

Research Methods in Linguistics

A comprehensive guide to conducting research projects in linguistics, this book provides a complete training in state-of-the-art data collection, processing, and analysis techniques. The book follows the structure of a research project, guiding the reader through the steps involved in collecting and processing data, and providing a solid foundation for linguistic analysis. All major research methods are covered, each by a leading expert. Rather than focusing on narrow specializations, the text fosters inter-disciplinarity, with many chapters focusing on shared methods such as sampling, experimental design, transcription, and constructing an argument. Highly practical, the book offers helpful tips on how and where to get started, depending on the nature of the research question. The only book that covers the full range of methods used across the field, this student-friendly text is also a helpful reference source for the more experienced researcher and current practitioner.

ROBERT J. PODESVA is an Assistant Professor in the Department of Linguistics at Stanford University.

DEVYANI SHARMA is a Senior Lecturer in Linguistics at Queen Mary University of London.

Cambridge University Press
978-1-107-01433-6 - Research Methods in Linguistics
Edited by Robert J. Podesva and Devyani Sharma
Frontmatter
[More information](#)

Cambridge University Press
978-1-107-01433-6 - Research Methods in Linguistics
Edited by Robert J. Podesva and Devyani Sharma
Frontmatter
[More information](#)

Research Methods in Linguistics

EDITED BY
ROBERT J. PODESVA
Stanford University
AND
DEVYANI SHARMA
Queen Mary University of London



Cambridge University Press
978-1-107-01433-6 - Research Methods in Linguistics
Edited by Robert J. Podesva and Devyani Sharma
Frontmatter
[More information](#)

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Published in the United States of America by Cambridge University Press, New York
Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781107696358

© Cambridge University Press 2013

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 2013

Printing in the United Kingdom by TJ International Ltd. Padstow Cornwall

A catalogue record for this publication is available from the British Library

ISBN 978-1-107-01433-6 Hardback

ISBN 978-1-107-69635-8 Paperback

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication,
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Contents

<i>List of figures</i>	<i>page</i> vii
<i>List of tables</i>	xii
<i>List of contributors</i>	xiv
<i>Acknowledgments</i>	xvi
1 Introduction	
<i>Devyani Sharma and Robert J. Podesva</i>	1
PART I DATA COLLECTION	9
2 Ethics in linguistic research	
<i>Penelope Eckert</i>	11
3 Judgment data	
<i>Carson T. Schütze and Jon Sprouse</i>	27
4 Fieldwork for language description	
<i>Shobhana Chelliah</i>	51
5 Population samples	
<i>Isabelle Buchstaller and Ghada Khattab</i>	74
6 Surveys and interviews	
<i>Natalie Schilling</i>	96
7 Experimental research design	
<i>Rebekha Abbuhl, Susan Gass, and Alison Mackey</i>	116
8 Experimental paradigms in psycholinguistics	
<i>Elsi Kaiser</i>	135
9 Sound recordings: acoustic and articulatory data	
<i>Robert J. Podesva and Elizabeth Zsiga</i>	169

vi	Contents	
10	Ethnography and recording interaction <i>Erez Levon</i>	195
11	Using historical texts <i>Ans van Kemenade and Bettelou Los</i>	216
	PART II DATA PROCESSING AND STATISTICAL ANALYSIS	233
12	Transcription <i>Naomi Nagy and Devyani Sharma</i>	235
13	Creating and using corpora <i>Stefan Th. Gries and John Newman</i>	257
14	Descriptive statistics <i>Daniel Ezra Johnson</i>	288
15	Basic significance testing <i>Stefan Th. Gries</i>	316
16	Multivariate statistics <i>R. Harald Baayen</i>	337
	PART III FOUNDATIONS FOR DATA ANALYSIS	373
17	Acoustic analysis <i>Paul Boersma</i>	375
18	Constructing and supporting a linguistic analysis <i>John Beavers and Peter Sells</i>	397
19	Modeling in the language sciences <i>Willem Zuidema and Bart de Boer</i>	422
20	Variation analysis <i>James A. Walker</i>	440
21	Discourse analysis <i>Susan Ehrlich and Tanya Romaniuk</i>	460
22	Studying language over time <i>Hélène Blondeau</i>	494
	<i>Index</i>	519

