to look at the processes of lexical access and selection, in particular the potential biasing effects of prior semantic context on the real-time comprehension of words with multiple meanings (e.g., *bug, bank*). Before the use of this method by Swinney (1979), researchers had employed *probe recognition* and *phoneme monitoring* to investigate questions relating to the processes of lexical access as well as anaphoric relations. The probe recognition method involves the presentation of a probe at the end of a stimulus sentence, and for phoneme monitoring, comprehenders must attend to the input and indicate the appearance of a phoneme. In both techniques, the critical decision which provides the dependent measure takes place sometime after the presentation of the critical word under investigation (or even at the end of the sentence, in the case of probe recognition). Therefore, as Swinney (1979) noted, such tasks cannot be maximally informative about the process of lexical access itself, but rather can tell us something about postaccess processes. Swinney and colleagues devised the CMP task to get around such a problem. In Swinney (1979), for instance, the participants heard texts containing ambiguous words (e.g., *bugs* in [1]) and were asked to make lexical decisions simultaneously on visually presented probe words that were either semantically related to one or more of the meanings of the ambiguous word (SPY, ANT), or semantically unrelated (SEW).

(1) Rumor had it that, for years, the government building had been plagued with problems. The man was not surprised when he found several **bugs #** in the corner of his room.

The word requiring the lexical decision was presented immediately at the offset of the ambiguous word (*bugs*), thus overcoming the problems of the late presentation of the target found with probe recognition and phoneme monitoring. In this pivotal study, Swinney found that participants' RTs were faster to items that were semantically related to both meanings of the ambiguous item (e.g., *bugs*), and there was no difference in RTs for these two meanings of the word. Thus, in neutral sentence contexts (as in [1]), all possible meanings of a word appear to be activated. However, this may be the case only for the very initial stages of lexical processing, because when the prior context was semantically biased towards one of the meanings (e.g., *The man was not surprised when he found spiders, roaches, and other bugs*), the authors found that downstream of the critical position, the facilitation effect was only in evidence for the semantically appropriate target probe word (ANT). Thus, at the earliest stages, all meanings of a lexical item may be activated, and semantic context effects come into play at later stages of lexical processing, at the selection phase. It is uncommon, however, for multiple meanings of a word to be balanced in terms of frequency, and research has shown that in semantically neutral contexts, the dominant meaning of a lexical item will be activated at an earlier point in time than any subordinate meanings (e.g., Duffy, Morris, & Rayner, 1988). This is an important factor to take into account when selecting probes for a CMP experiment, as will be explained in greater detail below.

In sum, the CMP technique can be used to investigate theoretical questions relating to lexical access and word recognition processes during the automatic and unconscious comprehension of the input, and the results from CMP studies are therefore able to inform models of lexical processing and the mental lexicon.

Grammatical Processing

As well as addressing questions relating to the structure of the mental lexicon and word recognition processes in sentence and discourse contexts, the CMP technique has been described as an ideal method to investigate the dependencies between constituents in a sentence (Love & Swinney, 1996; Swinney, 1979). This is because it is possible to look at *re*activation of previously processed elements, rather than merely continued activation. That is, experimenters can map out the changing activation levels of constituents throughout the processing of the sentence.[3] CMP has been particularly useful in the study of the processing of moved elements, such as how the processor establishes the link between a fronted *wh*- item and its subcategorizer, as in *wh*- questions or relative clauses. Other topics investigated include the processes underlying the linking of a pronoun or other anaphor to its antecedent in the earlier discourse. Such work on grammatical processing has been of interest to both linguists and psycholinguists, and many early studies of sentence comprehension aimed to test whether grammatical constraints described by formal linguists were put to use during sentence processing. For instance, Nicol (1988) asked whether the processor implemented *binding constraints* (e.g., Chomsky, 1981) during the on-line computation of structural dependencies. She looked at the processing of reflexive (2a) versus object pronouns (2b) and their antecedents and compared lexical decision RTs to probes identical to the target noun phrases (NPs) (e.g., BOXER, SKIER, DOCTOR) presented at the off-set of the pronominal (*himself* or *him*). In the reflexive condition (2a), she found that only the processing of the binding-relevant antecedent (DOCTOR) was facilitated, when RTs were compared to the other NPs in the sentence (BOXER, SKIER):

(2)

a. The boxer told the skier that the **doctor**$_i$ for the team would blame **himself**$_i$ for the recent injury.

b. The **boxer**$_i$ told the **skier**$_j$ that the doctor for the team would blame **him**$_{i/j}$ for the recent injury.

One possibility for the finding of facilitated processing for the NP probe DOCTOR in (2a) is that it is the closest to the reflexive in linear terms, and therefore the most recent NP to be processed. This ease of processing could therefore be caused by continued activation, rather than reactivation of the critical NP at the grammatically relevant position in the sentence. However, this explanation is unlikely given the findings for the object pronoun constructions like (2b), in which activation levels were excited only by the grammatically appropriate, although linearly more distant, NPs *the boxer* and *the skier.*

Such data suggest that comprehenders respect grammatical rules during on-line processing, and similar findings are reported for languages other than English, as well as in the study of the processing of sentences with purported empty elements such as traces and null pronouns (e.g., Nakano, Felser, & Clahsen, 2002; Clahsen & Featherston, 1999). Another question that has recently been addressed by use of this technique and which is of importance to sentence processing researchers as well as those interested in grammatical theory is what specifically is reactivated at the elision (or *gap*) site during the processing of sluicing (3) and verb phrase (VP)-ellipsis (4) constructions (e.g., Poirier, Wolfinger, Spellman, & Shapiro, 2010; Shapiro & Hestvik, 1995; Shapiro, Hestvik, Lesan, & Garcia, 2003). The comprehension of this type of construction involves filling in the missing part of the sentence (e.g., in [4], *did Ø too*), established from the contents of the antecedent context.

(3) *Sluicing*

The handyman threw a book to the programmer but I don't know which book [~~the handyman threw to the programmer~~] and no one else seems to know.

(4) *VP-ellipsis*

The mailman [bought a new tie] and the fireman did Ø too.

Evidence, mainly from CMP experiments, shows that as soon as the processor encounters the elided site, a link is established between this site and the antecedent clause, and the material required to resolve the ellipsis is recovered (e.g., Shapiro et al., 2003). One of the enormous benefits of the CMP method is illustrated by this type of study: one is able to test exactly which parts of the antecedent clause are reactivated at the elision site. Shapiro et al. tested whether processing of both the subject (*the mailman*) and the object (*the tie*) would be facilitated at the elision site. The authors found that the only NP that was primed was that contained in the VP; that is, probes that were semantically related to the direct object were primed (e.g., *tie*) but not those related to the subject NP (e.g., *mailman*). This result suggests that the processor reactivates only structurally relevant material at the elision site, going against theories that propose that all elements in the antecedent clause may be reactivated (e.g., Koeneman et al. 1998), and in line with evidence from the study of anaphor resolution noted above, in which the pronominal links only to the grammatically relevant antecedent in the preceding discourse (Nicol, 1988). Given that CMP studies have provided evidence that syntactic constraints may guide the resolution process such that only the appropriate antecedent is linked with the referring expression or elision site, such questions are of great importance to SLA researchers interested in the acquisition of grammar, as well as those interested in L2 sentence processing. Running similar studies with L2 learners could inform debates in the field, such as whether or not L2 learners beyond puberty are able to make use of abstract grammatical categories in on-line comprehension (c.f., the Shallow Structure Hypothesis, Clahsen & Felser, 2006; see Roberts, 2013, for a summary of this debate).

Issues in the Development and Presentation of Stimuli

Hardware and Software

The CMP technique requires a computer (a PC or laptop) on which to present the visual stimuli (either pictures or words) and the means to present the sentences auditorily. Software packages designed for psychological experiments such as E-Prime (Psychology Software Tools) and Presentation (Neurobehavioral Systems) are often used. The researcher can use such programs to present the experimental sentences via headphones, and to ensure that the visual prime (whether a picture or word) is precisely timed to coincide with the point in the presentation of the auditory sentence that is critical and/or theoretically important for the experiment (e.g., 500 ms from the offset of a particular word). These experimental packages also record participants' responses to the probe word or picture, either via buttons on a keyboard, or via a separately purchased push-button box. Therefore, both the speed of the response and the accuracy scores on the lexical/semantic decision as well as on any comprehension questions that may be presented throughout the experiment can be recorded, usually in a format that is compatible with statistical analysis (e.g., as a text file that can be imported into SPSS).

A Typical Trial

A typical trial in a CMP experiment involves the participant listening via headphones to a sentence, while seated in front of a computer monitor or laptop screen. The participant is told to listen to the sentence and to try to understand it. Before the sentence comes to an end, they will see a visual probe (a picture or word) on the monitor in front of them, and they must make a response to this probe as quickly as possible. The response can be a lexical decision (*Is this a real word?*) or a binary categorization response (*Is this alive or not alive?*). The picture or word probe will most often appear for an amount of time specified by the experimenter (e.g., 500 ms or 1 second). Once the response is made and the auditory input comes to an end, there might be a comprehension question that the participant hears, to which a yes/no response must be made. No feedback is usually given to the participant. Generally, the experimenter programs the experiment so that there is a visual indication on the screen that the trial has ended, and the participant can push a button (or the space bar, for instance) when they are ready for the next trial to begin.

Stimulus Sentences

As noted above, most experiments involve the participants' listening to sentences presented via headphones, and making a lexical or semantic decision to a critical target (either a picture or a word) presented on a screen in front of them at some critical point during its presentation. Therefore, the list of critical stimuli sentences needs to be constructed, recorded, and digitized so that they can be used with the relevant experimental software.

The critical experimental sentences should be of the same structural type (e.g., all are relative clauses with three-place predicates) and they should be matched in length as far as possible (number of words and syllables). Like most on-line sentence processing studies, the critical experimental sentences are mixed amongst fillers of various structural types (including some of the same type as the experimental items) to ensure that the participants cannot guess the aim of the study. The critical sentences should also be as semantically neutral as possible; that is, the experimenter aims to construct sentences that contain no material (other than the prime) that would semantically bias a participant's response towards the target or control probes. To ensure this as best as possible, it is suggested that there be a norming procedure in which the items are assessed for biases by naive native speakers who will participate no further in the experiment (see Sunderman, Chapter 8, this volume, for a more detailed discussion of norming).

