

Analysing Sociolinguistic Variation

Now in its second edition, this is an invaluable manual for teaching and learning variation analysis, the quantitative study of linguistic variation and change. Written by a leading scholar in the field with over thirty years of experience, it provides an insider's view of the methodology through practical, 'hands-on' advice, including straightforward instructions for conducting analyses using the R programming language, the new gold standard for analysis. It leads readers through each phase of a research study based on data gathered in a sociocultural context, beginning with the selection and sampling of a data source, to hints on successful project design, interview techniques, data management, analysis and interpretation, with systematic procedures provided at each step of the process. This edition has been fully updated, with new insights and explanations in line with recent discoveries in the field, making it essential reading for anyone embarking on their own sociolinguistic research project.

Sali A. Tagliamonte is Professor of Linguistics at the University of Toronto, Canada Research Chair in Language Variation and Change, and Fellow of the Royal Society of Canada. Her recent publications include *Making Waves* (2016) and *Teen Talk* (2016). She is also the editor of the book series *Studies in Language Variation and Change*.

Analysing Sociolinguistic Variation

SECOND EDITION

Sali A. Tagliamonte
University of Toronto





Shaftesbury Road, Cambridge CB2 8EA, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University’s mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009403085
DOI: 10.1017/9781009403092

© Sali A. Tagliamonte 2006, 2025

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 & Assessment.

When citing this work, please include a reference to the DOI 10.1017/9781009403092

First published 2006
Third printing 2009
Second edition 2025

Printed in the United Kingdom by CPI Group Ltd, Croydon CR0 4YY

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

Library of Congress Cataloging-in-Publication Data

Names: Tagliamonte, Sali, author.
Title: Analysing sociolinguistic variation / Sali Tagliamonte, University of Toronto.
Description: Second edition. | Cambridge ; New York, NY : Cambridge University Press, 2025. | Includes
bibliographical references and index.
Identifiers: LCCN 2024047969 | ISBN 978 1009403085 (hardback) | ISBN 978 1009403108 (paperback) |
ISBN 9781009403092 (ebook)
Subjects: LCSH: Language and languages – Variation. | Sociolinguistics – Research.
Classification: LCC P120.V37 T34 2025 | DDC 306.44072–dc23/eng/20250220
LC record available at <https://lcn.loc.gov/2024047969>

ISBN 978-1-009-40308-5 Hardback
ISBN 978-1-009-40310-8 Paperback

Cambridge University Press & Assessment 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.

For Dazzian, Freya, Shaman, Tara, Adrian
And then Emily, Craig, Melissa, Thomas
And now also Aurin, Kieran, Caliope
With love,
Mum/Salimum/Granna

Contents

List of Figures	<i>page</i> viii
List of Tables	xi
Preface	xiii
1 Introduction	1
2 Data Collection	16
3 The Sociolinguistic Interview	34
4 Data Handling	47
5 The Linguistic Variable	65
6 Formulating Hypotheses/Operationalising Claims	90
7 Why Use Statistics?	113
8 Distributional Analysis	153
9 Exploratory Modelling	193
10 Mixed Effects Modelling	225
11 Visualisation	250
12 Interpreting and Reporting Your Results	302
13 Finding the Story	327
Bibliography	336
Index	353

