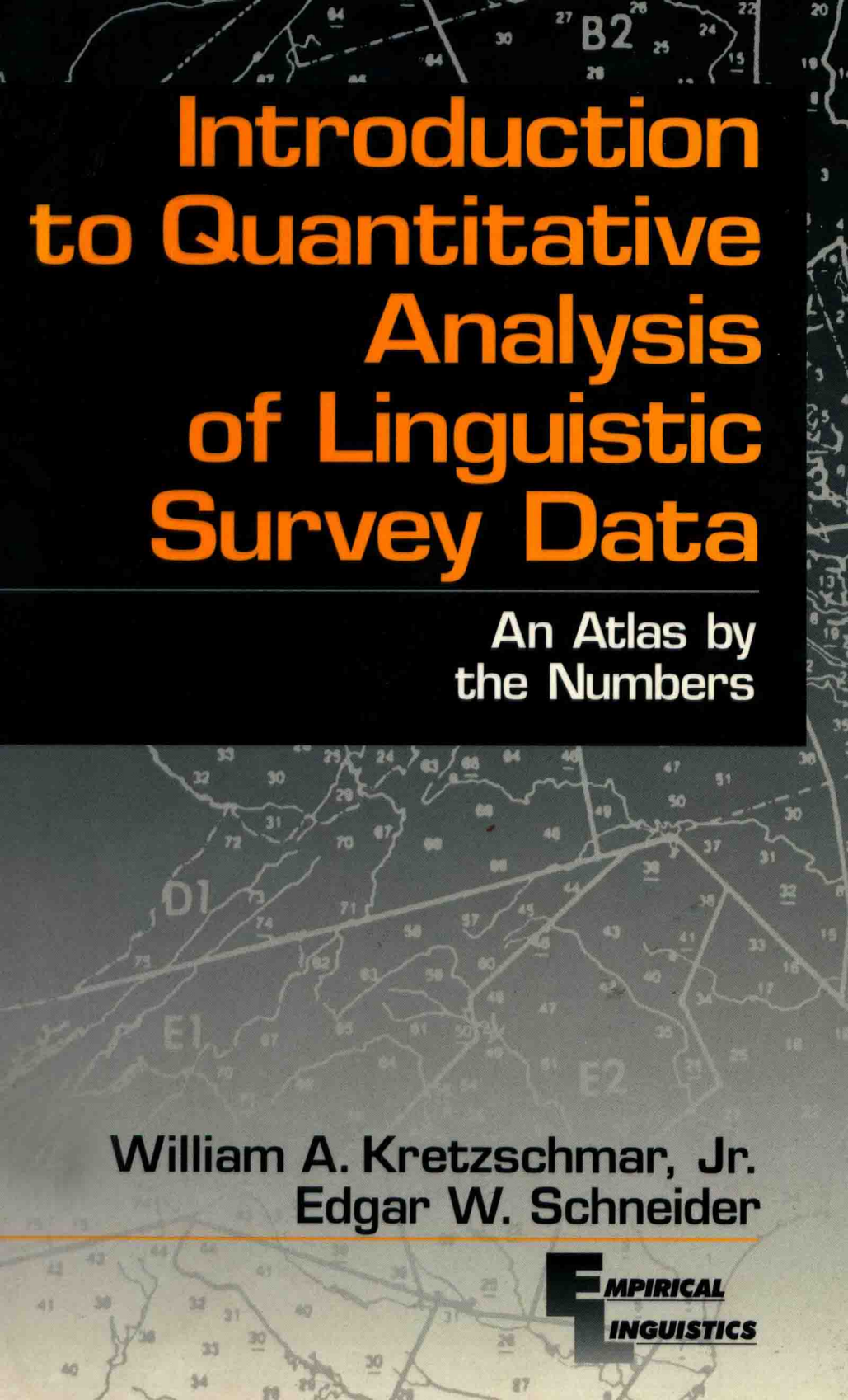


Introduction to Quantitative Analysis of Linguistic Survey Data

An Atlas by the Numbers

William A. Kretzschmar, Jr.
Edgar W. Schneider

EMPIRICAL LINGUISTICS

Introduction to Quantitative Analysis of Linguistic Survey Data

An Atlas by the Numbers

William A. Kretzschmar, Jr.
Edgar W. Schneider

EMPIRICAL LINGUISTICS

SAGE Publications
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information address:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

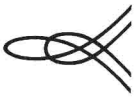
Library of Congress Cataloging-in-Publication Data

