CREATING AND DIGITIZING LANGUAGE CORPORA

VOLUME 2: DIACHRONIC DATABASES

Edited by
Joan C. Beal
Karen P. Corrigan
ann L. Moisl

ord by Shana Poplack

Creating a  ing Language

Volume 2: Diachronic Databases

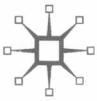
Edited by

Joan C. Beal
University of Sheffield

Karen P. Corrigan and Hermann L. Moisl
Newcastle University

Foreword by Shana Poplack
University of Ottawa

palgrave
macmillan



First published 2007 by
PALGRAVE MACMILLAN
Houndmills, Basingstoke, Hampshire RG21 6XS and
175 Fifth Avenue, New York, N. Y. 10010
Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN-13: 978–1–4039–4367–5
ISBN-10: 1–4039–4367–2

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

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

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

10 9 8 7 6 5 4 3 2 1
16 15 14 13 12 11 10 09 08 07

Printed and bound in Great Britain by
Antony Rowe Ltd, Chippenham and Eastbourne

Foreword

Only two or three decades ago, those of us who had the patience and the wherewithal to construct a computerized corpus of recorded speech, however clunky, were the envy of our colleagues. In those days, linguists interested in quantitative analysis simply slogged through their audio-tapes, extracting unfathomable quantities of data by hand. Cedergren, to name but one notable example, analyzed 53,038(!) tokens of phonological variables, culled individually from her tapes, in her 1973 analysis of Panamanian Spanish.

The gold standard for transcribed corpora at the time was the concordance, possessed by a fortunate few, and coveted by all who were doomed to manual extraction. Of course the vintage concordance was largely limited to lexically-based retrieval, but at least it was searchable. The papers that Joan Beal, Karen Corrigan and Hermann Moisl have assembled in these companion volumes are eloquent testimony to how far the field of corpus linguistics – now rife with electronic corpora – has come in so short a time.

Building a corpus arguably involves a greater investment in time, resources and energy than any other type of linguistic activity. Decisions are legion at every stage of the process: sampling, ensuring representativeness, collecting data, transcribing them, correcting, standardizing the transcription, correcting, tagging and markup, correcting, and facilitating retrieval. Adding to the challenge is the fact that at the outset of the project the researcher is often not even familiar enough with the materials to make the best decisions, and changing midstream is costly and time-consuming. What could possibly make such a huge front-end investment worthwhile? Dealing with corpora at every stage of development, from fledgling endeavours to large-scale, heavily exploited enterprises, these reports offer a state-of-the-art synthesis of the problems researchers have encountered and the solutions they have adopted to deal with them.

The focus of these volumes is on *unconventional* corpora, like the non-standard, regional and dialectal varieties of speech, creole texts, child language, and the correspondence, business transactions, prose and plays of past centuries discussed here. Each poses problems hardly imaginable to the early builders of more orthodox corpora based on written or standard materials. The unifying question is how to 'tame'

them, in the editors' terminology. Taming, as understood here, is largely a question of representation: How to represent forms for which there is no standard orthography, what to represent, how much to annotate, how much analysis to impose on the materials, how to represent ambiguities and indeterminacies, how to represent the finished product to the end-user. Noting the diversity, not only in the models underlying different corpora but also in their methods of encoding and analysis, the editors, themselves seasoned corpus builders, question whether it is reasonable or even feasible to aim for standardized protocols of the kind employed in traditional corpora for the collection, transcription, annotation and preservation of their less conventional counterparts.

Perhaps the first to grapple with the problem of taming unconventional data were the Sankoff-Cedergren team, whose *Montreal French Corpus* (Sankoff and Sankoff 1973) was built to elucidate a stigmatized variety previously widely believed to be an incorrect version of European French. Their goal was to show that the 'deviant' forms were part of a complex sociolinguistic structure, by tapping into different sources of speech variation: inter-individual, intra-individual and intra-linguistic. Chief among the problems inherent in such an endeavour was the issue of representativeness: How to guarantee representativeness of all the possible diversity in speech, while maintaining randomness in the selection of informants? They achieved this by implementing a detailed sampling frame, which, in contrast to their material procedures, has not yet been superseded. Their problems and solutions hark back to a simpler time, especially as compared with those corpus linguists face today. The transcription protocol – standard orthography – was dictated by the number of symbols on the punch keyboard for the IBM computer cards they used. Correction was effected by removing the card containing the error and inserting a correctly punched card in its place. The 100,000 cards containing the transcriptions then had to be converted into reams of computer printouts – and all without dropping a single card! In an era in which an entire corpus can be carried around on a memory stick or an iPod, it is worth noting that the print concordance of the 3.5 million-word *Ottawa-Hull French Corpus* (Poplack 1989), for example, occupies an entire wall – floor to ceiling – of the Ottawa Sociolinguistics Lab. The technology was primitive.