Figures

3.1	An example of a two-alternative forced-choice task	page 32
3.2	An example of the yes-no task	32
3.3	An example of a Likert scale task	33
3.4	An example of the magnitude estimation task	34
7.1	4 × 4 Latin squares design	121
8.1	Visual-world eye-tracking graph showing the probability of fixating objects on the screen (0 ms = onset of the critical word, e.g., <i>beaker</i>). Allopenna, Magnuson, and Tanenhaus 1998 (see Acknowledgments for full copyright information)	145
8.2	Examples of object-array displays. Allopenna, Magnuson, and Tanenhaus 1998; Trueswell et al. 1999; Brown-Schmidt and Konopka 2008; Sussman 2006 (see Acknowledgments for full copyright information)	146
8.3	Examples of clip-art displays. Kamide, Altmann, and Haywood 2003; Weber, Grice, and Crocker 2006; Kaiser 2011a; Arnold et al. 2000 (see Acknowledgments for full copyright information)	150
8.4	Poor man’s eye-tracking, “Tickle the frog with the feather.” Snedeker and Trueswell 2004 (see Acknowledgments for full copyright information)	154
9.1	Common microphone mounts: stand-mounted (left), head-mounted (middle), and lavalier (right)	174
9.2	Microphone jacks: XLR (left), mini-stereo (middle), and USB (right)	174
9.3	Solid state recorders: Marantz PMD660 (left) and Zoom H2n (right)	177
9.4	Range of data collection scenarios	180
9.5	Lip position for [ɸ] (left) and [sʷ] (right) in Sengwato	182
9.6	Palatogram (left) and linguogram (right) of American English /t/	183
9.7	Artificial palate with embedded electrodes (left); sample patterns for /s/ and /t/ (right). http://speech.umaryland.edu/epg.html (left); www.rds-sw.nihr.ac.uk/success_stories_lucy_ellis.htm (right)	184
9.8	Subject holding a sonograph transducer (top); sonograph image for the vowel /i/ (bottom). Gick 2002	185

viii	List of figures	
9.9	Example of an EGG waveform during modal voicing	187
9.10	Using a pressure/airflow mask (top); trace of pressure at the lips during [aɸa] (bottom)	188
9.11	Pictures of abducted (left) and adducted (right) vocal folds, taken via flexible endoscope. http://voicedoctor.net/media/normal-vocal-cord	189
9.12	MRI image of Portuguese [ã]. Martins et al. 2008	190
9.13	EMMA apparatus (top); ample movement trace (bottom). http://beckman.illinois.edu/news/2007/10/100307 (top); Fagel and Clemens 2004 (bottom)	191
9.14	EMG trace (solid line) shows a burst of activity in the cricothyroid muscle during pitch raising (dotted line) in Thai falling and rising tone. Erickson 1976	192
11.1	Demonstrative elements in dislocates. Los and Komen 2012	219
13.1	Markup in the TEI Header of file A01 in the XML Brown Corpus	265
13.2	The first sentence (and paragraph) in the text body of file A01 in the XML Brown Corpus (the tags beginning with p, s, and w mark the paragraph, sentence, and each word respectively)	266
13.3	The annotation of <i>in terms of</i> as a multi-word unit in the BNC XML	267
13.4	Two ways of representing the dispersion of a word (<i>perl</i>) in a file	276
13.5	Python session illustrating some functions in NLTK	281
13.6	R session to create a frequency list of a file from the Brown Corpus and the resulting plots	282
14.1	Stem-and-leaf plot of daily temperatures for Albuquerque in 2010	291
14.2	Histogram of Albuquerque temperatures in 2010	293
14.3	Histogram of men's and women's mean F0. Johnson, based on Peterson and Barney 1952	293
14.4	Histogram of men's and women's natural-log-transformed F0. Johnson, based on Peterson and Barney 1952	294
14.5	Three normal distributions: mean = 0, standard deviations = {0.5, 1, 2}	294
14.6	Histogram of 2009 household income, with central tendencies labeled	296
14.7	Dispersion of Peterson and Barney F0 for men and women	300
14.8	Plot of 2010 Albuquerque temperatures, by date	303
14.9	Plot of 2006–2010 Albuquerque temperatures, by date	304
14.10	F2 vs F0 for <i>heed</i> in Peterson and Barney data	304
14.11	The relationship between Pearson's r and Spearman's rho	306
14.12	Counts and proportions of quotative variants in 2006 York corpus	308