Figures

7.1	Typical R Studio opening screen.	<i>page</i> 127
7.2	Excel data file containing the initial extraction of variable (hwat).	133
7.3	Filepath to create an RMD file in R Studio.	134
7.4	Set-up chunk for an RMD file in R Studio.	135
7.5	Coding string to add in useful helper functions.	136
7.6	Coding string to add in varimp_helper.	137
7.7	Screenshot of a summary of the data file for variable (hwat).	141
7.8	Screenshot of a summary of the data file for variable (adj_pos).	143
7.9	Excel data file with preceding and following context and other details.	145
7.10	Excel data file for import to R.	146
9.1	Conditional inference tree – hwat by year of birth and community.	197
9.2	Conditional inference tree – hwat by social factors.	199
9.3	Conditional inference tree – <i>great</i> by year of birth.	200
9.4	Conditional inference tree – <i>great</i> by year of birth, minbucket 500.	201
9.5	Conditional inference tree – <i>great</i> , ctree partitions for age group.	202
9.6	Conditional inference tree – <i>awesome</i> by decade of birth.	203
9.7	Conditional inference tree – hwat by age.	205
9.8	Conditional inference tree – seven-way adj by gender.	206
9.9	Conditional inference tree – seven-way adj_pos by gender, adjustments.	208
9.10	Conditional inference tree – hwat, social factors.	209
9.11	Conditional inference tree – hwat adjusted with maxdepth 4.	210
9.12	How to save a ctree to your computer.	211
9.13	Screenshot of hwat summary.	215
9.14	Random forest analysis of hwat – social factor groups.	217
9.15	Random forest analysis of <i>great</i> cartesian co-ordinates, default.	221
9.16	Random forest analysis of <i>great</i> cartesian co-ordinates, 0–1.	221
9.17	Random forest analysis of <i>lovely</i> cartesian co-ordinates, 0–1.	222
10.1	Conditional inference tree – <i>cool</i> by social and linguistic factors.	246
10.2	Random forest analysis of <i>cool</i> after 1971.	247
11.1	Number of tokens of hwat by decade.	251

	List of Figures	ix
11.2	Number of adjectives by decade.	252
11.3	Number of adjectives by gender.	253
11.4	Number of adjective variants by gender – seven-way.	254
11.5	Number of adjective tokens overall.	255
11.6	Rate of hwat variants by decade.	256
11.7	Rate of hwat variants by community and decade.	257
11.8	Rate of adjective variants by decade – seven-way dep_var with lines.	258
11.9	Rate of individual adjective variants by decade – facet_wrap.	260
11.10	Rate of hwat variants by community and adjective type.	262
11.11	Rate of hwat variants by community and education.	263
11.12	Number of tokens – adjective variants by gender and education.	264
11.13	Rate of variants – adjectives and gender, seven-way.	265
11.14	Conditional inference tree – hwat by community and year of birth.	267
11.15	Conditional inference tree – adj_pos by year of birth and gender.	268
11.16	Random forest analysis – hwat, all factor groups.	270
11.17	Random forest analysis – hwat factor groups – coord_cartesian.	272
11.18	Random forest analysis – adj_pos ‘other’ – coord_cartesian 0–1.	273
11.19	Random forest analysis – adj_pos variants, cow plot comparison, coord_cartesian at 0.5.	277
11.20	Effect of age by adjective type – <i>lovely</i> .	281
11.21	Effects plot – hwat by gender and occupation.	283
11.22	Effects plot – hwat by grammatical category and community.	284
11.23	Effects plot – hwat by grammatical category and community – lines.	285
11.24	Effects plot – adjective <i>cool</i> by type, intensifier, and gender.	286
11.25	Effects plot – adjective <i>cool</i> by type, intensifier, and time period.	287
11.26	Effects plot – adjective <i>cool</i> by time period, intensifier, and type – alternative view.	288
11.27	Ribbon plot – hwat by gender and year of birth.	290
11.28	Ribbon plots – hwat by gender and year of birth, cow plot of different effects.	291
11.29	Ribbon plot – hwat by year of birth – facet_wrap community.	293
11.30	Ribbon plot – adjective <i>great</i> by type and age – facet_wrap gender.	295
11.31	Ribbon plot – adjective <i>great</i> by type and age – facet_wrap intensifier.	297
11.32	Box plot – rate of hwat by individual and community.	299
11.33	Box plot – rate of <i>awesome</i> by individual and decade.	300

x	List of Figures	
12.1	Conditional inference tree of hwat by community and year of birth.	310
12.2	Conditional inference tree of hwat without community.	310
12.3	Comparison of syntactic position in the UK and Canada – <i>obviously</i> .	312
12.4	Comparative random forest analysis of nine varieties of world Englishes – genitive.	316