Since then, striking advances, not only in terms of hardware, but also in the area of annotation systems, have revolutionized corpus linguistics. No protocol has yet emerged as standard, though – as observed by the editors in initiating this project. So it's no surprise that

the issue of annotation enjoys pride of place in these volumes, with researchers weighing in on what to annotate, how much detail to include, and whether it is preferable to replicate markup schemes of other corpora or tailor them to one's own. It is clear that the old problem of finding the right balance of quantity, recoverability and faithfulness is still with us. Faithfulness at every linguistic level to data with much inherent variability (i.e. all speech, and many older and/or nonstandard written texts) inevitably results in diminished recoverability and less quantity. Without sufficient quantity, statistical significance is impossible to establish and full cross-cutting conditioning yields mostly empty cells. Optimum recoverability comes at the expense of less faithfulness to the many variant realizations of what is underlyingly a single form.

Each of the contributors to these volumes grapples with these problems in their own way. Some prefer to abandon one or more of the principles, others respond with complicated interfaces. As a result, the corpora described in this collection illustrate the full gamut of possibilities, from an annotation system so rich and complex that it already incorporates a good deal of the linguistic analysis, at one extreme, to virtually no markup whatsoever at the other. Linkage of transcripts to (audio and video) recordings and syntactic parsing will no doubt the wave of the future.

The projected use of the corpus, as *end-product or tool*, is clearly the determining factor. Those for whom the corpus is a tool tend to advocate minimal annotation. These researchers are able to tolerate more indeterminacy and ambiguity, either because they have determined that it will not affect what they're looking for (e.g. a number of the corpora described here provide no detail on phonetic form or discourse processes), or because the sheer volume of data available allows them to omit the ambiguous cases or neutralize errors through large-scale quantitative analysis. Others, for whom the corpus is the end-product, tend to aim for consistency with guidelines for existing corpora, even if these do not seem immediately relevant to the proposed research. So what is the best annotation system? The amalgamated wisdom to be gleaned from these contributions: the one that works for you. At the moment, then, the answer to the editors' query regarding the feasibility of standardizing transcription protocols seems to be a qualified 'no'.

Comparatively less emphasis is placed on the issue of *representativeness*, the extent to which the sample of observations drawn from the corpus corresponds to the parent population. Achieving representativeness for (socio)linguistic purposes involves identifying the major

sources of variation in the population (of speakers and utterances) and taking them into account while constructing the sample. Few corpora in these volumes, by necessity or design, claim to be representative in the sense of Sankoff (1988). Rather, in most of these contributions, (as in much social science research more generally), the sample is opportunistic. This is an issue that every corpus must come to terms with, since even large numbers of observations cannot compensate for a sample frame from which the major sources of variation are missing. To the extent that the sample does not span the variant answers to the research question, pursuit of that question via that corpus can only be spurious.

Whether representativeness or annotation is more fundamental to the eventual utility of the corpus is a moot point. It is worth noting, however, that the awkward, and for some, simplistic, transcription protocols of early unconventional corpora did nothing to diminish their interest, value and current relevance. Hundreds of studies have been, and continue to be, based on them, perhaps because the research questions they were constructed to answer are still burning ones. The same is of course true of a number of the established corpora described in these volumes, and no doubt will be of the many more incipient ones as well. The good news is that these repositories have an enduring value that far transcends our automated treatment and handling of them.

I end this foreword by returning to the question I posed at the beginning. What could possibly make the huge front-end investment required to build a corpus worthwhile? Obvious answers include the enormously enhanced speed of data collection, enabling consideration of ever greater quantities of data with relatively little extra effort. This in turn increases the chances of locating rare tokens, achieving statistical significance and determining which factors condition the choice between alternating forms. All of these are inestimable boons for quantitative analysis, but they pale in comparison to what for me remains the most exciting aspect of corpus work: the opportunity it affords to serendipitously discover what one wasn't looking for, to characterize the patterned nature of linguistic heterogeneity, and in particular the hidden, unsuspected or 'irrational' constraints that are simply inaccessible to introspection or casual perusal.

How much closer are we to the goal of agreeing on a standardized annotation? Well, we aren't there yet, though only time will tell. In the interim, anyone who has ever considered building a corpus or is engaged in doing so now will want to have a copy of this book close at hand. The wide variety of contributions convey much of the excitement

of this burgeoning field. Despite inevitable differences in methods and projected end uses, the common thread is the shared goal of finding and implementing the best practices in corpus construction and preservation. These companion volumes, examining both synchronic and diachronic corpara, serve as a model for how to achieve them. For this, we can only be grateful to the editors, who encouraged such stimulating dialogue.