	List of figures	ix
14.13	Distribution of 335 ratings for “Mary has had more drinks than she should have done so” (0 = completely impossible, 10 = perfectly natural)	309
14.14	Distribution of 335 ratings for “Who did John see George and?” (0 = completely impossible, 10 = perfectly natural)	309
14.15	Mosaic plot of York quotative variants by grammatical person	312
15.1	Probability distributions for outcomes of equally likely binary trials. Three, six, and twelve trials (top row); twenty-five, fifty, and one hundred trials (bottom row)	319
15.2	A normal distribution (left panel); an exponential distribution (right panel)	323
15.3	Mosaic plot for the data in <i>walk</i>	326
15.4	Box plot of the Dice coefficients for the two subtractive word-formation processes	331
15.5	Graphical representation of the differences between <i>before</i> and <i>after</i>	334
16.1	A regression line (left) and a factorial contrast between a reference group mean <i>a</i> on the intercept and a group mean <i>b</i> . The difference between the two group means, the contrast, is equal to the slope of the line connecting <i>a</i> and <i>b</i> : 2. Both the regression line and the line connecting the two group means are described by the line $y = 1 + 2x$	340
16.2	Multiplicative interactions in the linear model	342
16.3	Example of an interaction of a factor and a covariate in an analysis of covariance	343
16.4	Tukey all-pairs confidence intervals for contrasts between mean pitch for different branching conditions across English tri-constituent compounds	346
16.5	Correlation of the by-word random intercepts and the by-word random slopes for Sex=male in the linear mixed-effects model fitted to the pitch of English tri-constituent compounds	353
16.6	Random effects structure for subject. Correlations of the BLUPs (upper panels); correlations of the by-subject coefficients (lower panels)	354
16.7	Fitted smooths (with 95 percent confidence intervals) for Pitch as a function of Time for the four branching conditions of the pitch dataset of English tri-constituent compounds	358
16.8	Tensor product for the interaction of Time by Danger Rating Score at channel FC2	359
16.9	The pronunciation distance from standard Dutch for different quantiles of word frequency	360
16.10	The probability of using <i>was</i> as a function of Age, Adjacency and Polarity	363