The ratio of experimental items to fillers depends on the number of experimental variables (with more conditions, the greater the number of both experimental and filler sentences), but as a general rule of thumb, a ratio of 1:3 is often used. If a study involves four conditions, for instance, four experimental lists would be created, each with the same set of perhaps 24 critical sentences and 72 fillers. Each critical sentence contains one word which is the prime and which is in the same structural position in all sentences (e.g., a *wh-* item, such as *which doctor,* in the fronted position). Identical (DOCTOR) or semantically related (SURGEON) target probes should elicit faster response times in comparison to the control probe, which can be a picture of an unrelated item (e.g., CARROT), or, in cross-modal lexical priming (CMLP), a pseudoword.[4] The critical sentence would be presented in one condition (and therefore in one stimulus list) with the target probe in the critical position (e.g., at the purported gap site or other theoretically important position, [5a]), and in a second condition, the probe appears in a control position (e.g., at the onset of the verb, or 500 ms earlier than the target, etc., [5b]). The two further conditions would involve the control/unrelated probe (e.g., a semantically unrelated item, CARROT) being presented during this same sentence. In one condition, it would appear in the critical position (e.g., the gap site [5c]) and in the other condition, it would be presented in the control position (5d). These positions in the experimental sentences remain the same across the whole set of experimental conditions, to ensure that reliable comparisons can be made. This is illustrated below.

(5)

a. *Identical probe, test position*

To which doctor did the woman give the present to [DOCTOR] in the hospital last week?

b. *Identical probe, control position*

To which doctor did the woman give the [DOCTOR] present to in the hospital last week?

c. *Unrelated probe, test position*

To which doctor did the woman give the present to [CARROT] in the hospital last week?

d. *Unrelated probe, control position*

To which doctor did the woman give the [CARROT] present to in the hospital last week?

Probes to which participants make the same type of lexical or semantic decisions should also appear during the presentation of all of the filler sentences, but at various different points during the sentences. This is to avoid the possibility of the participant anticipating the probe position, and again, to hide the experimental sentences. Given that CMP is a dual task, it is strongly advised to present a set of fillers as practice items at the beginning of the experimental list, to allow participants time to familiarize themselves with the task.

It is also a good idea to present yes/no comprehension questions after some or all of the experimental items and fillers. As with other on-line methods, this is an attempt to ensure that participants are paying attention to the meaning of the sentence, rather than employing strategies to do the task. This is particularly important for CMP, though, because it can deter the participants from merely waiting for the presentation of the probe without processing the critical experimental sentences, or focusing their attention solely on the lexical/semantic decision to be made with the button-push. Any errors in answering the comprehension questions are then taken as an indication of the extent to which the participants could or would do the task and whether or not they were paying attention.

Primes and Probes

As mentioned above, attempting to stop participants from integrating the probe into the sentence as it is being comprehended, and to avoid, as far as possible, any effects of end-of-sentence wrap-up processing, target and control probes should always be presented before the end of the sentence.

In studies using *repetition priming,* the target probe is identical to the prime (DOCTOR), whereas in other studies the target probe is semantically related (e.g., SURGEON), or associatively related (e.g., HOSPITAL). It has been argued that using identical primes and probes is the most direct way to tap into lexical activation (e.g., Clahsen & Featherston, 1999). There is little doubt that there is a measurable benefit from processing a target probe that is identical to the prime, and this could be an argument against using identical probes rather than semantically or associatively related items. However, because one looks at the facilitation of RTs to the identical probe in the test position in comparison to the control position, it can be argued that any general facilitation gained from processing the same lexical item is cancelled out. Thus it may indeed be a more direct measure of activation of the critical item than in studies using semantically related probes.

Both primes and probes require extremely careful selection. This is because lexical processing is highly sensitive to various factors including frequency, (semantic, phonological, orthographic) neighborhood density,[5] meaning dominance, length, and complexity. For instance, as noted earlier, activation occurs earlier for items that are shorter in length or more frequent, and for the dominant (and thus usually more frequent) meaning of a word with multiple meanings. To control for frequency effects, researchers often make use of resources such as COBUILD (Sinclair, 1987) or CELEX (Baayen, Piepenbrock, & Gulikers, 1995), which allow them to compare log frequencies for target and control probes. Ideally, one needs to be able to statistically compare log frequencies, and find that there is no significant difference among the probe items.

Probe items that are semantically related to the primes need to be tested independently to ensure that they are indeed semantically related to the prime. This is best done with piloting of the items with naive native speakers. For example, Shapiro et al. asked a group of participants (who participated no further in their experiment) to rate their target probes on a scale of semantic-relatedness to their prime items and selected only those items which fulfilled their criteria. This is of course a step one need not take if using identical, rather than (e.g., semantically) related probe items.

If participants need to make a category-based decision on, for instance, a picture target (e.g., *Is this alive or not alive?*), one needs to ensure that the probes used are clearly and unambiguously a member of one or other of the categories. Again, a group of naive native speakers is ideal to pretest the items. For instance, the researcher can ask the naive native speakers to judge whether the pictures are *alive* or *not alive,* and then select for use in the main experiment only those items that elicited a clear decision for all (or at least the majority) of the group.

In CMLP, the lexical decision task requires that half of the filler probes be pseudowords, and these should also be tested by a group of naive native speakers, to check that they are indeed pronounceable nonwords.

Training

As mentioned above, since CMP is a dual task, it is recommended that participants are presented with practice trials before the experiment begins. This is important for all participants, but especially so if one is testing children (e.g., Roberts et al., 2007).

Scoring, Data Analysis, and Reporting Results

Scoring

Two main types of data are elicited during a CMP study, which are the responses to the target and control probes in both speed (in milliseconds) and accuracy (for, for instance, the lexical decision). The dependent variable of greatest interest to the researcher is the speed of the participant's response to the target and control probes. That is, faster responses to the target probe as compared to those to the control probe is a measure of activation level of the critical item. Errors are taken as a reflection of a performance failure, or a lack of attention during the task, therefore, one should remove items which elicited an erroneous response from the RT analysis. Items for which the participant gave an erroneous response to the end-of-trial comprehension questions should also be removed. If both related and unrelated/control probes are balanced (as checked for in pretests), and sentence bias is controlled for, then accuracy for either responses to the probes or to the end-of-sentence comprehension questions should not differ across conditions.

Data Analysis

Once the data have been cleaned for performance errors, one is left with one response time per item per subject. As is the case for most behavioral experiments that elicit reaction times, one should screen the data for outliers before performing statistical analyses on aggregate means. As a general rule, extremely high RTs (for instance, over 2 seconds) are removed, and these are usually very few. After this, further screening usually entails computing the means per subject and item, and then the removal of individual reaction times that fall above two standard deviations from that mean.

Traditionally, ANOVAs and planned comparisons are then run on the means, separately for subjects and for items, created from these cleaned data. Such ANOVAs are usually 2 × 2, comprising two within-subject variables, each with two levels: Probe Type (related/unrelated) × Position (test/control). Priming is in evidence only if RTs to the related probe are faster in the test position both (a) in comparison to the unrelated probe in the same (test) position, and (b) in comparison to the related probe in the control position. Statistically speaking, this entails that there be an interaction between robe Type and Position. A main effect of Probe Type, if caused by faster RTs for the related probe in both positions, would reflect maintained activation of the prime, rather than *re*activation of the critical item in the (theoretically) critical (test) position.

More recently, researchers have been using mixed-effects models (Baayen, Davidson, & Bates, 2008), and this removes the need to clean outliers above two standard deviations or to create means tables. Either method of analysis will prevent the problem of the means potentially being skewed by outlying data points.

Reporting Results

Before reporting the results of the experiment proper, the norming procedures that need to be performed on the experimental probes must be described and the results of these reported in detail. Specifically, it is crucial that other researchers can see the procedure of each pretest clearly, and that the selection of experimental items is shown to be rigorous. This is important because lexical activation and selection is strongly influenced by a number of different factors, for instance frequency and length, and it should be clear that any observed facilitation effects are caused by priming (or not) of the items of interest.

As regards reporting the results of the experiment itself, as stated earlier, the critical measure of activation of the critical word/concept is that the participant was faster to make their (lexical decision/categorization/naming) response to the related probe in the critical position, in comparison to the other three conditions (related probe in control position, unrelated probe in both critical and control positions). Therefore, researchers often present the data in charts, as seen in Figure 9.1,

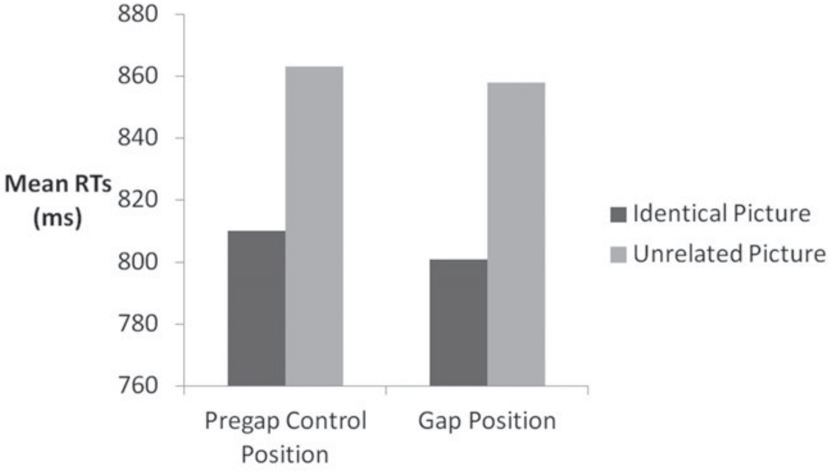


FIGURE 9.1 Example of a results graph with L2 learners' mean RTs in ms to picture targets (adapted from Felser & Roberts, 2007). This graph illustrates continued activation of the prime word, regardless of position in the stimulus sentence. This outcome was different from that of the L1 participants from Roberts et al. (2007), found in Table 9.1, where a high working memory span L1 group had shown activation of the prime word only at the gap position and a low working memory span L1 group had shown no activation of the prime word at either site (the L2 results did not vary according to working memory span).

TABLE 9.1 Example of a results table with the advantage (in ms) for related versus unrelated probes (adapted from Roberts et al., 2007)

	High memory span L1 adults	*Low memory span L1 adults*
Pregap position	-2	19
Gap position	31	-5

Note: The advantage for related picture probes was calculated by subtracting the mean RT for a related probe from a mean RT for an unrelated probe. The high working memory span L1 adults showed selective activation of the prime word only at the gap site, while the low working memory span L1 adults showed no activation of the prime word at either position in the stimulus sentence.

or tables with RTs to unrelated probes minus those to related probes, as seen in Table 9.1. Thus, the larger the difference, the greater the priming effect. As with other RT experiments, one must report the procedure employed to screen for outliers, as well as the proportion of data that this procedure affected. If an interaction between Probe Type and Position is found, one can follow up the analysis with planned comparisons (e.g., *t*-tests).