Kretzschmar, William A., Jr.
Introduction to quantitative analysis of linguistic survey data: an atlas by the numbers / William A. Kretzschmar, Jr., Edgar W. Schneider.
p. cm.—(Empirical linguistics)
Includes bibliographical references and index.
ISBN 0-7619-0111-6 (alk. paper).—ISBN 0-7619-0112-4 (pbk.: alk. paper).
1. Dialectology—Statistical methods. 2. Dialectology—Data processing. 3. Linguistic atlas of the Middle and South Atlantic states. 4. English language—Dialects—Middle Atlantic states—Data processing. 5. English language—Dialects—Southern states—Data processing. I. Schneider, Edgar W. (Edgar Werner), 1954– II. Title. III. Series.
P367.K67 1996
417'.021—dc20 96-10126

This book is printed on acid-free paper.

96 97 98 99 10 9 8 7 6 5 4 3 2 1

Sage Production Editor: Diana E. Axelsen Sage Typesetter: Marion S. Warren



Series Editor's Introduction

This volume is the first in the **Sage Empirical Linguistics Series.** It is a fitting volume to set the pattern for the series, for the hallmark of books in the series will be serious attention to empirical data, as here to survey research. Other books may take a somewhat different approach to "empirical data"—linguistic corpora, discourse, case studies—but every volume will accept the actual utterances of real people as the basis for study. In his *Course in General Linguistics* (as translated by Roy Harris, 1986, La Salle, IL: Open Court Classics), Ferdinand de Saussure distinguished what he called the "linguistics of speech" from linguistics proper, which for Saussure was the "science of linguistic structure" (18-20). This series begins with the "linguistics of speech."

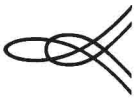
Saussure's strict dichotomy between speech and linguistic structure, *parole* and *langue,* has always had clay feet because it has always been necessary "to draw upon what the study of speech can tell us" (20). Nobody can dispense entirely with the empirical, because before we get to structure

we have to consult individual speakers about what they say (by noticing what they say, in more or less systematic ways), or to ask them for their judgments of the grammatically of examples of speech. The **Empirical Linguistics Series** fully embraces the "linguistics of speech," whereas many another linguistic forum admits as little empirical evidence as possible, preferring discussions that move as quickly as possible to linguistic structure.

Empirical linguists are not slow, not merely "dull cataloguers of data" as Robert Lees once wrote (see Chapter 1, this volume); they just prefer the careful methods of modern empirical science to reliance on the more romantic leaps of insight preferred by Lees. The series will showcase applications of such methods, and will also include textbooks that present the methods and results of different aspects of empirical linguistics. This inaugural volume proposes careful counting and statistical analysis of questionnaire data from survey research, as exemplified throughout by reference to the Linguistic Atlas of the Middle and South Atlantic States, still the largest single survey of its type. Collectively, the books of the series will build procedures for use of evidence in empirical linguistic study. Such procedures begin with detailed specification of where evidence comes from, how it is collected, and how it is categorized, and, as shown in this volume, these matters crucially affect the conclusions that can reasonably be drawn from particular evidence.

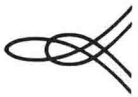
Empirical linguistic data cannot rise to achieve the status of evidence unless it is incorporated in an argument, unless it is applied to a purpose. One such goal is a better understanding of Saussure's linguistic structure, of how we might best think of language and linguistic structure, and of how language and linguistic structure work from the point of view of what people say. Another goal of linguistic evidence is to support arguments about the people who use a language, for instance their psychology or their cultural contexts. The **Empirical Linguistics Series** will not ignore social factors, as Saussure preferred to ignore them "for present purposes" (201). The **Empirical Linguistics Series** fully embraces the different purposes of the "linguistics of speech" as well as empirical study of speech itself.

William A. Kretzschmar, Jr.
University of Georgia

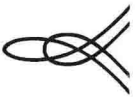


Acknowledgments

We would like to thank the Deutsche Forschungsgemeinschaft, which awarded Edgar Schneider a Heisenberg grant to support his stay in Athens in 1988–89, and the National Science Foundation and the National Endowment for the Humanities, whose grants to LAMSAS have provided equipment, materials, and research time for the work described here. We would also like to thank several anonymous readers for their suggestions, as well as the staff of the Linguistic Atlas of the Middle and South Atlantic States, especially Ellen Johnson and Rafal Konopka, for their help in making this volume possible.