x	List of figures	
16.11	Recursive partitioning tree for the Russian goal/theme data	365
16.12	The ndl network for the Finnish <i>think</i> verbs. Darker shades of grey indicate stronger positive connections, lighter shades of grey larger negative connections. For the abbreviations in the nodes, see Table 16.14	368
17.1	Waveform of several periods of the Dutch vowel /i/, illustrating glottal fold vibration	376
17.2	Waveform of several periods of the Dutch vowel /i/, illustrating the first formant	377
17.3	Waveform of several periods of the Dutch vowel /i/, illustrating the second formant	378
17.4	Waveform of several periods of the Dutch vowel /a/, illustrating mangled formants	378
17.5	Waveform of a whole Dutch /i/, illustrating duration and intensity	379
17.6	Waveform of the voiceless palatal plosive in [aca], illustrating silence and release burst	380
17.7	Waveform of the voiceless palatal fricative in [aça], illustrating the many zero crossings	380
17.8	Waveform of the alveolar trill in [ara], illustrating four passive tongue-tip closures	381
17.9	Determining the pitch of the sound in Figure 17.1 at a time of 0.3002 seconds (cross-correlation method). The top row shows parts of the sound just before that time, and the bottom row shows equally long parts just after. The two parts look most similar if they are 7.1 ms long	382
17.10	Pitch curve for the [i] vowel of Figure 17.5	383
17.11	Determining the intensity curve for the [i] vowel of Figure 17.5: (a) the original sound, as measured relative to the auditory threshold; (b) the square of this; (c) the Gaussian smoothing kernel, on the same time scale as the sound; (d) the intensity curve, computed as the convolution of the squared amplitude and the Gaussian; (e) the intensity curve along a logarithmic scale	385
17.12	Splitting up two periods of the [i] vowel of Figure 17.5 into six harmonics. At the top is the original sound. The rough features of the original sound are reconstructed by adding the first harmonic (1) and the second harmonic (2) to each other (1+2). When we add the 15th, 22nd, 23rd, and 24th harmonics to this, the original waveshape is approximated even more closely (bottom)	387
17.13	The Fourier spectrum of the two-period [i]-like sound of Figure 17.12	388
17.14	Spectrum of the vowel [i]	389

	List of figures	xi
17.15 Spectrogram of the vowels [a], [i], and [u]	390	
17.16 Spectrogram of sibilants	392	
17.17 Spectrogram of [aca], showing the four acoustic correlates of the plosive	392	
17.18 Spectrogram of [ara]	393	
17.19 Automated formant measurement in the vowels [a], [i], and [u], superimposed on the spectrogram of Figure 17.15	394	
19.1 Classes of representation of language	432	
20.1 Excerpt from coding instructions for the English future	448	
20.2 Fragment of an Excel coding sheet and GoldVarb token file for the coding of the English future	450	
22.1 Apparent-time distribution at Time 1	507	
22.2 Real-time distribution at Time 1 and Time 2: age-grading interpretation	507	
22.3 Real-time distribution at Time 1 and Time 2: community change interpretation	508	
22.4 Stability over time for two speakers	510	
22.5 Change over time for two speakers	510	

Tables

5.1	Relationship between sample size and sampling error. De Vaus 2001	<i>page</i> 82
5.2	The database for Spanish second language acquisition. Mitchell et al. 2008	84
13.1	A subset of the Uppsala Learner English Corpus. Adapted from Table 1 in Johansson and Geisler 2011	260
13.2	Four tagging solutions for English <i>rid</i>	267
13.3	Sub-corpora of the Brown written corpus	270
13.4	Sub-corpora of the ICE corpora	271
13.5	Sub-corpora of the MICASE spoken corpus	272
13.6	Sub-corpora of the BNC	272
13.7	Sub-corpora of the written component of COCA, as of April 2011	273
13.8	Frequency lists: words sorted according to frequency (left panel); reversed words sorted alphabetically (center panel); 2-grams sorted according to frequency (right panel)	275
13.9	Excerpt of a collocate display of <i>general/generally</i>	277
13.10	Excerpt of a concordance display of <i>alphabetic</i> and <i>alphabetical</i>	278
13.11	Examples of regular expressions	278
14.1	Frequency table of daily temperatures for Albuquerque in 2010	292
14.2	Cross-tabulations for survival vs sex and survival vs age on the <i>Titanic</i>	311
14.3	Cross-tabulation of York quotative variants by grammatical person, observed	311
14.4	Cross-tabulation of York quotative variants by grammatical person, expected (if no association)	312
15.1	All possible results from asking three subjects to classify <i>walk</i> as a noun or a verb	318
15.2	Fictitious data from a forced-choice part-of-speech selection task	325
15.3	Dice coefficients of source words for complex clippings and blends	328
16.1	A multivariate dataset with <i>n</i> cases (rows) and <i>k</i> variables (columns)	338