An Exemplary Study

CMP has not been used extensively in the study of L2 sentence processing or acquisition. However, one of the few studies to do so investigated how filler integration is established during real time comprehension study with Greek L2 learners of English (Felser & Roberts, 2007). This was based on an earlier CMPP experiment with adult English native speakers and 5- to 7-year-old children (Roberts, et. al., 2007). As with much recent L2 sentence processing research, the focus of the Felser and Roberts study was on the question of whether or not L2 learners perform like native speakers in their on-line grammatical processing. The authors investigated whether Greek L2 learners of English would reactivate a fronted indirect object (*peacock)* in object relative clauses like (6) at the structurally appropriate gap site (#1). In the earlier Roberts et al. study with native speakers, this type of construction was used because the purported gap site (after the direct object [#1]) does not coincide with the verb, as is the case when using sentences in which the gap site immediately follows the verb (e.g., *Which man did you see _?*). The Direct Association Hypothesis (e.g., Pickering, 1993; Pickering & Barry, 1991) predicts that facilitation effects found during the processing of such constructions is driven by the verb, whereas, according to the Trace Reactivation Hypothesis (e.g., Bever & McElree, 1988; Gibson & Hickok, 1993; Gorrell, 1993; Love & Swinney, 1996; Nicol & Swinney, 1989), the moved argument is reactivated only at the trace position. Thus the use of object relative clauses with three-place predicates allowed for the teasing apart of these two hypotheses.

(6) John saw the peacock to which the small penguin gave the #2 nice birthday present #1 in the garden.

In both the Roberts et al. and the Felser and Roberts (2007) studies, the target picture probes were either animals or inanimate objects and were either identical to the prime (a picture of a peacock*)* or unrelated (a picture of a carrot). The nouns denoting the pictures were matched for lemma frequency and syllable-length (Francis-Kucera, 1982). All pictures were taken from the Snodgrass and Vanderwart (1980) database, which provides a set of pictures that are normed for children between the ages of 5 and 15 years. The participants were asked to decide as quickly as possible whether the object depicted was *alive* or *not alive.* The speed and accuracy of their responses to the visual targets were recorded on a push-button box. The probes appeared either in the test position, which was at the indirect object gap site (i.e., at the offset of the final word of the direct object NP *present*) or the control position (#2), approximately 500 ms earlier in the input. Thus, four experimental lists were created: the first one contained the related target at the test (trace) position, the second one at the control position; the third one contained the unrelated target at the test position, and the fourth list at the control position. Four subject groups each saw counterbalanced presentation lists so that there were equal numbers of related and unrelated visual targets in each list. The experimental session always began with six familiarization sentences, followed by 20 experimental and 60 filler sentences of different construction types. Twelve of the fillers were also relative clauses with indirect objects, to further disguise the experimental items. The visual targets that accompanied the filler items appeared at various different positions during the presentation. The same 20 experimental sentences and picture targets were used in each of two 15-minute sessions with a short break between, but—to avoid repetition effects—the stimuli were distributed over different conditions. If, for example, in the first session a picture target was presented at the control position, in the second session it was presented at the gap position, and vice versa. Moreover, if for a given sentence participants saw a related target picture in the first session, they saw an unrelated one for the same sentence in the second session. To ensure that the participants paid attention to the task, they were also asked to respond to 38 yes/no comprehension questions asking for one of the main characters in the stimulus. For example, for the experimental sentence *Sue saw the hippo to which the tall giraffe gave the sweet tasty orange in the jungle yesterday afternoon,* we asked *Did Sue see the hippo?* The comprehension questions were presented aurally and were randomly interspersed throughout the experiment following both experimental and filler sentences.

Roberts et al. report that high working-memory (WM)-span native English adults and children showed priming effects only at the indirect object gap site. That is, their responses were faster to identical targets at the test position (#1) and at the control position (#2). Furthermore, there was no difference in response times to the identical and unrelated probes at the control position. Thus, it appears that for native speakers, at least those with high working memory, dislocated elements are reactivated during real-time processing only at grammatically defined positions (see also Nicol, 1993; Nakano et al., 2002; Clahsen & Featherston, 1999; but compare Traxler & Pickering, 1996). The effects observed for native speakers cannot be explained by residual activation, given that responses to the identical targets were not facilitated at a position earlier in the sentence, and, critically, closer to the antecedent. Furthermore, this would predict that the more distance between the antecedent and the test position, the more reduced the facilitation effect, and this was not observed. The authors argue that the results support a trace-based account of filler-gap dependency processing, however, it could be the case that the parser reactivates the moved constituent at the verb site as well as the indirect object gap site.

The Greek L2 learners of English tested by Felser and Roberts using the same materials and procedure performed differently from the native English speakers. Although the learners showed facilitation in the processing of the identical probes in comparison to the unrelated words like the natives, this effect was visible at both the gap and the control positions. Therefore, in contrast to the native speakers, the results suggest that the fronted argument (*peacock*) was *maintained* in memory throughout the processing of the sentence, rather than *reactivated* only at the gap position (#2). Interestingly, this L2-native speaker difference is unlikely to be caused by the fact that the learners had reduced working memory capacity relative to the native speakers, because they did not perform in the same way as the low working memory natives in the Roberts et al. study. The authors argue that the performance difference is best explained by appealing to differences in parsing procedures between the learners and the native speakers. Specifically, they argue that the learners were less able to make use of abstract grammatical features in the input, and that their processing is shallower and more verb-driven than that of native speakers, supporting the Shallow Structure Hypothesis (Clahsen & Felser, 2006). Of course, one would need to corroborate such findings with groups of learners from different language backgrounds and with different materials, but despite this, these results show that the CMP method can be used to investigate theoretically interesting questions in L2 acquisition and processing.

Conclusion

The cross-modal priming method can be used in the investigation of questions pertaining to lexical processing and sentence processing. One can ask what is activated at what points during the uninterrupted processing of auditory input. The researcher can find out potential biasing effects on lexical processing of prior discourse and can pinpoint specific positions during the input that may be of theoretical interest. Although a dual task and therefore more cognitively demanding than some other on-line techniques, this method has many advantages, including the ability to tap into comprehension processes during the uninterrupted presentation of input without requiring metalinguistic or other conscious responses, and thus can reduce the effects of the task. Although few researchers have yet to make use of this technique with L2 learners and bilinguals, it is hoped that this chapter illustrates the potential effectiveness of CMP to address highly interesting questions in L2 lexical and sentence processing as well as in acquisition theory.

Pros and Cons of Using the Method

Pros

- Cross-modal priming allows for a measure of activation of linguistic knowledge during the uninterrupted presentation of auditory stimuli. Therefore, in contrast to studies which segment the input into chunks, as in self-paced reading, one can examine processing in more naturalistic comprehension situations.
- Given that each experimental sentence is presented as an uninterrupted stream and the comprehender is asked to pay attention to the meaning of the sentence, there is no need for he or she to assess the input consciously as its being processed, unlike studies in which participants are required to monitor the input for certain stimuli or to make metalinguistic (e.g., grammaticality or plausibility) judgments while processing.
- The experimenter can place the probe at any point during the presentation of the stimuli, and therefore the cross-modal priming paradigm is arguably the best method for tapping into the rising and falling activation levels of a specific item throughout a sentence.
- It is comparatively inexpensive to run a cross-modal priming experiment: one needs only a computer (or laptop) and the means to present auditory stimuli. The most costly apparatus will be the experimental software package used to design and present the experimental materials and to record the participants' responses.
- Given that one can run a CMP experiment on a laptop, it is potentially a portable experiment.
- Cross-modal picture priming (probes are pictures rather than written word forms) can be used with children and/or participants who are nonliterate or not used to participating in experiments.

Cons

- Cross-modal priming is by definition a dual task, which means that it may place extra cognitive load on participants during the experiment. Despite this, it is less demanding than probe recognition, in which the participant makes their lexical decision at the end of the sentence. Therefore, the experimenter should train the participants on the task, and should assess objectively their participants' working memory capacity. Following Hestvik et al., a modified version of the CMPP task where participants name related picture probes as quickly as they can could be used with populations who may find the task with binary categorization responses too taxing.

- The cross-modal priming method measures activation of a lexical item, and therefore one cannot test for spillover effects during on-line processing.
- This method can only probe for activation of nouns and cannot readily investigate verbs, which limits the range of research questions that can be addressed.
- As stated above, the task involves the presentation of uninterrupted input for processing and is therefore more natural than tasks where the experimenter presents the input in chunks. Nevertheless, CMP is not as natural a comprehension task as is found in, for instance, eye-tracking during reading studies, because participants must perform an additional task (e.g., a lexical decision) during their comprehension of the experimental materials.

Discussion and Practice

Questions

1) What is the difference between maintained activation and reactivation in the context of a CMP experiment, and how can one ensure that a target probe item is reactivated?
2) What are the various possible relations between a prime and a probe item? Which type do you think is best to use to measure activation and why?
3) What are the benefits of using CMPP versus CMLP with different types of populations?
4) What are the potential pitfalls of using CMP, and how can researchers minimize these problems?
5) How would the design of a CMP investigating lexical processing differ between children, L2 learners, and adult native speakers?

Research Project Option A

According to some researchers, one of the ways in which L2 learners' processing differs from that of native speakers is that learners make less use of abstract grammatical information, instead relying on lexical-semantic and discourse information (see e.g., Clahsen & Felser, 2006). This predicts that L2 learners should be more sensitive than native speakers to discourse context during real-time processing. An interesting research project would therefore be to replicate a study (e.g., Swinney, 1979) on the effects of discourse context on the activation of ambiguous words, comparing L2 learners with native speakers.

Research Project Option B

Recently, there have been a number of studies (e.g., Poirier et al., 2010) investigating exactly what parts of an antecedent clause (e.g., the subject as well as the verb) are activated during the real-time processing of VP-ellipsis sentences (e.g., *The mailman [bought a new tie] and the fireman did Ø too*). An interesting project would involve examining the processing of such constructions by L2 learners, comparing learners in whose first language VP-ellipsis is or is not instantiated.

Notes

1. This chapter follows the convention in CMP research of using italics to refer to the stimulus word that serves as the prime and all capitals to refer to the probe word.
2. Given the fact that children and children with SLI have reduced working capacity in comparison to adults, Hestvik, Schwartz, & Tornyova (2007), modified the CMPP task such that their participants with SLI named the target picture as fast as possible, rather than having to make a (cognitively more demanding) binary decision on the target picture (e.g., *Is this alive or not alive?*).
3. It should be noted, however, that this specifically relates to arguments; verbs may remain active throughout the processing of a sentence (e.g., Callahan, Shapiro, & Love, 2010; De Goede, Shapiro, Wester, Swinney, & Bastiaanse, 2009).
4. Pseudowords are pronounceable nonwords.
5. Research has found that neighborhood density (i.e., the number of similar words) affects different aspects of lexical processing including lexical acquisition, word recognition, and naming (see e.g., Storkel, 2004; Vitevitch & Sommers, 2003). Relevant to the current discussion is the finding that the greater the number of semantic, phonological, and/or orthographic neighbors (neighborhood density) a word has, the slower lexical processing is for that word (e.g., Vitevitch & Rodríguez, 2005).