Contents



List of Tables

The Research Design of a Linguistic Atlas

This book describes how the authors and the staff of the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) have reconceived this sixty-year-old project in light of computerization and recent advances of sociolinguistics. Dialectology, the study of variation in language rather than an invariant, homogenous "grammar," lends itself quite naturally to computer assistance, both for entry and maintenance of a database and for quantitative analysis. Out of the welter of different words collected in an atlas-style survey, in all of their different morphological and phonetic shapes, it makes sense to find out which variants occur more or less frequently and which ones might be associated with particular extralinguistic circumstances, and, then, which of these distributions and associations may be due to chance and which ones might genuinely have something to say about language and culture. Serious quantification and large-scale application of statistics demand the ability to handle the large amounts of data elicited in a broad survey with the efficiency of automation. Labov's breakthrough study of New York's Lower East Side (1966) secured the central role of quantitative techniques in sociolinguistics and so provided a model for dialectologists, but traditional dialectology has been slow to institutionalize the benefits of good counting (although trials were indeed attempted) in part because appropriate computer tools were not available. We now have the computer means, and we now can explore the ways of serious quantitative analysis. We have assembled the ways and

means for LAMSAS, the largest of the traditional American Atlas projects, and thus LAMSAS has been renewed as a source for new analyses.

We were able to carry out such a reassessment for LAMSAS because it still had not reached final publication. The first two fascicles of LAMSAS, full phonetic data arranged in topical lists, were published in 1980 (McDavid and O'Cain 1980); camera-ready copy for two more fascicles was completed in the early 1980s, but these were not published because the editorial staff and the publisher mutually agreed to save them until editorial funding could be found and suitable computer methods developed to allow resumption of publication on a regular schedule. Raven McDavid first asked Kretzschmar to investigate computer-assisted word processing for fascicle production in 1982; a custom-made typographic system with Atlas phonetic symbols and diacritics for computer display and for the laser printer was finally completed in 1989 (Kretzschmar 1989; Kretzschmar et al. 1993). Significant funding for editing from the National Endowment for the Humanities was awarded the following year. Computer methods have constantly improved as editing continues, although major funding from the NEH has not been available for several years. About fifteen percent of LAMSAS data have now been entered and proofread, and we can hope for renewed major funding to complete computerization of the data set.

During the 1980s, the goal for use of the computer on LAMSAS expanded from text handling to include creation of a database, and finally to include automated methods of analysis. Creation of a database and preparation for automated analysis, however, demanded answers to questions that either had not seemed crucial or had not been asked before for LAMSAS. For the database, it was necessary to decide what information needed to be stored, with what kind of structure. For eventual analysis, it was necessary to decide what kinds of questions might be asked of the data. To an extent larger than is immediately apparent in the following discussions, the approach chosen for LAMSAS owes a great debt to the model of Lee Pederson's *Linguistic Atlas of the Gulf States* (LAGS; Pederson et al. 1986–92). Although Pederson has not carried out statistical analyses on the basis of LAGS data and, to our knowledge, does not plan to conduct such analyses, the structuring of LAGS in its conversion to a fully computerized database serves as a model for all future projects and, due to the clarity of the database's organization, LAGS data are susceptible almost immediately to statistical testing. Parallels of our approach with the earlier LAGS model have thus been deliberately intended all through the expansion of the use of computers for LAMSAS (Kretzsch-

mar 1988), and this is in part motivated by the vision of ultimately "realizing Kurath's original dream for an American Atlas" (Kretzschmar 1988, 216) in the form of an electronic atlas, uniting computerized regional projects that have adopted a comparable coding framework. That goal has now begun to be realized on the Linguistic Atlas Web page (URL as of publication: http://hyde.park.uga.edu), which contains data and materials from several regional atlas projects.

As the computer marketplace came eventually to develop hardware and software that could be adapted for LAMSAS purposes (key components became available in 1986–88), the time approached when final decisions needed to be made about the database and eventual analytical methods. After Kretzschmar invited Schneider to collaborate in the decision making, the authors applied for and received funding from the Deutsche Forschungsgemeinschaft and from the National Science Foundation during 1988–89 to study the basic issues of LAMSAS database construction and application of statistical methods. We have continued to pursue these issues until now—when we feel confident that we have achieved a set of ideas sufficient to support the future needs of the project and its users. This book provides a developmental and logical view of our database and analytical methods; it should serve to complement the exhaustive factual treatment of LAMSAS in Kretzschmar et al. (1993).