Tables

2.1	Template for sample design.	<i>page</i> 28
2.2	Sample design for L1 English vs L2 (or more) English.	29
2.3	Sample design for ethnicity and generation.	29
4.1	Automated retrieval of individual information and characteristics.	50
4.2	Excerpt of an index.	61
8.1a	Rate of [x] by polarity.	158
8.1b	Proportion of negative contexts over time.	158
12.1	Generalised linear model of factors contributing to the probability of zero plural.	304
12.2	Variable (-ing) York English – grammatical category.	307
12.3	Variable (-ing) York English – occupation.	307
12.4	Variable (t,d) York English – occupation.	307
12.5	Variable (t,d) York English – phonological segment and occupation.	308
12.6	Variable (t,d) York English – grammatical category and education.	309
12.7	Probability of zero plural in ANSE and NPE.	311
12.8	Relative strength of factors to the probability of stem weak verbs; range values.	314
12.9	Relative strength of factor groups to the probability of <i>going to</i> ; range values.	315
12.10	Linear mixed effects model for variable (hwat).	322

Preface

The variationist approach to sociolinguistics began during the 1960s, when William Labov, working with Uriel Weinreich and Martin Herzog, developed a theory of language change (Weinreich et al., 1968). Thereafter, Labov continued to advance the method and analysis of language variation and change (e.g. Labov, 1963, 1966, 1969b). In the 1970s, one of Labov's graduate students at the University of Pennsylvania was Shana Poplack. In 1981, Shana became a professor of sociolinguistics at the University of Ottawa's Department of Linguistics, the same year I entered the MA programme. I was fortunate to be Shana's student from 1981 to 1995. We produced many joint publications (e.g. Poplack & Tagliamonte, 1989, 2001), and our work together has had a lasting impact on my research. I also benefited tremendously from the influence of David Sankoff, who always had astute answers to my questions about method and analysis. When I wrote the 2006 edition of this book, everything in it had come directly from what had been passed on from this lineage – training, techniques, insights, knowledge, and sheer passion for the field.

At that time, knowledge and learning in variation analysis had been acquired through word of mouth, from one researcher to the next (see also Guy, 1988:124). It was often noted that the practical details of how to *actually do it* were arcane, largely unwritten, and for the most part, undocumented. That is why the 2006 book was conceived and written. The method needed to be recorded, systematically, thoroughly, and straightforwardly. By 2024 there is a lot of water under the bridge. A variationist approach to language science has changed substantially, and its rigorous, empirical, corpus-based, quantitative methods have spread around the world. Scholars everywhere have been training new generations of variationist linguists, and conferences and workshops focused on many burgeoning new aspects of language variation and change are flourishing.

In completing this second edition, I am indebted to Nathalie Dion, Bridget Jankowski, and Katharina Pabst who worked their way through the text and the code in superb detail, and to Bridget Jankowski and Elena Manzella for checking the final manuscript for me. I have taken everyone's comments into account and then some. My own method has evolved in just this way, changing from one research project to the next, one student or post-doc, one collaborator to the next in my perpetual efforts to do things more efficiently, more usefully, and more transparently (not to mention the joy of working with others!). As far as quantitative savvy is concerned, I have profited from the guidance and wise counsel of many people who are far better at statistics than I am, especially Harald Baayen, John Paolillo, Stefan Gries, Karlien Franco, and Jeremy Needle. Jeremy has been my workshopping

consultant over the past few years and is the expert behind the R-steps for variation linguistics that bring the content up to date with recent developments in the field.

This second edition is an entirely reworked version of the original book. No word, concept or idea has avoided scrutiny and re-evaluation. While the original work enshrined in Weinreich et al. (1968) and then built upon and elaborated by Labov (1982) are fundamental and pervasive, the enrichment and expansion in the field from other sources in the last twenty years should also be apparent and strong, including methods and ideas from the research acumen in variation analysis developing in the EU in historical linguistics, syntax, and cognitive and usage-based linguistics. I have also benefited from the input of many scholars, collaborators, and friends who are too numerous to name. They know who they are. The tracks of their influence are woven indelibly through my life, often in the lines of my CV, but sometimes less conspicuously in citations and acknowledgements, or in the odd social media posts. In sum, this new edition is a thoroughly updated, tried-and-true manual of best practice over the many years I have been working and learning in the ever-blossoming field of language variation and change.