Suggested Readings

Brien, C., & Sabourin, L. (2012). Second language effects on ambiguity resolution in the first language. *EUROSLA Yearbook, 12,* 191–217.

Clahsen, H., & Featherston, S. (1999). Antecedent-priming at trace positions: Evidence from German scrambling. *Journal of Psycholinguistic Research, 28,* 415–437.

Love, T., Maas, E., & Swinney, D. (2003). The influence of language exposure on lexical and syntactic language processing. *Experimental Psychology, 50,* 204–216.

Poirier, J., Wolfinger, K., Spellman, L., & Shapiro, L. P. (2010). The real-time processing of sluiced sentences. *Journal of Psycholinguistic Research, 39,* 411–427.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bever, T., & McElree, B. (1988). Empty categories access their antecedents during comprehension. *Linguistic Inquiry, 19,* 35–43.

Brien, C., & Sabourin, L. (2012). Second language effects on ambiguity resolution in the first language. *EUROSLA Yearbook, 12,* 191–217.

Callahan, S. M., Shapiro, L. P., & Love, T. (2010). Parallelism effects and verb activation: The sustained reactivation hypothesis. *Journal of Psycholinguistic Research, 39,* 101–118.

Caplan, D. (1972). Clause boundaries and recognition latencies for words in sentences. *Perception and Psychophysics, 12,* 73–76.

Chomsky, N. (1981). *Lectures on Government and Binding.* Dordrecht, Netherlands: Foris.

Clahsen, H., & Featherston, S. (1999). Antecedent-priming at trace positions: Evidence from German scrambling. *Journal of Psycholinguistic Research, 28*(4), 415–437.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27,* 3–42.

Sinclair, J. M. (Ed.). (1987). *Collins COBUILD English Language Dictionary*. London/Glasgow: Collins.

De Goede, D., Shapiro, L. P., Wester, F., Swinney, D. A., & Bastiaanse, Y. R. M. (2009). The time course of verb processing in Dutch sentences. *Journal of Psycholinguistic Research, 38*(3), 181–199.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language, 27,* 429–446.

E-Prime [Computer Software]. Pittsburgh, PA: Psychology Software Tools.

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research, 23,* 9–36.

Francis, W. N., & Kucera, H. (1982). *Frequency Analysis of English Usage.* Boston: Houghton Mifflin.

Gibson, E., & Hickok, G. (1993). Sentence processing with empty categories. *Language and Cognitive Processes, 8,* 147–61.

Gorrell, P. (1993). Evaluating the direct association hypothesis: a reply to Pickering and Barry, 1991. *Language and Cognitive Processes, 8,* 129–46.

Hestvik, A., Schwartz, R. G., & Tornyova, L. (2007). Gap-filling and Sentence Comprehension in Children with SLI. Paper presented at the BUCLD 31: The 31st Annual Boston University Conference on Language Development, Boston, MA.

Koeneman, O., Baauw, S., & Wijnen, F. (1998). Reconstruction in VP-Ellipsis: Reflexive vs. Nonreflexive predicates. Poster presented at the 11th Annual CUNY Conference on Human Sentence Processing, New Brunswick, NJ.

Love, T., & Swinney, D. (1996). Coreference processing and levels of analysis in object-relative constructions: Demonstration of antecedent reactivation with the cross-modal priming paradigm. *Journal of Psycholinguistic Research, 20*(1), 5–24.

Love, T., & Swinney, D. (1997). Real time processing of object relative constructions by preschool children. Poster presented at the 10th Annual CUNY Conference on Human Language Processing, Santa Monica, CA.

Marinis, T., & Van der Lely, H. (2007). On-line processing of wh-questions in children with G-SLI and typically developing children. *International Journal of Language & Communication Disorders, 42*(5), 557–582.

McKee, C., Nicol, J., & McDaniel, D. (1993). Children's application of binding during sentence processing. *Language and Cognitive Processes, 8,* 265–290.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. G. (1975). Loci of contextual effects on visual word recognition. In P.M. A. Rabbitt & S. Dornic (Eds.), *Attention and Performance V* (pp. 98–118). San Diego, CA: Academic Press.

Nakano, Y., Felser, C., & Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research, 31,* 531–571.

Neely, J. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys. (Eds.), *Basic Processes in Reading: Visual Word Recognition* (pp. 264–336). Mahwah, NJ: Erlbaum.

Nicol, J. (1988). *Coreference processing during sentence comprehension* (Unpublished doctoral dissertation). MIT, Cambridge, MA.

Nicol, J. (1993). Reconsidering reactivation. In G. Altmann & R. Shillcock (Eds.), *Cognitive Models of Speech Processing: The Second Sperlonga Meeting* (pp. 321–350). Hove, UK: Erlbaum.

Nicol, J., & Swinney, D. (1989). The role of structure in coreference assignment during sentence comprehension. *Journal of Psycholinguistic Research, 18,* 5–20.

Pickering, M. (1993). Direct Association and sentence processing: A reply to Gibson & Hickok. *Language and Cognitive Processes, 8,* 163–196.

Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes, 6,* 229–259.

Poirier, J., Wolfinger, K., Spellman, L., & Shapiro, L. P. (2010). The real-time processing of sluiced sentences. *Journal of Psycholinguistic Research, 39*(5), 411–427.

Presentation [Computer Software]. Albany, CA: Neurobehavioral Systems.

Roberts, L. (2013). Sentence processing in bilinguals. In R. van Gompel (Ed.). *Sentence Processing. Current Issues in the Psychology of Language* (pp. 221–246). London: Psychology Press.

Roberts, L., Marinis, T., Felser, C., & Clahsen, H. (2007). Antecedent priming at trace positions in children's sentence processing. *Journal of Psycholinguistic Research, 36,* 175–188.

Shapiro, L. P., & Hestvik, A. (1995). On-Line Comprehension of VP-Ellipsis: Syntactic Reconstruction and Semantic Influence. *Journal of Psycholinguistic Research, 24*(6), 517–532.

Shapiro, L. P., Hestvik, A., Lesan, L., & Garcia, A. R. (2003). Charting the time-course of VP-ellipsis sentence comprehension: Evidence for an initial and independent structural analysis. *Journal of Memory and Language, 49,* 1–19.

Snodgrass, J., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, Familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6,* 174–215.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics, 25*, 201–221.

Swinney, D. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning & Verbal Behaviour, 18,* 645–659.

Swinney, D., Onifer, W., Prather, P., & Hirshkowitz, M. (1979). Semantic facilitation across sensory modalities in the processing of individual words and sentences. *Memory and Cognition, 7,* 159–165.

Traxler, M., & Pickering, M. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language, 35,* 454–475.

Vitevitch, M. S., & Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders, 3,* 64–73.

Vitevitch, M. S., & Sommers, M. S. (2003). The facilitative influence of phonological similarity and neighborhood frequency in speech production in younger and older adults. *Memory & Cognition, 31,* 491–504.

10

ON PSYCHOLINGUISTIC METHODS

Michael J. Leeser

The chapters in this book demonstrate the ways in which second language (L2) research has moved toward the use of behaviorally sensitive measures and on-line methodologies in order to investigate critical issues related to L2 learners' mental representations and how learners process the language they hear and read. More than a decade ago, Juffs (2001) suggested that it was "embarrassing" that, with the exception of research on lexical processing in L2 learners and bilinguals, commonly used psycholinguistic methodologies such as reaction time measures had hardly been utilized in L2 research. Since that time, however, there has been substantial growth in the use of psycholinguistic methods, most notably in the area of L2 sentence processing. Given the increase in L2 studies using on-line methods, this volume offers graduate students, as well as beginning and more seasoned researchers, a timely and much needed resource for understanding the various methodologies, research designs, and data analysis techniques used in psycholinguistic approaches to L2 research. Each chapter provides specific guidelines regarding issues of calibration, counterbalancing stimuli, the proportion of target stimuli to fillers, and cleaning up and organizing data in preparation for analysis. In addition, the pros and cons of each method are presented so that readers can make informed decisions regarding which method(s) may be the most appropriate to their own research questions and the research environment in which they work.

Keeping in mind that readers of this book may be unfamiliar with or still learning about some of methods discussed in this volume, I have three goals in this concluding chapter. First, drawing on the exemplary studies presented in previous chapters, I briefly synthesize the ways in which the psycholinguistic methods presented in this volume address some of the critical areas of inquiry that VanPatten outlined in Chapter 1. Second, I highlight methodological issues that researchers need to consider carefully, regardless of which psycholinguistic method is used. Third, I explore future directions for using psycholinguistic methods in

L2 research by focusing on one area that could benefit greatly by employing these methods: instructed second language acquisition (SLA).

Psycholinguistic Methods and Critical Areas of L2 Processing

In the first chapter of this volume, VanPatten presented some of the areas of inquiry within second language processing research, including:

- the role of the native language in non-native processing;
- whether native-like processing in the L2 is attainable;
- how processing can provide insights into the nature of linguistic representation; and
- the role of explicit learning in L2 processing and/or acquisition.

The exemplary studies presented in the chapters focus primarily on the role of the native language and whether L2 learners can achieve native-like processing procedures. These studies are summarized in Table 10.1.

Native Language Influence

As the table illustrates, there are a number of aspects that can be investigated that relate to the role of the first language (L1) while processing the L2. These include (but are not limited to) the influence of L1 processing routines (Jackson, 2010), the presence or absence of specific L1 morphosyntactic structures (Dussias et al., 2013), and L1 lexical activation (Linck Kroll, & Sunderman, 2009). Although not all models of processing posit a fundamental role for the native language in L2 processing (see, e.g., VanPatten, Chapter 1, this volume), the role of the L1 continues to be a fruitful avenue of research, particularly given the ample evidence that, for bilinguals and multilinguals (i.e., those that speak three or more languages), all languages are activated and interact at some level (e.g., de Groot, 2011). A more recent direction of research involves the role of cross-linguistic transfer in multilingual speakers. As a result, the methods in this volume will give researchers the tools to address current debates surrounding the influence of first and second languages (or source languages) on the third (target language), as well as the role of the following factors in determining cross-linguistic influence in processing languages beyond the L2: typological distance between target and source languages, context and recent use of source and target languages, source and target language proficiency, and order of acquisition of each language.