We also explicitly intend that our discussion of LAMSAS should be applicable to other atlas and atlas-style surveys, and to other questionnaire-based linguistic studies (such as, for instance, postal, telephone, or computer-mail questionnaire surveys). LAMSAS was originally planned and executed before the era of modern survey research, and so our discussion must treat some basic notions of sampling and statistical inference to determine their applicability to LAMSAS. These basic issues, always presented in the context of LAMSAS or other concrete examples, are the same issues that must confront any researcher who wishes to use survey questions. In particular, we believe our treatment of the logic of analysis of multiple-response questions—as opposed to allowance for only a single answer for a given question—to be of value to our colleagues and students. Sources other than this volume will offer better coverage of work in the field (for instance, Pederson et al. 1972; Pederson 1996). This volume is most relevant to work in the research office and classroom, for both planning and analysis. We hope that the issues we raise and solutions we propose can help other investigators to use their resources more fruitfully.

Chapter 1 of our book is devoted to the relation of sociolinguistics to the original conception of the Atlas and to fundamental preliminary considerations. Chapter 2 will describe in detail the adaptation of LAMSAS to the needs of computerization and the research methods envisioned. The third chapter discusses some of the mechanics involved in computerizing LAMSAS, how to handle and to analyze the data in the database management system chosen for this purpose, R:Base, and the creation of categories for analysis. The fourth chapter discusses the logic of statistical testing, and illustrates how to submit partial data sets to tests for statistical significance. The fifth and final chapter provides model analyses of two lexical files, one already treated in Kurath (1949) and one that was not, to discover and to substantiate certain conditions of linguistic usage. It goes without saying that these limited presentations by no means exhaust the possibilities, but they do provide some orientation about which directions such analyses can and should follow.

The initial discussion of research design falls naturally into two parts. The first is a foreshortened historical view of how the American Linguistic Atlas program came to be established (for more detailed information, see O'Cain 1979; Kretzschmar 1988) and the relation of the Atlas effort to the development of modern sociolinguistics. The second part consists of commentary on the nature of atlas procedures. Specific procedures will be treated in other chapters; the theme of this section is not the actions of computing, statistical testing, and categorization but instead the essential interaction of these topics with the conception of atlas work. The first part leads into the second part because a comparison of American Atlas work and sociolinguistics necessarily reveals differences in basic assumptions. As Ralph Fasold's comprehensive two-volume text survey clearly demonstrates (1984, 1990), sociolinguistics is understood and practiced in many different ways by many different people. Not a few of those different ways can be related to atlas-style research, from studies of diglossia to perceptions of language variation to language planning. For our purposes, though, sociolinguistics is identified with the work initiated in the 1960s by William Labov on Martha's Vineyard and in New York City (1963, 1966) and across the Atlantic by Peter Trudgill in Norwich (1974). We do not intend to produce a programmatic comparison of Labovian sociolinguistics with traditional dialectology but instead intend to isolate and assess what we think to be key differentiating factors between them. This kind of characterization is all the more important because the four

textbooks of dialectology from the early 1980s all take partisan positions: Francis's *Dialectology* (1983) concentrates on the methods of traditional dialectology; Petyt's *The Study of Dialect* (1980) is most noteworthy for its exploration of the relationship between dialectology and linguistic theory; Chambers and Trudgill's *Dialectology* (1980) privileges sociolinguistics, and often takes a dim view of the "largely superannuated" methods (206) of what they call "dialect geography"; and Davis's *English Dialectology* (1983) defends traditional dialectology and finds numerous flaws in an assessment of the statistics presented by sociolinguists in several early works. The second part of this chapter asserts basic principles that in large part emerge from contrasts between sociolinguistics and traditional dialectology, and shape how we now think about atlas work.

Sociolinguistics and the Atlas Design

Hans Kurath established the methods for the Linguistic Atlas effort in the United States, and these same methods have been used for the separate autonomous regions of the American Atlas. In brief, as these methods were applied to the earliest of the regional studies along the Atlantic Coast, New England (LANE; see Kurath et al. 1939), and the Middle and South Atlantic States (LAMSAS; see Kretzschmar et al. 1993), communities were selected to form a representative grid of the region; informants were selected to be representative of the communities; interviewers were specially trained to record responses in fine phonetic notation; and a questionnaire including items selected from everyday speech was administered in as informal a situation as could be obtained. Communities (normally counties) were chosen with regard to culture, settlement, and demographic characteristics so as to include historically important places and cultural groups within an even spread of area and population. Within the communities, two speakers were normally selected as representative of the community because of lifelong residence there, one a member of the oldest living generation with little education or compensating experience, one younger and better educated with a less insular outlook. In twenty percent of the communities, a member of the local elite was interviewed; in urban areas, a greater number of informants was selected to accord with population density. Interviews required an average of eight hours to complete, divided into multiple sessions, often in