SHANA POPLACK

References

Cedergren, Henrietta. 1973. 'Interplay of social and linguistic factors in Panama'. PhD dissertation, Cornell University.

Poplack, Shana. 1989. 'The care and handling of a mega-corpus'. *Language Variation and Change* (Current Issues in Linguistic Theory, 52), ed. by R. Fasold and D. Schiffrin, pp. 411–451. Philadelphia: John Benjamins.

Sankoff, David. 1988. 'Problems of representativeness'. *Sociolinguistics. An International Handbook of the Science of Language and Society*, Vol. 2, ed. by U. Ammon, N. Dittmar and K. J. Mattheier, pp. 899–903. Berlin: Walter de Gruyter.

Sankoff, David and Sankoff, Gillian. 1973. 'Sample survey methods and computer-assisted analysis in the study of grammatical variation'. *Canadian Languages in their Social Context*, ed. by R. Darnell, pp. 7–63. Edmonton: Linguistic Research Inc.

Notes on the Contributors

Will Allen worked from 2001 to 2005 as a Research Associate on the AHRC-funded Newcastle Electronic Corpus of Tyneside English (NECTE) project in the School of English Literature, Language and Linguistics at Newcastle University. Since then he has been working as a Consultant Trainer for Netskills, Newcastle University, delivering and developing internet-related training.

Joan C. Beal was a Senior Lecturer in English Language at Newcastle University until moving to the University of Sheffield in 2001. She was Co-investigator on the NECTE project at Newcastle and is currently Professor of English Language and Director of the National Centre for English Cultural Tradition. Recent publications include *English in Modern Times* (2004) and *Language and Region* (2006).

Karen P. Corrigan has held lectureships at University College Dublin and the Universities of Edinburgh and York (UK). She was Principal Investigator on the NECTE project and is currently a Reader in Linguistics and English Language at Newcastle University. She was awarded a Leverhulme Trust Research Fellowship (2000–02) and has recently published *Syntax and Variation* (2005) (with Leonie Cornips).

David Denison is Professor of English Linguistics at the University of Manchester and has held visiting posts in Amsterdam, Vancouver, Santiago de Compostela and Paris. Recent jointly edited publications include *Fuzzy Grammar* (2004) and *A History of the English Language* (2006). He is a founding editor of the journal *English Language and Linguistics*.

Susan Fitzmaurice has held academic posts at the Universities of Cape Town, Cambridge and Northern Arizona. She is currently Professor of English Language at the University of Sheffield. She publishes widely on socio-historical linguistics and pragmatics using the Network of Eighteenth Century English Texts as a major data source.

Elizabeth Gordon taught at the University of Canterbury from 1967 until she retired in 2003 as an Associate Professor. She is co-leader of

the University of Canterbury research team on Origins of New Zealand English (ONZE) and one of the authors of *New Zealand English: Its Origins and Evolution* (2004).

Jennifer Hay is a Senior Lecturer in the Department of Linguistics at the University of Canterbury and is also a member of the ONZE team. Recent publications include *Causes and Consequences of Word Structure* (2003), *New Zealand English: Its Origins and Evolution* (co-author, 2004) and *Probabilistic Linguistics* (co-editor, 2003).

Raymond Hickey studied for postgraduate degrees at Trinity College, Dublin and at Kiel, Germany. He completed his German Habilitation in Bonn in 1985 and has held professorial appointments at the universities of Bonn, Munich, Bayreuth and, currently, Essen. His recent publications include *A Source Book for Irish English* (2002), *Corpus Presenter* (2003) and *Legacies of Colonial English* (2005).

Francis Jones is a literary translator and Senior Lecturer in Applied Linguistics at Newcastle University. His recent book *Prevoditeljev put* (2004) examines ideology, identity and literary translation studies against the break-up of Yugoslavia. Aided by a British Academy grant, he is currently researching poetry translation processes.

Margaret Maclagan is a Senior Lecturer in Communication Disorders at the University of Canterbury, NZ. She is another member of the ONZE research team and was one of the authors of *New Zealand English: Its Origins and Evolution* (2004).

Warren Maguire is in the final stages of his PhD research on vocalic mergers in Tyneside English at Newcastle University and was formerly a Research Associate on the NECTE project. He currently works as a Research Associate on another AHRC-funded project, namely, 'Sound Comparisons: Dialect and Language Comparison and Classification by Phonetic Similarity' at Edinburgh University.

Anneli Meurman-Solin is a Lecturer in English Philology at Helsinki University. She has published widely in the fields of historical dialectology/stylistics and corpus linguistics. She is currently a Fellow of the Helsinki Collegium for Advanced Studies and acts as a domain leader for a research strand at the Research Unit for Variation, Contacts and Change in English.