Attainability of Native-Like Processing

In terms of whether native-like processing is attainable, Table 10.1 shows that different methodologies have been used to examine whether advanced L2 learners

TABLE 10.1 Summary of the areas of inquiry of exemplary studies

Area of inquiry	*Method*	*Research question(s) in exemplary studies*
Role of the L1 in processing the L2	Self-paced reading	Do L2 learners of German transfer a lexical verb-based strategy from L1 English? (Jackson, 2010)
	Visual world	Does the presence grammatical gender in the L1 (Italian vs. English) affect the degree to which gender processing is nativelike in the L2 (Spanish)? (Dussias et al., 2013)
	Translation Recognition	Does immersion in L2 Spanish (study abroad) affect the activation/suppression of L1 English during lexical processing? (Linck, Kroll, & Sunderman, 2009)
Differences between native and non-native processing	Self-paced reading	Do L2 learners of German (L1 English and L2 Dutch) process sentences containing subject-object ambiguities like L1 German speakers? (Jackson, 2010)
	Self-paced listening	Do L1 Russian L2 Greek children process grammatical voice morphology (active vs. passive), subject-verb agreement, and subject-object ambiguities like L1 Greek children and adults? (Papangeli, 2010)
	Eye-tracking	Do L2 learners make use of structural constraints (Binding Principle A) to interpret sentences in real time or do they primarily utilize nonstructural information (discourse-level constraints)? (Felser & Cunnings, 2012)
	ERPs	How much instruction is required for native English learners to develop native-like lexical processing patterns in L2 French, with differences between pseudowords, words preceded by semantically unrelated primes, and words preceded by semantically related primes? (McLaughlin, Osterhout, & Kim, 2004)
	Cross-modal Priming	Do L2 English learners (L1 Greek) make use of lexical information or abstract grammatical features to process complex sentences containing object relative clauses? (Felser & Roberts, 2007)
	fMRI	What areas of the brain are activated in early and late French-German bilinguals during sentence processing in French and German? (Saur et al., 2009)

make use of structural information (e.g., syntactic properties, implicit knowledge of agreement phenomena) or whether they rely primarily on lexical-semantic information during input processing. Much of this line of research is motivated by the Shallow Structures Hypothesis (SSH) (Clahsen & Felser, 2006), which claims that native and non-native processing are qualitatively different. On the one hand, native speakers occasionally rely on shallow processing (reliance on nonstructural information such as lexical, semantic, and pragmatic information) but also engage in structural processing (computation of syntactic structure) during comprehension. Non-native speakers, however, rely exclusively on shallow processing (see, e.g., Keating, 2009; Papadopoulou & Clahsen, 2003; and VanPatten, Chapter 1, this volume, for further discussion of the SSH). Although the SSH continues to generate a good deal of research, there are a number of issues that researchers must take into account whenever comparing native and non-native processing. Among these include the characteristics of the native speakers themselves (whether they are monolinguals, bilinguals, or multilinguals), as well as the variability of L2 data. These issues, and others, are explored in the next section.

Theoretical and Methodological Considerations

Linking Theoretical Constructs to On-Line Measures

Within the "What is Looked at and Measured" section of each chapter, the authors outline and describe the measures or types of data obtained from the method discussed. Table 10.2 provides an overview of these commonly used on-line measures and their definitions.

The fact that several methodologies have been and can be used to investigate these issues may leave novice researchers wondering if there is a "best" method. Obviously, no methodology is perfect, and each has its own pros and cons, which the authors of the previous chapters have pointed out. In fact, the choice of which method to use will depend not only on researchers' specific questions, but also on access to and familiarity with the technology associated with the different experimental techniques.

What is important to note in the definitions provided in Table 10.2 is that none suggests a direct reflection of a specific linguistic or cognitive process. For example, no definition states something along the lines of "morphosyntactic anomaly" or "conceptual sensitivity." Rather, as Carreiras and Clifton (2004) point out, "an increase in comprehension time or a disruption in the eye-movement record is just an increase in time. It does not by itself carry any sort of signature about what processes gave rise to the disruption" (p. 2). For this reason, it is necessary to understand the *linking assumptions* that connect behavioral responses and changes in brain activity to a theoretical construct of interest.

To illustrate this point, I'll use an example of how processing is normally used to provide insights into linguistic representation (i.e., the third of the four areas of

TABLE 10.2 Definitions of commonly used measures of various on-line methods

Method	*Representative measures*	*Definition*
Self-paced reading	Reading time	Duration (in milliseconds) between subsequent button presses
Self-paced listening	Listening time	Duration between subsequent button presses
Eye-tracking	Fixation duration	Duration of the first fixation on a word or region
	Gaze duration	Summed duration of all fixations before the eyes leave a word or region within the first pass
	Rereading times	Summed duration of all fixations before the eyes leave the word or region after the first pass
	Total reading time	Summed duration of all fixations on a word or region
	Proportion of regressions	Number of regressive saccades divided by total number of sentences in a condition
	Proportion of fixations	Number of fixations divided by total number of trials in a condition
ERP	N400	A negative waveform peaking approximately 400 ms after stimulus onset
	LAN	A left anterior negative waveform peaking approximately 200 ms after stimulus onset
	P600	A positive waveform peaking approximately 600 ms after stimulus onset
fMRI	Activation	Difference in signal (i.e., blood oxygenation levels) between different conditions
Translation Recognition	Reaction Times	Duration between word presentation and participant response
	Accuracy	Proportion of correct decisions regarding translation equivalency over total number of trials
Cross-Modal Priming	Reaction Times	Duration between a probe (picture or word) and participant response

inquiry described by VanPatten in Chapter 1 of this volume). A researcher investigating whether L2 learners possess implicit knowledge or automatic competence of subject agreement in English might conduct an experiment in which participants read or listen to a set of sentences, some of which are grammatically correct (1) and others that are not (2).

(1)

a. The detective finds the evidence in the house.
b. The detectives find the evidence in the house.

(2)

a. *The detective find the evidence in the house.
b. *The detectives finds the evidence in the house.

In the case of a self-paced reading task, a researcher could examine whether participants' reading times increase upon encountering a grammatical violation at the target region (e.g., the verb *find/s*) and at the spillover regions or the words immediately following the target (see, e.g., Jegerski, Chapter 2, this volume). If statistically significant differences are observed in reading times between target regions of grammatical and ungrammatical sentences, the researcher might conclude that participants possess implicit knowledge of subject agreement. The justification of this conclusion is based on a series of assumptions. The first is that sentence comprehension (parsing) occurs on a word by word basis, in which each word is assigned to an attachment site within the current syntactic structure of the sentence (e.g., Just & Carpenter, 1980). The second assumption is that during this process, agreement features associated with a given word, such as number, person, case, and gender, are computed and checked based on the reader/listener's underlying grammar (Pritchett, 1992). Finally, if a segment from an utterance does not match the structural description in the underlying grammar, a difficulty and/or delay in processing can occur, resulting in increased reading times. Using other methods, the mismatch between an incoming utterance and the participant's mental representation could manifest itself as longer listening times, longer eye gaze durations, a greater number of regressive eye movements (saccades), LAN or P600 effects, and differential brain region activation in fMRI studies.

A clear understanding of the linking hypotheses and assumptions behind on-line measures is vital not only for the proper interpretation of data, but also in considering what else can contribute to the data patterns observed in on-line studies. The discussion that follows illustrates how what we instruct participants to do in a study can impact their on-line processing.

The Role of Tasks

In addition to collecting on-line processing data, such as the measures listed in Table 10.2, most sentence processing studies also include data concerning participants' final interpretation or acceptability of an utterance. That is, experiments utilizing the on-line methods also include an additional task to ensure that participants are not passively reading or listening to stimuli (Jegerski, Chapter 2; Keating, Chapter 4; Newman, Chapter 7, this volume). To take an example from sentence processing research, in some studies participants are prompted to make a grammaticality or acceptability judgment after reading or listening to an utterance. In others, participants answer a YES/NO comprehension question. These tasks are included to ensure that the on-line processing data reflect what participants are doing when processing sentences correctly, that is, accurately comprehending and/or judging the grammaticality of a sentence. Thus, researchers typically only analyze data from sentences for which participants correctly answered the grammaticality judgment prompt or comprehension question. Little is known, however, about the ways in which these offline tasks may interact with and even alter learners' on-line processing performance. Put another way, could the data gathered be mediated by what participants are instructed to do? In both the L1 and L2 language processing literature, evidence exists from ERP, fMRI, and self-paced reading studies to suggest that this is indeed the case. Three of these studies are discussed below.

Hahne and Friederici (2002) collected ERP data from L1 German speakers as they listened to the following types of sentences in their native language: (a) semantically plausible and syntactically correct, (b) semantically implausible and syntactically correct, (c) semantically plausible and syntactically incorrect, or (d) semantically implausible and syntactically incorrect. In one experiment, participants were instructed to judge the overall acceptability of the sentences, and in a second experiment they were instructed to judge only the semantic plausibility of the sentences. Among their findings was that syntactically incorrect sentences elicited a P600 in the experiment in which participants made acceptability judgments (Experiment 1), but this effect was greatly attenuated when participants were instructed to focus exclusively on semantic acceptability (Experiment 2). These findings suggest that the observed P600 effects were dependent on task type and instructions.

In an fMRI study, Love, Swinney, Haist, Nicol, and Buxton (2003) investigated the effects of syntactic complexity and task condition on the activation of Broca's area, which is an area of the brain associated with language production and other language tasks. Motivated by conflicting findings from fMRI and PET research regarding the role of Broca's area in linguistic processing, native English speakers listened to English sentences of differing complexity in three task conditions of differing processing/memory demands: (a) listening and understanding the sentences (the least demanding condition), (b) deciding whether a word was part of the sentence, and (c) deciding whether the first noun phrase was the sentential agent (the most demanding condition). They found no differential activation in Broca's area based on the structural complexity of the sentences; however, increased activation in Broca's area was observed during sentence processing when the off-line task engaged higher processing demands. Although the researchers do not rule out a role for Broca's area in language comprehension, they suggest that the processing demands of the off-line task may account for activation of Broca's area in previous studies.