the informant's home. A questionnaire item corresponds to a particular page and line of the questionnaire; some items actually constitute more than one question, because the informant's response yields more than one term or form of interest (see A. Davis et al. 1969). Typical questioning styles avoided "how do you say . . ." questions in favor of less direct approaches to reduce the formality of the interview situation (discussed in Pederson et al. 1972). In this age before tape recorders, interviewers took down responses in Kurath's phonetics (see Kurath et al. 1939; Kretzschmar et al. 1993) during the interview, indicated any special circumstances of responses, and captured informants' comments. Interviewers made a detailed biographical sketch for each informant.

Kurath also set the analytical standard in his landmark *Word Geography of the Eastern United States* (1949) and with Raven McDavid in the somewhat more controversial *Pronunciation of English in the Atlantic States* (1961). The primary analytical tools in these works were his well-known isoglossic and symbolic maps. In the *Word Geography,* Kurath employed weighting adverbs such as "regularly," "frequently," or "occasionally" with reference to the intensity of a form's use, but no quantifying statements in the narrow sense. Indications of frequency of occurrence appear for the first time in an atlas analysis, in a fairly simple fashion, in Atwood 1953, where one can commonly read that a certain number or proportion of the informants of a certain type in a certain area use a particular form (e.g., "Driv /drɪv/ is used by 24/176 informants in n.e. N. Eng. . . . About one third of the Md., Va., and N.C. Negro informants use driv" (11). Other modifications of Kurath's methods and a few innovations in data acquisition and analysis have been well described by Harold Allen (1977) for the period following the publication of Kurath's *Linguistic Atlas of New England* (1939–43) until the publication of Allen's own *Linguistic Atlas of the Upper Midwest* (1973–76).

While Kurath carried certain basic procedures of linguistic geography across the Atlantic from the European models of Gilliéron and Edmont's French Atlas (1902–10) and Jaberg and Jud's Italian-Swiss Atlas (1928–40), he also went beyond the traditional reliance on folk speakers to include interviews on three social levels, and he collected biographical information about his informants just for the purpose of social analysis (1939, 44). Kurath's *Studies in Area Linguistics,* his last word on methods, contains a chapter titled "The Social Dimension of Area Linguistics" (1972, 164–84) in which he talked about urban dialectology (see also Kurath 1970) and other

social questions, and in which he discussed particularly the work of Labov in New York City (he also cited Labov's Martha's Vineyard study). Raven McDavid, Kurath's successor as editor of LAMSAS, had a similar interest in social variation in American English and, in a string of landmark articles after World War II (McDavid 1946, 1948, 1950, 1951, 1953, 1955, 1956, 1960, 1961; McDavid and McDavid 1951), he reified the potential for social dialectology that was implicit in Kurath's plans (see Allen 1977, 233–34).

Even granted that areal dialectology in the United States has always had at least some concern for a social dimension in language variation, it has been necessary for those working on Linguistic Atlas projects to react to the challenge of Labovian sociolinguistics. From its origins in the 1960s, sociolinguistics as practiced in America by Labov, Shuy, Wolfram, and others emerged as a separate discipline with separate goals, not as a lineal development out of the social aspects of traditional dialectology. In a 1980 article, Raven McDavid suggested that

> one should consider linguistic geography and sociolinguistics not as competing but as complementary. . . . The linguistic geographer cannot expect the sociolinguist to replicate his previous findings, but he can expect his findings to be utilized when they deal with communities in question. The sociolinguist in turn cannot expect the linguistic geographer to use his methods, but he can expect a recognition that various aims and techniques are appropriate to various situations. Neither should let theoretical orientation stand in the way of the data. (275–76)

McDavid further suggested that

> the difference in aims between linguistic geography and sociolinguistic studies are reflected in two principal ways: the selection of informants and the presentation of data. Since the regional surveys are concerned with indigenous benchmark cultural groups, their investigators rely on local intermediaries who know the community and its people. . . . With the interest in statistics common to the social sciences, sociolinguistic studies try to employ objective techniques of selecting and classifying informants. . . . Linguistic geographers generally identify informants and their responses . . . sociolinguistic studies, emphasizing the usage of groups, generally submerge the individual in statistical frequencies. . . . As with other differences, each kind of investigation can learn from the other, and both would probably benefit from a better knowledge of statistics. (277)