Hermann L. Moisl is a Senior Lecturer in Computational Linguistics at Newcastle University and he was Co-investigator on the NECTE project. His interests and publications are in natural language processing, neural modelling of language, and multivariate analysis of corpora.

Terttu Nevalainen is Professor of English Philology at the University of Helsinki and the Director of the Research Unit for Variation, Contacts and Change in English. Her publications include: 'Early Modern English lexis and semantics', in *The Cambridge History of the English Language* (1999), *Historical Sociolinguistics* (2003, with H. Raumolin-Brunberg) and *An Introduction to Early Modern English* (2006).

Helena Raumolin-Brunberg is a Senior Scholar in the Research Unit for the Study of Variation, Contacts and Change in English at Helsinki University. Her interests include historical sociolinguistics, language change and corpus linguistics. She has recently published *Historical Sociolinguistics* (2003, with Terttu Nevalainen).

Naomi Standen has been Lecturer in Chinese History at the University of Newcastle since 2000, having previously worked at the University of Wisconsin-Superior and St John's College, Oxford. She is co-editor of *Frontiers in Question: Eurasian Borderlands, 700–1700* (1999), and author of *Unbounded Loyalty: Frontier Crossing in Liao China* (forthcoming 2007).

Ann Taylor is currently a Research Fellow at the University of York. In cooperation with colleagues at the Universities of York, Pennsylvania and Helsinki, she has been instrumental in creating syntactically annotated corpora for Old, Middle, and Early Modern English, as well as publishing on historical variation in English and Greek.

Linda van Bergen obtained her PhD from the University of Manchester in 2000 and subsequently held a British Academy Postdoctoral Fellowship at the University of York. She is now a lecturer in English Language at the University of Edinburgh. Her publications include *Pronouns and Word Order in Old English* (2003).

List of Abbreviations

AHDS	Arts and Humanities Data Service
AHRB	Arts and Humanities Research Board
AHRC	Arts and Humanities Research Council
ASCH	American National Standard Code for Information Exchange
ASP	Active Server Pages
BNC	British National Corpus
CC	Canterbury Corpus
CEEC	Corpus of Early English Correspondence
CEECE	Corpus of Early English Correspondence Extension
CEECS	Corpus of Early English Correspondence Sampler
CELEX	Centre for Lexical Information
CLAWS	Constituent Likelihood Automatic Word-tagging System
CLEP	Corpus of Late Eighteenth-Century Prose
CONCE	Corpus of Nineteenth-century English
CS	Code-switching
CSC	Corpus of Scottish Correspondence
DAT	Digital audio tape
DOE	*Dictionary of Old English*
DTD	Document Type Definition
ECOS	Edinburgh Corpus of Older Scots
ESRC	Economic and Social Research Council
FTP	File Transfer Protocol
HC	Helsinki Corpus of English Texts
HTML	HyperText Markup Language
IA	Intermediate Archive
ICAME	International Computer Archive of Modern and Medieval English
ICLAVE	International Conference on Language Variation in Europe
IHD	Institute for Historical Dialectology
IPA	International Phonetic Alphabet
LAEME	Linguistic Atlas of Early Middle English
LALME	Linguistic Atlas of Late Medieval English
LAOS	Linguistic Atlas of Older Scots
LIDES	Language Interaction Data Exchange System

MU	Mobile Unit
NAS	National Archives of Scotland
NECTE	Newcastle Electronic Corpus of Tyneside English
NEET	Network of Eighteenth-Century English Texts
OED	*Oxford English Dictionary*
ONZE	Origins of New Zealand English
OTA	Oxford Text Archive
OTP	Orthographic Transcription Protocol
OU	Overall Unit
PCEEC	Parsed Corpus of Early English Correspondence
PDV	Putative Diasystemic Variable
PPCME2	Penn–Helsinki Parsed Corpus of Middle English II
PVC	Phonological Variation and Change in Contemporary English
SCOTS	Scottish Corpus of Texts and Speech
SGML	Standard Generalized Mark-up Language
SPSS	Statistical Package for the Social Sciences
SQL	Structured Query Language
SS15	Fifteenth Sociolinguistics Symposium
SSRC	Social Science Research Council
TACT	Text Analysis Computing Tools
TEI	Text Encoding Initiative
TLS	Tyneside Linguistic Survey
UCREL	University Centre for Corpus Research on Language
UKLVC	UK Language Variation and Change Conference
VARIENG	Research Unit for Variation and Change in English
XHTML	Extensible HyperText Markup Language
XML	Extensible Markup Language
XSLT	Extensible Stylesheet Language Transformations
YCOE	York–Toronto–Helsinki Parsed Corpus of Old English Prose

Contents

List of Tables

List of Figures