Within L2 research, Leeser, Brandl, and Weissglass (2011) conducted a self-paced reading study and investigated the effects of secondary off-line tasks (comprehension questions vs. grammaticality judgments) on L1 and L2 Spanish readers' on-line sensitivity to morphosyntactic violations. A primary motivation for the Leeser et al. study was the mixed findings reported in previous research investigating L2 learners' sensitivity to agreement violations. On the one hand, evidence from ERP studies (e.g., Tokowicz & MacWhinney, 2005; Osterhout, McLaughlin, Pitkänen, Frenck-Mestre, & Molinaro 2006) suggested that L2 learners at the beginning stages of acquisition can demonstrate on-line sensitivity to agreement violations. On the other hand, studies using eye movement and self-paced reading methods reported that L2 learners demonstrate sensitivity to these violations only at the advanced level, if at all (e.g., Jiang, 2004, 2007; Keating, 2009). In all of these studies, the on-line data (reaction times, eye movements, ERPs) served as the primary dependent measure. However, these studies also included a secondary offline task, in which participants were given some kind of prompt once they had finished reading a sentence, such as answering a comprehension question (Jiang, 2004, 2007; Keating, 2009) or judging the grammaticality or acceptability of a sentence (Osterhout et al., 2006; Tokowicz & MacWhinney, 2005). A rationale for the Leeser et al. (2011) study, therefore, was to investigate whether the conflicting findings in previous studies could be attributed to off-line task effects. Native Spanish speakers and intermediate L2 Spanish learners read sentences containing nouns followed by a modifying adjective. In half of the sentences, the adjective agreed with the noun and in half they did not. In one task, learners answered yes/no comprehension questions after each sentence; in the other task, they made grammaticality judgments. The results revealed that intermediate L2 learners did not demonstrate on-line sensitivity to agreement violations when they were prompted with comprehension questions following each sentence. However, when they were prompted with grammaticality judgments, the intermediate learners were sensitive to the agreement violations at the regions of interest (e.g., the target adjective and the spillover regions).

Another interesting finding of this study is that the secondary task not only mediated L2 learners' sensitivity at the regions of interest, but task effects were also observed in the sentence regions prior to and immediately after the regions of interest, as can be seen in Figure 10.1. Prior to the target region, participants in the grammaticality judgment condition demonstrated significantly slower reading times than those in the comprehension condition. In the segments following the target region, however, the participants' reading times were significantly faster.

These results suggest that when readers are instructed to make grammaticality judgments, they process sentences at a significantly slower rate, perhaps drawing upon metalinguistic knowledge until they locate the violation. Once they locate it, they know that they have achieved what they were instructed to do and may simply rush through the rest of the sentence in order to answer the grammaticality prompt. Keeping in mind the discussion of the linking assumptions from the previous section, the question that arises is whether this sort of strategic processing behavior is guided by the same knowledge that is assumed to guide sentence comprehension.

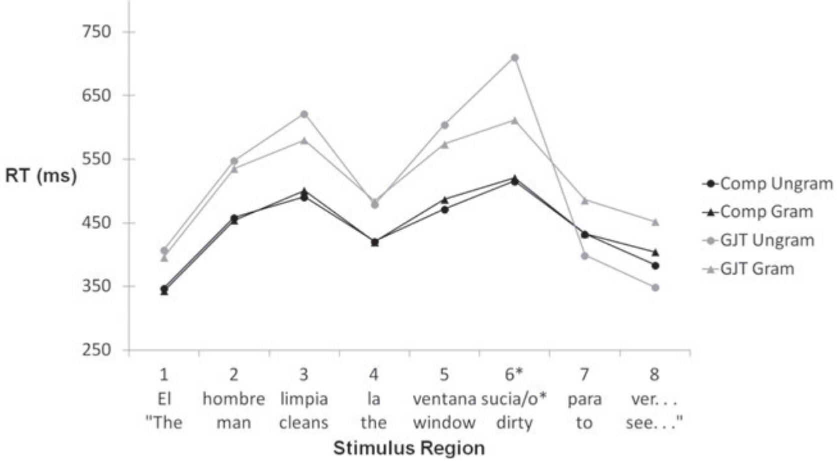


FIGURE 10.1 Intermediate-level L2 Spanish reading times by region. Adapted from Leeser et al. (2011).

Although the knowledge utilized in making grammaticality judgments may include implicit information regarding syntax, semantics, or pragmatics, the process of making grammaticality judgments also draws upon metalinguistic knowledge, particularly if participants possess explicit knowledge of target structures due to formal instruction (e.g., Birdsong, 1989; Jiang, 2007). In other words, the problem with participants drawing upon explicit, metalinguistic knowledge during an on-line task is that it involves a different kind of knowledge than the implicit grammar or underlying competence used to guide sentence comprehension (Schütze, 1996). To be sure, the role of explicit knowledge in SLA continues to be the subject of much debate, particularly regarding (a) whether explicit knowledge can contribute to the development of and/or become implicit knowledge (see, e.g., DeKeyser, 2003; Ellis, 2005) and (b) the role of explicit information in instructed SLA (VanPatten, Borst, Collopy, Qualin, & Price, 2013). These debates aside, however, there is general agreement that acquisition involves the development of implicit knowledge (Ellis, 2005), and this implicit knowledge guides on-line sentence comprehension. Consequently, if the issue under investigation is comprehension, then every effort must be made to minimize the activation and use of explicit knowledge.

The findings of the studies reviewed here on task effects underscore the fact that comprehension is a goal-driven process, and the goals vary and range from searching for specific information to comprehending for general understanding (Grabe & Stoller, 2002). Furthermore, these goals affect *how* people comprehend in their L1 or their L2 (e.g., Taraban, Rynearson, & Kerr, 2000). Therefore, the way in which participants process sentences during an on-line experiment is likely to be influenced by the goals of the tasks that they are given (Swinney, 2000). For this reason, given that obtained on-line data is used to support or falsify theoretical positions about L2 learners' underlying linguistic knowledge and real time processing, it is important to consider carefully how tasks used in on-line sentence processing research (and all L2 research) can affect the very processes under investigation.

The Variability of L2 Data

That L2 learning is variable in its outcome is one of the hallmark differences between first and second language development. Children L1 learners are uniformly successful at acquiring their native language, provided that they are not cognitively impaired. Most second language learners, however, do not achieve the same level of implicit knowledge and skills as native speakers. What is more, even L2 learners under the same conditions of exposure and/or learners grouped together in the same general proficiency level exhibit a good deal of variability in terms of representation and language use. In fact, in on-line L2 studies, L2 within-group variation is generally much larger than L1 within-group variation (Jiang, 2012; Juffs, 2005). This variability in L2 performance can pose a challenge when attempting to interpret L2 data gathered from on-line experiments. In the chapter sections on data analysis and reporting results, the authors explain how to organize data (i.e., response times, temporal and spatial eye movement data, ERP measures, and fMRI signals) in order to perform statistical analyses by subject and by item, comparing different participant groups, such as native speakers, advanced learners, and so forth. The data analysis procedures described have been standard procedures in L1 processing research for decades. However, keeping in mind the large within-group variability in L2 data, it is often important to look beyond statistical results of participant groups (e.g., "advanced" learners) and consider the data of individual participants, particularly when investigating whether L2 learners can attain native-like processing.

To illustrate the issue of L2 variability and comparisons of native and nonnative processing, Keating's (2009) eye-tracking study investigated native and nonnative Spanish speakers' sensitivity to gender agreement violations in three different conditions: (1) within a determiner phrase (DP) (i.e., **la cerveza frío* "the cold beer"), (2) across phrase boundaries (**la cerveza está bien frío* "the beer is quite cold"), and (3) across clause boundaries (**la cerveza sabe mejor cuando se sirve frío* "beer tastes better when it is served cold"). In terms of fixation times and regressive saccades, beginning and intermediate L2 learners showed no difference in how they processed grammatical and ungrammatical sentences in all three conditions. However, both advanced L2 Spanish learners and native speakers of Spanish demonstrated longer fixation times on ungrammatical adjectives than on grammatical adjectives when the gender agreement violation occurred within the DP. But only native speakers demonstrated sensitivity to violations across phrase and clause boundaries. The findings lend support to the Shallow

Structures Hypothesis in that advanced L2 learners only demonstrated native-like processing in local domains (i.e., within the DP) but not when agreement violations occurred across phrase or clause boundaries. Closer inspection of the reading time data for the advanced learners may provide a clue as to why this might be the case. The descriptive statistics reported revealed that the standard deviations for the advanced learners' total reading time ranged from 50 to 100% greater than the standard deviations for the L1 Spanish group. In other words, the variability of the advanced learners' data was up to twice as much as that of the native speakers. Given that tests of statistical significance rely heavily on measures of variability (i.e., the greater the variability, the less likely it is to obtain statistically significant results), it is not surprising that statistically significant findings were not observed for the advanced learners as a group. When looking at the total reading times of ungrammatical adjectives versus grammatical adjectives of individual advanced learners, however, one third of the participants demonstrated reading time differences between ungrammatical and grammatical adjectives by at least 161 ms in the cross-clausal condition, and one half had a difference of at least 117 ms in the cross-phrasal condition. This suggests that a number of advanced learners were demonstrating differential processing between grammatical and ungrammatical sentences similar to the native speakers, even though the difference in reading times for advanced learners as a group did not reach statistical significance. Therefore, because of the variability of L2 data, it would be worthwhile to supplement standard statistical procedures with an examination of individual participants when examining issues of ultimate attainment.

Native Speaker Characteristics

Another issue to keep in mind when comparing native and non-native processing is the characteristics of the native speakers themselves. That is, are they monolinguals, bilinguals, or multilinguals? This issue is an important one given that the languages of bilinguals and multilinguals are activated and interact with each other during language processing. As a result, knowledge of more than one language may alter linguistic representation and processing, as well as the acquisition of subsequent languages. As de Groot (2011) points out, a consequence of this activation and interaction is that "the [processing of] linguistic utterances of bilinguals and multilinguals differs from those of monolingual speakers of the languages involved. In other words, multilingualism does not equal multiple monolingualism" (p. 401).

One example of this interaction and activation of more than one language in the sentence processing literature can be found in the self-paced reading experiment reported in Dussias (2003). In this study, Spanish monolinguals and Spanish-English bilinguals (L1 Spanish-L2 English) read sentences containing complex noun phrases followed by relative clauses, in which disambiguating information in the relative clause forced an interpretation in which the relative clause modified either the first noun phrase (high attachment) or the second noun phrase (low attachment), as shown in examples (3a) and (3b), respectively:

(3)

a. El perro mordió a la cuñada del maestro / que vivió en Chile / con su esposo.

"The dog bit the sister-in-law of the teacher$_{\text{MASC}}$ / who lived in Chile / with her husband."

b. El perro mordió al cuñado de la maestra / que vivió en Chile / con su esposo.

"The dog bit the brother-in-law of the teacher$_{\text{FEM}}$ / who lived in Chile / with her husband."

Although monolingual Spanish speakers demonstrated faster reading times when the disambiguating segment forced high attachment (3a) than if it forced low attachment (3b), the opposite pattern was found for the L1 Spanish-L2 English bilinguals: they were faster for the segments forcing low attachment than those forcing high attachment. In other words, even though Spanish is a language in which monolinguals have consistently demonstrated a high attachment interpretation when ambiguous relative clauses follow a complex noun phrase (NP; Cuetos & Mitchell, 1988), the Spanish-English bilinguals, who were native speakers of Spanish, favored low attachment instead.