	List of tables	xiii
16.2	An example of treatment dummy coding for two-way analysis of variance	341
16.3	Predicted group means given the dummy coding in Table 16.2 and regression equation (2)	341
16.4	Predicted group means for the data in Table 16.2 given regression equation (4)	342
16.5	An example of treatment dummy coding for an analysis of covariance with an interaction	343
16.6	Four kinds of compound stress patterns in English tri-constituent compounds	344
16.7	Coefficients of an analysis of covariance model fitted to the pitch of English tri-constituent compounds	345
16.8	Sequential model comparison for Pitch in English tri-constituent compounds	347
16.9	A repeated measures dataset with <i>gn</i> cases with observations on <i>k</i> variables collected for <i>n</i> items and <i>g</i> subjects	350
16.10	Notation for adjustments to intercept and predictors	351
16.11	Standard deviations and correlation parameter for the random-effects structure of the mixed-effects model fitted to the pitch of English tri-constituent compounds	352
16.12	Model comparison for a series of models with increasing nonlinear structure fitted to the pitch dataset	356
16.13	Log odds for four Finnish near-synonyms meaning <i>think</i>	364
16.14	Naive discrimination learning weights for four Finnish near-synonyms for <i>think</i>	368
20.1	Factors contributing to the occurrence of the alveolar variant <i>-in</i> ' in Toronto English	452
22.1	Indirect and direct approaches to time	496
22.2	The pseudo-longitudinal effect in SLA	498

Contributors

- REBEKHA ABBUHL
California State University, Long Beach, US
- R. HARALD BAAYEN
Eberhard Karls University, Tübingen, Germany, and University
of Alberta, Canada
- JOHN BEAVERS
The University of Texas at Austin, US
- HÉLÈNE BLONDEAU
University of Florida, US
- PAUL BOERSMA
University of Amsterdam, Netherlands
- ISABELLE BUCHSTALLER
Leipzig University, Germany
- SHOBHANA CHELLIAH
University of North Texas, US
- BART DE BOER
Vrije Universiteit Brussel, Belgium
- PENELOPE ECKERT
Stanford University, US
- SUSAN EHRLICH
York University, Canada
- SUSAN GASS
Michigan State University, US
- STEFAN TH. GRIES
University of California, Santa Barbara, US
- DANIEL EZRA JOHNSON
Lancaster University, UK
- ELSI KAISER
University of Southern California, US

- GHADA KHATTAB
Newcastle University, UK
- EREZ LEVON
Queen Mary University of London, UK
- BETTELOU LOS
Radboud University, Netherlands
- ALISON MACKEY
Georgetown University, US
- NAOMI NAGY
University of Toronto, Canada
- JOHN NEWMAN
University of Alberta, Canada
- ROBERT J. PODESVA
Stanford University, US
- TANYA ROMANIUK
Portland State University, US
- NATALIE SCHILLING
Georgetown University, US
- CARSON T. SCHÜTZE
University of California, Los Angeles, US
- PETER SELLS
University of York, UK
- DEVYANI SHARMA
Queen Mary University of London, UK
- JON SPROUSE
University of Connecticut, US
- ANS VAN KEMENADE
Radboud University, Netherlands
- JAMES A. WALKER
York University, Canada
- WILLEM ZUIDEMA
University of Amsterdam, Netherlands
- ELIZABETH ZSIGA
Georgetown University, US

Acknowledgments

This book has been a truly collaborative enterprise. It could never have been produced without the expertise and dedication of our contributing authors, to whom we owe our greatest debt. In our effort to foster dialogue across the subdisciplines of our field, we have asked contributors to take a broad perspective, to reflect on issues beyond their areas of particular specialization, and to neatly package their ideas for a diverse readership. In rising to meet this challenge, authors have consulted scholars and readings that interface with their own areas of expertise, endured a lengthy external review and extensive revision process, and in all cases produced chapters that we think will be useful to wide swaths of researchers. We thank the authors for their significant contributions.

The initial impetus for this book came from our students, who ask all the right questions about data (who? when? how much?) and analysis (why? how?). We hope they find answers and new questions in these pages. We are also indebted to five anonymous reviewers who, at an early stage of this project, affirmed the usefulness of the proposed collection and made crucial recommendations regarding its scope, structure, and balance of coverage.