One explanation offered for these findings is that the L1 Spanish-L2 English bilinguals were residing in a predominantly English-speaking environment in the United States. Because these participants had been living in a linguistic environment in which local attachment is the predominant parsing routine when relative clauses follow complex NPs, this greater exposure to English may have altered the parsing processes of the L1. A second possible reason for the L1 Spanish-L2 English bilinguals demonstrating a preference for low attachment is related to the cognitive demands of the bilingual processor. Assuming that local attachment involves a lesser processing cost given that new material can be immediately integrated, the processing demands associated with housing two languages may "constrain the bilingual parser to use operations such as late closure, which ensure that new material is immediately integrated with prior material (by way of local attachment) and minimize the chances of exceeding the memory limits of the sentence-processing mechanism" (Dussias, 2003, pp. 552–553). In other words, the parser may adopt one strategy for both languages in order to function more efficiently (O'Grady, in press).

Regardless of whether the explanation for the difference in parsing strategies between Spanish monolinguals and L1 Spanish-L2 English bilinguals can be attributed to the processing demands associated with housing two languages or to extended exposure to English, these studies illustrate the point that bilinguals may process their native language differently than do monolinguals. A reasonable question, therefore, is whether the apparent failure of high proficiency L2 learners to parse sentences like monolingual native speakers can be attributed to an incomplete parser (or an incomplete grammar guiding the parser), or whether it is actually a consequence of bilingualism, in which the parser adopts a more efficient strategy. For this reason, if researchers seek to investigate whether L2 learners can attain native-like processing, comparing the processing of bilingual native speakers and L2 learners may be more informative than comparing L2 learners with monolinguals only.

Future Directions in L2 Research: The Role of Instruction

Using psycholinguistic methods to investigate issues of L2 representation and processing will continue to be a fruitful avenue of research, as the field seeks to gain a deeper understanding of the role of the native language and whether native-like processing/representation is attainable. Of the four areas of inquiry presented in VanPatten's chapter, the one in which the use of psycholinguistic methods has been largely absent is the role of explicit learning in L2 processing and/or acquisition. Therefore, in this final section, I would like to focus on ways in which psycholinguistic methods can advance a related area: instructed SLA. More specifically, I consider issues related to measuring acquisition in effects of instruction studies, attention and awareness, and processing instruction.

Measuring Acquisition

In their groundbreaking meta-analysis of effects of instruction research, Norris and Ortega (2000) concluded more than a decade ago that, "focused L2 instruction results in large gains over the course of an intervention. Specifically, L2 instruction of particular language forms induces substantial *target-oriented change,* whether estimated as pre-to-post change within experimental groups [. . .] or as differences in performance between treatment and control groups on *post-test measures* . . ." (p. 500, emphasis added). Yet, two questions that arise from this conclusion are the following: (a) what do the (pre and post) tests actually *measure* in the effects of instruction studies and (b) what specifically has *changed* as the result of an instructional intervention? Considering the first question, Norris and Ortega acknowledge that more than 90% of the outcome measures consisted of discrete point, grammar-based paper and pencil tests. For this reason, they suggest that "observed instructional effectiveness within primary research to date has been based much more extensively on the application of explicit declarative knowledge under controlled conditions, without much requirement for fluent, spontaneous use of contextualized language" (p. 486). A logical conclusion, then, is that the kind of change that has taken place refers to acquired explicit knowledge about language or the ability to deploy it rather than any change in representation or implicit knowledge. In fact, this bias toward explicit outcome measures within instructed SLA research has led Doughty (2003) to argue that "the case for explicit instruction has been overstated" (p. 274).

Although there has been an increase in research designs incorporating spontaneous, communicative use of the L2, a majority of instructed SLA studies still rely on controlled, explicit outcome measures. To provide one example, a recent meta-analysis conducted by Spada and Tomita (2010) on the effects of explicit and implicit instruction on the acquisition of simple and complex English constructions revealed that 50% of the studies employed free production tasks to measure learning. However, they note that even the measures coded as free production "may not represent 'pure' measures of spontaneous ability tapping into exclusively implicit knowledge" (p. 287).

Clearly, as noted earlier in this chapter, no psycholinguistic method can measure linguistic competence directly. Also, I am not advocating that researchers abandon production-based tasks, as production data provide important insights into how learners map abstract features onto surface morphology. However, the interpretation of speech production data can be a challenge given that learners may fail to produce target-like forms due to communicative pressure or memory limitations, and/or their target-like production may be due to monitoring or utilizing explicit knowledge (Jiang, 2004, 2007). Therefore, a substantial advantage of the techniques described in this book is that they can minimize the use of metalinguistic knowledge and provide a more accurate insight into L2 learners' implicit knowledge and processing (provided, of course, that any distracter or secondary tasks do not prompt learners to draw upon metalinguistic knowledge). Unfortunately, however, with the exception of a handful of studies (e.g., Morgan-Short, Sanz, Steinhauer, & Ullman, 2010; Smith & VanPatten, in press), the use of on-line methods to investigate the effects of instruction has been mostly absent. A fruitful area of investigation, therefore, could be whether instruction type can impact L2 learners' on-line sensitivity to violations of a particular form or structure, as measured by one of the methods described in this volume. Obviously, as with any instructed SLA study, research employing these methods will also need to adequately motivate the targeted linguistic structure, as well as the type and length of treatment.

Attention and Awareness

Assuming that input is the primary ingredient for L2 development, one of the issues confronting instructed SLA research for over two decades is how instruction can manipulate input and push learners to process input correctly in order to maximize acquisition. To this end, a variety of input enhancement techniques have been proposed, all of which attempt to direct learners' attention to a relevant grammatical form while their attention is also on processing meaning from the input (Sharwood Smith, 1993). These techniques include input flood, text enhancement, structured input, and grammar consciousness-raising tasks (see Wong, 2005 for a full description of each). Theoretical perspectives on the role of attention in SLA have provided much of the motivation for input enhancement research. For example, Schmidt (1990, 1995, 2001) argues that the crucial construct for the conversion of input to intake is *noticing,* which he defines as a "conscious registration of some event" such as "surface level phenomena" (Schmidt, 1995, p. 29). In this sense, attention is equated with conscious awareness, permitting verbalization about an experience.

In contradistinction to the view that attention at the level of conscious awareness is necessary for input processing, Tomlin and Villa (1994) argue that of the three attentional components—alertness, orientation, and detection—*detection,* which does not require conscious awareness, is the crucial component involved in acquisition. The other attentional components of alertness and orientation serve to increase the chances of a form being detected. Along the same lines, Carroll (1999) has argued that "we detect, encode and respond to properties of sounds that correspond to distinctions that we are not aware of and cannot comment on" (p. 354; see also Truscott, 1998).

The majority of SLA studies investigating issues of attention and awareness in SLA have focused on learners' conscious awareness of linguistic form and utilize concurrent or retrospective verbal reports to measure learners' noticing. In concurrent reports, learners are instructed to say aloud whatever comes to mind (in their L1 or L2) about what they are attending to *while* reading a text in the L2 or engaging in some kind of language task. In retrospective reports, participants verbalize their thoughts *after* a given task. For both types of reports, the protocols are transcribed, coded, and analyzed in order to determine whether learners demonstrate a particular level of awareness about targeted linguistic forms. The data generated from the protocols is often compared with pretest-posttest data to examine the role of awareness in learning the targeted forms (see Bowles, 2010, for an in-depth look at the validity, uses, and implications of verbal reports in L2 research).

The use of concurrent and retrospective reports has contributed valuable understanding into learners' cognitive processes during input processing, in spite of criticisms regarding the reactivity of concurrent verbal reports (i.e., instructing participants to verbalize their thoughts may alter the original task). A major limitation, however, of these reports is that they only reflect what learners choose to verbalize. In other words, data gathered from verbal reports may only represent a subset of what they notice in the input. Furthermore, they can only measure learners' conscious awareness of forms, but not what they cognitively register or detect outside of awareness.

The psycholinguistic methods presented in this volume can overcome two important limitations of verbal reports. First, they can measure what learners attend to, regardless of whether they are consciously aware of it. Second, by not providing the additional task of verbalizing one's thoughts while processing input, they can avoid the reactivity criticism. Any of the techniques examining processing times can reveal which forms in the input learners attend to more or less often under different input conditions (flooded, enhanced, structured, and so forth), and whether a relationship exists between processing time on specific forms and subsequent learning of those forms.

Processing Instruction

Another promising area for which psycholinguistic methods can be applied to instructed SLA research is with processing instruction. Processing instruction (PI) is an instructional technique that seeks to alter or improve learners' nonoptimal processing strategies so that they are more likely to make correct form-meaning connections during comprehension (e.g., VanPatten 1996, 2004). PI is based on the principles of input processing (VanPatten, 1996, 2004, 2007), which is a model that attempts to capture (a) why learners make some form-meaning connections and not others, and (b) how learners assign syntactic roles to nouns in sentences. For example, according to the Lexical Preference Principle, learners "will process lexical items for meaning before grammatical forms when both encode the same semantic information" (VanPatten, 2007, p. 118). In other words, in a sentence such as "*John studied in the library last night*," learners are more likely to understand that the action occurred in the past from the temporal adverb *last night* than from the tense morphology on the verb.

Keeping in mind principles such as lexical preference, PI was developed to help learners overcome this and other nonoptimal processing strategies, so that learners are more likely to make correct form-meaning connections in the input. PI aims to achieve this by providing learners with explicit information about the particular grammatical form. Next, learners receive information about processing itself, such as reminding learners that although adverbials like *last night, yesterday,* and so forth are good indicators that an action has taken place in the past, they are not always present in sentences; therefore, it is important to recognize past tense verb forms. Finally, learners complete structured input (SI) activities. The input in these activities is "structured" or manipulated in such a way that the learner must correctly interpret the target form in order to understand/complete the activity. Learners do not produce the target form; instead, they are provided with a series of activities in which comprehension of the target form is necessary to interpret the meaning. In the case of past tense morphology, activities are created in which adverbials of time are omitted from sentences so that learners must understand present, past, and future time reference from verb endings. For example, learners might hear "*My sister called me,*" and they would have to select a word that best fits with the sentence such as *last night, these days,* or *tomorrow.* (see, e.g., Lee & VanPatten, 2003; VanPatten, 1996).