For their expert advice and truly generous contributions, we thank an army of reviewers and advisors, none of whom of course bears any responsibility for the choices ultimately made: David Adger, Paul Baker, Joan Beal, Claire Bower, Kathryn Campbell-Kibler, Charles Clifton, Paul De Decker, Judith Degen, Susanne Gahl, Cynthia Gordon, Matthew Gordon, Tyler Kendall, Roger Levy, John Moore, Naomi Nagy, Jeanette Sakel, Rebecca Scarborough, Morgan Sonderegger, Naoko Taguchi, Marisa Tice, Anna Marie Trester, and Alan Yu.

We would also like to acknowledge our departments: the Department of Linguistics at Stanford University and the Department of Linguistics at Queen Mary University of London. The range of methods represented in the work of our closest colleagues continues to inspire us and push our field forward. Thanks also to the Department of Linguistics at Georgetown University and the Department of English at the National University of Singapore, where we spent significant time during the production of this volume.

Helena Dowson, Fleur Jones, Gnanadevi Rajasundaram, Christina Sarigiannidou, Alison Tickner and the team at Cambridge University Press provided efficient and very patient support throughout the production schedule. Finally, special thanks to Andrew Winnard at Cambridge University Press for his encouragement and support. Like us, he recognized the many challenges of

developing such a project, but also shared our enthusiasm for its potential uses in a fast-developing field. We hope this book represents a proof of concept.

The editors and publisher acknowledge the following sources of copyright material reproduced in Chapter 8 and are grateful for the permissions granted:

Figure 8.1 Reprinted from *Journal of Memory and Language* 38, Allopenna, Magnuson, and Tanenhaus, Tracking the time course of spoken word recognition: evidence for continuous mapping models. Copyright 1998, with permission from Elsevier.

Figure 8.2 reprinted from:

(a) *Journal of Memory and Language* 38, Allopenna, Magnuson, and Tanenhaus, Tracking the time course of spoken word recognition: evidence for continuous mapping models. Copyright 1998, with permission from Elsevier.

(b) *Cognition* 73, Trueswell, Sekerina, Hill, and Logrip, The kindergarten-path effect: studying on-line sentence processing in young children, 89–134. Copyright 1999, with permission from Elsevier.

(c) *Cognition* 109, Brown-Schmidt and Konopka, Little houses and casas pequeñas: message formulation and syntactic form in unscripted speech with speakers of English and Spanish, 274–80. Copyright 2008, with permission from Elsevier.

(d) *Verb-Instrument Information During On-line Processing*, Rachel Sussmann, Copyright 2006, with permission from the author.

Figure 8.3 reprinted from:

(a) *Journal of Memory and Language* 49, Kamide, Altmann, and Haywood, Prediction and thematic information in incremental sentence processing: evidence from anticipatory eye movements, 133–56. Copyright 2003, with permission from Elsevier.

(b) *Cognition* 88, Weber, Grice, and Crocker, The role of prosody in the interpretation of structural ambiguities: a study of anticipatory eye movements, B63–B72. Copyright 2006, with permission from Elsevier.

(c) *Language and Cognitive Processes* 26, Kaiser, Consequences of subjecthood, pronominalisation, and contrastive focus, 1625–66. Copyright 2011, reprinted by permission of Taylor & Francis Ltd, www.tandf.co.uk/journals.

(d) *Cognition* 76, Arnold, Eisenband, Brown-Schmidt, and Trueswell, The rapid use of gender information: eyetracking evidence of the time-course of pronoun resolution, B13–B26. Copyright 2000, with permission from Elsevier.

Figure 8.4 Reprinted from *Cognitive Psychology* 49, Snedeker and Trueswell, The developing constraints on parsing decisions: the role of lexical-biases and referential scenes in child and adult sentence processing, 238–99. Copyright 2004, with permission from Elsevier.