Classroom PI studies generally include a sentence-level interpretation task that is based on a structured input activity performed by the PI group (such as "select the word that best fits the sentence you hear" from the previous example). In addition, participants complete a production test based on an activity performed by a production-based instruction group (e.g., "complete each sentence with the correct form of the verb"). For more than twenty years, research has consistently demonstrated that learners who receive PI show significant improvement in their interpretation and production of targeted forms, even though they are never required to produce those forms during the treatment. Furthermore, PI research to date has shown that it is consistently superior to traditional production-based instruction (TI) that uses mechanical, meaningful, and communicative drills. Although the effectiveness of PI has been well-established in a variety of languages (see, e.g., Benati & Lee, 2008, for an overview of these studies), one important direction for PI research is to use methods presented in this book to investigate the effects of PI on learners' moment by moment processing and their processing of input that is not structured. The rationale for examining this stems from the goal of PI itself: to alter inefficient processing strategies and to instantiate more optimal strategies for processing input. Furthermore, PI is informed by an input processing model, which "is a model of moment-by-moment sentence processing during comprehension and how learners connect or don't connect particular forms with particular meanings" (VanPatten, 2007, p. 116). To date, however, PI research has examined learners' final interpretation of utterances using off-line interpretation tasks. If we return to the PI treatment example targeting tense morphology, we can pose the following questions: What happens when learners are exposed to input containing adverbials of time? Do they still attend to the tense morphology or do they go back to relying primarily on lexical items (last week, next year, etc.) to understand temporal reference? An experiment using self-paced listening/reading or eye-tracking as a pre- and posttest measure could compare participants' processing times on sentence conditions in which adverbials of time match temporal morphology, as in (4a), and those for which there is a mismatch, as in (4b).

(4)

a. My sister called me yesterday to wish me a happy birthday.
b. *My sister calls me yesterday to wish me a happy birthday.

If learners are not processing the temporal information encoded on the verb, we would expect no differences in processing times on the adverbial (or on the verbs if the adverbials precede the verbs). However, if learners are no longer relying on a lexical preference strategy and processing the temporal morphology, posttest measures should reveal longer processing times on those critical regions. Studies of this kind could represent an important step forward in investigating the ways in which PI alters learners' on-line processing of input.

Conclusion

In this chapter, I have pointed out the primary areas of L2 research in which psycholinguistic methods have been used, and I have suggested other ways these techniques can be employed to address other critical areas of L2 processing. I have also discussed methodological issues that should be considered regardless of the method used. Given the recent growth of psycholinguistic methods in L2 research, this volume should serve not only as a guide for students and researchers utilizing these methods but also as a catalyst for researchers who have not yet used them to consider how they can apply these methods to their own research on L2 acquisition and processing.

Discussion Questions

1) If you were to investigate the role of L1 influence in L2 processing, what structures or processing problems would you examine? Select two and present them for discussion, explaining why you would use those two structures/problems and how they would elucidate the question of L1 influence.
2) The question of native-like attainment begs the question of whether L1 speakers process language in uniform ways (both as individuals and as a group). Unlike studies of representation, where native speakers seem to converge, some studies on language processing show wider variation among native speakers. Without knowing any of the literature, can you think of reasons why L1 speakers might perform more similarly to each other on tests of representation as opposed to measures of processing? Keep in mind some of the methods you have read about in this book.
3) An important issue raised in this chapter is the nature of tasks and the instructions given to participants and how such things affect experimental outcomes. In reading original research, then, it is critical to be able to discern exactly what the researchers did and how the participants understood the task. Select two of the exemplary studies presented in this volume and read in detail the methods and procedures section.
   a. Upon reading these sections, would you have enough information to conduct a replication study? What additional information do you need?
   b. Can you make a determination, based on these methods and procedures sections, about what changes you might make in a replication study to see if task or instructions make a difference in research outcomes?
4) Explain why, in psycholinguistic and neurolinguistic research, we would want to minimize the influence of explicit knowledge during processing experiments.
5) One suggestion in this chapter regarding instructed SLA research is that investigators consider using some of the methods used in the present volume as a way to avoid issues related to production, introspection, judgment tasks, verbal

reports, and other data collection devices. Two studies were cited that do this: Morgan-Short, Steinhauer, Sanz, and Ullman (2010) and Smith and VanPatten (in press). Find an instructed SLA study that uses an off-line measure of knowledge and develop a replication in which you use an on-line method. How would you justify this change if you were writing up the study?

References

Benati, A. G., & Lee, J. F. (2008). *Grammar Acquisition and Processing Instruction: Secondary and Cumulative Effects.* Buffalo, NY: Multilingual Matters.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence.* New York: Springer-Verlag.

Bowles, M. A. (2010). *The think-aloud controversy in second language research.* New York: Routledge.

Carreiras, M., & Clifton, C. (2004). On the on-line study of language comprehension. In M. Carreiras & C. Clifton (Eds.), *The on-line study of sentence comprehension* (pp. 1–14). Brighton, UK: Psychology Press.

Carroll, S. E. (1999). Putting 'input' in its proper place. *Second Language Research, 15*(4), 337–388.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*(1), 3–42.

Cuetos, F., & Mitchell, D.C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition, 30*(1), 73–105.

de Groot, A.M. B. (2011). *Language and Cognition in Bilinguals and Multilinguals: An introduction.* New York: Taylor and Francis.

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Malden, MA: Blackwell.

Doughty, C. J. (2003). Instructed SLA: Constrains, compensation, and enhancement. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 256–310). Malden, MA: Blackwell.

Dussias, P. (2003). Syntactic ambiguity resolution in second language learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition, 25*(4), 529–557.

Dussias, P., Valdés Kroff, J., Guzzardo Tamargo, R. E., & Gerfen, C. (2013). Looking into comprehension of Spanish-English code-switched sentences: Evidence from eye movements. *Studies in Second Language Acquisition 35,* 353–387.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141–172.

Felser, C., & Cunnings, I. (2012). Processing reflexives in a second language: The role of structural and discourse-level constraints. *Applied Psycholinguistics, 33*(4), 571–603.

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research, 23*(1), 9–36.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading.* London, UK: Pearson Education Longman.

Hahne, A., & Friederici, A.D. (2002). Differential task effects on semantic and syntactic processes as revealed by ERPs. *Cognitive Brain Research, 13*(3), 339–356.

Jackson, C. N. (2010). The processing of subject-object ambiguities by English and Dutch L2 learners of German. In B. VanPatten & J. Jegerski (Eds.), *Second language processing and parsing: issues in theory and research* (pp. 207–230). Amsterdam: John Benjamins.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*(4), 603–634.

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning, 57*(1), 1–33.

Jiang, N. (2012). *Conducting reaction time research in second language studies.* New York: Routledge.

Juffs, A. (2001). Psycholinguistically oriented second language research. *Annual Review of Applied Linguistics, 21,* 207–220.

Juffs, A. (2005). The influence of first language on the processing of *wh*-movement in English as a second language. *Second Language Research, 21*(2), 121–151.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixation to comprehension. *Psychological Review, 87,* 329–354.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning, 59*(3), 503–535.

Lee, J. F., & VanPatten, B. (2003). *Making communicative language teaching happen* (2nd ed.). New York: McGraw-Hill.

Leeser, M., Brandl, A., & Weissglass, C. (2011). Task effects in second language sentence processing research. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching, and research: Insights from psycholinguistics* (pp. 179–198). Amsterdam: John Benjamins.

Linck, J., Kroll, J. F., & Sunderman, G. (2009). Losing access to the native language while immersed in a second language: Evidence for the role of inhibition in second language learning. *Psychological Science, 20*(12), 1507–1515.

Love, T., Swinney, D., Haist, F., Nicol, J., & Buxton, R. (2003). Task-demand modulation of activation in Broca's area. *Brain and Language, 87*(1), 77–78.

McLaughlin, J., Osterhout, L., & Kim, A. (2004). Neural correlates of second-language word learning: Minimal instruction produces rapid change. *Nature Neuroscience, 7,* 703–704.

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullman, M. (2010). Acquisition of gender agreement in second language learners: An event-related potential study. *Language Learning, 60*(1), 154–193.

Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning, 50*(4), 417–528.

O'Grady, W. (in press). The illusion of language acquisition. *Linguistic Approaches to Bilingualism.*

Osterhout, L., McLaughlin, J., Pitkänen, I., Frenck-Mestre, C., & Molinaro, N. (2006). Novice learners, longitudinal designs, and event-related potentials: A means for exploring the neurocognition of second language processing. *Language Learning, 56*(1), 199–230.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition, 24,* 501–528.

Papangeli, A. (2010). *Language development and on-line processing in L1 and L2 children.* (Unpublished doctoral dissertation). University of Reading, Reading, UK.

Pritchett, B. L. (1992). *Grammatical competence and parsing performance.* Chicago: The University of Chicago Press.

Saur, D., Baugmgaertner, A., Moehring, A., Büchel, C., Bonnesen, M., Rose, M., Musso, M., & Meisel, J. M. (2009). Word order processing in the bilingual brain. *Neuropsychologia, 47*(1), 158–168.

Schmidt, R. (1990). The role of consciousness in L2 learning. *Applied Linguistics, 11*(2), 129–158.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu: HI: University of Hawaii.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology.* Chicago: University of Chicago Press.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition, 15*(2), 165–179.

Smith, M., & VanPatten, B. (in press). Instructed SLA as parameter setting: Evidence from earliest-stage learners of Japanese as L2. In C. Laval, M. J. Arche, & A. Benati (Eds.), *The grammar dimension in instructed second language acquisition: theory, research, and practice.* London: Continuum Press.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263–308.

Swinney, D. (2000). Understanding the behavioral-methodology/language-processing interface. *Brain and Language, 71*(1), 241–244.

Taraban, R., Rynearson, K., & Kerr, M. (2000). College students' academic performance and self-reports of comprehension strategy use. *Reading Psychology, 21*(4), 283–308.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition, 27*(2), 173–204.

Tomlin, R., & Villa, H. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition, 16*(2), 183–203.

Truscott, J. (1998). Noticing in second language acquisition: A critical review. *Second Language Research, 14*(2), 103–135.

VanPatten, B. (1996). *Input processing and grammar instruction: Theory and research.* Norwood, NJ: Ablex.

VanPatten, B. (2004). Input processing in SLA. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–32). Mahwah, NJ: Erlbaum.

VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 115–135). Mahwah, NJ: Erlbaum.

VanPatten, B., Borst, S. Collopy, E., Qualin, A., & Price, J. (2013). Explicit information, grammatical sensitivity, and the First-noun Principle: A cross-linguistic study in processing instruction. *The Modern Language Journal, 97,* 506–527.

Wong, W. (2005). *Input enhancement: From theory and research to the classroom.* New York: McGraw-Hill.

This page intentionally left blank

INDEX